

# Analyse des Correspondances

Recherches au cœur de l'analyse des données

Brigitte Escofier–Cordier (1941-1994)

Presses Universitaires de Rennes

## 12.5. Conclusion

In this paper we have discussed the ways for treating missing data that were most important in our opinion. We related these options for the treatment of missing data to the types of missing data that can occur. Table 3 indicates which option for the treatment of missing data will most often be preferable, given that one knows what type of missing data one is dealing with. We agree with [Nishisato, 1980], see also section 12.3.3) that if the reason for being missing is unclear, it seems wise to assume that the data missing not at random and start with *missing single*. Thus one is able to study the behavior of the extra categories.

For most of the types of missing data we pointed out that often, in our opinion, more than one option for the handling of missing data was acceptable. We never preferred *missing multiple* due to the fact that is likely to produce outliers. *Missing insertion* is never preferred because we think that it is only useful in very specific circumstances: in ordinary circumstances there are always better candidates, in our opinion. When more than one option is preferable, a choice between these options should be made if there are specific ideas about the properties of MCA that are thought to be important.

# Analyse Factorielle lissée et Analyse Factorielle des différences locales

H. Bénali, B. Escofier<sup>1</sup>

## Résumé

Nous proposons ici deux méthodes qui permettent de tenir compte, dans une analyse en composantes principales, d'une structure de proximité ou de contiguïté définie sur l'ensemble des lignes d'un tableau de données. La première, l'analyse lissée permet d'analyser les tendances générales des données en éliminant les fluctuations locales. La deuxième, l'analyse des différences locales permet, au contraire, d'analyser ces fluctuations en éliminant les variations générales liées aux positions spatiales. Nous comparons cette dernière à l'analyse locale (cf. [Lebart, 1984]) et à l'analyse en composantes principales pour des unités statistiques corrélées (cf. [Causinus, Aragon, 1980]). Des exemples d'applications illustrent ces techniques. Ces méthodes permettent aussi d'étudier les relations entre un ensemble de variables et une structure de contiguïté.

## 13.1. Les données

Nous traitons ici le cas de tableaux de données traités habituellement par l'Analyse en Composantes Principales (ACP) c'est-à-dire des tableaux croisant des individus et des variables numériques. Nous exposerons ultérieurement la généralisation des méthodes proposées à d'autres types de tableaux.

On note  $I$  l'ensemble des individus et  $n$  leur nombre. On note  $J$  l'ensemble des variables et  $p$  leur nombre. On note  $X$  le tableau de données, son terme général  $x_{ij}$  représente la valeur de la variable  $j$  pour l'individu  $i$  et  $x_j$  représente la variable  $j$ . La plupart du temps, les variables sont centrées et réduites ; pour simplifier les notations,  $X$  désigne la matrice des données ainsi transformées.

Sur l'ensemble des individus existe, de plus, une structure de contiguïté. Il peut s'agir d'une proximité plane si, par exemple, les individus sont des zones géographiques ou

to our discussion of *missing not categorizable*, we do not dismiss *missing insertion* because it is a possible way to insert the missing objects into the category they would have had if they were not missing.

### 12.4.3 Missing completely at random

If we deal with data that are missing completely at random, it is perhaps best to *not* use the option *missing single* since it treats all objects missing in the same way whereas they do not have anything in common, due to the definition of *missing completely at random*. On the other hand, *missing multiple* seems elegant since it gives a category point for each missing score, but on the other hand it is very susceptible to producing outliers. Good choices are the options *missing passive* and *missing passive modified margin*, since they only use the available information for each object. Using the fuzzy options can also be defended, since this corresponds to imputing the average for *missing fuzzy average* or subgroup averages for *missing fuzzy subgroup*. Obviously, the latter of these two fuzzy options is more elegant. *Missing insertion* will also be a reasonable option in this context, since it can lead to inserting an object into its original category.

Our conclusions correspond roughly with findings from stability studies performed by [Chan, 1978]; see also [Nishisato, 1980], for a summary), [Meulman, 1982] and [Bénali, 1985]. In these studies data sets are used with a known structure, and in these data sets cells are defined completely at random as missing. It is evaluated how well the various options recover the properties of the complete data from the analysis of the incomplete data.

[Chan, 1978] uses a three-parameter logistic latent trait model for binary items. The generated data are one-dimensional. She compares, among others, *missing passive* and *missing single*, and finds that the recovery of information in the complete data from the incomplete data declines rapidly for *missing single* when the proportion of missing goes from 15 to 20%. On the other hand *missing passive* works relatively good, even with 25% missing values.

[Meulman, 1982] compares the options *missing passive*, *missing single* and *missing multiple* using many criteria. Missing data are randomly inserted into three types of data sets, namely an algebraic data set, data generated under the multinormal distribution, and a set of real data that we do not discuss here further. All data sets are one-dimensional. The first data set has a rather strong first dimension. In general *missing single* performs worse than *missing passive* and *missing multiple* when the number of missing values increases from 7 via 14 to 20%. There are two data sets generated with the multinormal distribution. The first has a rather weak first dimension:  $r = .30$  for seven variables each discretized into five categories, with  $n=80$ . For 9% missing, *missing passive* performs best; for 16% missing, *missing passive* still performs best, though *missing single* still performs acceptable, and *missing multiple* is not acceptable anymore. The second data set is drawn from a multinormal distribution with  $r = .20$ , for  $n=1500$ ; 14% missing values are imputed. In this case *missing single* performs slightly better than *missing passive*, whereas *missing multiple* should be

three options behave rather well; when the first dimension is less strong, then missing passive and missing single are superior.

[Bénali, 1985] compares *missing passive modified margin* with at random inserting the objects missing in any of the categories. He uses real data that are clearly more-dimensional in which he defined missing values completely at random. He finds that *missing passive modified margin* gives somewhat more stable results than the alternative option. However, the evidence provided by this study is relatively small, since only one real data set is used.

Important as these studies may be, it should be noticed that it only describes the behavior of some of the options under completely random missing data, and it concentrates mainly on the behavior of options in one-dimensional data sets. *Missing passive* and *missing passive modified margin* seem to perform adequately. Contradictory results are found by Chan and Meulman for *missing single*: Meulman explains the acceptable performance of *missing single* in her study from the fact the data generated by her have more categories, and she concludes *the smaller number of categories, the worse missing single is expected to behave* (Meulman, 1982, p. 164). The reason for this is that, with a smaller number of categories, the relative increase of the total inertia is larger (see section 12.3.3).

### 12.4.4 Missing at random

Given the definition of *missing at random* the option *missing fuzzy subgroup* seems a very elegant alternative since it specifies probabilities that an object falls into any of the categories given its other scores. For the same reason *missing fuzzy average*, and *missing insertion* are suboptimal options. We could use *missing passive* and *missing passive modified margin* if we only want to make use of the scores an object has. Conceptually, *missing multiple* is also useful, but probably produces outliers. *Missing single* is advised against since the objects missing do not necessarily stem from the same (unknown) class.

### 12.4.5 Created missing

Lastly, we might be dealing with data made missing on purpose because we want to eliminate the influence of some category. In this case we could best use *missing fuzzy average*, since this option minimizes the total inertia, and thus eliminates the influence of this category completely. Other possible candidates are *missing passive* and *missing passive modified margin*, in which no attempt is made to extract information from the fact that someone is missing. A drawback of *missing fuzzy average* is that it does not work for binary variables: if one of both categories is eliminated, the average proportion for the other category becomes 1, and we have a column with all values equal to 1.

*missing not at random*. This may be the case if the observed salary distribution of the higher educated persons differs from the non-observed salary distribution for the higher-educated persons that refused to specify their salary.

Another instance in which we deal with missing data in MCA is when we create them on purpose. We coin these missing data *created missing* (CM). The general idea of creating missing data is that we do not want certain categories to have an effect on the solution; therefore we define objects falling into these categories as missing. For example, we might eliminate categories with very low frequencies because they have produced outliers or they have dominated the solution in an earlier analysis. Another reason is that we consider objects in a specific class to be in that class for heterogeneous reasons, for example, some persons do not agree with an item because it is to extreme whereas other do not agree because it is not extreme enough. Typical examples of this last phenomenon are *pick any out of m-data*, *seriation data*, or *voting data*. For example, for *pick any out of m-data* it is possible that only persons that pick some item are thought to be identical, since the persons who did not pick the item might have done so for different reasons. For *seriation data*, objects found in a tomb might indicate an identical period, whereas objects not found might indicate a period before or after the period in which these objects were found. And lastly, voters for some law are likely to have similar objectives, whereas voters against some law might do so for opposite reasons. For a thorough discussion of MCA of this type of data, we refer to [Hamrouni, Benzécri, 1976], [Heiser, 1981] and [Meulman, 1982].

We will now try to draw some conclusions from the properties of the different options for the treatment of missing data, by relating these properties to the different types of missing data. Our conclusions are summarized in table 3, having types of missing data in the columns, and the options for treating them in the rows. These conclusions will be helpful if we want to make a justified choice for any of the different options, given that we deal with a specific type of missing data. We will discuss table 3 columnwise.

#### 12.4.1 Missing not categorizable

If an object having a missing value is not classifiable in the other categories, this object defines a new class. There are two possibilities: either we are interested in the relation of being missing with the other categories, or we are not interested in this relation. We assume here that in first instance we are interested in this relation. For the case that we are not interested in this relation we refer to section 12.4.5 where we discuss *created missing data*.

**Table 3:** Types of missing data (columns) with options for their treatment (rows). The entries of the table indicate whether a combination can reasonably be defended or not. Options that seems preferable are in bold. Abbreviations are MNC for 'missing not categorizable', MNAR for 'missing not at random', MCAR for 'missing completely at random', MAR for 'missing at random' and CM for 'created missing'. For details, see text.

	MNC	MNAR	MCAR	MAR	CM
Missing passive (modified margin)	no	no	yes	yes	yes
Missing single	yes	yes	no	no	no
Missing multiple	yes	yes	yes	yes	no
Missing insertion	no	yes	yes	yes	no
Missing fuzzy average	no	no	yes	yes	yes
Missing fuzzy subgroup	no	no	yes	yes	no

Given that we are interested in being missing, we should extract information from it, and hence we can only use *missing single* or *missing multiple* (see section 12.3.9). In the other options the state of being missing does not play an active role in the analysis. *Missing single* is to be preferred if we assume that those objects missing have something in common, for example, some persons omit a question about children because it is irrelevant for them since they have no children. *Missing multiple* is to be preferred if it is assumed that the objects do not have something in common, for example, they omit a question for different reasons. However, this option is likely to produce outliers.

#### 12.4.2 Missing not at random

If we deal with data that are missing not at random, most often we do also best to extract information from the fact that someone is missing. In this way we can study, firstly, in which way *being missing* is related to the categories of the variable for which the objects are missing, and, secondly, in which way *being missing* is related to the categories of the other variables. As for *missing not categorizable* we conclude therefore that we should use *missing simple* or *missing multiple*, depending on the fact

*variance* do not hold anymore, however, due to the fact that the indicator matrix contains proportions.

Notice that by filling in the average it is not tried to extract information from the fact that someone is missing. On the contrary, a solution is sought in which, in terms of the total inertia, the influence of the missing entries is completely eliminated. It resembles the approach used in the context of missing quantitative data, where sometimes means are imputed for the missing values (compare [Little, Rubin, 1987]).

### 12.3.7 Missing fuzzy subgroup

Missing fuzzy subgroup is a natural extension of *missing fuzzy average*: now we fill in proportions that are found for some subgroup, for example, when an older male has a missing value on some category, we calculate the proportions on this variable for the non-missing older males, and fill these in for the missing older male. For the missing entries in table 2a this implies the following: for row 3 we fill in the distribution of objects having scored  $q$   $w$ : this distribution is  $1/2 \ 1/2 \ 0$ . For row 4 we fill in the distribution having scored  $v$ : this is  $2/3 \ 1/3 \ 0 \ 2/3 \ 0 \ 1/3$  (see table 2b). Though not precisely identical, this approach is suggested in [Greenacre, 1984], p. 155. In this vein it is also possible to use more sophisticated methods to estimate the missing data (compare [Little, Rubin, 1987], chapters 9 and 11). It is tried to find a best guess of the score of an object, given its scores on the other variables. So, again, it is not tried to extract information from the fact that someone has a missing value on some variable.

*Chi-squared distances* can again be simplified since  $x_{r++} = x_{r++} = m$ . The *transition formulas* apply, however, compared to missing fuzzy average, an individual is not anymore in the average of his non-missing variables only, since we have made a best guess for his missing variable. The *total inertia* is smaller than  $(k/m) - 1$  (see [Greenacre, 1984]; [van Rijckevorsel, 1987]) but larger than the total inertia for *missing fuzzy average*. In the *quantified data matrix Q* the cells that correspond with missing values have a score that is in the average of the scores that the objects in its subgroup have. As in *missing fuzzy average*, MCA is not optimal in terms of *principal component analysis, generalized canonical analysis or analysis of variance*.

### 12.3.8 Concluding remarks

[Nishisato, 1980] distinguishes approaches that extract information from the fact that some object has a missing score, and approaches that do not do this. Obviously, *missing single* and *missing multiple* are options that extract information from being missing by defining *being missing* explicitly as alternatives. Approaches that do not try to extract information from missing values are the available case methods *missing passive, missing passive modified margin*, the fuzzy options *missing fuzzy average, missing fuzzy subgroup*, and *missing insertion*. In *missing passive* an object score is only based on the non-missing scores, and in this sense no information is extracted from the fact that it has scores missing. The same holds for *missing passive modified margin* that can be considered as an option that tries to remedy some MCA properties that are lost in

only on the non-missing information, so in this case we also do not extract information from the fact someone is missing. *Missing insertion* does not use the fact that someone is missing for the determination of the category into which one is inserted. So we conclude that in this respect the main difference is between *missing single* and *missing multiple* on the one hand, and the other options. We have the impression that if the number of missing values is small, the methods that do not extract information from being missing will give very similar results. First of all, the procedures *missing passive* and *missing passive modified margin* are almost identical. Compared with the fuzzy procedures and the inserting procedures, the indicator matrices only differ somewhat for the objects having missing values.

We have only discussed the approaches to missing data that are most distinct. We have omitted some approaches from our study. For example, a set of combinations is found in [Greenacre, 1984], who combines in various ways *missing single* together with the two fuzzy options. We have chosen to discuss these options separately in their *pure form*. In [Bastin *et al.*, 1980] it is suggested to use reconstitution of order zero (see *missing fuzzy average*) to end up with an appropriate choice of zeros and ones for the missing values, but this is not worked out. A last option, discussed in [Greenacre, 1984] and [Bénali, 1985] is to create one extra column, indicating the number of missing values each object has. This has the objective to have again an indicator matrix with constant row margins. Compared to missing single, this comes to the same, geometrically, as merging all missing categories into one point, their centroid ([Greenacre, 1984]).

## 12.4. Types of missing data

Here we will discuss distinct types of missing data that can appear in object by variable data (compare [Little, Rubin, 1987]). A first distinction is between data values that are really missing, and data values that are not really missing. We speak of data values that are really missing if in reality an object falls into one of the categories but we do not know into which category. As an example of data values that are not really missing, we might think of attitude items to which a person doesn't know the answer. In this case such a person cannot be classified in any of the prescribed categories. In the sequel we will refer to this type of missing data as *missing not categorizable*, abbreviated to MNC.

For data values that are really missing, [Little, Rubin, 1987] distinguish different classes of missing values, namely missing completely at random (MCAR), *missing at random* (MAR), and *missing not at random* (MNAR). Consider a categorical variable *Salary* with one or more missing values, and the categorical variable *Education* without missing. If the mechanism to be missing on *Salary* does not depend on *Salary* nor on *Education*, we speak of *missing completely at random*. In this case the observed contingency table *Salary* × *Education* should be approximately equal to the not observed (due to the missing values on *Salary*) contingency table. If the mechanism to be missing on *Salary* does depend on *Education* but not on *Salary*, we speak of "missing at random". This implies that the conditional distribution of *Salary* given *Education* for

it receives multiple categories. This option is discussed in [Gifi, 1981], and studied in detail by [Meulman, 1982].

As in missing single, missing multiple also provides us with a complete indicator matrix, and therefore all equations in section 2 can be applied in a straightforward way. Similar remarks as for *missing single* can be made for the *chi-squared distances*. Here, when objects have one or more missing values, they will be placed very far from the other objects in full-dimensional space. A striking point is that the *total inertia* increases considerably when the number of missing categories is large. Therefore this option is likely to produce outliers, given the tendency of MCA to place categories with small marginal frequencies in the periphery of the solution. The *transition formulas* show that, using  $(R, C^*)$ , the missing categories receive the same scores as the objects for which they are missing. For more details we refer to [Meulman, 1982].

We conclude that this option is easy to interpret, like *missing single*. Conceptually it is attractive in the case that it is assumed that the missing objects are missing for different reasons. However, this option is likely to produce outliers, especially if there are objects that have more than one missing value.

### 12.3.5 Missing insertion

In *missing insertion* objects having missing values are inserted in some way into one of the original categories. [Nishisato, 1980] (see also [Nishisato, Levine, 1975]) discusses two possible options to insert, namely "insertion of most consistent responses" and "insertion of least consistent responses". In the former approach objects with missing values are assigned to categories in such a way that the first eigenvalue is *maximized*, in the latter approach objects are assigned to categories in such a way that the first eigenvalue is *minimized* (for more details, see [Nishisato, 1980]). Van Buuren (Van Buuren, 1988) applies the k-means algorithm to assign an object with missing values to categories in such a way that the total inertia is maximized. In these options we deal with complete indicator matrices, and hence the properties of MCA all hold. They require a relatively large computational effort. He discusses the use of the above techniques for the determination of the range of the first eigenvalue, and the loss of information due to missing data. His aim is to indicate empirically when an analysis of the data should be given up, which is the case when the gap between the best and the worst configurations generated is too large to ignore (Nishisato, pers. comm.).

In the program SPADN ([Lebart et al., 1987]) a procedure is used that is closely related to inserting: there objects in categories with very low marginal frequencies are inserted at random into the other categories. Inserting at random seems to give good results if the main interest is in the configuration of category points.

### 12.3.6 Missing fuzzy average

In this option we fill in proportions for the missing categories. Much attention has been given to the following approach: when variable  $j$  has  $k_j$  prespecified categories, and some object is missing on this variable, it will receive a value  $1/k_j$  for each category. So its value is distributed uniformly over the categories. We call this *fuzzy* because fuzzy coding is the name given to indicator matrices with values between zero and one (see [Greenacre, 1984]; [van Rijkevorsel, 1987]). This option is discussed in [Benzécri et al., 1973], p. 310, [Hamrouni, Benzécri, 1976], and properties of this option are studied in [Cazes, 1977] (republished in [Bastin et al., 1980]), and [Greenacre, 1984]. [Benzécri et al., 1973] already warned for the inherent danger in this approach that when some category has a very low marginal frequency, an object with missing information receives a relatively large score on this category using this approach. This objection seems so serious to us, that we will not consider this possibility any further, but instead go to its natural generalization, that we coin *missing fuzzy average*.

In *missing fuzzy average* we fill in the marginal proportions for the cells corresponding with a missing value. So for the first variable in table 2a this is  $4/8 \ 2/8 \ 2/8$ , since 4 out of 8 non-missing objects fall into  $a$ , and so on; for the second variable the proportions are  $4/9 \ 3/9 \ 2/9$  (see table 2b). We call this *average*, because the average distribution for the non-missing objects is filled in into the indicator matrix. This approach was suggested by [Benzécri et al., 1973], p. 327, and [Hamrouni, Benzécri, 1976], and can be found in modified form in [Greenacre, 1984]. These proportions can also be found using a procedure for missing data in ordinary contingency tables, namely *reconstitution of order zero* (see [Tallur, 1973]; [Mutumbo, 1973]; [Nora, 1975]; [Greenacre, 1984]; [de Leeuw, van der Heijden, 1988]). When we apply reconstitution of order zero to the cells of an indicator matrix that correspond to the missing elements in the matrix of objects by variables, this procedure will iteratively find values that are independent given the new margins. These independent values are the average proportions. For *reconstitution of order zero* it is known that the cells for which independent values are fitted do not contribute to the inertia (compare [de Leeuw, van der Heijden, 1988]). We will now discuss some other properties.

Due to the fact that the row margins  $x_{j+} = m$ , some of the MCA properties are still intact. *Chi-squared distance* formulas apply in a straightforward way. The total inertia will be lower than  $(k/m) - 1$ : the cells in which we find the marginal proportions do not contribute to the total inertia since  $x_{ij} = x_{i+}x_{+j} / x_{++}$  (compare (3a)). In this sense *missing fuzzy average* provides us with the *lowest possible total inertia* when we try to find an indicator matrix starting from an ordinary object by variable matrix as in table 2a. As far as the *transition formulas* are concerned, they can be applied straightforwardly: using  $(R^*, C)$ , an object is in the average of the categories that are non-missing; however, using  $(R, C^*)$ , a category is not in the average of the non-missing objects only, since the missing objects also contribute somewhat. In the *quantified data matrix Q* we find for the missing values the score 0, since  $\sum_j x_{ij}c_{j\alpha} = 0$  for these cells. This clarifies

what we do when we fill in average proportions into the indicator matrix: in terms of the quantified data matrix it comes down to filling in the average value (namely, zero)

and both objects fall into the category  $l$  of variable  $j$ , the distance  $\delta^2(i, i')$  increases. The total inertia is not simply  $(k/m) - 1$  anymore, but larger. This follows from (3b): since in missing passive  $x_{i, l} \leq m$ , it follows that  $1/x_{i, l} \geq 1/m$ , and hence the total inertia becomes larger. The general equations (4) and (5a) cannot be simplified since  $D_r$  is not proportional to  $I$  anymore. However, the transition formulas define MCA still in a very simple way: using normalization  $(R, C^*)$ , each category is still in the centroid of the objects that fall into it, and using normalization  $(R^*, C)$ , each object (also an object with missing values) is in the centroid of the categories it falls into. A drawback of this approach is that, since  $D_r$  is not proportional to the identity matrix anymore, the interpretations of MCA in terms of principal component analysis, generalized canonical analysis and analysis of variance do not hold anymore (see [Meulman, 1982], for a proof).

This approach has the advantage of being very simple, and the transition formulas can still be interpreted in the same way. However, many other important MCA properties are lost. Missing passive resembles so-called "available case" methods (compare [Little, Rubin, 1987], ch. 3) that are sometimes used for data analysis with missing values: the resemblance is that only the non-missing values play an active role in the analysis.

### 12.3.2 Missing passive modified margin

This approach, proposed by Escoufier ([Escoufier, 1981], [Escoufier, 1987]), is concerned with the property of the indicator matrix for missing passive that the row margins  $x_{i, \dots}$  are not all equal to  $m$ , the number of variables. The option missing passive modified margin solves this by artificially using constant row margins  $1/n$  in the MCA calculations. So, compared to (4), we can write this approach in terms of generalized CA decomposition (see [Escoufier, 1983]; [van der Heijden, de Leeuw, 1985]).

$$X / x_{i, \dots} = S_r I S_c + S_r R A C S_c \quad (6)$$

where  $X$  is the same matrix as in missing passive,  $S_c$  is a diagonal matrix with the elements  $x_{i, \dots} / x_{i, \dots}$ , but  $S_r$  is a diagonal matrix with values  $1/n$ . Notice that the sum of the elements in  $X / x_{i, \dots} = 1$  whereas it is  $nm / x_{i, \dots}$  for  $S_r I S_c$ . The objective of fixing the row margins artificially is to obtain a solution in which many of the MCA properties still hold. So missing passive modified margin is meant to remedy the weak points of missing passive.

Like in MCA without missing data, but unlike missing passive, the metric for the category points is the identity metric. Chi-squared distances between the rows have the property that, when  $x_{i, j l} = x_{i', j l} = 1$ , this does not increase  $\delta^2(i, i')$ , since artificially  $x_{i, \dots} = x_{i', \dots}$ . The total inertia is equal to (1d) (up to a proportionality factor  $nm / x_{i, \dots}$ ), to remedy the fact that the total number of observed values  $x_{i, \dots}$  is not  $nm$ ). Transition formula (4b) holds as usual, i.e. using normalization  $(R, C^*)$  the categories are still in the averages of the objects that used them; however, (4a) has a slightly different form, since the row margins of  $X / x_{i, \dots}$  are not equal to the row margins of  $S_r I S_c$  (compare [van der Heijden, de Leeuw, 1985]). Another nice property of missing passive modified

it falls beyond the scope of this paper to discuss this option in more detail, we refer instead to [Escoufier, 1987].

We conclude that, compared to missing passive, missing passive modified margin has many nice MCA properties that are lost in missing passive. As for missing passive only the non-missing scores play an active role in the analysis, and as such this approach resembles available case methods also. The approach is computationally more difficult to use since a special program for generalized CA is needed.

### 12.3.3 Missing single

In the option missing single for each variable having missing values a single extra category is created for the missing objects (see table 2b). This option for the treatment of missing data is very popular. Since it comes down to an ordinary MCA, all formulas in section 2 can be used in a straightforward way. Equation (1d) shows that, by adding columns to the indicator matrix, the total inertia increases with  $1/m$  for each extra column. This shows that the missing categories take part actively in the solution (as compared to missing passive). Problems might result from this, for example, when the number of missing objects for some variable is very small, then the new categories may dominate the solution ([Nishisato, 1980]). Notice that the increase of the inertia (and hence of the influence on the solution) is relatively larger if the original number of categories is smaller (see (1d)). So if all variables have few categories, then missing single will have a greater influence on the solution than if all variables have many categories. As far as the transition formulas are concerned, we find that, as usual, categories are in the centroid of the objects that chose them (this also holds for the categories for missing); objects are now in the average of all their categories, i.e. not only the categories of the variables on which they were not missing. In the quantified data matrix  $Q$  the objects missing on variable  $j$  all have identical scores, that may very well be extreme.

We conclude that this approach is very simple to apply. It might sometimes lead to unwanted results: for example, when we are not really interested in the missing values, we might very well end up with a solution that distinguishes persons having missing values from persons not having missing values. In terms of general missing data procedures, this approach is not comparable to available case methods like missing passive, but instead to methods where optimal values are filled in into the data matrix. Notice that all objects that are missing are inserted into a single extra category. This is useful when it is assumed that these objects have something in common. It is generally suggested to use missing single when the mechanism behind the missing data is not random, i.e. has a specific meaning (see, a.o., [Hamrouni, Benzécri, 1976]; [Bastin, 1980], p. 310; [Meulman, 1982]; [Greenacre, 1984]; [Bénali, 1985]; [Bénali, Escoufier, 1987]), or when one is not sure about the mechanism behind the missing data ([Nishisato, 1980]).

### 12.3.4 Missing multiple

In this approach, for each missing value in the original data matrix a separate extra

scores) only. The relation between MCA and generalized canonical analysis runs along the same line, but now each of the columns of  $Q$  is viewed as an optimal linear combination of the columns of their corresponding part of the indicator matrix. The relation between MCA and analysis of variance is that it is tried to make the values in each row of  $Q$  as similar as possible while at the same time the averages of the rows (which correspond to object scores) are made as dissimilar as possible. This is done under the restriction that each column of  $Q$  has average 0 and variance 1. It comes to the same as maximizing the average of the ratio of the between variance and the total variance for each variable, hence the relation with analysis of variance. For more details on these approaches we refer to [Tenenhaus, Young, 1985] and the references given there. Here we only emphasize the importance of the quantified matrix  $Q$  in many interpretations of MCA, since it makes some of the missing data approaches more easily understandable.

### 12.3. Missing data procedures in multiple correspondence analysis

In section 2 we discussed MCA without missing data. We approached MCA as a CA of an indicator matrix. Now we can introduce most of the missing data procedures described so far in the literature by simply showing how the presence of missing value in the original data leads to the adjustment of the indicator matrix. Consider again our small example in table 1a, in which we have inserted three missing values into row 3 and 4, see table 2a. We proceed by discussing the missing data procedures separately.

Table 1: An object by variable matrix with its indicator matrix

Table 1a: An object by variable matrix

	1	2	3
1	a	p	v
2	b	q	w
3	b	q	w
4	a	r	v
5	c	r	w
6	b	p	v
7	a	p	w
8	c	p	w

Table 1b: The corresponding indicator matrix

	1			2			3		
	a	b	c	p	q	r	v	w	
1	1	0	0	1	0	0	1	0	
2	0	1	0	0	1	0	0	1	
3	0	1	0	0	1	0	0	1	
4	1	0	0	0	0	1	1	0	
5	0	0	1	0	0	1	0	1	
6	0	1	0	1	0	0	1	0	
7	1	0	0	1	0	0	0	1	
8	0	0	1	1	0	0	0	1	

Table 2: Row 3 and 4 of table 1a with missing values, and the corresponding indicator matrices

Table 2a: Row 3 and 4 of table 1a with missing entries on variable 1 and 2	?	q	w
	?	?	v

Table 2b: Row 3 and 4 of indicator matrices, illustrating distinct options to deal with missing entries.

	row	a	b	c	?	?	p	q	r	?	v	w
Missing passive	3	0	0	0			0	1	0		0	1
	4	0	0	0			0	0	0		1	0
Missing single	3	0	0	0	1		0	1	0	0	0	1
	4	0	0	0	1		0	0	0	1	1	0
Missing multiple	3	0	0	0	1	0	0	1	0	0	0	1
	4	0	0	0	0	1	0	0	0	1	1	0
Missing fuzzy average	3	1/2	1/4	1/4			0	1	0		0	1
	4	1/2	1/4	1/4			4/9	3/9	2/9		1	0
Missing fuzzy subgroup	3	1/2	1/2	0			0	1	0		0	1
	4	2/3	1/3	0			2/3	0	1/3		1	0

#### 12.3.1 Missing passive

This is a very popular approach to missing data. The idea is simply that  $x_{ij} = 1$  if object  $i$  falls into category  $l$  of variable  $j$ , and  $x_{ij} = 0$  if not. This implies that if the information for object  $i$  on variable  $j$  is missing,  $x_{ij} = 0$  for each category  $l$  of variable  $j$  (see table 2b). As a second step the centroid of objects that are missing on variable  $j$  can be calculated passively (i.e. it does not influence the other analysis results) as a point for missing on variable  $j$ , hence the name *missing passive* [Gifi, 1981]. This approach is suggested in [Benzécri et al., 1973], p. 327), [Hamrouni, Benzécri, 1976], [Nishisato, 1980] and [Gifi, 1981] and the properties of this approach are studied in detail by [Meulman, 1982]. It is the default option for missing data in the program HOMALS [van de Geer, 1985].

Notice that the row margins are not all equal to  $m$ , the number of variables (see table 2b). This affects many of the MCA properties. First of all the metric for the category

frequency. If object  $i$  falls into both  $jI$  and  $j'I'$ , then the increase in  $\delta^2(jI, j'I')$  is  $n(1/x_{+jI} - 1/x_{+j'I'})^2$  which will be larger the more the marginal frequencies of  $jI$  and  $j'I'$  differ. Another way to show the influence of marginal frequencies on the distance between two columns is the following: if we work out (2b), using the fact that  $x_{ijI}^2 = x_{ijI}$  and  $x_{ij'I'}^2 = x_{ij'I'}$ , we find

$$\delta^2(jI, j'I') = \frac{n \left( 1 - \sum_i x_{ijI} x_{ij'I'} \right)}{x_{+jI}} + \frac{n \left( 1 - \sum_i x_{ijI} x_{ij'I'} \right)}{x_{+j'I'}} \quad (2c)$$

which shows that the distance between  $jI$  and  $j'I'$  is proportional to the sum of the proportion of objects having  $jI$  but not  $j'I'$  and the proportion of objects having  $j'I'$  but not  $jI$ . It also shows that the distance between two categories  $I$  and  $I'$  of the same variable  $j$  is proportional to  $1/x_{+jI} + 1/x_{+j'I'}$ .

When we use (2a) to study the distance to the average column profile  $O$ , we find

$$\delta^2(jI, O) = n \sum_i \left( \frac{x_{ijI}}{x_{+jI}} - \frac{1}{n} \right)^2 \quad (2d)$$

This shows that, when object  $i$  does not fall into  $jI$ , then  $x_{ijI} = 0$ , and  $\delta^2(jI, O)$  increases with  $1/n$ . When object  $i$  falls into  $jI$ ,  $x_{ijI} = 1$ , then  $\delta^2(jI, O)$  increases  $(n/x_{+jI}^2) - (2/x_{+jI}) + (1/n)$ . The total distance  $\delta^2(jI, O)$  can easily be calculated as the sum of  $(n - x_{+jI})/n$  (i.e.  $n - x_{+jI}$  objects do not fall into  $jI$ ) and  $x_{+jI}[(n/x_{+jI}^2) - (2/x_{+jI}) + (1/n)]$  (since there are  $x_{+jI}$  objects falling into  $jI$ ), and hence we find  $\delta^2(jI, O) = n/x_{+jI} - 1$ . This shows that the distance of  $jI$  to the average column profile is larger for a category with a smaller marginal frequency.

When we take the sum over all  $jI$ , weighted with the masses  $x_{+jI}/x_{+++}$ , we find as total inertia (1d), showing that the total inertia of the configuration of row points is identical to the total inertia of the configuration of the column points.

We now give another way to derive the total inertia, that will be useful when we discuss the total inertia for different missing value approaches. This way is

$$\text{Total inertia} = \sum_{i, jI} \left( \frac{x_{ijI} - \frac{x_{+jI} x_{+i}}{x_{+++}}}{x_{+++} - \frac{x_{+jI}^2}{x_{+++}}} \right)^2 \quad (3a)$$

and by working out the term in the numerator, using the fact that  $x_{ijI}^2 = x_{ijI}$ , we find

$$\text{Total inertia} = \sum_{i, jI} \frac{x_{ijI}}{x_{+++} x_{+jI}} - 1 \quad (3b)$$

and since  $x_{i..} = m$ , we can reduce (3b) further to (1d).

full dimensional space as possible. One generalized singular value decomposition (SVD) can be used to project the full dimensional space for the rows and the full dimensional space for the columns onto lower dimensional subspaces (see, for example, [Greenacre, 1984]). This corresponds to the following decomposition of  $X$ :

$$X / x_{+++} = D_r (I + R \Lambda C) D_c \quad (4)$$

where  $D_r$  is diagonal with margins  $x_{i..} / x_{+++}$ ,  $D_c$  is diagonal with margins  $x_{+jI} / x_{+++}$ ,  $R$  is a matrix with object scores  $r_{i\alpha}$  for row  $i$  on dimension  $\alpha$ , and  $C$  is a matrix with category scores  $c_{jI\alpha}$  for category  $I$  of variable  $j$  on dimension  $\alpha$ , and  $\Lambda$  is a diagonal matrix with singular values  $\lambda_\alpha$ .  $R$  is normalized so that  $R'D_r R = I$  and  $C$  is normalized similarly as  $C'D_c C = I$ . Since  $D_r = I/n$ , and  $x_{+++} = nm$ , (4) can be simplified, and the normalization of  $R$  is  $R'R = nI$  and  $I'R = 0$ .

The row scores  $R$  and column scores  $C$  are related by the transition formulas.

$$R^* = R \Lambda = D_r^{-1} X C \quad (5a)$$

$$C^* = C \Lambda = D_c^{-1} X'R \quad (5b)$$

These properties are the key concepts in *reciprocal averaging*, one of the approaches to MCA. Since  $D_r = I/n$ , (5a) can be simplified. Equation (5a) shows that object scores can be derived as averages of the scores for the categories that they use, and (5b) shows that category scores can be derived as averages of the scores for the objects that use these categories. Distances between the rows are equal to chi-squared distances when  $R^*$  is used as coordinates for the rows, and distances between the columns are equal to chi-squared distances when  $C^*$  are used as coordinates for the columns. Often the pair  $(R, C^*)$  is used for a joint representation in which the category points are in the weighted average of the objects falling into them. Notice that  $R^* D_r R^* = \Lambda^2 = C^* D_c C^*$ , where trace  $\Lambda^2$  equals the total inertia. This shows that the total inertia is spread out over the row points and dimensions, and the column points and dimensions.

MCA is related to many techniques for the analysis of quantitative data, such as analysis of variance, principal components analysis and generalized canonical analysis. We will not discuss this in detail here, but refer instead to [Tenenhaus, Young, 1985]. Here we will only discuss this relation verbally. Central in the relation between MCA and these techniques is the notion of the *quantified data matrix* that we will denote as  $Q$ . Starting from a non-quantified data matrix  $P$  (see table 1a), we can use the scores found for the categories in MCA to construct a quantified data matrix  $Q$  of the same order as  $P$ . In principle such a quantified data matrix can be constructed for each MCA dimension, but here we only describe this for the first dimension. This matrix  $Q$  has values  $q_{ij} = \sum_l x_{ijl} c_{jI}$ , i.e. we replace the labels in  $P$  by the quantification of these labels on dimension 1.

The relation between MCA and *principal components analysis* is that the matrix  $Q$  has the property that the average of the squared correlation between each of the columns of  $Q$  and the object scores on dimension 1 is maximized and equal to the first eigenvalue

In a typology of missing data, Little and Rubin ([Little, Rubin, 1987]) distinguish in the first place data that are really missing from data that are not really missing. An example of the latter class is that objects cannot be classified into the original categories, for example, when a question is irrelevant for some persons. As an extra type of missing data we define missing data that are created on purpose, for example, in order to eliminate the influence of dominating categories. We will try to come to conclusions that in some specific missing data situation some specific missing data procedure seems better suited than another. We start with a description of MCA for non-missing data.

12.2. Multiple correspondence analysis

There are many ways to introduce MCA. We will introduce MCA here shortly as simple correspondence analysis (CA) of an indicator matrix (see, for example, [Greenacre, 1984], ch. 5). Many other approaches to MCA are possible, see, for example, [Nishisato, 1980], [Gifi, 1981], [Lebart, Morineau, Warwick, 1984]. For a summary we refer to [Tenenhaus, Young, 1985]. Introductions to simple CA are also provided by the above authors. We will emphasize CA properties that result from the fact that an indicator matrix is analyzed. We only discuss properties needed for our discussion of the various approaches to missing values in MCA: this paper is not meant as an introduction to MCA. For more details we refer to the authors mentioned above.

An indicator matrix is defined in the following way. Consider a data matrix  $P$  with qualitative measures of  $n$  objects indexed by  $i$ , on  $m$  variables indexed by  $j$ . See table 1a. Each variable  $j$  has  $k_j$  categories, indexed by  $l$ , and the total number of categories is  $k = \sum_j k_j$ . We transform the matrix  $P$  into a binary indicator matrix  $X$  of order  $n \times k$ ,

where each category  $l$  of variable  $j$  has its own column. In this matrix  $x_{ijl} = 1$  if object  $i$  falls into category  $l$  of variable  $j$ , and  $x_{ijl} = 0$  else. See table 1b. Notice that  $x_{i++} = 1$ , hence  $x_{i++} = m$  and  $x_{+++} = nm$ , and  $x_{+jl}$  is the marginal frequency for category  $l$  of variable  $j$ . We will now discuss CA of this matrix, emphasizing geometrical properties (see also [Meulman, 1982], [Meulman, 1986], [Greenacre, 1984] [Carroll, Green, Schaffer, 1986]).

CA gives two geometrical representations of a matrix, one for the rows and one for the columns. First consider the representation for the rows. Each row is represented by a point in  $k$ -dimensional euclidean space, and has coordinates equal to  $x_{ijl} / x_{i++}$ . The metric in this space is defined by  $D_c^{-1}$ , where  $D_c$  is diagonal with elements the column margins  $x_{+jl} / x_{+++} = x_{+jl} / nm$ . The vector of coordinates  $x_{ijl} / x_{i++}$  is defined as the profile for row  $i$ . With each point  $i$  a mass  $x_{i++} / x_{+++} = 1/n$  is associated. The distance between the profiles for row  $i$  and  $i'$  is called the chi-squared distance, and is defined as

$$\delta^2(i, i') = \sum_{jl} \frac{x_{i++}}{x_{+jl}} \left( \frac{x_{ijl}}{x_{i++}} - \frac{x_{i'jl}}{x_{i'++}} \right)^2 \quad (1a)$$

$$\delta^2(i, i') = \frac{1}{n} \sum_{jl} \frac{1}{x_{+jl}} (x_{ijl} - x_{i'jl})^2 \quad (1b)$$

Since  $x_{i++} = x_{i'++} = m$  and  $x_{+++} = nm$ , this simplifies to

This shows that combination  $jl$  does not increase  $\delta^2(i, i')$  when objects  $i$  and  $i'$  either both fall or both do not fall into the same category  $l$  of variable  $j$ : in both cases  $x_{ijl} - x_{i'jl} = 0$ . When only one of the two objects falls into  $jl$ , the increase of  $\delta^2(i, i')$  for  $jl$  is proportional to  $(1/x_{+jl})$ . Therefore objects  $i$  and  $i'$  will be close together when they have many categories in common; this also shows that  $\delta^2(i, i')$  will become larger when  $i$  and  $i'$  differ by falling into distinct categories with lower marginal frequencies.

We will now study the distance of object  $i$  to the average row profile  $O$ , which is the profile of the column margins. This average row profile is the weighted average of all row profiles, using masses  $x_{i++} / x_{+++}$  as weights, and has values  $x_{+jl} / x_{+++}$ . So using (1a)  $\delta^2(i, O)$  can be found to be

$$\delta^2(i, O) = \frac{1}{m} \sum_{jl} \frac{1}{x_{+jl}} \left( x_{ijl} - \frac{x_{+jl}}{n} \right)^2 \quad (1c)$$

This shows that when object  $i$  does not fall into  $jl$ , then  $x_{ijl} = 0$ , and  $\delta^2(i, O)$  increases with  $x_{+jl} / nm$  for column  $jl$ ; when object  $i$  falls into  $jl$ , then  $x_{ijl} = 1$ , and  $\delta^2(i, O)$  increases with  $\frac{1}{m} - \frac{2}{nm} + \frac{x_{+jl}}{nm}$  for column  $jl$ . We conclude that, when object  $i$  falls into categories with lower marginal frequencies  $x_{+jl}$ ,  $\delta^2(i, O)$  increases more.

We can also derive the weighted sum of squared distances of all objects to the origin. Since there are  $x_{+jl}$  objects falling into  $jl$ , and  $(n - x_{+jl})$  objects not falling into  $jl$ , it follows that for  $jl$  the sum of all squared distances is  $\frac{n}{m} - \frac{x_{+jl}}{m}$ . Weighting this with the masses of the objects, and summing over all possible  $jl$  (which is allowed due to Pythagoras theorem), we find the so-called total inertia:

$$\text{total inertia} = (k/m) - 1 \quad (1d)$$

Now we discuss the chi-squared distances between the columns of  $X$ . Columns are represented as points in  $n$ -dimensional euclidean space, with coordinates for column  $jl$   $x_{ijl} / x_{+jl}$ . The metric of the space is defined by the  $D_r^{-1}$ , with  $D_r$  diagonal having elements  $x_{i++} / x_{+++} = 1/n$ , so we deal with a metric that is proportional to the identity metric. Each column  $jl$  has a mass  $x_{+jl} / x_{+++} = x_{+jl} / nm$  associated with it. The profile for category  $l$  of variable  $j$  has values  $x_{ijl} / x_{+jl}$ , and we find as the chi-squared distance between  $jl$  and  $j'l'$ :

$$\delta^2(jl, j'l') = \sum_i \frac{x_{+++}}{x_{+jl}} \left( \frac{x_{ijl}}{x_{+jl}} - \frac{x_{ij'l'}}{x_{+j'l'}} \right)^2 \quad (2a)$$

Since  $x_{i++} = m$  and  $x_{+++} = nm$ , (2a) can be simplified to

$$\delta^2(jl, j'l') = n \sum_i \left( \frac{x_{ijl}}{x_{+jl}} - \frac{x_{ij'l'}}{x_{+j'l'}} \right)^2 \quad (2b)$$

If object  $i$  does not fall into either of  $x_{ijl}$  or  $x_{ij'l'}$ , then the increase of  $\delta^2(jl, j'l')$  for

## Multiple Correspondence Analysis with missing data

P. G. M. Van Der Heijden, B. Escofier<sup>1</sup>

### Summary

For multiple correspondence analysis many procedures are known for the handling of missing data. In this paper some properties of the most important procedures are discussed. As a second step distinct types of missing data are distinguished. Apart from a typology presented by [Little, Rubin, 1987], we give attention to the fact that in multiple correspondence analysis regularly missing data are created on purpose, in order to eliminate the influence of specific categories. Given the properties of the different procedures for handling missing data, it is discussed which types of missing data can best be handled with which procedures for missing data.

### 12.1. Introduction

Multiple correspondence analysis (MCA) is a tool for data description that received much attention in the last decade. Many reasons for this can be given, one of it being that it is a very flexible tool that can be applied to many types of data (see for examples [Nishisato, 1980] [Gifi, 1981] [Greenacre, 1984]). In this paper we discuss MCA of categorical data that are coded into an object by variable matrix. Such data are very common in the social sciences, for example as a result of using questionnaires. We concentrate on the case that the object by variable matrix contains missing values.

For object by variable data many different missing data procedures exist for MCA. Major references in which these procedures are discussed are [Hamrouni, Benzécri, 1976], [Nishisato, 1980], [Meulman, 1982], [Greenacre, 1984] and [Bénali, 1985]. In our paper we give an overview of the most important missing data procedures. In our opinion this overview is much needed, since the above authors often do not seem to be aware of each others work. We will compare these procedures systematically by studying their properties.