# Missing Data in Hierarchical Classification – a study*

Ana Lorga da Silva[1], Helena Bacelar-Nicolau[2], Gilbert Saporta[3]

[1]ISEG, Universidade Tecnica de Lisboa
e-mail:aigcls@iseg.utl.pt
[2]LEAD-FPCE, Universidade de Lisboa
e-mail:hbacelar@fc.ul.pt
[3]Chaire de Statistique Appliquée
Conservatoire National des Arts et Métiers, Paris, France
e-mail:saporta@cnam.fr

**Abstract**: In this work we analyse the effect of missing data in hierarchical classification of variables according to the following factors: amount of missing data, imputation techniques, similarity coefficient, and aggregation criterion. We have used two methods of imputation, a regression method using an ordinary-least squares method and an EM algorithm. For the similarity matrices we have used the basic affinity coefficient and the Pearson's correlation coefficient. As aggregation criteria we apply average linkage, single linkage and complete linkage methods. To compare the structure of the hierarchical classifications the Spearman's coefficient between the associated ultrametrics has been used. We present here simulation experiments in two multivariate normal cases.

**Keywords**: Missing Data, Hierarchical Cluster Analysis, Affinity Coefficient, Pearson's Coefficient, Spearman's Coefficient, Ultrametric, OLS method, EM Algorithm.

## 1 Introduction

The missing data problem has been dealt in a large number of papers and books where several methods to minimise missing data effect have been developed (Rubin(1974), Rubin(1987), Little and Rubin(1987), Dempster, Laird and Rubin(1977), Orchard and Woodbury(1972), Beale and Little(1975) among others).

When one wants to classify variables, for instance in marketing analysis and social sciences, one frequently finds missing data. We are interested in analysing the effect of missing data in some particular (originally complete) hierarchical classification structures of variables, as well the results of imputation methods in those cases.

In the present work we consider hierarchical clustering models based on two similarity coefficients – basic affinity (Matusita(1955),Bacelar-Nicolau(1988)) and Pearson's correlation - and three classical aggregation criteria. We use two types of imputation methods in simulation studies with different percentage of missing data at random. The data are issued from multinormal populations (Saporta(1990)).

## 2 Hierarchical cluster analysis

In this work we are interested in the classification of variables. We use the following hierarchical aggregation criteria as defined in Anderberg(1973):

Average linkage (AL): $C(A,B)=\frac{1}{(\# A)\times(\# B)}\sum c(X_j, X_{j'}),\ X_j \in A, X_{j'} \in B$

Single linkage (SL): $C(A,B)=max\{c(X_j, X_{j'}), X_j \in A, X_{j'} \in B\}$

and

Complete linkage (CL): $C(A,B)=min\{c(X_j, X_{j'}), X_j \in A, X_{j'} \in B\}$

where *A* and *B* represent two clusters and *c* is a similarity coefficient between two variables ($X_j$ , $X_{j'}$ are $(n\times 1)$ variables ) which can be one of the two following:

The (unweighted) basic affinity coefficient $c_a=\sum_{i=1}^{n}\sqrt{\frac{x_{ij}}{x_{.j}}\frac{x_{ij'}}{x_{.j'}}}$ , where $x_{.j}=\sum_{i=1}^{n}x_{ij}$ and $x_{.j'}=\sum_{i=1}^{n}x_{ij'}$, as defined for instance in Bacelar-Nicolau(1988, 2000).

* This work has been partially supported by the Franco-Portuguese Scientific Programme VECEMH (Embassy of France and Portuguese Ministry of Science and Technology – ICCTI) and the Multivariate Data Analysis research team of CEAUL/FCUL.

The Bravais-Pearson correlation coefficient $c_p = \frac{\sum_{i=1}^{n}\left(x_{ij} - \bar{x}^{j}\right)\left(x_{ij'} - \bar{x}^{j'}\right)}{s_{x^j} s_{x^{j'}}}$.

In order to compare hierarchical classification models, we will use the Spearman's coefficient between the ultrametic matrices, based on pairs of observations with the usual correction for ties.

## 3 The missing data – MAR

The data are said that missing at random if its missingness does not depend of the values assumed on the variables having missing values, but depends on the values observed in other completely observed variables. The expression of the general notion of MAR can be then written as: $Pr\,ob\left(R \middle| X_{obs}, X_{mis}\right) = Pr\,ob\left(R \middle| X_{obs}\right)$, where $X_{obs}$ represents the observed values of $\mathbf{X}_{n\times p}$, $X_{mis}$ the missing values of $\mathbf{X}_{n\times p}$ and $R = \lfloor R_{ij} \rfloor$ is a missing data indicator, $R_{ij} = \begin{cases} 1, & \text{if } x_{ij} \text{ observed} \\ 0, & \text{if } x_{ij} \text{ missing} \end{cases}$

## 4 The imputation methods

An ordinary least square regression method (OLS) is used: $\mathrm{Y} = b_1 + b_2\mathrm{X}$ is defined as usually, $b_1, b_2$ are estimated over the observed values of the dependent variable, $X_{obs} = b_1 + b_2 X'_{obs}$ ( $X'_{obs}$ is a "sample" of **X** corresponding to the observed values of **X** – $X_{obs}$ ) and then the missing values of **X** ( $X_{mis}$ ) are imputed by the regression on $X'_{mis}$ (those are observed values corresponding to the missing dependent values of $X_{mis}$ ) under the estimated model $X_{mis} = b_1 + b_2 X'_{mis}$ .

An EM algorithm has been used as follows:

At the E step of the algorithm (at the *t*th iteration),

$$x_{ij}^{t} = \begin{cases} x_{ij} & \text{if } x_{ij} \text{ is not missing} \\ \hat{x}_{ij} & \text{if } x_{ij} \text{ is missing} \end{cases}$$

"The E step imputes the best linear predictors of the missing values, using current estimates of the parameters available so that a suitable choice can be made. It also calculates the adjustments $c_{jk}$ to the estimated covariance matrix needed to allow for imputation of missing values" Little and Rubin(1987)

At the M step

$$m^{(t+1)} = \left[ n^{-1} \sum_{i=1}^{n} x_{ij}^{t} \right],\ s^{(t+1)} = (n-1)^{-1} \sum_{i=1}^{n} \left[ \left( x_{ij}^{t} - m_j^{(t+1)} \right) \left( x_{ik}^{t} - m_k^{(t+1)} \right) + c_{jki}^{t} \right],\ j,k = 1,\dots, p$$

We consider the missing values only over a dependent variable.

## 5 Numerical Experiments

In order to study the performance of the affinity and the Pearson's correlation coefficients as measures of similarity between variables, in hierarchical classification and in presence of missing data, we use here the three classical hierarchical clustering methods AL, SL and CL: in the cases of complete data; in MAR case - 10%, 15% and 20% (over the total of the data – each 1000×5 matrix); and when the missing data are filled-in using the two imputation methods as mentioned in **4.**.

One hundred samples have been generated of each type of simulated data set, from two normal multivariate populations:

Case A: $\boldsymbol{X}_1 \sim$ **À**$\left(\mathbf{m}_1, \boldsymbol{s}_1\right)$ and Case B: $\boldsymbol{X}_2 \sim$ **À**$\left(\mathbf{m}_2, \boldsymbol{s}_2\right)$, such as $X_1$ and $X_2$ are 1000×5 matrices .

The values of the variance-covariance matrices have been chosen with the aim of obtaining specific hierarchical structures as following:

$\boldsymbol{m}_1 = [4 \quad 4 \quad 4 \quad 4 \quad 4]$, $\boldsymbol{S}_1 = \begin{bmatrix} 1 & 0.8 & 0.7 & 0.2 & 0.3 \\ 0.8 & 1 & 0.6 & 0.3 & 0.2 \\ 0.7 & 0.6 & 1 & 0.4 & 0.3 \\ 0.2 & 0.3 & 0.4 & 1 & 0.6 \\ 0.3 & 0.2 & 0.3 & 0.6 & 1 \end{bmatrix}$ associated to the dendrogram:

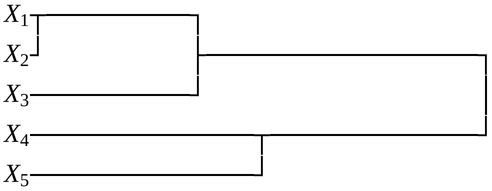

Fig. 1

and

$\boldsymbol{m}_2 = [4 \quad 4 \quad 4 \quad 4 \quad 4]$, $\boldsymbol{S}_2 = \begin{bmatrix} 1 & 0.8 & 0.6 & 0.5 & 0.2 \\ 0.8 & 1 & 0.5 & 0.6 & 0.2 \\ 0.6 & 0.5 & 1 & 0.7 & 0.2 \\ 0.5 & 0.6 & 0.7 & 1 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 1 \end{bmatrix}$ associated to the dendrogram:

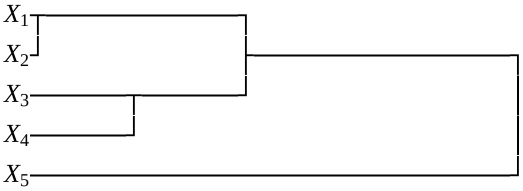

Fig.2

In order to have missing data at random MAR we have deleted (10%, 15% and 20%) values at random from variables $X_1$ and $X_2$ .

In the following we present the results of the simulations, respectively in cases A and B, by increasing order of the percentages of missing data (MD), according to the similarity coefficients, the agglomerative methods and the imputation methods. In each case we compare the ultrametrics associated to the originally complete data with the ultrametric matrices associated to the incomplete and the reconstructed data respectively.

The comparison between ultrametrics is obtained using a 5% Spearman's bilateral test (the critical value is $c'_S = 0{,}648$, see for instance Saporta (1990)).

In presence of MD, the classification is obtained by a pairwise method i.e. we have only considered for the analysis the complete rows (by eliminating the rows with MD).

In analysis of cases A and B, $c_S = 1$, $|c_S| > c'_S$ , $|c_S| < c'_S$ mean that:

$c_S = 1$, the general "structure" of the two hierarchical classifications being compared is the same, that is the two associated ultrametrics are "ordinal equivalent" (each pair of ranked trees give the same "ordinal" structure).

$|c_S| > c'_S$ the two hierarchical classification structures are not the same, but the two ultrametrics are "significantly correlated" (at 5%).

$|c_S| < c'_S$ the two hierarchical classification structures are "significantly different"

The percentages of cases $c_S = 1$, $|c_S| > c'_S$ and $|c_S| < c'_S$ are also indicated in each cell of the tables.

**Case A**

All the simulated complete data reproduced the same general hierarchical structure (see dendrogram above, Fig. 1) using both coefficients - Affinity and Pearson's correlation – and the three hierarchical methods, AL, SL and CL.

Next table describes the results obtained in presence of MD:

| | Affinity coefficient | | | Pearson’s coefficient | | |
|---|---|---|---|---|---|---|
| **MD** | **AL** | **SL** | **CL** | **AL** | **SL** | **CL** |
| **10%** | **100%** $c_S = 1$ | **99%** $c_S = 1$<br>**1%** $\lvert c_S \rvert > c'_S$ * | **100%** $c_S = 1$ | **5%** $c_S = 1$<br>**95%** $\lvert c_S \rvert < c'_S$ | **49%** $c_S = 1$<br>**51%** $\lvert c_S \rvert < c'_S$ | **100%** $\lvert c_S \rvert < c'_S$ |
| **15%** | **100%** $c_S = 1$ | **96%** $c_S = 1$<br>**4%** $\lvert c_S \rvert > c'_S$ * | **99%** $c_S = 1$<br>**1%** $\lvert c_S \rvert > c'_S$ | **5%** $c_S = 1$<br>**95%** $\lvert c_S \rvert < c'_S$ | **41%** $c_S = 1$<br>**59%** $\lvert c_S \rvert < c'_S$ | **100%** $\lvert c_S \rvert < c'_S$ |
| **20%** | **100%** $c_S = 1$ | **87%** $c_S = 1$<br>**13%** $\lvert c_S \rvert > c'_S$ * | **100%** $c_S = 1$ | **2%** $c_S = 1$<br>**97%** $\lvert c_S \rvert < c'_S$ | **21%** $c_S = 1$<br>**69%** $\lvert c_S \rvert < c'_S$ | **100%** $\lvert c_S \rvert < c'_S$ |

*with chain effect

**Table 1.**

After using <u>both</u> imputation methods, results are:

| | Affinity coefficient | | | Pearson’s coefficient | | |
|---|---|---|---|---|---|---|
| **ID**** | **AL** | **SL** | **CL** | **AL** | **SL** | **CL** |
| **10%** | **100%** $c_S = 1$ | **100%** $c_S = 1$ | **100%** $c_S = 1$ | **98%** $c_S = 1$<br>**2%** $\lvert c_S \rvert < c'_S$ | **100%** $c_S = 1$ | **98%** $c_S = 1$<br>**2%** $\lvert c_S \rvert < c'_S$ |
| **15%** | **100%** $c_S = 1$ | **100%** $c_S = 1$ | **100%** $c_S = 1$ | **94%** $c_S = 1$<br>**5%** $\lvert c_S \rvert > c'_S$<br>**1%** $\lvert c_S \rvert < c'_S$ | **96%** $c_S = 1$<br>**4%** $\lvert c_S \rvert > c'_S$ | **95%** $c_S = 1$<br>**4%** $\lvert c_S \rvert > c'_S$<br>**1%** $\lvert c_S \rvert < c'_S$ |
| **20%** | **94%** $c_S = 1$<br>**2%** $\lvert c_S \rvert > c'_S$<br>**4%** $\lvert c_S \rvert < c'_S$ | **94%** $c_S = 1$<br>**2%** $\lvert c_S \rvert > c'_S$<br>**4%** $\lvert c_S \rvert < c'_S$ | **94%** $c_S = 1$<br>**2%** $\lvert c_S \rvert > c'_S$<br>**4%** $\lvert c_S \rvert < c'_S$ | **21%** $c_S = 1$<br>**58%** $\lvert c_S \rvert > c'_S$<br>**21%** $\lvert c_S \rvert < c'_S$ | **21%** $c_S = 1$<br>**58%** $\lvert c_S \rvert > c'_S$<br>**21%** $\lvert c_S \rvert < c'_S$ | **21%** $c_S = 1$<br>**58%** $\lvert c_S \rvert > c'_S$<br>**21%** $\lvert c_S \rvert < c'_S$ |

**imputed data

**Table 2**

**Case B**

As in case A all the simulated complete data, using both coefficients and the three hierarchical classification methods produced the same general hierarchical structure (see dendrogram above, Fig. 2).

Next table describes the results obtained in presence of MD.

| | Affinity coefficient | | | Pearson’s coefficient | | |
|---|---|---|---|---|---|---|
| **MD** | **AL** | **SL** | **CL** | **AL** | **SL** | **CL** |
| **10%** | **100%** $c_S = 1$ | **100%** $c_S = 1$ | **100%** $c_S = 1$ | **100%** $c_S = 1$ | **89%** $c_S = 1$<br>**11%** $\lvert c_S \rvert > c'_S$ | **100%** $c_S = 1$ |
| **15%** | **100%** $c_S = 1$ | **100%** $c_S = 1$ | **100%** $c_S = 1$ | **100%** $c_S = 1$ | **50%** $c_S = 1$<br>**50%** $\lvert c_S \rvert > c'_S$ | **100%** $c_S = 1$ |
| **20%** | **100%** $c_S = 1$ | **100%** $c_S = 1$ | **100%** $c_S = 1$ | **100%** $c_S = 1$ | **69%** $c_S = 1$<br>**31%** $\lvert c_S \rvert > c'_S$ | **100%** $c_S = 1$ |

**Table 3**

Results after using the **imputation methods are**:

**Ordinary least squares method**

| | Affinity coefficient | | | Pearson’s coefficient | | |
|---|---|---|---|---|---|---|
| **ID** | **AL** | **SL** | **CL** | **AL** | **SL** | **CL** |
| **10%** | **100%** $c_S = 1$ | **69%** $c_S = 1$<br>**31%** $\lvert c_S \rvert > c'_S$ | **100%** $c_S = 1$ | **99%** $c_S = 1$<br>**1%** $\lvert c_S \rvert > c'_S$ | **92%** $c_S = 1$<br>**8%** $\lvert c_S \rvert > c'_S$ | **99%** $c_S = 1$<br>**1%** $\lvert c_S \rvert > c'_S$ |
| **15%** | **100%** $c_S = 1$ | **17%** $c_S = 1$<br>**83%** $\lvert c_S \rvert > c'_S$ | **100%** $c_S = 1$ | **99%** $c_S = 1$<br>**1%** $\lvert c_S \rvert > c'_S$ | **41%** $c_S = 1$<br>**39%** $\lvert c_S \rvert > c'_S$ | **99%** $c_S = 1$<br>**1%** $\lvert c_S \rvert > c'_S$ |
| **20%** | **79%** $c_S = 1$<br>**21%** $\lvert c_S \rvert > c'_S$ | **3%** $c_S = 1$<br>**97%** $\lvert c_S \rvert > c'_S$ | **99%** $c_S = 1$<br>**17%** $c_S = 1$ | **12%** $c_S = 1$<br>**88%** $\lvert c_S \rvert > c'_S$ | **3%** $c_S = 1$<br>**97%** $\lvert c_S \rvert > c'_S$ | **12%** $c_S = 1$<br>**88%** $\lvert c_S \rvert > c'_S$ |

**Table 4**

**EM method**

| | Affinity coefficient | | | Pearson's coefficient | | |
|---|---|---|---|---|---|---|
| **ID** | **AL** | **SL** | **CL** | **AL** | **SL** | **CL** |
| **10%** | **98%** $c_S = 1$<br>**2%** $\lvert c_S \rvert > c'_S$ | **81%** $c_S = 1$<br>**19%** $\lvert c_S \rvert > c'_S$ | **98%** $c_S = 1$<br>**2%** $\lvert c_S \rvert > c'_S$ | **93%** $c_S = 1$<br>**7%** $\lvert c_S \rvert > c'_S$ | **70%** $c_S = 1$<br>**30%** $\lvert c_S \rvert > c'_S$ | **93%** $c_S = 1$<br>**7%** $\lvert c_S \rvert > c'_S$ |
| **15%** | **58%** $c_S = 1$<br>**42%** $\lvert c_S \rvert > c'_S$ | **14%** $c_S = 1$<br>**86%** $\lvert c_S \rvert > c'_S$ | **95%** $c_S = 1$<br>**5%** $\lvert c_S \rvert > c'_S$ | **91%** $c_S = 1$<br>**9%** $\lvert c_S \rvert > c'_S$ | **49%** $c_S = 1$<br>**51%** $\lvert c_S \rvert > c'_S$ | **94%** $c_S = 1$<br>**6%** $\lvert c_S \rvert > c'_S$ |
| **20%** | **75%** $c_S = 1$<br>**25%** $\lvert c_S \rvert > c'_S$ | **8%** $c_S = 1$<br>**92%** $\lvert c_S \rvert > c'_S$ | **80%** $c_S = 1$<br>**20%** $\lvert c_S \rvert > c'_S$ | **31%** $c_S = 1$<br>**69%** $\lvert c_S \rvert > c'_S$ | **1%** $c_S = 1$<br>**99%** $\lvert c_S \rvert > c'_S$ | **31%** $c_S = 1$<br>**69%** $\lvert c_S \rvert > c'_S$ |

**Table 5**

## 6 Conclusions

In both studied cases A and B the affinity coefficient performs better than Pearson's correlation coefficient in presence of data missing at random.

Better results are obtained in presence of MD, than after the application of both imputation methods.

When using the imputation methods in case A both imputation methods gave the same results, and also that the affinity coefficient performs better than the Pearson's coefficient. In case B the results are different when using the two imputation methods, and the least squares method performs better for the AL and CL models.

In the "worst" situation of 20% of missing data and filled-in data, the affinity coefficient performs always better than the Pearson's coefficient.

The following developments of this work are related to other similarity coefficients and hierarchical structures, namely concerning a probabilistic classification approach, and different types of missing data and imputation methods.

## References

ANDERBERG, M. R. (1973) Cluster analysis for applications, Academic Press, New York.

BACELAR-NICOLAU(1988) Two probabilistic models for classification of variables in frequency tables in Classification and related methods of data analysis, H. H. Bock (ed.), Elsiever Sciences, Publishers B. V., North Holland, 181-186.

BACELAR-NICOLAU(2000) The Affinity Coefficient in Analysis of Symbolic Data Exploratory Methods for Extracting Statistical Information from Complex Data. H.H. Bock and E.Diday (Eds.), Springer,160-165.

BEALE, E. M. L. and LITTLE, R. J. A.(1975) Missing values in multivariate data analysis. J. R. *Statist. Soc.* B, **37**, 129-145.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B.(1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1-38.

LITTLE, R. J. A. and RUBIN, D. B.(1987) Statistical Analysis With Missing Data, John Wiley & Sons, New York.

MATUSITA,K., Decision rules, based on distance for problems of fit, two samples and estimation. Ann. Math. Stat., vol26, nº4, 631-640.

ORCHARD, T. and WOODBURY, M. A.(1972) A missing information principle: theory and applications. Proceedings 6th Berkley Symposium on Mathematical Statistic and Probability, **1**, 697-715.

RUBIN, D. B.(1974) Characterising the estimation of parameters in the estimation of parameters in incomplete-data problems. *JASA*, **69**,467-474.

RUBIN, D. B(1987) Multiple Imputation for Nonresponse in Surveys, Willey, New York

SAPORTA, G.(1990) Probabilités, Analyse des Données et Statistique, Editions Technip, Paris