Missing Data and Imputation Methods in Partition of Variables

Ana Lorga da Silva, Gilbert Saporta, Helena Bacelar-Nicolau

Missing Data and Imputation Methods in Partition of Variables*

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

[1] Universidade Lusofona de Humanidades e Tecnologias,
Department of Economics and Management, Sala I.0.5
Av. do Campo Grande, 376,
1749 - 024 Lisboa, Portugal
ana.lorga@ulusofona.pt
[2] Statistics Department, CNAM, Paris, France saporta@cnam.fr
[3] FPCE, LEAD, Lisbon University, Lisbon, Portugal hbacelar@fpce.ul.pt

Abstract We deal with the effect of missing data under a "Missing at Random Model" on classification of variables with non hierarchical methods. The partitions are compared by the Rand's index.

1 Introduction

The missing data problem in some classical hierarchical methods has been studied using the affinity coefficient (Bacelar-Nicolau, 2002) and the Bravais-Pearson correlation coefficient, e.g. in Silva, Saporta et al. (2001) and also in Silva et al. (2003), where we have been studying the missing data under a "Missing at Random Model" - MAR - as described for instance in Little and Rubin(2002), which we and other authors consider a more "realistic model" - which means that it occurs more often in the real situations - of missing data. Missing data can be found in data from marketing analysis and social sciences, among others.
So, when we do classification, we must be prepared to "interpret" the results. In most papers on missing data, they deal mainly with estimation of the parameters of the population such as mean, standard deviation, among others, or in estimating regression models (more frequent in Economics studies).

* This work has been partially supported by the Franco-Portuguese Scientific Programme "Modeles Statistiques pour le Data Mining" MSPLLDM-542-B2 (Embassy of France and Portuguese Ministry of Science and Superior Education - GRICES) co-directed by H. Bacelar-Nicolau and G. Saporta and the Multivariate Data Analysis research team of CEAUL/FCUL directed by H. Bacelar-Nicolau.

Partition methods are sometimes used as a complement of hierarchical classification methods for choosing the best level where "cut the structure"; here we analyse the performance of one of those methods when missing data are present.

2 Methodology

2.1 The Partition Method

The partition method we use in this work is composed of two algorithms: a hierarchical algorithm followed by the partitioning algorithm.
Our method[5] is derived from Vigneau and Qannari(2003) and closely related to principal component analysis. We may also say that this approach consists in clustering variables around latent components.
More precisely, the aim is to determine simultaneously k clusters of variables and k latent components so that the variables in each cluster are related to the corresponding latent component.
This method leads us to choose the adequate number groups in a partition.

The method:
Let,

- p variables measured on n objects - $x_1, x_2, ..., x_p$ (the variables are centered),
- K clusters of the p variables - $G_1, G_2, ..., G_k$ (composig a partition wich we can design by P_1)
- K latent variables - $l_1, l_2, ..., l_k$ - associated with each of the K groups.
- The criterium $S = \sqrt{n} \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} cov(x_j, l_k)$, under the constraint,

$$l_k . l'_k = 1$$

where,

$$\delta_{kj} = \begin{cases} 1 \, if \;\; x_j \in G_k \\ 0 \, if \;\; x_j \notin G_k \end{cases}$$

We optimize S rather than a criterion based on squared correlation because, for us, in many situations, the sign of the correlation coefficient makes sense, see Vigneau and Qannari(2003): " p consumers are asked to rate their acceptability of n product. A negative covariance between the scores of two consumers emphasizes their different views of the products."
The partition algorithm used in this work:

[5] No work about missing data nor imputation methods is known for this method.

i) we start with K groups obtained by a hierarchical cluster method[6]
ii) In cluster G_k $(k = 1, 2, ..., K)$, $l_k = \frac{\bar{x}_k}{\sqrt{\bar{x}_k \bar{x'}_k}}$, where, $\bar{x}_k$ is the centroid of G_k, such as, $\bar{x}_k = \frac{\sum_{j=1} x_{kj}}{p_k}$; p_k is the number of elements in G_k.
iii) "New clusters are formed by moving each variable to a new group if its covariance with the standardised centroid of this group is higher than any other standardised centroid" (Vigneau and Qannari, 2003).

2.2 The Imputation Methods

We consider the folowing cases:

a) listwise method,
b) NIPALS algorithm adapted to a regression method (as described in Silva et al. (2002) and Tenenhaus(1998)),
c) EM imputation method,
d) OLS (ordinary least squares) regression method, i.e. one estimates missing values by standard multiple regression - e.g. as described in Silva et al. (2002),
e) PLS2 regression method used as an imputation method; PLS2 stand for a particular application of PLS regression when one has to predict simultaneously q variables (PLS or PLS1 stand for $q = 1$),
f) Multiple Imputation - (MI) - a Bayesian Method based on an OLS regression (Rubin, 1987).

It is to be noted that, usually, neither the PLS2 regression method nor NIPALS algorithm is used as an imputation method.
The version of PLS2 we use here comes naturally from NIPALS algorithm which has as main feature, the possibility of allowing us to work with missing data without suppressing observations that have missing data (and even without estimating the missing data).
In the simulation studies, we shall deal with two variables having missing data.
For the MI imputation method, the results will be combined in two ways as described at section 2.4.

2.3 The missing data - MAR

The missing data is said to be MAR - Missing at Random - if it can be written as:

$$Prob(M|X_{obs}, X_{miss}) = Prob(M|X_{obs})$$

[6] This hierarchical cluster method is based on the same criterion S described for the partition see Vigneau and Qannari(2003).

,
Where,

- X_{obs} represents the observed values of $X_{n\times p}$,
- X_{miss} the missing values of $X_{n\times p}$
  and
- $M = [M_{ij}]$ is a missing data indicator,

$$M_{ij} = \begin{cases} 1, \text{ if } x_{ij} \text{ observed} \\ 0, \text{ if } x_{ij} \text{ missing} \end{cases}$$

2.4 Multiple Imputation, Correlation Matrices and Partitions

By using imputation methods m matrices $X^1, X^2, ..., X^m$ are obtained, (usually $m = 5$ is enough) , it is necessary to combine the results in order to apply the partition method.

First Method: Combination of the correlation matrices (MIave).

i) First, we obtain $X^1, X^2, ..., X^m$ to which are associated m correlation matrices $(R_k, k = 1, ...m)$,
ii) We determine the average of the m correlation matrices:

$$\overline{R} = \frac{\sum_{k=1}^{m} R_k}{m}$$

,
iii) We apply the partition method to $\overline{R}$.

Second Method:"Consensus between the Partitions" (MIcons).

With this method we try to establish a "consensus between m partitions", in order to find a representative partition. To compare the partitions we use, the Rand's index, see Saporta and Youness(2002), that gives us the proportion of agreements between the partitions.
Suppose we have two partitions P_1, P_2 of p variables, with the same number of classes k.
We will find four types of pairs:

- **a** - number of pairs belonging simultaneously to the same classes of P_1 and
  P_2
- **b** - number of pairs belonging to different classes of P_1 but to the same of P_2
- **c** - number of pairs belonging to different classes of P_2 but to the same of P_1
- **d** - number of pairs belonging to different classes of P_1 and P_2

So we have:

- A=a+d represents the total number of agreements
- D=b+c represents the total number of discordances

and $A + D = \frac{p(p-1)}{2}$.

The classical Rand's index is given then by the expression:

$$I_R = \frac{2A}{p(p-1)}.$$

But we will use the version modified by Marcotorchino[7]:

$$I_R = \frac{2\sum_{i=1}^{k}\sum_{j=1}^{k}(n_{ij})^2 - \sum_{i=1}^{k}(n_{i.})^2 - \sum_{j=1}^{k}(n_{.j})^2 + n^2}{n^2},$$

n_{ij} are the elements of the contingency table crossing the two partitions, $n_{i.}$ is the row total and $n_{.j}$ the column total.
In this paper we will deal with five variables and two partitions. In this case 0.5 is the expectation of Rand's index under the independence hypothesis.

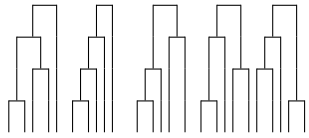
We start with each of the m matrices $X^1, X^2, ..., X^m$. The explained methodology is applied to each one, then m partitions are obtained.
We do a "consensus between the partitions", that consists in:

i) determining if there are $n_i \in]\frac{m}{2}, m]$, partitions for which I_R=1 and those partitions are the representative, and also all equal, so we find a representative partition;
ii) If the first condition is not satisfied, we reapply the imputation method with $m = 10$; then new partitions in the referred conditions are searched. If they are not found, it means that there is no consensus between the partitions and no partition is representative of the 10 partitions obtained.

3 Simulation studies

In order to study the performance of imputation methods in the presence of missing data we use the Rand's index as described above. First, one hundred samples of each type of simulated data set were generated from five normal multivariate populations (Saporta and Bacelar-Nicolau, 2001)), and $X_i \sim N(\mu_i, \Sigma_i)$, $i = 1, ..., 5$, (1000 observations, 5 variables) with, $(\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5)$. The values of the variance and co-variance matrices have been chosen in order to obtain specific hierarchical structures:

[7] Note: In spite of the modification of the coefficient, we use the same notation for the index.



(order of variables: x_1, x_2, x_3, x_4, x_5)

Here we use twenty five matrices for each one of the five normal multivariate cases.
The partition method gives one of the two following partitions:

i) $\{x_1, x_2, x_3\}$, $\{x_4, x_5\}$
ii) $\{x_1, x_2, x_3, x_4\}$, $\{x_5\}$

Therefore, it performs quite well on the original structures.
For the missing data, we consider different percentages: 10%, 15% and 20% of missing data (over the total of the data - each 1000×5 matrices). Missing data are estimated with the referred imputation methods.
Then we evaluate the effect of missing data and imputation methods on the partition method by comparing each obtained partition with the corresponding partition obtained with the original (complete) data.
In the next tables, we present the results of the comparisons (where the first line represents the mean and the second line - in brackets - the standard deviation).
$I_R = 1$, $I_R > 0,5$ and $I_R <= 0,5$, means:

- $I_R = 1$ the obtained partitions are the same
- $I_R > 0,5$ the obtained partitions are correlated
- $I_R <= 0,5$ the obtained partitions are independent

Table 1. 10% of missing data

	listwise	EM	OLS	NIPALS	PLS	MIave	MIcons
$I_R = 1$	23.4	15	24.8	25	22.9	15	12.2
	(3.6)	(13.7)	(0.4)	(0)	(3.9)	(13.7)	(11.9)
$I_R > 0.5$	1.6	0.6	0.2	0	2.1	10	4
	(4)	(1.3)	(0.4)	(0)	(3.9)	(13.7)	(8.9)
$I_R <= 0.5$	0	9.4	0	0	0	0	8.8
	(0)	(12.9)	(0)	(0)	(0)	(0)	(2.2)

Table 2. 15% of missing data

	listwise	EM	OLS	NIPALS	PLS	MIave	MIcons
$I_R = 1$	21.6	15	24.2	23.8	19.6	14.2	16
	(7.1)	(13.5)	(1.8)	(1.3)	(9.3)	(13.1)	(12.5)
$I_R > 0.5$	3.4	4.2	0.8	0.6	0.6	6	0
	(7.1)	(9.4)	(1.8)	(0.9)	(1.3)	(10.7)	(0)
$I_R <= 0.5$	0	5.8	0	0.6	4.8	4.8	5
	(0)	(10.8)	(0)	(1.3)	(9.8)	(10.2)	(11.2)

Table 3. 20% of missing data

	listwise	EM	OLS	NIPALS	PLS	MIave	MIcons
$I_R = 1$	21	5.4	23.6	19.4	20.5	14.8	16
	(7.9)	(10.9)	(3.1)	(9.8)	(6.1)	(13.5)	(12.5)
$I_R > 0.5$	3.8	10	1.4	2.4	1.2	5.4	0
	(7.9)	(13.7)	(3.1)	(3.9)	(1.8)	(10.9)	(0)
$I_R <= 0.5$	0.2	9.8	0	3	2.3	4.8	9
	(0.4)	(13.4)	(0)	(6.2)	(6)	(10.7)	(12.5)

4 Conclusions

Surprisingly (as compared to previous results for hierarchical methods) multiple imputation does not perform well.
The best results are obtained with OLS regression method and NIPALS algorithm as an imputation method. On the whole, we can say that the OLS regression method performs better than the others but not significantly.
This study shows that the partition method we have analysed to help us in finding the "best" partition in a hierarchical classification, does not perform well when missing data are present and we are using imputation methods. A probabilistic approach based on the affinity coefficient for finding the best "cut-off" appears to be a good solution.

References

1. Bacelar-Nicolau, H. (2002). On the Generalised Affinity Coefficient for Complex Data. *Byocybernetics and Biomedical Engineering*, **22**, 1, 31–42.

2. Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data, New York: John Wiley and Sons, Inc.
3. Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys, New York, John Wiley and Sons.
4. Saporta, G., Youness, G.(2002). Comparing two partitions: Some Proposals and Experiments. In *Proceedings in Computational Statistics*, eds. Hardle, W. and Ronz, B.,pp. 243–248, Physica Verlag, Berlin
5. Silva, A. L., Bacelar-Nicolau, H., Saporta, G.(2002). Missing Data in Hierarchical Classification of Variables - a Simulation Study. In *Classification Clustering and Data Analysis*, eds. Jajuga,K. and Bock, H.-H, pp. 121-128, Springer, Crakow.
6. Silva,A. L., Saporta, G. and Bacelar-Nicolau, H.(2002). Dados omissos em Classificaco Hierarquica de Variaveis e o Algoritmo Nipals In *Proceedings of IX Jornadas de Classificao e Analise de Dados*, ed. ESCS-IPL, pp.42–43, Lisbon.
7. Silva, A. L., Saporta, G., Bacelar-Nicolau, H. (2003). Classification hierarchique ascendante avec imputation multiple des donnees manquantes. In *Methodes et Perspectives en Classification (10emes Rencontre de la Societe Francophone de Classification)*, eds. Yadolah Dodge and Giuseppe Melfi, pp. 59–62 Neuchatel.
8. Tenenhaus, M. (1998). La regression PLS, Editions Technip, Paris.
9. Vigneau, E., Qannari, E.M. (2003). Clustering of Variables Around Latent Components. *Communications in Statistics: Simulation and Computation* **32**, Issue 4, pp. 1131–1150.