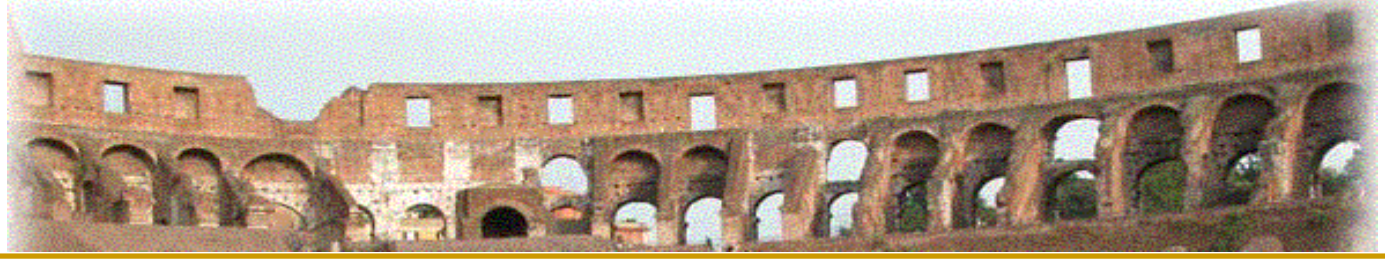


Dealing with missing data in a k-means method – – A simulation based approach[1,2,3]

1-Ana Lorga da Silva, ana.lorga@ulusofona.pt
Department of Economics and Management – ULHT, Portugal
2-Gilbert Saporta – CNAM, Paris
3-Helena Bacelar-Nicolau – UL, Portugal

Outline

- Introduction
- Partition Method
- Simulated Data
- Missing Data
- Three Imputation Methods (IM)
- Multiple Imputation (details)
- Combining Partitions in MI
- Comparing Original Partitions and Partitions with Imputations
- Results
- Conclusions & Perspectives of Work
- References

Introduction

- Evaluate the effect of imputing missing data in a partition method,
- Context: Classification of variables,
- Simulation study.

Partition Method

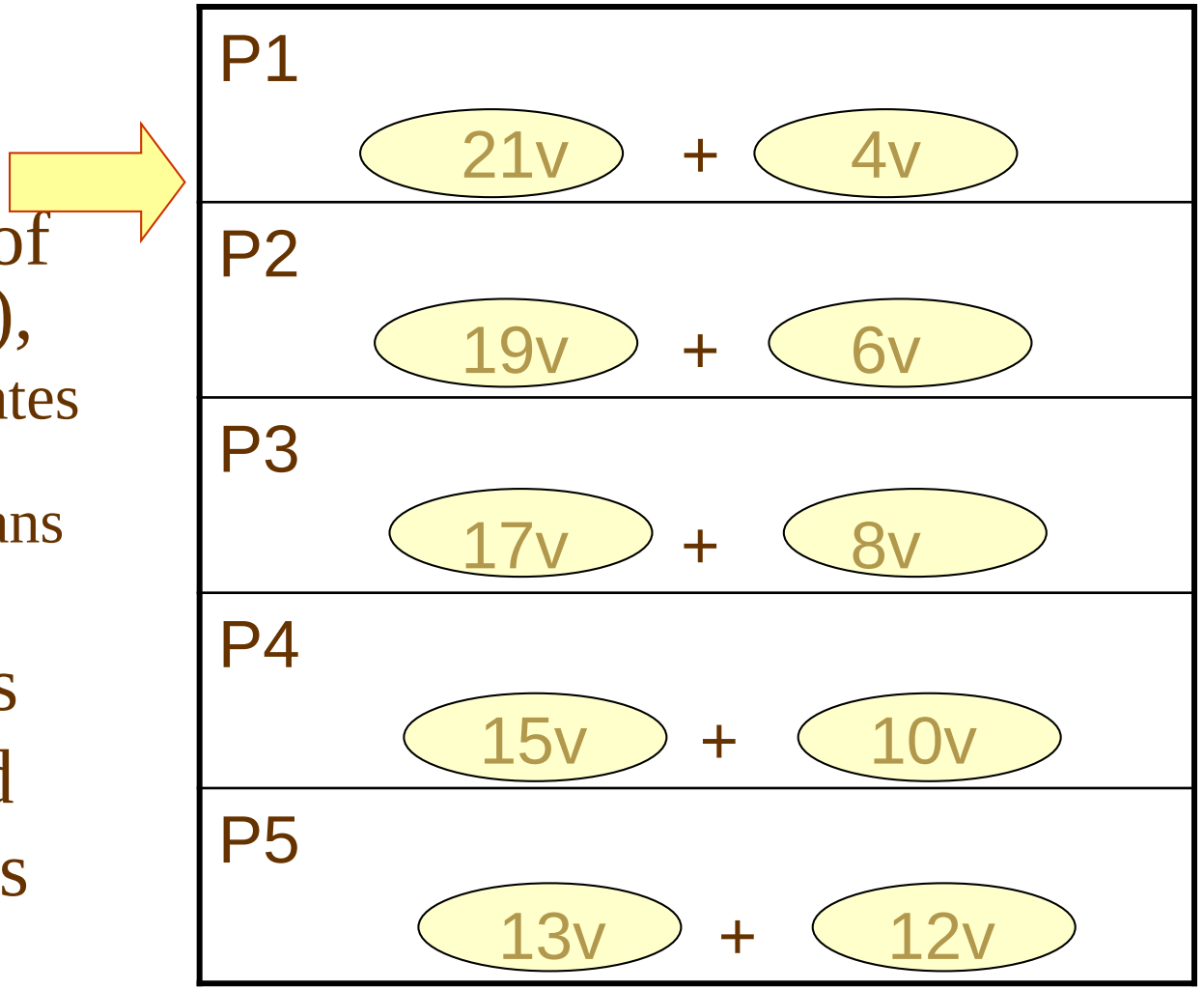
(i) A distance measure is used: $d=1-r^2$,

(ii) Multidimensional Scaling (MDS) Method (Borg & Groenen, 2005), with the aim of use "any distance measure" (an euclidian distance in this work) to apply:

(iii) Forgy's K-means method (Nacache & Confais, 2005), is used over the coordinates obtained by the MDS method .

Simulated Data

- We generate matrices **five partitions** – of variables (25),
  - Two coordinates obtained by MDS→k-means method
- The partitions are composed by two groups



Simulated Data (cont.)

- We generate matrices:
  - X≡X*obs* → type: 1000x25
- Each partition is simulated 100 times →500 matrices,
- Each group obtained:

- $G_1 \rightarrow C^1+\varepsilon^{1i}$, $C^1 \sim N(0, 1)$, $\varepsilon^{1i} \sim N(0, \varepsilon_i)$; $0.1< \varepsilon_i <0.9$
- $G_2 \rightarrow C^2+\varepsilon^{2i}$, $\varepsilon^{2i} \sim N(0, \varepsilon_i)$; $0.1< \varepsilon_i <0.9$

$C^2=\rho C^1+\varepsilon$, $\varepsilon \sim N(0, 1)$

Missing Data (MD)

- 10%, 15% and 20% of MD over each of the 1000x25 matrices - $X \equiv (X_{obs}, X_{mis})$,
- MD on 10 variables,
- Data is Missing at Random – MAR (Little & Rubin, 2002)

$$\Pr ob\ \mathbf{M}|\mathbf{X}_{obs}, X_{mis} = \Pr ob\ \mathbf{M}|X_{obs}$$ such as, $\mathbf{M} = \left[M_{ij}\right]$

is a missing data indicator

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

Three Imputation Methods

1. Implicit Imputation: **Listwise** (Little & Rubin, 2002) ,
2. Single Imputation – **EM** algorithm (Dempster, Laird & Rubin, 1977),
3. Multiple Imputation – $m>1$ (usually $m=5$) imputations (with the aim to reflect certain variability) - A **Data Augmentation** (DA) algorithm is used→this algorithm is based on Monte Carlo Markov Chain Methods (Schaffer, 1997, Fraley,1999 & others).

Multiple Imputation (details)

- It's a Bayesian approach,"*Markov Chain Monte Carlo is a collection of techniques for creating pseudorandom draws from probability distribution*" (Schaffer, 1997)
- *m* independent draws, from the posterior predictive distribution: $P(X_{mis} \mid X_{obs}) = \int P(X_{mis} \mid X_{obs}, \theta) P(\theta, X_{obs}) d\theta$
- I-step X_{mis}^{t+1} is drawn with density $P\left(X_{mis} \mid X_{obs}, \theta^{t}\right)$;
- P-Step draw θ^{t+1} from it's complete data posterior $P\left(\theta \mid X_{obs}, X_{mis}^{t+1}\right)$ this is an iterative process that converges to the posterior distribution of (θ, X_{mis}) given X_{obs}.

Combining Partitions in MI

- To obtain the Partititions when MI are used (five matrices are obtained) we combine the distances d_i (i=1,2,…,5), issued from each one:

$$d = \sum_{i=1}^{5} d_i$$

to which we apply the Partition method described before.

Comparing Original Partitions and Partitions with Imputations

- The Rand index – modified version is used (Youness & Saporta, 2004),

$$R' = \frac{2\sum_u\sum_v n^2_{uv} - \sum_u n^2_{u.} - \sum_v n^2_{v.} + n^2}{n^2}$$

- The Ochiai coefficient (Bacelar-Nicolau, 2000)

$$Och = \frac{\sum_u\sum_v n^2_{uv} - n}{\sqrt{\sum_u n^2_{.u}}\sqrt{\sum_v n^2_{.v}}}$$

We have a contingency table, where two partitions P_1 and P_2 are crossed:
n – number of variables,
n_{uv} – the effective of the cell (u,v),
In this Rand Index the pairs (j,j) are considered.

Comparing Original Partitions and Partitions with Imputations (cont.)

- $0 \leq R' \leq 1$ and $0 \leq Och \leq 1$,
- R'=1 $\Leftrightarrow$ Och=1 $\leftrightarrow$ Identical Partitions,
- The hypothesis of independence for the two Partitions is rejected if R'>0.65 at a 5% significance level (Youness & Saporta, 2004),
- The hypothesis of independence for the two Partitions is rejected if Och>0.797 at a 5% significance level (Sousa, 2006).

Results – Percentage of identical Partititions

	Listwise	EM	MI
10%	42.6 (52.5)	86 (30.7)	63.4 (24)
15%	49.4 (44.5)	85.8 (31.7)	14.6 (28.8)
20%	18.2 (25.9)	86.6 (29.9)	13.2 (11.7)

The Results – Percentage of partitions with 0.65<R'<1

	Listwise	EM	MI
10%	53.6 (50.9)	13.8 (30.8)	35.8 (35)
15%	56.2 (42.8)	13.6 (30.4)	41 (33.3)
20%	66.8 (19.5)	13.4 (29.9)	65.6 (17.3)

The Results – Percentage of partitions with 0.797<Och<1

	Listwise	EM	MI
10%	37.4 (42.2)	0 (0)	8.2 (9)
15%	25.6 (29.2)	0 (0)	26.6 (33.6)
20%	37.8 (20.4)	0 (0)	32.8 (25.7)

The Results – Percentage of partitions with R'≤0.65

	Listwise	EM	MI
10%	3.8 (7.9)	0.2 (0.4)	15.4 (4.8)
15%	6.8 (6.9)	0.6 (1.3)	38.8 (35)
20%	15 (17.7)	0 (0)	21.2 (12.6)

The Results – Percentage of partitions with Och ≤0.797

	Listwise	EM	MI
10%	20.0 (31.9)	14 (30.7)	27.6 (26.6)
15%	25.0 (25.7)	14.2 (31.7)	53 (38.9)
20%	44 (39.7)	13.4 (29.9)	52 (35.1)

Conclusions & Perspectives of Work

- There are differences between the comparisons when we use the Rand index and the Ochiai coefficient,
- Better and “good” results are obtained when EM algorithm is used,
- Worst results are obtained with MI, similar to early works from the authors using another partition method and another MI method,

Conclusions & Perspectives of Work (cont.)

- This MDS method allows us to introduce the five different distance matrices and combine them to obtain the coordinates – that's one of the methods that we shall try in a future work,
- We also intend to use the PLS regression method as an imputation method.

References

- Bacelar-Nicolau, H. (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, pp. 160-165.
- Borg, I. & Groenen, P.J.F. (2005). *Modern Multidimensional Scaling*, 2nd edition. New York: Springer.
- Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion), *J. R. Statist. Soc*. B, **39**, pp. 1-38.
- Fraley, C. (1999),On Computing the Largest Fraction of Missing Information for the EM Algorithm and the Worst Linear for Data Augmentation. *Computational Statistics and Missing Data Analysis,* 31(1), pp.13-26.
- Little, R. J. A. & Rubin, D. B. (2002), *Statistical Analysis With Missing Data*, 2ª Ed. John Wiley & Sons, New York.
- Nacache, J-P & Confais, J. (2005), *Approche Pragmatique de la Classification*, Editions Technip, Paris.
- Schaffer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, Chapman & Hall.
- Sousa, A. (2006), Phd Thesis, University of Azores.
- Youness, G. & Saporta, G., (2004), Une Méthodologie pour la Comparaison de Partitions, *Revue de Statistique Appliquée*, vol. 52(1), pp. 97-120.