COMPSTAT'2008

Multiple Imputation and Multidimensional Scaling Applied to a *k*-means Method[1,2,3]

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

COMPSTAT'2008

Outline

1. Introduction
2. Partition Method
3. Imputation Methods & Multidimensional Scaling (PROXSCAL)
4. Simulations
5. Results (comparing partitions)
6. Conclusions & Perspectives
7. References

1. Introduction

- Evaluate the effect of imputing missing data in a partition method - particularly the use of PROXSCAL when a multiple imputation method is used,
- Context – Classification of variables,
- Simulation Study.

2. Partition Method

1. Define a dissimilarity between variables: $d=1-r^2$,
2. Perform a Multidimensional Scaling – PROXSCAL – PROXimity SCALing (Commandeur & Heiser, 1993), in order to get coordinates (two dimensions in this work),
3. Use Forgy's *K*-means method, with these coordinates.

3. Imputation Methods & Multidimensional Scaling (PROXSCAL)

1. **Implicit Imputation**: **Listwise** (Little & Rubin, 2002) ,
2. **Single Imputation** – **EM** algorithm (Dempster, Laird & Rubin, 1977),
3. **Multiple Imputation** – $m>1$ (usually and in this work $m=5$) imputations - **Data Augmentation** (**DA**) algorithm → based on Monte Carlo Markov Chain Methods (Schaffer, 1997; Fraley,1999 & others).

Imputation (cont.)

- 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 $\theta^{(t+1)}$ 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 given $X_{obs.}$

Imputation (cont.)

- To obtain the Partititions when MI is used (*m* matrices are obtained) the distances d_i (i=1,2,…,5), issued from each one are combined in two different ways that we will call respectively:

IMd and IMp

Imputation (cont.)

- IMd:

$$d = \sum_{i=1}^{m} d_i / m$$

the Partition method – described before - is applied.

Imputation (cont.)

IMp

- Based on the PROXSCAL procedure where we have matrices from different sources – distance matrices in this work,
- The PROXSCAL procedure:
  - Aims at minimizing Kruskal's Stress (a loss function):

$$f(X_1,\ldots,X_k)=\frac{1}{m}\sum_{k=1}^{m}\sum_{i<j}^{n} w_{ijk}\left[\delta_{ijk}-d\left(X_k\right)\right]^2$$

which is the weighted mean square error between the transformed proximities and the distances between *n* variables (usually objects) within *m* sources.

Imputation (cont.)

n – number of variables

m – number of sources (five for IMp, one in the other cases)

δ_{ijk} – the given dissimilarities between the n variables

$d_{i,j}(X_k)$ – Euclidean distances between the variable points, with the coordinates in the rows of X_k (matrix with individual space coordinates for source k) – unrestricted $X_k = ZA_k$, Z-common space and A_k contain the weights.

w_{ijk} – weights (with diferent purposes)

$$i,j=1,\ldots,n;\ k=1,\ldots,p$$

Imputation (cont.)

- The main algorithm in order to minimize the Stress function:
  1. Choose an initial X_k^0 and evaluate, $f\left(X_k^0\right)$
  2. Update X_k using the Guttman transformation (Commandeur & Heiser 1993) - δ_{ijk} is also updated,
  3. Evaluate the loss function,
  4. Repeat 2 if the difference between 1 and 3 larger than a predefined criterion, otherwise stop.

Imputation (cont.)

- For unconditional transformations, all proximities are allowed to be compared with each other irrespective of the source – a new distance matrix is obtained as the "associated final" coordinates - the Partition method is applied.

Imputation (cont.)

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

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

- Ochiai coefficient (Bacelar-Nicolau, 2000)

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

To compare the partitions

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.

Imputation (cont.)

- $0 \leq R' \leq 1$ and $0 \leq Och \leq 1$,
- **R'=1⇔ Och=1 ↔ Identical Partitions**,
- **Independence rejected** when, **R'>0.65** at a **5%** significance level (Youness & Saporta, 2004),
- **Independence rejected** when, **Och>0.797** at a **5%** significance level - based on the work of Sousa (2006).

COMPSTAT'2008

4. Simulations

- **Five partitions** – of variables (25) ↔ two groups each one

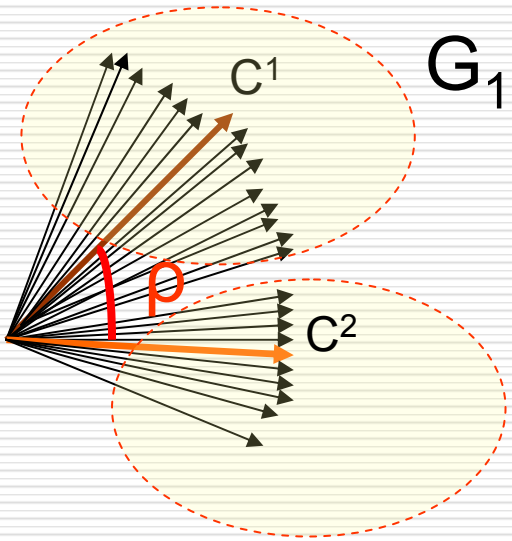
P1	21v	+	4v
P2	19v	+	6v
P3	17v	+	8v
P4	15v	+	10v
P5	13v	+	12v

Simulations (cont.)

- Each Partition:
  - $G_1 \rightarrow C^1$; $C^{1i}=C^1+\varepsilon^{1i}$, $C^1 \sim N(0, 1)$, $\varepsilon^{1i} \sim N(0, \varepsilon_i)$;
  - $G_2 \rightarrow C^2=\rho C^1+\varepsilon$; $C^{2i}=C^2+\varepsilon^{2i}$, $\varepsilon \sim N(0, 1)$, $\varepsilon^{2i} \sim N(0, \varepsilon_i)$;

$0.1 \leq \varepsilon_i \leq 0.9$; i=1,…,23,

e.g.



C^1, C^{1i}, C^2, C^{2i} , i=1,..23 $\rightarrow X_j$=1,…25

Simulations (cont.)

- **Data:**
  - Five hundred matrices: 1000X25

$$X \equiv X_{obs}$$

(one hundred for each case)

Simulations (cont.)

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

$$P\left(\mathbf{M}\middle|X_{obs}, X_{mis}\right) = P\left(\mathbf{M}\middle|X_{obs}\right)$$

such as, $\mathbf{M} = [M_{ij}]$, is a missing data indicator

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

5. Results (comparing partitions)

Percentage of identical Partitititions*

	Listwise	EM	IMd	IMp
10%	42.6 (52.4)	86 (30.7)	47.2 (34.2)	67.6 (25,3)
15%	34.6 (47.6)	85.8 (31.7)	25.6 (28.9)	27.8 (31.9)
20%	18.6 (26.3)	86.6 (29.9)	13.4 (11.1)	11.6 (13.3)

*between brackets standard deviations

Results (cont.)

Percentage of partitions with 0.65<R'<1

	Listwise	EM	IMd	IMp
10%	53.2 (42.5)	13.4 (29.9)	38 (40.7)	20.2 (26.1)
15%	57 (42.5)	13.4 (29.9)	38.6 (34.8)	42.8 (38.4)
20%	65.8 (20.2)	13.4 (29.9)	66.6 (15.7)	73.2 (18.6)

COMPSTAT'2008

Results (cont.)

Percentage of partitions with 0.797<Och<1

	Listwise	EM	IMd	IMp
10%	36.8 (41.8)	0 (0)	26.8 (41.5)	7.6 (8.2)
15%	37.8 (30.2)	0 (0)	27.4 (32.7)	29.8 (36.5)
20%	40 (19.2)	0 (0)	34.6 (27.3)	42.5 (35.2)

Results (cont.)

Percentage of partitions with $R' \leq 0.65$

	Listwise	EM	IMd	IMp
10%	4.2 (9.7)	0.6 (1.7)	14.8 (8.2)	12.2 (3.4)
15%	8.4 (9.7)	0.8 (1.7)	35.8 (36)	29.4 (39.6)
20%	15.6 (19.6)	0 (0)	20 (10.7)	15.2 (5.6)

Results (cont.)

Percentage of partitions with Och ≤0.797

	Listwise	EM	IMd	IMp
10%	20.6 (29.9)	14 (30.7)	26 (28)	24.8 (27.7)
15%	27.6 (29.3)	14.2 (31.7)	47 (37.4)	42.4 (41.4)
20%	41.4 (39.3)	13.4 (29.9)	52 (38.2)	48.2 (39.1)

6. Conclusions & Perspectives

- There are differences between the comparisons : Rand index \ Ochiai coefficient,
- Better and "good" results are obtained with EM algorithm ,
- Comparing Imd and Imp (introduced in this work) the results have improved with Imp, mainly to 10% and 20%,
- Also better results are obtained with Imp than Listwise to 10% and 20%,

Conclusions & Perspectives (cont.)

- The obtained results with MI, similar to early works from the authors using another partition method and another MI method are not globaly the best,

- We intend to use 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.
- Commandeur, J. & Heiser, W.J.(1993) Mathematical Derivations in the Proximity Scaling (PROXSCAL) of Symetric Data Matrices. Research Report RR-93-04, Departement of Data Theory, Leiden University.
- 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.
- Nakache, 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), Contribuições à Metodologia VL e índices de validação para dados de natureza complexa, Phd Thesis, Universidade dos Açores.
- Youness, G. & Saporta, G., (2004), Une Méthodologie pour la Comparaison de Partitions, *Revue de Statistique Appliquée*, vol. 52(1), pp. 97-120.