Dados omissos em Classificação hierárquica de variáveis e o algoritmo NIPALS

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

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

1. Introdução

O problema dos dados omissos tem sido abordado em diversos artigos e nalguns livros, onde se encontram vários métodos para minimizar o efeito dos dados em falta:

- Orchard and Woodbury(1972)
- Rubin(1974)
- Beale and Little(1975)
- Dempster, Laird and Rubin(1977)
- Rubin(1987)
- Little and Rubin(1987)

entre outros

- Estamos interessados em analisar o efeito de **dados omissos** nalgumas estruturas particulares (originalmente completas) – **classificação hierárquica ascendente de variáveis**, bem como os resultados obtidos com os **dados imputados** nesses casos.

- Neste trabalho consideramos modelos de classificação hierárquicos baseados em dois coeficientes de semelhança:
  - **coeficiente de afinidade básico**
    (Matusita(1955), Bacelar-Nicolau(1988))
  - **Coeficiente de correlação de Pearson**
    e três coeficientes de agregação clássicos

- Imputamos os dados recorrendo ao algoritmo NIPALS com diferentes percentagens de dados omissos – MAR (Missing at Random)

- Utilizámos matrizes de dados com distribuição multinormal (Saporta (1990)).

2. Classificação Hiérarquica Ascendente

Utilizaram-se os seguintes critérios de agregação:
Average linkage (AL) / Single linkage (SL) / Complete linkage (CL)

Os coeficientes de semelhança entre duas variáveis ($n\times1$) (Xj , Xj'):

- O coeficiente de afinidade básico $c_a = \sum_{i=1}^{n} \sqrt{\frac{x_{ij}}{x_{.j}} \frac{x_{ij'}}{x_{.j'}}}$, onde $x_{.j} = \sum_{i=1}^{n} x_{ij}$ e $x_{.j'} = \sum_{i=1}^{n} x_{ij'}$, (e.g. Bacelar-Nicolau(1988, 2000)).

- O coeficiente de correlação de **Bravais-Pearson** $c_p = \frac{\sum_{i=1}^{n} \left(x_{ij} - \bar{x}^{j}\right)\left(x_{ij'} - \bar{x}^{j'}\right)}{s_{xj}\, s_{xj'}}$.

Para comparar os modelos de classificação hierárquica utilizámos o coeficiente de Spearman entre as matrizes das semelhanças ultramétricas associadas aos métodos de agregação referidos, com as correcções usuais para os empates.

3. Os dados omissos – MAR

A expressão geral do conceito de **MAR** pode escrever-se:

$$\Pr ob\left(R \middle| X_{obs}, X_{mis}\right) = \Pr ob\left(R \middle| X_{obs}\right)$$

X_{obs} representa os valores observados de $\mathbf{X}_{n\times p}$

X_{mis} os dados omissos de $\mathbf{X}_{n\times p}$

e

$R = \left[R_{ij}\right]$ é um indicador dos dados omissos, $R_{ij} = \begin{cases} 1, \ se \ x_{ij} & \text{e' observado} \\ 0, \ se \ x_{ij} & \text{e' omisso} \end{cases}$

4. O algoritmo NIPALS

O Algoritmo NIPALS apresentado por Wold (Wold et al.(1969),Wold(1973)) permite realizar uma análise em componentes principais com dados em falta sem suprimir as linhas que contêm dados omissos nem estimar os dados em falta, tal como descrito em Tenenhaus(1998):

O algoritmo NIPALS em presença de dados omissos:

Etapas:

A) $X_0 = X$

B) para $h = 1, 2, ..., a$:

Bi) $t_h = (X_{h-1})_{.1}$

Bii) repete-se até à convergência de p_h :

$$\text{para } j = 1, ...p, \qquad p_{hj} = \frac{\sum_{\{j: x_{ij} \text{ e } t_{hi} \text{ existem}\}} x_{h-1,ij} t_{hi}}{\sum_{\{j: x_{ij} \text{ e } t_{hi} \text{ existem}\}} t_{hi}^2} \rightarrow \|p_h\| = 1$$

$$\downarrow$$

$$\text{para } i = 1, ...n, \qquad t_{hi} = \frac{\sum_{\{j: x_{ij} \text{ existe}\}} x_{h-1,ij} t_{hi}}{\sum_{\{j: x_{ij} \text{ existe}\}} t_{hi}^2}$$

C) $X_h = X_{h-1} - t_h p_h'$

5. A imputação (**após aplicação do algoritmo NIPALS**)

Reconstituímos a matriz incompleta do seguinte modo:

$$x_{ij} = \begin{cases} x_{ij} \quad se \quad x_{ij} \quad \text{e' observado} \\ \hat{x}_{ij} \quad se \quad x_{ij} \quad \text{e' omisso} \end{cases}$$

$$\hat{x}_{ij} = \sum_{l=1}^{n_c} t_{li}\, p_{lj}$$

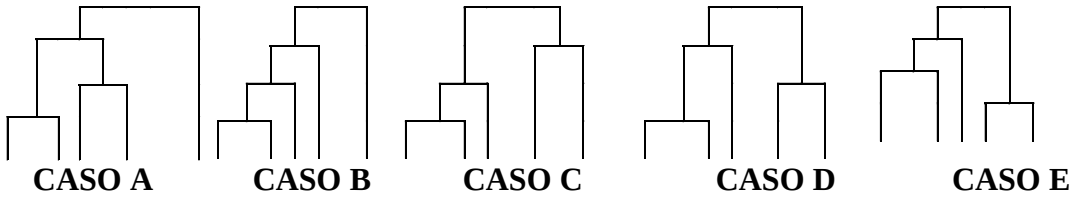
$t_{li}\,, p_{li}$ são respectivamente as coordenadas das componentes principais e dos vectores directores dos eixos principais

6. Experiências numéricas

Casos **(A), (B), (C), (D) e (E):** $X_i \sim \aleph\left(\mu_i, \Sigma_i\right)$ tais que $\boldsymbol{X_i}$ são matrizes 1000×5

$i = 1,...,5$

Os valores das matrizes de variância-covariância foram escolhidas com objectivo the obter estruturas hierarquicas específicas:



Com o objectivo de obter dados omissos do tipo MAR retiraram-se dados a duas variáveis X_1 e X_2 – 10%,15% e 20% de dados sobre o total da população (matriz $X_{1000\times5}$).

Avaliámos os resultados das simulações por ordem crescente da percentagem de dados omissos (MD - missing data), de acordo com os coeficientes de semelhança, de acordo com os métodos de classificação e após a imputação dos dados (ID).

Em cada caso comparamos as ultramétricas associadas aos dados originalmente completos com as matrizes das ultramétricas associadas aos dados incompletos e reconstituídos respectivamente.

A comparação entre as ultramétricas obtém-se utilizando um teste bilateral de Spearman (a 5%)

Em presença de dados omissos, utiliza-se o método “listwise”, i.e. considera-se na análise apenas as linhas completas (eliminando as linhas com dados omissos).

Caso A

	Coeficiente de afinidade			Coeficiente de Correlação		
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%** $c_S = 1$ **11%** $\|c_S\| > c'_S$	**100%** $c_S = 1$
15%	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$	**50%** $c_S = 1$ **50%** $\|c_S\| > c'_S$	**100%** $c_S = 1$
20%	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$	**69%** $c_S = 1$ **31%** $\|c_S\| > c'_S$	**100%** $c_S = 1$

Tabela 1 descreve os resultados na presença de dados omissos

	Coeficiente de afinidade			Coeficiente de Correlação		
ID	**AL**	**SL**	**CL**	**AL**	**SL**	**CL**
10%	**100%** $c_S = 1$	**98%** $c_S = 1$ **2%** $\|c_S\| > c'_S$	**100%** $c_S = 1$	**100%** $c_S = 1$	**75%** $c_S = 1$ **25%** $\|c_S\| > c'_S$	**100%** $c_S = 1$
15%	**98%** $c_S = 1$ **2%** $\|c_S\| > c'_S$	**58%** $c_S = 1$ **42%** $\|c_S\| > c'_S$	**100%** $c_S = 1$	**99%** $c_S = 1$ **1%** $\|c_S\| > c'_S$	**61%** $c_S = 1$ **39%** $\|c_S\| > c'_S$	**100%** $c_S = 1$
20%	**56%** $c_S = 1$ **44%** $\|c_S\| > c'_S$	**24%** $c_S = 1$ **76%** $\|c_S\| > c'_S$	**94%** $c_S = 1$ **6%** $\|c_S\| > c'_S$	**97%** $c_S = 1$ **3%** $\|c_S\| > c'_S$	**61%** $c_S = 1$ **39%** $\|c_S\| > c'_S$	**99%** $c_S = 1$ **1%** $\|c_S\| > c'_S$

Tabela 2 descreve os resultados com os dados imputados após a utilização do algoritmo NIPALS

Caso B

	Coeficiente de afinidade			Coeficiente de Correlação		
MD	**AL**	**SL**	**CL**	**AL**	**SL**	**CL**
10%	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$
15%	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$	**98%** $c_S = 1$ **2%** $\lvert c_S \rvert > c'_S$	**100%** $c_S = 1$	**100%** $c_S = 1$
20%	**100%** $c_S = 1$	**99%** $c_S = 1$ **1%** $\lvert c_S \rvert > c'_S$	**100%** $c_S = 1$	**94%** $c_S = 1$ **6%** $\lvert c_S \rvert > c'_S$	**97%** $c_S = 1$ **3%** $\lvert c_S \rvert > c'_S$	**81%** $c_S = 1$ **18%** $\lvert c_S \rvert > c'_S$ **1%** $\lvert c_S \rvert < c'_S$

Tabela 3 descreve os resultados na presença de dados omissos

	Coeficiente de afinidade			Coeficiente de Correlação		
ID	**AL**	**SL**	**CL**	**AL**	**SL**	**CL**
10%	**100%** $c_S = 1$	**99%** $c_S = 1$ **1%** $\lvert c_S \rvert > c'_S$	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$
15%	**100%** $c_S = 1$	**99%** $c_S = 1$ **1%** $\lvert c_S \rvert > c'_S$	**99%** $c_S = 1$ **1%** $\lvert c_S \rvert > c'_S$	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$
20%	**84%** $c_S = 1$ **16%** $\lvert c_S \rvert > c'_S$	**82%** $c_S = 1$ **18%** $\lvert c_S \rvert > c'_S$	**87%** $c_S = 1$ **13%** $\lvert c_S \rvert > c'_S$	**85%** $c_S = 1$ **15%** $\lvert c_S \rvert > c'_S$	**84%** $c_S = 1$ **16%** $\lvert c_S \rvert > c'_S$	**87%** $c_S = 1$ **13%** $\lvert c_S \rvert > c'_S$

Tabela 4 descreve os resultados com os dados imputados após a utilização do algoritmo NIPALS

Caso C

	Coeficiente de afinidade			Coeficiente de Correlação		
MD	**AL**	**SL**	**CL**	**AL**	**SL**	**CL**
10%	**100%** $c_S = 1$	**99%** $c_S = 1$ **1%** $\|c_S\| > c'_S$ *	**100%** $c_S = 1$	**5%** $c_S = 1$ **95%** $\|c_S\| < c'_S$	**49%** $c_S = 1$ **51%** $\|c_S\| < c'_S$	**100%** $\|c_S\| < c'_S$
15%	**100%** $c_S = 1$	**96%** $c_S = 1$ **4%** $\|c_S\| > c'_S$ *	**99%** $c_S = 1$ **1%** $\|c_S\| > c'_S$	**5%** $c_S = 1$ **95%** $\|c_S\| < c'_S$	**41%** $c_S = 1$ **59%** $\|c_S\| < c'_S$	**100%** $\|c_S\| < c'_S$
20%	**100%** $c_S = 1$	**87%** $c_S = 1$ **13%** $\|c_S\| > c'_S$ *	**100%** $c_S = 1$	**2%** $c_S = 1$ **97%** $\|c_S\| < c'_S$	**21%** $c_S = 1$ **69%** $\|c_S\| < c'_S$	**100%** $\|c_S\| < c'_S$

Tabela 5 descreve os resultados na presença de dados omissos

*** efeito de cadeia**

	Coeficiente de afinidade			Coeficiente de Correlação		
ID	**AL**	**SL**	**CL**	**AL**	**SL**	**CL**
10%	**74%** $c_S = 1$ **24%** $\|c_S\| > c'_S$ **2%** $\|c_s\| < c'_s$	**74%** $c_S = 1$ **22%** $\|c_S\| > c'_S$ **4%** $\|c_s\| < c'_s$	**74%** $c_S = 1$ **25%** $\|c_S\| > c'_S$ **1%** $\|c_s\| < c'_s$	**74%** $c_S = 1$ **26%** $\|c_S\| > c'_S$	**74%** $c_S = 1$ **26%** $\|c_S\| > c'_S$	**74%** $c_S = 1$ **26%** $\|c_S\| > c'_S$
15%	**88%** $c_S = 1$ **11%** $\|c_S\| > c'_S$ **1%** $\|c_s\| < c'_s$	**88%** $c_S = 1$ **6%** $\|c_S\| > c'_S$ **6%** $\|c_s\| < c'_s$	**88%** $c_S = 1$ **11%** $\|c_S\| > c'_S$ **1%** $\|c_s\| < c'_s$	**88%** $c_S = 1$ **12%** $\|c_S\| > c'_S$	**88%** $c_S = 1$ **11%** $\|c_S\| > c'_S$ **1%** $\|c_s\| < c'_s$	**88%** $c_S = 1$ **12%** $\|c_S\| > c'_S$
20%	**99%** $c_S = 1$ **1%** $\|c_S\| > c'_S$	**99%** $c_S = 1$ **1%** $\|c_s\| < c'_s$	**99%** $c_S = 1$ **1%** $\|c_S\| > c'_S$	**99%** $c_S = 1$ **1%** $\|c_s\| < c'_s$	**99%** $c_S = 1$ **1%** $\|c_s\| < c'_s$	**99%** $c_S = 1$ **1%** $\|c_S\| > c'_S$

Tabela 6 descreve os resultados com os dados imputados após a utilização do algoritmo NIPALS

Caso D

	Coeficiente de afinidade			Coeficiente de Correlação		
MD	**AL**	**SL**	**CL**	**AL**	**SL**	**CL**
10%	**99%** $c_S = 1$ **1%** $\lvert c_S\rvert < c'_S$	**100%** $c_S = 1$	**100%** $c_S = 1$	**97%** $c_S = 1$ **3%** $\lvert c_S\rvert < c'_S$	**96%** $c_S = 1$ **4%** $\lvert c_S\rvert > c'_S$	**99%** $c_S = 1$ **1%** $\lvert c_S\rvert > c'_S$
15%	**98%** $c_S = 1$ **1%** $\lvert c_S\rvert > c'_S$ **1%** $\lvert c_S\rvert < c'_S$	**99%** $c_S = 1$ **1%** $\lvert c_S\rvert > c'_S$	**99%** $c_S = 1$ **1%** $\lvert c_S\rvert > c'_S$	**95%** $c_S = 1$ **5%** $\lvert c_S\rvert < c'_S$	**90%** $c_S = 1$ **10%** $\lvert c_S\rvert > c'_S$	**95%** $c_S = 1$ **5%** $\lvert c_S\rvert > c'_S$
20%	**99%** $c_S = 1$ **1%** $\lvert c_S\rvert < c'_S$	**100%** $c_S = 1$	**100%** $c_S = 1$	**92%** $c_S = 1$ **8%** $\lvert c_S\rvert < c'_S$	**85%** $c_S = 1$ **15%** $\lvert c_S\rvert > c'_S$	**93%** $c_S = 1$ **7%** $\lvert c_S\rvert > c'_S$

Tabela 7 descreve os resultados na presença de dados omissos

	Coeficiente de afinidade			Coeficiente de Correlação		
ID	**AL**	**SL**	**CL**	**AL**	**SL**	**CL**
10%	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$
15%	**66%** $c_S = 1$ **34%** $\lvert c_s\rvert < c'_s$	**66%** $c_S = 1$ **34%** $\lvert c_s\rvert < c'_s$	**66%** $c_S = 1$ **34%** $\lvert c_s\rvert < c'_s$	**66%** $c_S = 1$ **34%** $\lvert c_s\rvert < c'_s$	**66%** $c_S = 1$ **34%** $\lvert c_s\rvert < c'_s$	**66%** $c_S = 1$ **34%** $\lvert c_s\rvert < c'_s$
20%	**5%** $c_S = 1$ **2%** $\lvert c_S\rvert > c'_S$ **93%** $\lvert c_s\rvert < c'_s$	**4%** $c_S = 1$ **3%** $\lvert c_S\rvert > c'_S$ **93%** $\lvert c_s\rvert < c'_s$	**5%** $c_S = 1$ **2%** $\lvert c_S\rvert > c'_S$ **93%** $\lvert c_s\rvert < c'_s$	**6%** $c_S = 1$ **1%** $\lvert c_S\rvert > c'_S$ **93%** $\lvert c_s\rvert < c'_s$	**5%** $c_S = 1$ **2%** $\lvert c_S\rvert > c'_S$ **93%** $\lvert c_s\rvert < c'_s$	**6%** $c_S = 1$ **1%** $\lvert c_S\rvert > c'_S$ **93%** $\lvert c_s\rvert < c'_s$

Tabela 8 descreve os resultados com os dados imputados após a utilização do algoritmo NIPALS

Case E

	Coeficiente de afinidade			Coeficiente de Correlação		
MD	**AL**	**SL**	**CL**	**AL**	**SL**	**CL**
10%	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$
15%	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$	**97%** $c_S = 1$ **3%** $\|c_S\| < c'_S$	**98%** $c_S = 1$ **2%** $\|c_S\| < c'_S$
20%	**100%** $c_S = 1$	**98%** $c_S = 1$ **1%** $\|c_S\| > c'_S$ **1%** $\|c_S\| < c'_S$	**99%** $c_S = 1$ **1%** $\|c_S\| > c'_S$	**100%** $c_S = 1$	**87%** $c_S = 1$ **13%** $\|c_S\| < c'_S$	**92%** $c_S = 1$ **8%** $\|c_S\| < c'_S$

Tabela 9 descreve os resultados na presença de dados omissos

	Coeficiente de afinidade			Coeficiente de Correlação		
ID	**AL**	**SL**	**CL**	**AL**	**SL**	**CL**
10%	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$	**100%** $c_S = 1$
15%	**91%** $c_S = 1$ **9%** $\|c_s\| < c'_s$	**91%** $c_S = 1$ **9%** $\|c_s\| < c'_s$	**91%** $c_S = 1$ **9%** $\|c_s\| < c'_s$	**91%** $c_S = 1$ **9%** $\|c_s\| < c'_s$	**91%** $c_S = 1$ **9%** $\|c_s\| < c'_s$	**91%** $c_S = 1$ **9%** $\|c_s\| < c'_s$
20%	**78%** $c_S = 1$ **8%** $\|c_s\| > c'_s$ **14%** $\|c_s\| < c'_s$	**80%** $c_S = 1$ **8%** $\|c_s\| > c'_s$ **14%** $\|c_s\| < c'_s$	**78%** $c_S = 1$ **8%** $\|c_s\| > c'_s$ **14%** $\|c_s\| < c'_s$	**86%** $c_S = 1$ **14%** $\|c_s\| < c'_s$	**86%** $c_S = 1$ **14%** $\|c_s\| < c'_s$	**868%** $c_S = 1$ **14%** $\|c_s\| < c'_s$

Tabela 10 descreve os resultados com os dados imputados após a utilização do algoritmo NIPALS

7. Conclusões

- Nas simulações feitas o comportamento do algoritmo NIPALS é melhor no que se refere ao coeficiente de Pearson.
- Nestes casos estudados obtêm-se melhores resultados com o coeficiente de afinidade quando se utiliza o método listwise.
- Quer em relação a este estudo quer em relação a trabalhos anteriores (e.g. Silva (2001)) obtêm-se melhores resultados utilizando o método listwise do que utilizando os métodos de imputação.
- Em todos os trabalhos por nós desenvolvidos os melhores resultados obtiveram-se com a utilização simultânea do coeficiente de afinidade e do método listwise.

Os próximos desenvolvimentos deste trabalho estão relacionados com:

- Outras estruturas hierárquicas (nomeadamente uma aproximação probabilística)
- Classificação de indivíduos
- Métodos de imputação múltipla – aplicação de técnicas de consensus

(...)

Bibliografia

BACELAR-NICOLAU(1988) Two probabilistic models for classification of variables in frequency tables - 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 – 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.,(1955) 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

SILVA,A.L, BACELAR-NICOLAU, SAPORTA, G. and GEADA, M.(2001) Missing Data in Hierarchical Classification – a study with Personality development data, – 32nd European Mathematical Psycology /EMPG 2001, pp.109-110.

TENENHAUS, M. (1998) La Régression PLS, Théorie et Pratique, , Editions Technip, Paris

WOLD,H., LYTTKENS,E., (1969) "Nonlinear Iterative Partial Least Squares (NIPALS) Estimation Procedures", Bull. Intern. Statist. Inst: Proc. 37th Session, London, pp.1-15

WOLD,H., (1973), "Nonlinear Iterative Partial Least Squares (NIPALS) modelling: some current developments", Multivariate Analysis III, Proc. 3rd international Symposium, Dayton, pp. 383-407.