Efeito de um Método de Imputação Múltipla em Classificação Hierárquica de Variáveis

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

Efeito de um Método de Imputação Múltipla em Classificação Hierárquica de Variáveis*

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

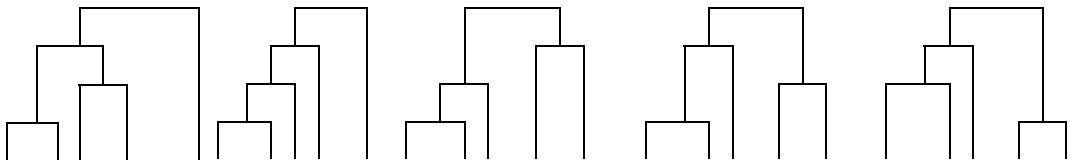
[1]ISEG, Universidade Tecnica de Lisboa, Conservatoire National des Arts et Métiers, CEDRIC
e-mail:aigcls@iseg.utl.pt
[2]LEAD-FPCE, Universidade de Lisboa
e-mail:hbacelar@fpce.ul.pt
[3]Chaire de Statistique Appliquée, CEDRIC
Conservatoire National des Arts et Métiers
e-mail:saporta@cnam.fr

Palavras Chave: Dados omissos, Imputação Múltipla, Classificação Hierárquica.

Compara-se resultados obtidos em classificação hierárquica ascendente - em particular na classificação de variáveis - após a reconstituição de dados em falta recorrendo a um método de imputação múltipla baseado num modelo de regressão OLS, com estruturas resultantes de matrizes de dados incompletas.

Tal como em trabalhos anteriores (**[4]**, **[5], [6]** e **[7]**) utilizam-se matrizes de dados, originalmente completos, com distribuição multinormal (**[3])**, às quais são retirados dados de acordo com a condição MAR - "Missing at Random" - $\Pr ob\left(R\middle|X_{obs},X_{mis}\right)=\Pr ob\left(R\middle|X_{obs}\right)$, onde X_{obs} representa os valores observados da matriz de dados $X_{n\times p}$, X_{miss} representa os valores omissos em $X_{n\times p}$ e $R=\left[R_{ij}\right]$ é um indicador dos dados em falta, $R_{ij}=\begin{cases}1,\ se\ x_{ij} & \text{e´ observado}\\ 0,\ se\ x_{ij} & \text{e´ omisso}\end{cases}$.

Os dados gerados, consistem em matrizes 1000×5 - cem simulações para cada caso - com o objectivo de obter estruturas específicas representadas pelos seguintes dendrogramas:



Utiliza-se como coeficientes de semelhança 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 $x_{.j'}=\sum_{i=1}^{n}x_{ij'}$, tal como definido por exemplo em **[1]** e o coeficiente de correlação de Bravais-Pearson $c_p=\frac{\sum_{i=1}^{n}\left(x_{ij}-\overline{x}^{j}\right)\left(x_{ij'}-\overline{x}^{j'}\right)}{s_{x^j}s_{x^{j'}}}$.

* Este trabalho foi suportado parcialmente pelo Programa Científico Franco-Português MSPLDM-542-B2 (Embaixada de França e Ministério Português de Ciência e tecnologia - ICCTI) e pela equipa de ***Multivariate Data Analysis*** do CEAUL/FCUL.

Como métodos de agregação utilizam-se aqui três critérios de agregação clássicos: "average linkage", "single linkage" e "complete linkage".

O procedimento comporta três fases gerais:

1. Retiram-se dados a duas variáveis (às duas primeiras variáveis - as variáveis apresentam-se pela ordem X_1, X_2, X_3, X_4, X_5, nos cinco dendrogramas representados) – 10%,15% e 20% de dados sobre o total da população (matriz $X_{1000\times5}$). Os dados omissos apresentam um padrão maioritariamente monótono, apenas com uma pequena percentagem de dados omissos representados por padrão não monótono.
2. Faz-se o estudo dos resultados obtidos utilizando os métodos "listwise" (Tabela 1), "pairwise"(Tabela 2) "pairwise após retirados os dados omissos correspondentes ao padrão não monótono de dados omissos"(Tabela 3) e de imputação múltipla quer na presença de todos os dados omissos (Tabela 4) quer após retirados os dados omissos que não satisfazem ao padrão monótono (Tabela 5).
3. Para comparar os modelos de classificação hierárquica utilizou-se o coeficiente de Spearman entre as matrizes das ultramétricas.

O método de imputação múltipla utilizado é um método baseado sobre um modelo preditivo de regressão OLS (**[2]**, **[8]**).

Nos métodos de imputação múltipla o dado omisso é substituído por mais do que um valor (m>1) plausível (dando origem a m matrizes de dados), para que representem uma certa incerteza sobre o dado imputado.

O modelo aqui utilizado é baseado na teoria Bayesiana, como descrito por exemplo em **[2].** Em primeiro lugar o modelo preditivo de regressão OLS é estimado a partir dos dados completamente observados, como habitualmente. Utiliza-se esse modelo para construir outros, em que os parâmetros são retirados aleatoriamente da sua distribuição à posteriori. "The randomly drawn values are used to generate imputations, wich include random deviations from the model's predictions" (**[8])**. Deste modo garante-se uma incerteza suplementar devido ao facto dos parâmetros serem estimados a posteriori, e não determinados a partir dos dados observados.

Neste trabalho, considera-se m=5. Reconstitui-se a matriz incompleta (de cinco modos distintos), utilizando o referido modelo de regressão, do seguinte modo:

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

Depois de obtidas, para cada caso (cada m), as cinco matrizes resultantes do método de imputação múltipla determina-se, para cada matriz, a matriz de semelhança correspondente S_k, k=1,2,...,5, calculando-se a média das matrizes de semelhança, S, tal que, $S = \left(\sum_{k=1}^{5} S_k\right) \Big/ 5$. Aplica-se em seguida a cada matriz S, cada um dos métodos de agregação.

Compara-se então (fase 3) as estruturas hierárquicas obtidas com matrizes reconstituídas desta forma, com as estruturas obtidas na presença de dados omissos (em relação às estruturas originais com os dados completos).

Método "listwise" / complete cases

MD		Coef. de Afinidade AL	SL	CL	Coef. de Pearson AL	SL	CL
	$c_s = 1$	99,8	99,4	99,6	86	91,0	83,6
		0,4	0,9	0,5	29,1	12,1	34,4
10%	$\|c_s\| > c'_s$	0	0	0,2	13,2	5,8	15,6
		0	0	0,4	29,5	12,9	34,8
	$\|c_s\| < c'_s$	0,2	0,6	0,2	0,8	3,2	0,8
		0,4	0,9	0,4	1,3	4,8	0,8
	$c_s = 1$	99,6	99,2	98,8	84	83,6	79,8
		0,9	2,8	1,8	30,3	16,9	35,3
15%	$\|c_s\| > c'_s$	0,2	0,2	1	14	8,4	16,6
		0,4	0,4	0,4	31,3	18,8	37,1
	$\|c_s\| < c'_s$	0,2	1,6	0,2	2	8	3,6
		0,4	2,5	0,4	2,5	10	4,3
	$c_s = 1$	99,2	96,2	97,8	80,8	76,0	76,4
		0,4	5,8	3,3	30,8	19,7	35
20%	$\|c_s\| > c'_s$	0,4	0,2	1,8	14,8	10,6	17
		0,5	0,4	3,5	33	23,7	38
	$\|c_s\| < c'_s$	0,4	3,6	0,4	4,4	13,4	6,6
		0,5	5,8	0,5	4	13,5	7,8

Tabela 1

Método "pairwise" / available cases

MD		Coef. de Afinidade AL	SL	CL	Coef. de Pearson AL	SL	CL
	$c_s = 1$	99	94,2	99,4	84	95,2	73,6
		0,7	10,3	0,9	22,3	5,6	29,9
10%	$\|c_s\| > c'_s$	0,4	0,4	0	11	0	15,6
		0,9	0,9	0	24,6	0	34,8
	$\|c_s\| < c'_s$	0,6	5,4	0,6	5	4,8	10,8
		0.9	10,4	0,9	5,3	5,6	10,1
	$c_s = 1$	99,4	86,4	98,8	79,8	91,2	60
		3	23,8	2,2	22	7,6	28,2
15%	$\|c_s\| > c'_s$	1	2,2	0,8	11,2	3	15,8
		2	4,9	1,3	25	6,7	35,3
	$\|c_s\| < c'_s$	0,6	11,4	0,4	9	5,8	24,2
		0,9	24,3	0,9	10,5	7,6	22,5
	$c_s = 1$	98,2	88,8	98,2	94,6	97,8	81,2
		4	17,5	2,5	7,2	1,3	32,5
20%	$\|c_s\| > c'_s$	1,4	2,8	1,4	4,8	1,6	16,2
		3,1	6,3	1,9	7,5	1,5	32,9
	$\|c_s\| < c'_s$	0,4	8,4	0,4	0,6	0,6	2,6
		0,9	17,7	0,9	0,9	1,3	1,7

Tabela2

Método "pairwise" / available cases/padrão monótono

MD		Coef. de Afinidade AL	SL	CL	Coef. de Pearson AL	SL	CL
	$c_s = 1$	99	90,4	99	67,8	84,2	38,6
		1,7	17,7	1,7	26,7	20,4	35,1
	$\|c_s\| > c'_s$	0,4	0,8	0,4	14	10	3,6
10%		0,9	1,8	0,9	31,3	22,4	1
	$\|c_s\| < c'_s$	0,6	8,8	0,6	18,2	5,8	45,8
		0,9	18	0,9	19,2	7,9	42,4
	$c_s = 1$	96,8	76,4	99,2	47	79,8	26,2
		4,4	36,5	1,7	30,8	28	41,6
	$\|c_s\| > c'_s$	1,4	6,6	0,4	15,4	13,6	16,6
15%		3,1	14,7	0,8	34,4	30,4	37,1
	$\|c_s\| < c'_s$	1,8	17	0,4	37,6	6,6	57,2
		3	36,9	0,9	34,8	9,4	52,2
	$c_s = 1$	92	68	97	46,4	83,2	25,8
		11,2	44,6	3,4	28,8	29,8	39,5
	$\|c_s\| > c'_s$	5	12,4	2,6	15	14,6	16,4
20%		10,6	27,2	3,1	31,8	30,9	35,0
	$\|c_s\| < c'_s$	3	19,6	0,2	38,6	2,2	57,8
		6,2	42,7	0,4	35,5	3	51,9

Tabela 3

Dados imputados

MD		Coef. de Afinidade AL	SL	CL	Coef. de Pearson AL	SL	CL
	$c_s = 1$	100	100	96,8	100	99,8	96,4
		0	0	7,2	0	0,4	1
	$\|c_s\| > c'_s$	0	0	3,2	0	0	3,6
10%		0	0	7,2	0	0	1
	$\|c_s\| < c'_s$	0	0	0	0	0,2	0
		0	0	0	0	0,4	0
	$c_s = 1$	100	100	96	100	100	96
		0	0	8,9	0	0	9
	$\|c_s\| > c'_s$	0	0	4	0	0	4
15%		0	0	8,9	0	0	9
	$\|c_s\| < c'_s$	0	0	0	0	0	0
		0	0	0	0	0	0
	$c_s = 1$	100	100	91,4	100	100	93,6
		0	0	19,2	0	0	10
	$\|c_s\| > c'_s$	0	0	8,6	0	0	6,4
20%		0	0	19,2	0	0	10
	$\|c_s\| < c'_s$	0	0	0	0	0	0
		0	0	0	0	0	0

Tabela 4

Dados imputados / Padrão Monótono

MD		Coef. de Afinidade AL	SL	CL	Coef. de Pearson AL	SL	CL
	$c_s = 1$	100	100	98,4	100	100	96,2
		0	0	3,58	0	0	8,49
	$\|c_s\| > c'_s$	0	0	1,6	0	0	3,8
10%		0	0	3,58	0	0	8,49
	$\|c_s\| < c'_s$	0	0	0	0	0	0
		0	0	0	0	0	0
	$c_s = 1$	100	100	98,2	100	100	97,6
		0	0	4,02	0	0	5,37
	$\|c_s\| > c'_s$	0	0	1,8	0	0	2,4
15%		0	0	4,02	0	0	5,37
	$\|c_s\| < c'_s$	0	0	0	0	0	0
		0	0	0	0	0	0
	$c_s = 1$	99,6	99,8	95,6	99,0	98,8	95,8
		0,55	0,45	8,74	1	0,84	8,29
	$\|c_s\| > c'_s$	0	0	4	0,4	0,4	3,8
20%		0	0	8,99	0,89	0,89	8,49
	$\|c_s\| < c'_s$	0,4	0,2	0,4	0,6	0,8	0,4
		0,55	0,45	0,55	0	0,84	0,55

Tabela 5

Na análise dos cinco casos $c_s = 1, |c_s| > c'_s$ e $|c_s| < c'_s$ significa que:

$c_s = 1$, a estrutura geral das duas classificações hierárquicas a serem comparadas é a mesma; isto quer dizer que as duas ultramétricas associadas são "ordinalmente equivalentes" (cada par de arvores de classificação hierárquica, têm a mesma estrutura "ordinal").

$|c_s| > c'_s$, a estrutura geral das duas classificações hierárquicas não é a mesma, mas as duas ultramétricas são "significativamente correlacionadas" (a 0,1%).

$|c_s| < c'_s$, as duas estruturas da classificação hierárquica são "significativamente diferentes".

Em cada uma das tabelas e para cada caso $c_s = 1$, $|c_s| > c'_s$, $|c_s| < c'_s$, os valores numéricos da primeira linha representam as médias das ocorrências e os da segunda linha representam os correspondentes desvios padrão.

Conclui-se, nos casos estudados, que o coeficiente de afinidade tem um comportamento mais robusto do que o coeficiente de correlação e que (globalmente) são obtidos melhores resultados após a utilização deste método de imputação múltipla, do em presença de dados omissos. Os melhores resultados são obtidos utilizando o método de imputação múltipla associado ao coeficiente de afinidade e com os critérios de agregação "average linkage" e "single linkage".

Bibliografia

[1]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, p.p.160-165.

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

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

[4]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, INE, pp.109-110.

[5]SILVA, A. L., BACELAR-NICOLAU, H. and SAPORTA, G. (2001) "Missing Data in Hierarchical Classification- a study" in G. Govaert, J. Janssen and N. Limnios (Eds.): *Applied Stochastic Models and Data Analysis,* UTC, p.p.697-698.**[6]**SILVA,A.L, SAPORTA, G. and BACELAR-NICOLAU (2002) "Dados omissos em Classificação Hierárquica de Variáveis e o Algoritmo Nipals" - IX Jornadas de Classificação e Análise de Dados/JOCLAD 2002, ESCS-IPL, pp. 42-43.

[7]SILVA,A.L, BACELAR-NICOLAU and SAPORTA, G.(2002)"Missing Data in Hierarchical Classification of Variables - a Simulation Study" in Classification Clustering and Data Analysis, Springer, p.p.121-128.

[8]STATISTICAL SOLUTIONS, Lda. (2001) *SOLAS for Missing Data Analysis*, 3.0. Cork, Ireland: Statistical Solutions.