Dados omissos em Classificação hierárquica de variáveis e o algoritmo NIPALS*

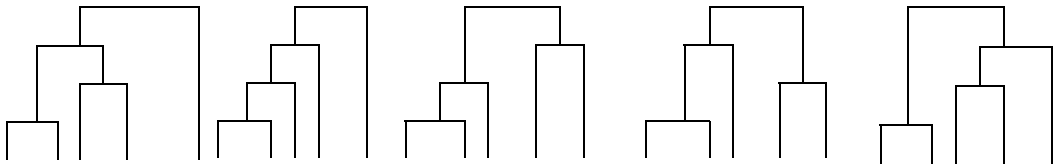
Ana Lorga da Silva[1], Gilbert Saporta[2], Helena Bacelar-Nicolau[3]

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

Palavras Chave: Dados omissos, Classificação Hierárquica, NIPALS

Neste trabalho comparamos os resultados obtidos em classificação hierárquica ascendente (particularmente classificação de variáveis) após a reconstituição de dados em falta recorrendo ao algoritmo NIPALS(Wold(1969)) (Nonlinear estimation by partial least squares), com outros métodos de reconstituição de dados que utilizamos anteriormente.
O Algoritmo NIPALS permite realizar uma análise em componentes principais com dados em falta sem suprimir as linhas que contêm dados omissos; e sem estimar os dados em falta, tal como descrito em Tenenhaus(1998).
Em trabalhos anteriores (por exemplo Silva(2001)) utilizámos matrizes de dados (originalmente completos) com distribuição multinormal (Saporta(1990)), às quais retiramos dados satisfazendo à 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 $X_{n\times p}$ e X_{miss} representa os valores omissos em $X_{n\times p}$ e $R=\lfloor R_{ij}\rfloor$ 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 foram gerados (matrizes 1000×5 - cem simulações para cada caso) com o objectivo de obter estruturas específicas representadas pelos dendrogramas seguintes:



Utilizou-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 Bacelar-Nicolau (2000) e o coeficiente de correlação de Bravais-Pearson

* 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 Análise Multivariada de Dados do CEAUL/FCUL.

$c_p = \frac{\sum_{i=1}^{n} \left(x_{ij} - \bar{x}^j\right)\left(x_{ij'} - \bar{x}^{j'}\right)}{s_{x^j} s_{x^{j'}}}$. Como métodos de agregação utilizaram-se critérios de agregação clássicos "average linkage", "single linkage", "complete linkage".

Retiraram-se dados a duas variáveis – 10%,15% e 20% de dados sobre o total da população (matriz $X_{1000\times5}$). Fez-se o estudo dos resultados obtidos utilizando os métodos "listwise" e de imputação – aplicando um algoritmo EM e um método OLS. Para comparar os modelos de classificação hierárquica utilizou-se o coeficiente de Spearman entre as matrizes das semelhanças ultramétricas associadas aos métodos de agregação referidos.

Concluiu-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 em presença de dados omissos do que após a imputação dos dados.

O algoritmo NIPALS permite também estimar os dados omissos utilizando a posteriori a formula de reconstituição habitual $\hat{x}_{ij} = \sum_{l=1}^{n_c} t_{li}\, p_{lj}$, tal que n_c é o número de componentes principais escolhido e t_{li}, p_{li} são respectivamente as coordenadas das componentes principais e dos vectores directores dos eixos principais. Utilizando o referido algoritmo reconstitui mos a matriz incompleta do seguinte modo:

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

Comparamos então as estruturas hierárquicas obtidas com matrizes reconstituídas desta forma, com as estruturas obtidas nos casos estudados anteriormente.

Nas simulações feitas até agora o comportamento do algoritmo NIPALS parece equivalente ao do método listwise, no que se refere ao coeficiente de afinidade mas em relação ao coeficiente de correlação no caso de 10% de dados omissos são obtidos melhores resultados com este algoritmo.

Bibliografia

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,160-165.

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