Método de detecção de outliers baseado em clustering: escolha do número de nuvens pelo critério AIC

C. M. Santos-Pereira
CEMAT/IST e Departamento de Engenharia Civil da Faculdade de Engenharia da Universidade do Porto - carlasp@fe.up.pt

A. M. Pires
Departamento de Matemática e CEMAT, IST, UTL -apires@math.ist.utl.pt

Palavras–chave: AIC, outliers, estimadores robustos, clustering

Em Santos-Pereira e Pires [8] apresentamos um método genérico de classificação com rejeição por indecisão e observações atípicas (outliers). Em Santos-Pereira e Pires [7] propusemos um método de detecção de outliers baseado em análise de clusters e utilização de distâncias tipo Mahalanobis com estimadores clássicos e robustos. Por forma a avaliar o desempenho deste último método, conduziu-se um estudo de simulação, com várias situações distribucionais, que envolveu a utilização de três métodos de clustering: K-means, pam (Kaufman e Rousseeuw [2]) e mclust (Banfield e Raftery [1]), e três pares de estimadores de localização-dispersão: clássicos, RCMD (Rousseeuw [5]) e OGK (Maronna e Zamar [3]). Uma das dificuldades encontradas na implementação deste método foi a escolha do número de clusters, k, do método de clustering e do estimador de localização-dispersão. Nesse momento para a escolha da melhor combinação eram experimentados vários valores para k (no mínimo 1 e no máximo uma valor que dependeria do número de observações e do número de variáveis) e escolhia-se aquele que minimizava (ver Sakamoto, Ishiguro e Kitagawa [6] e Ronchetti [4])

$$AIC = -2\sum_{i=1}^{n} \log \hat{f}(\mathbf{x}_i) + 2k\left(p + \frac{p(p+1)}{2}\right), \tag{1}$$

com

$$\hat{f}(\mathbf{x}) = \sum_{j=1}^{k} \frac{n_j}{n_T} f_N(\mathbf{x}; \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) \;\text{ e }\; n_T = \sum_{j=1}^{k} n_j, \tag{2}$$

onde

$$f_N(\mathbf{x}; \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) \text{ representa a f.d.p. da } N_p(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}). \tag{3}$$

Nesta comunicação discute-se a robustez do critério AIC aqui descrito, para a escolha do número de clusters, com base num conjunto de situações distribucionais.

Referências

[1] Banfield, J. e Raftery, A. (1992). Model-Based Gaussian and non-Gaussian. *Biometrics*, 49, 803-822.

[2] Kaufman, L. e Rousseeauw (1990). *Finding Groups in data: An introduction to Cluster Analysis*. New York: Wiley.

[3] Maronna, R. e Zamar, R. (2002). *Robust estimates of location and dispersion for high dimensional data sets*. Em *Technometrics*, 44, 307-317.

[4] Roncheti, E. (1997). Robusteness aspects of model choice. *Statistica Sinica*, 7, 327-338.

[5] Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. Em *Mathematical Statistics and Applications* (W. Grossman, G.Plufg, I. Vincze e W. Werz, eds.), Vol.B, 283-297. Dordrecht, Reidel.

[6] Sakamoto, Y., Ishiguro, M. e Kitagawa, G. (1988). *Akaike Information Criterion Statistics*. New York: Kluwer Academic Publishers.

[7] Santos-Pereira, C. e Pires, A. (2002). Detection of outliers in multivariate data: a method based on clustering and robust estimators. Em *Computational Statistics* (Härdle, W. e Rönz, B., eds.), 291-296, Heidelberg,Physica-Verlag.

[8] Santos-Pereira, C. e Pires, A. (2004). Método de classificação com rejeição por indecisão e observações atípicas. Em *Actas do XI Congresso da SPE* (Rodrigues, P., Rebelo, E. e Rosado, F., eds.), 595-604, Funchal.