Variable selection in finite mixture of linear mixed models using the EM and CEM algorithms

Luísa Novais[1] **and Susana Faria**[1]

[1] *University of Minho, Portugal*

Abstract

Variable selection is an important problem of any modeling study, involving the search for the simplest model that adequately describes the data, which assumes a great importance in the context of mixture models. However, the technological advances of the last decades have led to the use of data of large dimensions and of great complexity. As such, the classic variable selection methods become impracticable with the increasing size of the data, being computationally too demanding to be used in practice.

Therefore, in order to deal with the computational complexity, the need to develop new methods for variable selection has emerged in recent years. Among the new methods, methods based on penalizing functions have received great attention. These methods, unlike the classic methods, can be used in complex data problems since they allow the identification of the subset of the most relevant explanatory variables, by estimating the effect of the non-significant variables to be zero and, consequently, removing them from the model, thus drastically reducing the computational burden.

In this work we analyse the problem of variable selection in finite mixture of linear mixed models in the presence of a large number of explanatory variables. In order to do this, we compare the performance of a penalized likelihood approach for variable selection via the Expectation-Maximization (EM) and the Classification Expectation-Maximization (CEM) algorithms through a simulation study and a real data application.

Keywords

EM algorithm, CEM algorithm, Mixture models, Penalized likelihood, Variable selection.

Acknowledgements

The research of L. Novais was financed by FCT - Fundação para a Ciência e a Tecnologia, through the PhD scholarship with reference SFRH/BD/139121/2018.

References

[1] Celeux, G., Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14(3), 315–332.

[2] Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, 39(1), 1–38.

[3] Du, Y., Khalili, A., Nešlehová, J. G., Steele, R. J. (2013). Simultaneous fixed and random effects selection in finite mixture of linear mixed-effects models. *Canadian Journal of Statistics*, 41(4), 596–616.

[4] Khalili, A., Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, 102(479), 1025–1038.