Unsupervised clustering of upper level units in multilevel linear models*

Leonardo Grilli
Department of Statistics, Informatics, Applications 'G. Parenti'
University of Florence

Agnese Panzera
Department of Statistics, Informatics, Applications 'G. Parenti'
University of Florence
e-mail: a.panzera@disia.unifi.it

Carla Rampichini
Department of Statistics, Informatics, Applications 'G. Parenti'
University of Florence

1 Abstract

We discuss an extension of the unsupervised density-based approach of Azzalini and Torelli (2007), and Menardi and Azzalini (2014) to deal with the issue of clustering level 2 units in a simple multilevel linear model. In a density-based approach, clusters are identified by high-density regions of the density underlying the data. The proposed extension exploits the level 2 residuals as predictions of the unobserved random effects, and then identifies clusters by regions with high *prediction-based* kernel density estimate.

The performance of the proposed approach is evaluated by means of a simulation study and compared with the popular nonparametric maximum likelihood approach, where random effects are allowed to follow an arbitrary discrete distribution specified by a set of mass points and weights (Aitkin, 1999). Here, the crucial task of the selection of the number of clusters (which corresponds to the number of mass points) is addressed by adapting the Integrated Classification Likelihood criterion (McLachlan and Peel, 2000) to the multilevel setting.

The simulation experiment is performed for different scenarios, corresponding to different values of both the separation degree of level 2 units (defined as the proportion of the between-cluster variance of random effects) and the reliability (defined as the ratio between the variance of the random effects and the variance of the sample means of level 2 units).

The simulations show that the performance of the density-based approach heavily depends on the reliability of the level 2 residuals as predictions of random effects. In particular, simulations

*Presented at the second internal meeting of the FIRB ("Futuro in ricerca" 2012) project "Mixture and latent variable models for causal-inference and analysis of socio-economic data", Rome (IT), January 23-24, 2015

suggest to use the density-based approach when the reliability is high. However, beyond this case, the density-based approach could be still valuable as an exploratory tool.

References

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55:117–128.

Azzalini, A. and Torelli, N. (2007). Clustering via nonparametric density estimation. *Statistics and Computing*, 17:71–80.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

Menardi, G. and Azzalini, A. (2014). An advancement in clustering via nonparametric density estimation. *Statistics and Computing*, 24:753–767.