Classification Based on Mixed Type Numeric, Ordinal and Binary Longitudinal Data

Arnošt Komárek, Jan Vávra

Dept. of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

In different types of studies data are nowadays routinely gathered repeatedly over time on the same units leading to *longitudinal* or *panel* data. On top of that, multiple outcomes, both *numeric* and *categorical*, i.e., of a *mixed type*, are recorded at each measurement occasion leading to *multivariate mixed type longitudinal data*. An example of such a dataset which also motivates our research is *The European Union Statistics on Income and Living Conditions database* (EU-SILC). This is an instrument with the goal to collect timely and comparable cross-sectional and longitudinal multidimensional microdata on income, poverty, social exclusion and living conditions in the European Union, Iceland, Norway and Switzerland. The reference population includes all private households of respective countries and variables, which are collected annually via questionnaires, refer both to households and to individuals from the household.

Within the talk, we show two approaches towards unsupervised classification (clustering) based on such a type of data. Both approaches will utilize ideas of the Model Based Clustering (MBC, Fraley and Raftery, 2002) and a method developed earlier by us (Komárek and Komárková, 2013). That is, the model behind the clustering procedure will evolve from a multivariate (generalized) linear mixed model. Said differently, a sort of a mixed model will be assumed for longitudinal evolution of each outcome. A common joint distribution will be assumed for all random effects to model dependencies between different longitudinal outcomes obtained on a single subject. For numeric outcome, a classical linear mixed model (Laird and Ware, 1982) will be assumed. The two approaches will differ in a model being assumed for categorical outcomes. In the first approach, we will use a thresholding concept (Albert and Chib, 1993) to link a categorical (ordinal) outcome again to the linear mixed model. In the second approach, a certain form of a generalized linear mixed model will be assumed (for both nominal and ordinal categorical outcomes). Both estimation of unknown parameters, as well as the clustering procedure proceeds within the Bayesian framework and Markov chain Monte Carlo computation. The proposed methods will be illustrated on the analyzis of the Czech part of the EU-SILC database.

References

Albert J. H., Chib S. (1993) Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**(422), 669–679, DOI 10.2307/2290350.

Fraley C., Raftery A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**(458), 611–631, DOI 10.1198/016214502760047131.

Komárek A., Komárková L. (2013) Clustering for multivariate continuous and discrete longitudinal data. *The Annals of Applied Statistics*, **7**(1), 177–200, DOI 10.1214/12-AOAS580.

Laird N. M., Ware J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**(4). 963–974, DOI 10.2307/2529876.