

Clustering for multivariate continuous and discrete longitudinal data

Arnost Komarek

Charles University Prague

Multiple outcomes, both continuous and discrete are routinely gathered on subjects in longitudinal studies. We propose a model-based statistical method for clustering (classification) of subjects into a prespecified number of groups with a priori unknown characteristics on basis of repeated measurements of all longitudinal outcomes and show its implementation in the R package `mixAK`. We start by modeling the evolution of each outcome using the classical generalized linear mixed model (GLMM). Possible dependence between the values of different outcomes is captured by specifying a joint distribution of all random effects involved in the GLMM for each outcome. The basis for subsequent clustering is provided by assuming a heteroscedastic mixture of multivariate normal distributions in the random effect distribution. Mainly for computational reasons we base the inference on a Bayesian specification of the model and simulation based MCMC methodology. We also discuss evaluation of un certainty in the classification and adapt recently proposed methodologies for model comparison to explore optimal number of clusters. Everything will be exemplified on the analysis of real data using the routines from the contributed R package `mixAK`.