MODEL BASED CLUSTERING PROCEDURES FOR MULTIVARIATE MIXED TYPE LONGITUDINAL DATA

Arnošt Komárek [1]

[1] Dept. of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (e-mail: komarek@karlin.mff.cuni.cz)

ABSTRACT: The talk will present model based clustering methods to classify units based on so called multivariate mixed type longitudinal or panel data. The multivariate aspect of data points to the situation when more than one outcome is observed for each unit within a longitudinal study at each measurement occassion. The mixed type data then arise in situations when such multivariate outcomes are not necessarily of the same type, e.g., some of them are numeric, some of them categorical. The talk will provide an overview of clustering approaches for such data developed by author over past about 10 years, partly in cooperation with Lenka Komárková, Jan Vávra, Bettina Grün and Gertraud Malsiner-Walli.

KEYWORDS: classification, finite mixture, generalized linear mixed model, panel data.

1 Introduction

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 longitudinal data of a mixed type. An important area of interest is how to suitably model and analyze this kind of data if *unobserved heterogeneity* is suspected. In this case a statistical method is frequently required which forms homogeneous groups of similar units in the study population and develops a classification rule on how to classify not only available but perhaps also future units into those groups using the same type of data.

2 Notation

We are assuming that a dataset suitable for analysis by methods presented in this talk is composed of N units which we want to classify into $K > 1$ groups,

where K, the number of groups, is not necessarily known in advance. We are further assuming that the aim is to classify the units into the groups on basis of $R \geq 1$ of longitudinally gathered outcomes (being possibly of a mixed type). Let for $i = 1, \ldots, N$ and $r = 1, \ldots, R$, $\mathbf{Y}_{i,r} = \left(Y_{i,r,1}, \ldots, Y_{i,r,n_i}\right)$ denote a vector of the values of the rth outcome of the ith unit obtained at n_i occasions at times $\mathbf{t}_i = \left(t_{i,1}, \ldots, t_{i,n_i}\right)$. Further, let $\mathbf{v}_{i,r,1}, \ldots, \mathbf{v}_{i,r,n_i}$ be vectors of additional covariates that may explain random fluctuation of the outcomes $\mathbf{Y}_{i,r}$ we may want to adjust for in the classification procedure. These additional covariates are also allowed to be both numeric and categorical. They may include characteristics that are constant over time for a given unit but may also be time dependent. Furthermore, let $\mathcal{C}_{i,r} = \left\{\mathbf{t}_i, \mathbf{v}_{i,r,1}, \ldots, \mathbf{v}_{i,r,n_i}\right\}$ denote both the measurement times and the covariate values for the rth outcome of the ith panel member. Finally, let

$$\mathbf{Y}_i = \left(\mathbf{Y}_{i,1}, \ldots, \mathbf{Y}_{i,R}\right), \qquad \mathcal{C}_i = \left\{\mathcal{C}_{i,1}, \ldots, \mathcal{C}_{i,R}\right\}$$

denote all information (outcomes and covariate values) available for the ith unit that can be used in the data analysis and exploited for the classification of the units into one of the K groups.

By the fact that the outcomes are of a mixed type, we consider a situation that for different values of r, the elements of the vectors $\mathbf{Y}_{i,r}$ are possibly of a different type. Some of them might be *numeric*, some of them *counts*, *binary*, *ordinal* or general *multinomial*. This reflects a common practical situation of gathering multiple outcomes of different nature in one longitudinal study. Finally, it is obvious that the elements of the vectors $\mathbf{Y}_i$ cannot be assumed to be independent and some modelling of the dependence structure is a must with any realistic modelling approach.

3 Model based clustering

In the talk, several model based clustering approaches developed in Komárek & Komárková, 2013, Komárek & Komárková, 2014, Vávra & Komárek, 2022 and Vávra *et al.*, 2023 will be presented for data having the structure outlined in Section 2. The model behind all clustering procedures is a sort of the mixture of (generalized) linear mixed models. Unknown model parameters are estimated using the Bayesian approach and the Markov chain Monte Carlo (MCMC) methodology. Furthermore, related R software routines will also be discussed. Finally, we show on how to perform not only clustering of units based on available data but also how to classify a new observation into one of the clusters.

References

KOMÁREK, A., & KOMÁRKOVÁ, L. 2013. Clustering for multivariate continuous and discrete longitudinal data. *The Annals of Applied Statistics*, **7**(1), 177–200.

KOMÁREK, A., & KOMÁRKOVÁ, L. 2014. Capabilities of R package mixAK for clustering based on multivariate continuous and discrete longitudinal data. *Journal of Statistical Software*, **59**(12), 1–38.

VÁVRA, J., & KOMÁREK, A. 2022. Classification based on multivariate mixed type longitudinal data with an application to the EU-SILC database. *Advances in Data Analysis and Classification (early access)*.

VÁVRA, J., KOMÁREK, A., GRÜN, B., & MALSINER-WALLI, G. 2023. Clusterwise multivariate regression of mixed-type panel data. *Preprint on Research Square*.