Classification based on a joint modelling of multivariate mixed type longitudinal data

Jan Vávra[1], Arnošt Komárek[1]

[1] Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

E-mail for correspondence: vavraj@karlin.mff.cuni.cz

Abstract: At IWSM 2019, a poster was presented (Komárek et al., 2019) showing initial ideas towards modelling of mixed type panel or longitudinal data and usage of the model for unsupervised classification or clustering. For this contribution, the methodology has been enriched in several directions and applied to identify poverty and social exclusion temporal patterns of Czech households using data from The European Union Statistics on Income and Living Conditions database.

Keywords: Multivariate panel data; Ordinal response; Classification probabilities.

1 Data and research problems

Our research is motivated by *The European Union Statistics on Income and Living Conditions* database (EU-SILC) that collects annually gathered multidimensional data from European households (and also individuals living there). Primarily targeted income, poverty, social exclusion and living conditions obtained via questionnaire are described and represented by outcomes of various nature: *numeric* (e.g., income), *binary* (e.g., affordability of week holiday) and *ordinal* (e.g., level of a financial burden of housing). It is our primary aim to use such longitudinally gathered outcomes towards segmentation of households according to typical patterns of their temporal evolution.

To this end, we propose a statistical model capable of joint modelling of longitudinal outcomes of various nature (*numeric, binary, ordinal*) while taking potential dependencies as well longitudinal as among different outcomes obtained at each occasion into account. Consequently, we use the model within a Bayesian model based clustering (MBC) procedure to perform unsupervised classification of study units (households).

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

2 Joint model for mixed type longitudinal data

In general, we have data on n units/panel members (e.g., households) at our disposal containing $R \geq 1$ longitudinally gathered outcomes (being possibly of a mixed type). Let $\mathbf{Y}_i = \left(\mathbf{Y}_{i,1}^\top, \ldots, \mathbf{Y}_{i,R}^\top\right)^\top$ stand for a vector consisting of vectors of the values $\mathbf{Y}_{i,r}^\top$ of the rth outcome ($r = 1, \ldots, R$) of the ith unit ($i = 1, \ldots, n$) obtained at n_i occasions. Let $\mathcal{C}_i$ stand for available covariates (the times of measurements, possibly other explanatory variables) of i-th unit. Finally, let $g(\mathbf{y}_i; \mathcal{C}_i, \boldsymbol{\theta})$ represent the assumed distribution of the outcome vector $\mathbf{Y}_i$ which possibly depends on the covariates $\mathcal{C}_i$ and also on a vector $\boldsymbol{\theta}$ of unknown parameters. This distribution is built from the following model.

First, if the rth, $r = 1, \ldots, R$, longitudinal outcome vector $\mathbf{Y}_{i,r}$ is *ordinal* or *binary*, we will take a natural thresholding approach and will assume that each element of $\mathbf{Y}_{i,r}$, $Y_{i,r,j}$, $j = 1, \ldots, n_i$, is determined by corresponding element of a latent continuous variable $Y_{i,r,j}^\star$, which is covered by one of the intervals given by the set of thresholds $\boldsymbol{\gamma}_r$. In the following, denote these latent continuous counterparts by $\mathbf{Y}_{i,r}^\star$. In case the rth longitudinal outcome is directly observed as *continuous*, we set $\mathbf{Y}_{i,r}^\star = \mathbf{Y}_{i,r}$.

Further, for each $\mathbf{Y}_{i,r}^\star$, $r = 1, \ldots, R$, a classical linear mixed model (LMM) is assumed. That is, $\mathbf{Y}_{i,r}^\star = \mathbb{X}_{i,r}\boldsymbol{\beta}_r + \mathbb{Z}_{i,r}\mathbf{B}_{i,r} + \boldsymbol{\varepsilon}_{i,r}$, where $\mathbb{X}_{i,r}$ and $\mathbb{Z}_{i,r}$ are model matrices derived from the covariate information $\mathcal{C}_i$, $\boldsymbol{\beta}_r$ is a vector of unknown parameters. Further, $\mathbf{B}_{i,r}$ is a vector of random effects related to the rth longitudinal outcome and $\boldsymbol{\varepsilon}_{i,r}$ is an error term vector for which a classical normality assumption is exploited, i.e., $\boldsymbol{\varepsilon}_{i,r} \sim \mathcal{N}_{n_i}\left(\mathbf{0}, (\tau_r)^{-1}\mathbf{I}_{n_i}\right)$. The residual variance $(\tau_r)^{-1}$ is unknown.

Dependencies among the R longitudinal outcomes $\mathbf{Y}_{i,1}, \ldots, \mathbf{Y}_{i,R}$ are taken into account by considering a joint distribution for the random vector $\mathbf{B}_i = \left(\mathbf{B}_{i,1}^\top, \ldots, \mathbf{B}_{i,R}^\top\right)^\top$ which joins the random effect vectors from the mixed models for all R longitudinal measurements. Namely, a multivariate normal distribution is assumed here, i.e., $\mathbf{B}_i \sim \mathcal{N}_q\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$, where both the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ are unknown parameters.

Let $\boldsymbol{\zeta}$ be the set of unknown parameters of interest, i.e. $\boldsymbol{\zeta} = \{\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, where $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ stand for sets of parameters $\boldsymbol{\beta}_r$ and τ_r across all outcomes $r = 1, \ldots, R$. Then, the density of (latent) continuous outcomes of the i-th individual is given by integration of product of a multivariate normal density related to the LMM and a density of $\mathcal{N}_q\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$, which is known to lead to the density $g^\star\left(\mathbf{y}_i^\star; \mathcal{C}_i, \boldsymbol{\zeta}\right)$ of multivariate normal distribution. To obtain the density of actually observed outcomes $g(\mathbf{y}_i; \mathcal{C}_i, \boldsymbol{\theta})$ we need to separate $\mathbf{y}_i^\star$ into numeric (N) and ordinal (O) parts (including binary):

$$g\left(\mathbf{y}_i; \mathcal{C}_i, \boldsymbol{\zeta}, \boldsymbol{\gamma}\right) \;=\; \int t\left(\mathbf{y}_i^{\mathsf{O}} \,\middle|\, \mathbf{y}_i^{\mathsf{O},\star}; \boldsymbol{\gamma}\right) g^\star\left(\mathbf{y}_i^\star; \mathcal{C}_i, \boldsymbol{\zeta}\right) \mathrm{d}\mathbf{y}_i^{\mathsf{O},\star}, \tag{1}$$

where $t(\cdot|\cdot)$ represents the thresholding procedure.

3 Model based clustering

We first assume that K (the number of groups into which we intend to classify the units) is known in advance and $K \geq 2$. The classification proceeds by using the model outlined in Section 2 within the Bayesian model based clustering procedure. Hence, it is assumed that the overall model, f, is given as a finite mixture of certain group-specific models f_k, $k = 1, \ldots, K$. That is, $f(\mathbf{y}_i; \mathcal{C}_i, \boldsymbol{\theta}) = \sum_{k=1}^{K} w_k\, f_k(\mathbf{y}_i; \mathcal{C}_i, \boldsymbol{\psi}, \boldsymbol{\psi}^k)$, where $\mathbf{w} = \big(w_1, \ldots, w_K\big)^\top$ are the mixture weights (proportions of the groups in the population), $\boldsymbol{\psi}$ is a vector of unknown parameters common to all groups and $\boldsymbol{\psi}^k$, $k = 1, \ldots, K$, are vectors of group-specific unknown parameters. Hence the vector $\boldsymbol{\theta}$ of all unknown parameters is $\boldsymbol{\theta} \equiv \big\{\mathbf{w}, \boldsymbol{\psi}, \boldsymbol{\psi}^1, \ldots, \boldsymbol{\psi}^K\big\}$. Using the notation from previous section we set the group-specific density f_k to be the density g, however, depending on parameter $\boldsymbol{\zeta}^k$ elements of which ($\boldsymbol{\beta}_r^k$, τ_r^k, $\boldsymbol{\mu}^k$, $\boldsymbol{\Sigma}^k$) may (or may not) be group-specific, i.e. different value of the parameter is considered to be in different groups. For example, if we suppose that the groups differ only in the effects, then

$$f(\mathbf{y}_i; \mathcal{C}_i, \boldsymbol{\theta}) = \sum_{k=1}^{K} w_k\, g\Big(\mathbf{y}_i; \mathcal{C}_i, \underbrace{\boldsymbol{\tau}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}}_{\boldsymbol{\psi}}, \underbrace{\boldsymbol{\beta}^k, \boldsymbol{\mu}^k}_{\boldsymbol{\psi}^k}\Big).$$

Further, let $U_i \in \big\{1, \ldots, K\big\}$ be the unobserved allocation of the ith unit into one of the K groups. As it is usual with the mixture models, the group-specific distribution $f_k(\mathbf{y}_i; \mathcal{C}_i, \boldsymbol{\psi}, \boldsymbol{\psi}^k)$, $k = 1, \ldots, K$, can be viewed as a conditional distribution of the outcome $\mathbf{Y}_i$ given $U_i = k$ while the mixture weights $\mathbf{w}$ determine the distribution of the allocations, i.e., $\mathsf{P}(U_i = k) = w_k$, $k = 1, \ldots, K$. Classification of the ith unit can then be based on suitable estimates of the conditional individual allocation probabilities $p_{i,k}(\boldsymbol{\theta})$, $k = 1, \ldots, K$, calculated by the Bayes rule:

$$p_{i,k}(\boldsymbol{\theta}) = \mathsf{P}\big(U_i = k \,\big|\, \mathbf{Y}_i = \mathbf{y}_i; \mathcal{C}_i, \boldsymbol{\theta}\big) = \frac{w_k\, g\left(\mathbf{y}_i; \mathcal{C}_i, \boldsymbol{\psi}, \boldsymbol{\psi}^k\right)}{f(\mathbf{y}_i; \mathcal{C}_i, \boldsymbol{\theta})}. \qquad (2)$$

Calculation of such probabilities requires performing the integration (1), which is in fact the integration of multivariate normal density over an ($n_i \times$# ordinal outcomes)-dimensional interval, bounds of which are determined by the measured levels of ordinal outcomes $\mathbf{y}_i^{\mathsf{O}}$ and threshold parameter $\boldsymbol{\gamma}$. A method for computing such possibly highly dimensional integrals needs to be chosen carefully with respect to not only the precision but the computation time as well, since for one set of parameters $\boldsymbol{\theta}$ we need to use it at least $(n \times K)$-times. Moreover, we can limit ourselves to first j observations only, $j = 1, \ldots, n_i$, to capture the evolution of classification probability as the amount of available information increases.

To infer on the model parameters and to perform related classification a Bayesian approach was adopted and implemented in the R software in

combination with the C language and routines from the R package `mvtnorm` to calculate integrals (1). Monte Carlo Markov chain (MCMC) methods were used to obtain a sample from posterior distribution of $\boldsymbol{\theta}$ and consequently also from the posterior distribution of each of classification probabilities $p_{i,k}(\boldsymbol{\theta})$. Not only their posterior means but also their credible intervals were used for classification to quantify uncertainty in allocation of the study units into the groups.

4 Application

The methodology was applied to Czech households from the EU-SILC data while considering jointly a numeric outcome (disposable income), three binary outcomes – ability to afford (1) a weekly holiday, (2) regularly meat meals, (3) unexpected expenses, and also three ordinal outcomes – level of possession of (1) computer, (2) car, (3) financial burden of housing. Relevant results (sample of which is shown in Figure 1) will be presented during the conference.

References

Komárek, A., Vávra, J., Bína, V. (2019). Model based clustering using multivariate mixed type panel data. *Proceedings of the 34th International Workshop on Statistical Modelling*, **II**, 291 – 294.

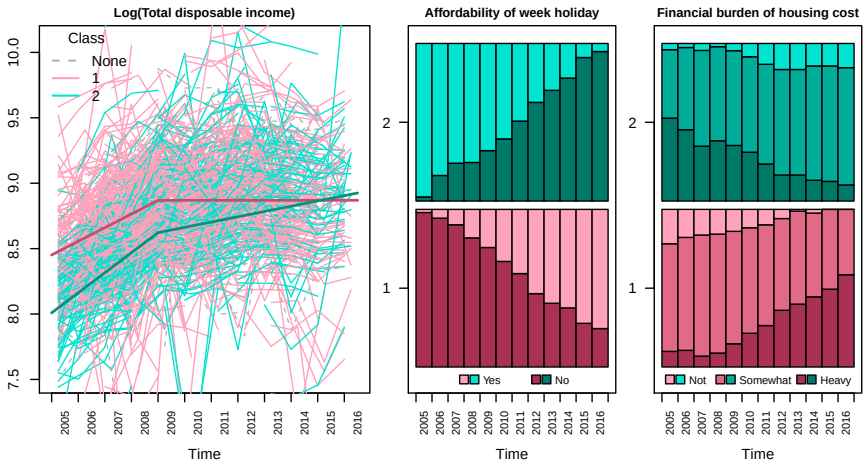

FIGURE 1. Longitudinal profiles of numeric, binary and ordinal outcomes of $n = 1000$ randomly selected Czech households. Bold curves on the left represent the estimated conditional expectation with change point in 2009 of response within $K = 2$ discovered groups. Categorical outcomes are presented by the proportions of categories in each year separately for the two discovered groups. Some households remain unclassified.