Identification of Temporal Patterns in Income and Living Conditions of Czech Households: Clustering Based on Mixed Type Panel Data from the EU-SILC Database

Jan Vávra[1], Arnošt Komárek[2]

Abstract. The EU-SILC database contains annually gathered rotating-panel data on a household level covering indicators of monetary poverty, severe material deprivation or low work household intensity. Data are obtained via questionnaires leading to outcome variables of diverse nature: *numeric*, *binary*, *ordinal* being gathered at each occasion in each household. Only limited number of approaches exist in the literature to analyze such *mixed-type* panel data. We present a statistical model for such type of data which is built on a thresholding approach to link binary or ordinal variables to their latent numeric counterparts. All, possibly latent, numeric outcomes are then jointly modelled using a multivariate version of the linear mixed-effects model. A mixture of such models is then used to model heterogeneity in temporal evolution of considered outcomes across households. A Bayesian variant of the Model Based Clustering (MBC) methodology is finally exploited to classify households into groups with similar evolution of indicators of monetary poverty, material deprivation or low work household intensity. The method is applied to socially-economic focused dataset of Czech households gathered in a time span 2005 – 2016.

Keywords: Multivariate panel data, Mixed type outcome, Model based clustering, Classification.

JEL Classification: C33, C38
AMS Classification: 62H30

1 Data and research problem

Throughout the EU states the poverty and social exclusion is measured using indicators of monetary poverty, severe material deprivation and very low work household intensity. Relevant data are gathered within *The European Union Statistics on Income and Living Conditions* project (EU-SILC, https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions. 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. Data are obtained via questionnaires leading to outcome variables of diverse nature: *numeric* (e.g., income), *binary* (e.g., affordability of paying unexpected expenses) and *ordinal* (e.g., level of ability to make ends meet). 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 diverse nature (*numeric, binary, ordinal*) while taking potential dependencies as well longitudinal as among different outcomes obtained at each occasion into account. This is a topic of Section 2. Consequently, we use the model within a Bayesian model based clustering (MBC) procedure to perform unsupervised classification of study units (households) into groups whose characteristics are not known in advance. This part of methodology is described in Section 3. The final Section 4 describes in detail the use of this methodology on the Czech subset of the EU-SILC dataset. The paper is finalized by conclusions in Section 5

2 Joint model for mixed type panel data

In general, we have data on n units/panel members (e.g., households) at our disposal containing $R \geq 1$ longitudinally gathered outcomes (e.g., income, affordability of week holiday and level of a financial burden of housing). Let $\mathbf{Y}_i = \left(\mathbf{Y}_{i,1}^\top, \ldots, \mathbf{Y}_{i,R}^\top\right)^\top$ stand for a vector consisting of all the values $\mathbf{Y}_{i,r} = (Y_{i,r,1}, \ldots, Y_{i,r,n_i})^\top$ of the rth

[1] Charles University, Faculty of Mathematics and Physics, Dept. of Probability and Mathematical Statistics, Sokolovská 83, 186 75 Praha 8, Czech Republic, vavraj@karlin.mff.cuni.cz
[2] Charles University, Faculty of Mathematics and Physics, Dept. of Probability and Mathematical Statistics, Sokolovská 83, 186 75 Praha 8, Czech Republic, komarek@karlin.mff.cuni.cz

outcome ($r = 1, \ldots, R$) of the ith unit ($i = 1, \ldots, n$) obtained at n_i occasions. Let C_i stand for available covariates (the times of measurements, possibly other explanatory variables) of i-th unit. Finally, let $g(\mathbf{y}_i; C_i, \boldsymbol{\theta})$ represent the assumed distribution of the outcome vector $\mathbf{Y}_i$ which possibly depends on the covariates C_i and also on a vector $\boldsymbol{\theta}$ of unknown parameters. It is assumed that this distribution is built from the following hierarchical model.

First, if the rth, $r = 1, \ldots, R$, longitudinal outcome vector $\mathbf{Y}_{i,r}$ is composed of *ordinal* or *binary* variables, we will take a natural thresholding approach (see, e.g., Dunson [1]) and will assume that each element of $\mathbf{Y}_{i,r}$, $Y_{i,r,j}$, $j = 1, \ldots, n_i$, is determined by corresponding latent continuous variable $Y^{\star}_{i,r,j}$, which is covered by one of the intervals given by the set of thresholds $\boldsymbol{\gamma}_r$. That is,

$$Y_{i,r,j} = l, \qquad \text{iff} \quad \gamma_{r,l} < Y^{\star}_{i,r,j} \leq \gamma_{r,l+1}, \qquad l = 0, \ldots, L_r, \tag{1}$$

where $\boldsymbol{\gamma}_r = \left(\gamma_{r,1}, \ldots, \gamma_{r,L_r}\right)^{\top}$ are unknown thresholds such that $-\infty = \gamma_{r,0} < \gamma_{r,1} < \cdots < \gamma_{r,L_r} < \gamma_{r,L_r+1} = \infty$. In the following, denote these latent continuous counterparts by $\mathbf{Y}^{\star}_{i,r}$. In case the rth longitudinal outcome is directly observed as *continuous*, we set $\mathbf{Y}^{\star}_{i,r} = \mathbf{Y}_{i,r}$.

Further, for each $\mathbf{Y}^{\star}_{i,r}$, $r = 1, \ldots, R$, a classical linear mixed model (LMM) is assumed. That is,

$$\mathbf{Y}^{\star}_{i,r} = \mathbb{X}_{i,r}\boldsymbol{\beta}_r + \mathbb{Z}_{i,r}\mathbf{B}_{i,r} + \boldsymbol{\varepsilon}_{i,r}, \tag{2}$$

where $\mathbb{X}_{i,r}$ and $\mathbb{Z}_{i,r}$ are model matrices derived from the covariate information 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 effects vector $\mathbf{B}_i = \left(\mathbf{B}^{\top}_{i,1}, \ldots, \mathbf{B}^{\top}_{i,R}\right)^{\top}$ 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.

Finally, 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 $g^{\star}\left(\mathbf{y}^{\star}_i; C_i, \boldsymbol{\zeta}\right)$ of multivariate normal distribution. To obtain the density of actually observed outcomes $g(\mathbf{y}_i; C_i, \boldsymbol{\theta})$ we need to separate $\mathbf{y}^{\star}_i$ into numeric (N) and ordinal (O) parts (including binary):

$$g\left(\mathbf{y}_i; C_i, \boldsymbol{\zeta}, \boldsymbol{\gamma}\right) = \int t\left(\mathbf{y}^{\mathsf{O}}_i \,\middle|\, \mathbf{y}^{\mathsf{O},\star}_i; \boldsymbol{\gamma}\right) g^{\star}\left(\mathbf{y}^{\star}_i; C_i, \boldsymbol{\zeta}\right) \mathrm{d}\mathbf{y}^{\mathsf{O},\star}_i, \tag{3}$$

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 (MBC, Fraley and Raftery [2]). 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, $f(\mathbf{y}_i; C_i, \boldsymbol{\theta}) = \sum_{k=1}^{K} w_k\, f_k(\mathbf{y}_i; C_i, \boldsymbol{\psi}, \boldsymbol{\psi}^k)$, where $\mathbf{w} = \left(w_1, \ldots, w_K\right)^{\top}$ 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 $\boldsymbol{\theta} \equiv \left\{\mathbf{w}, \boldsymbol{\psi}, \boldsymbol{\psi}^1, \ldots, \boldsymbol{\psi}^K\right\}$.

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 ($\boldsymbol{\beta}^k_r$, τ^k_r, $\boldsymbol{\mu}^k$, $\boldsymbol{\Sigma}^k$) 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 covariate effects, then

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

Further, let $U_i \in \left\{1, \ldots, K\right\}$ denote 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 $f_k(\mathbf{y}_i; C_i, \boldsymbol{\psi}, \boldsymbol{\psi}^k)$, $k = 1, \ldots, K$, 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:

$$p_{i,k}(\boldsymbol{\theta}) \;=\; \mathsf{P}\big(U_i = k \,\big|\, \mathbf{Y}_i = \mathbf{y}_i;\, C_i,\, \boldsymbol{\theta}\big) \;=\; \frac{w_k\, g\left(\mathbf{y}_i;\, C_i,\, \boldsymbol{\psi},\, \boldsymbol{\psi}^k\right)}{f(\mathbf{y}_i;\, C_i,\, \boldsymbol{\theta})}. \tag{4}$$

Calculation of such probabilities requires performing the integration (3), which is in fact the integration of multivariate normal density over an (n_i×# 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 (3). 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 to EU-SILC data

The methodology was applied to Czech households from the EU-SILC data while considering jointly

- logarithm of lowest income to make ends meet (to pay for its usual necessary expenses),
- affordability of paying unexpected expenses (required expense faced without help of anybody - only own resources used, e.g. surgery, a funeral, major repairs in the house, replacement of durables like washing machine, car) - binary variable:
  - Yes (household can afford unexpected expenses),
  - No (household cannot afford unexpected expenses),
- ability to make ends meet (respondent's feeling with respect to his household's income) - ordinal variable with six levels:
  1. With great difficulty,
  2. With difficulty,
  3. With some difficulty,
  4. Fairly easily,
  5. Easily,
  6. Very easily.

Each of the $n = 20\,299$ Czech households had been followed for $n_i = 4$ years induced by rotational design which replaces households that would exceed the 4-year limit with newly entering households. The set of households included in the following analysis consists of households entering the study after 2005 and leaving the study before 2016.

This year span covers also the economical crisis in 2009 that is suspected to influence the prosperity and social status of Czech households. In order to capture the possible change in the evolution of outcomes of interest we were discouraged from using simple linear trend and, therefore, were forced to use more flexible parametrization of time. Numeric (and the latent numeric counterparts of categorical outcomes) were modelled using B-spline parametrization of order 3 with knots at years 2008 and 2011. This piecewise-cubic parametrization leads to 6 coefficients (including the intercept term) and forms the crucial part of the fixed effect part of the model. Furthermore, it is extended by the weighted family size[1] that potentially could affect the outcomes of interest. Thus, the structure of fixed effects is in the form

$$\mathbf{X}_{i,j,r}^{\top}\boldsymbol{\beta}_r = \beta_{0,r} + \beta_{1,r}s_1(t_{i,j}) + \cdots + \beta_{5,r}s_5(t_{i,j}) + \beta_{6,r}w_{i,j},$$

where t_{ij} is the time (in years) that has passed since 2005 at which the i-th household was interviewed for j-th time, $s_1(t), \ldots, s_5(t)$ then corresponds to above mentioned spline parametrization at time t and w_{ij} is the corresponding weighted family size at that time.

[1] Each member of the household contributes to the family size by the following values: 1 for adult person in the role of the head of the family, 0.5 for other person older than 14 and 0.3 for person younger 14 years.

All three outcomes are linked through household-specific random intercept term $\mathbf{B}_{0,i}$ that follows trivariate zero-mean normal distribution with general covariance matrix $\mathbf{\Sigma}$. Combining fixed and random effects part we obtain the supposed mixed-effects model for each (latent) numeric outcome r:

$$Y^{\star}_{i,j,r} = B_{0,i,r} + \mathbf{X}^{\top}_{i,j,r}\boldsymbol{\beta}_r + \varepsilon_{i,j,r} = B_{0,i,r} + \beta_{0,r} + \beta_{1,r}s_1(t_{i,j}) + \cdots + \beta_{5,r}s_5(t_{i,j}) + \beta_{6,r}w_{i,j} + \varepsilon_{i,j,r}.$$

Moreover, each cluster $k = 1, \ldots, K$ is defined by its own set of fixed effects $\boldsymbol{\beta}^{(k)}$ for clustering purposes. Other parameters like matrix $\mathbf{\Sigma}$ or precisions of the error terms τ_r are considered to be the same among clusters and, therefore, do not help to differentiate them. Hence, the discovered clusters could be distinguished only by interpretation of differences between β coefficients and the shape of the spline curve they correspond to.

4.1 Results

Gibbs sampling procedure was applied to data on Czech households for several values of the total number of clusters K to determine which number will be the most suitable one. The choice of $K \in \{2, 3, 4, 5, 6\}$ can be supported by low values of deviance of the model defined as $D(\boldsymbol{\theta}; \mathbf{Y}_1, \ldots, \mathbf{Y}_n) = -2 \sum_{i=1}^{n} \log g\left(\mathbf{Y}_i; C_i, \boldsymbol{\theta}\right)$, which involves integration of all latent variables. Figure 1 presents estimates of the posterior distribution of deviance (viewed as a parametric function of $\boldsymbol{\theta}$). The deviance appears to decrease with higher value of K, with the exception of $K = 5$. Although, the choice of $K = 6$ seems to be the most beneficial, we should not blindly believe it. Let us explore the behaviour (and interpretation) of these clusters first.

Focusing on the numeric variable only we plot the estimated spline curves for each choice of K including $K = 1$ which corresponds to general evolution when no clustering was applied. Curves in Figure 2 are plotted for households of unit size (just one adult member) since the weighted family size should not be ignored as its effect may differ among clusters. In general it seems that the need for higher income has been increasing till 2009, after which this increasing trend has slowly stabilized or even begun to decrease.

As we begin to distinguish hidden clusters we always find a pair of clusters sharing the same shape of the evolution described above differing only in the level. It seems that sorting in more than two clusters in similar way is inefficient since curves of other clusters follow completely different shape. These clusters usually represent a very low percentage of households that behave extremely in some specific sense. For example violet cluster groups households with extreme growth of lowest income to make ends meet in years 2005 – 2008. However, the same cluster groups households with rapid decrease in the time span 2008 – 2010. Analogously, we could interpret even the blue and the brown cluster. This is not caused just by our chosen spline parametrization but mainly by the nature of the gathered data. We have to keep in mind, that households were questioned only four times in consecutive years. Therefore, for $K > 3$ we discover clusters of extreme behaviour which is usually limited to a certain time span. Moreover, we should not forget that discovered clusters differ in the evolution of categorical outcomes as well which may be even more extreme. Unless we are interested in these extremes, we should limit ourselves to lower count of clusters which still represent a considerable fractions of households and can characterize households in

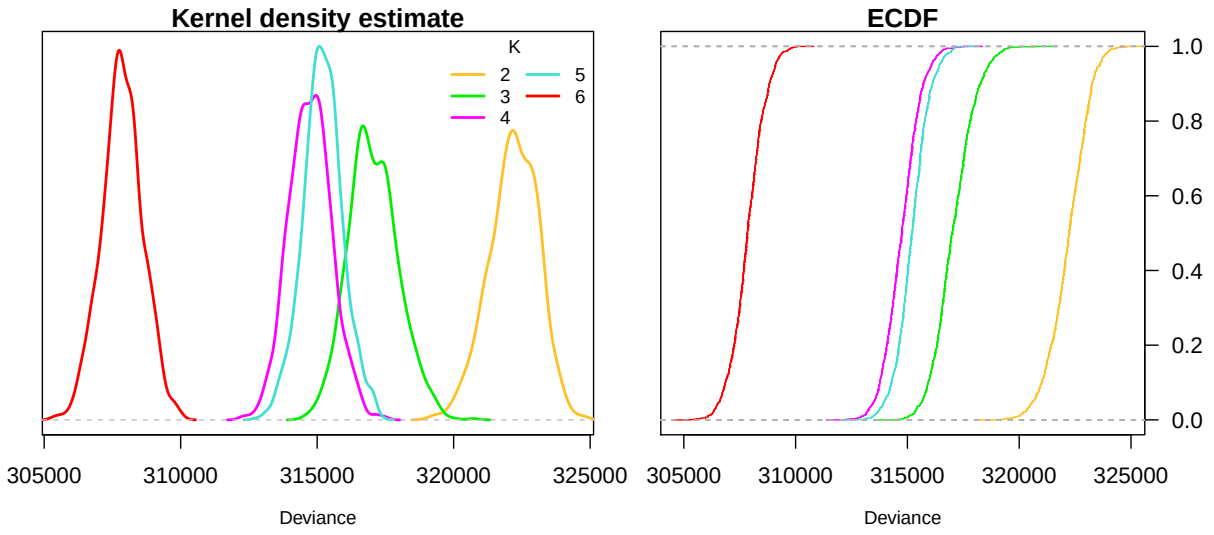

Figure 1: Estimated posterior distribution of deviances under several choices of number of clusters $K = 2, 3, 4, 5, 6$.

the whole time span. For this reason, we would recommend to use either $K = 2$ or $K = 3$.

Let us examine more the case of $K = 3$ which seems to be the most reasonable. Using Figures 2 (unit weighted family size) and 3 (weighted family size of 2) we can describe the found clusters in the following way:

2 This cluster (blue) represents about 7 % of households that used to have very high living standard until the crisis in year 2009 came after which households in this cluster have the lowest lowest income to make ends meet. The need for higher income does not rise with enlarging the family size as fast as in other two clusters, since the estimated $\beta_{6,1}^{(2)} \doteq 0.36$ corresponding to weighted family size is the lowest among all $\beta_{6,1}^{(k)}$ parameters. This is supported by a gap between blue and other spline curves in Figure 3 compared to 2. On the other hand, about one third of households in this cluster could not afford to pay unexpected expenses in 2005 – 2010. After 2010 almost all households in this cluster were prepared to unfavourable circumstances. Similarly, after 2010 they were more able to make ends meet as the proportion of households easily making ends meet increased which relates to the low living standard.

1,3 Cluster 1 (red) shares the same evolution of the lowest income to make ends meet as the cluster 3 (green). In both clusters it seems to have increased until the crisis came and after which the actually needed amount of income stabilizes (maybe slightly decreases). Households in this cluster differ from the third one in the much higher increase of the needed income for an additional family member: $\beta_{6,1}^{(1)} \doteq 0.61 > 0.41 \doteq \beta_{6,1}^{(3)}$, which is supported by the switch of the red and the green spline curves in Figures 2 and 3. Clusters 1 and 3 can be further distinguished by the evolution of proportions of categorical variables. In 2005 – 2011, the cluster 1 represents households with increasing probability of being able to pay for unexpected expenses, whereas, this probability decreases in cluster 3. After 2011, clusters 1 and 3 switched the monotonicity in the evolution of this probability. Similarly, cluster 1 represents households with increasing difficulties to make ends meet until 2011, after which year these difficulties fade away. The cluster 3 reflects this behaviour in the completely opposite way as around 2011 it consists of households having no difficulties to make ends meet.

Households were classified by the rule based on highest posterior density intervals (HPD). If the lower bound for the maximal posterior probability of belonging to cluster is higher than upper bounds for all other probabilities,

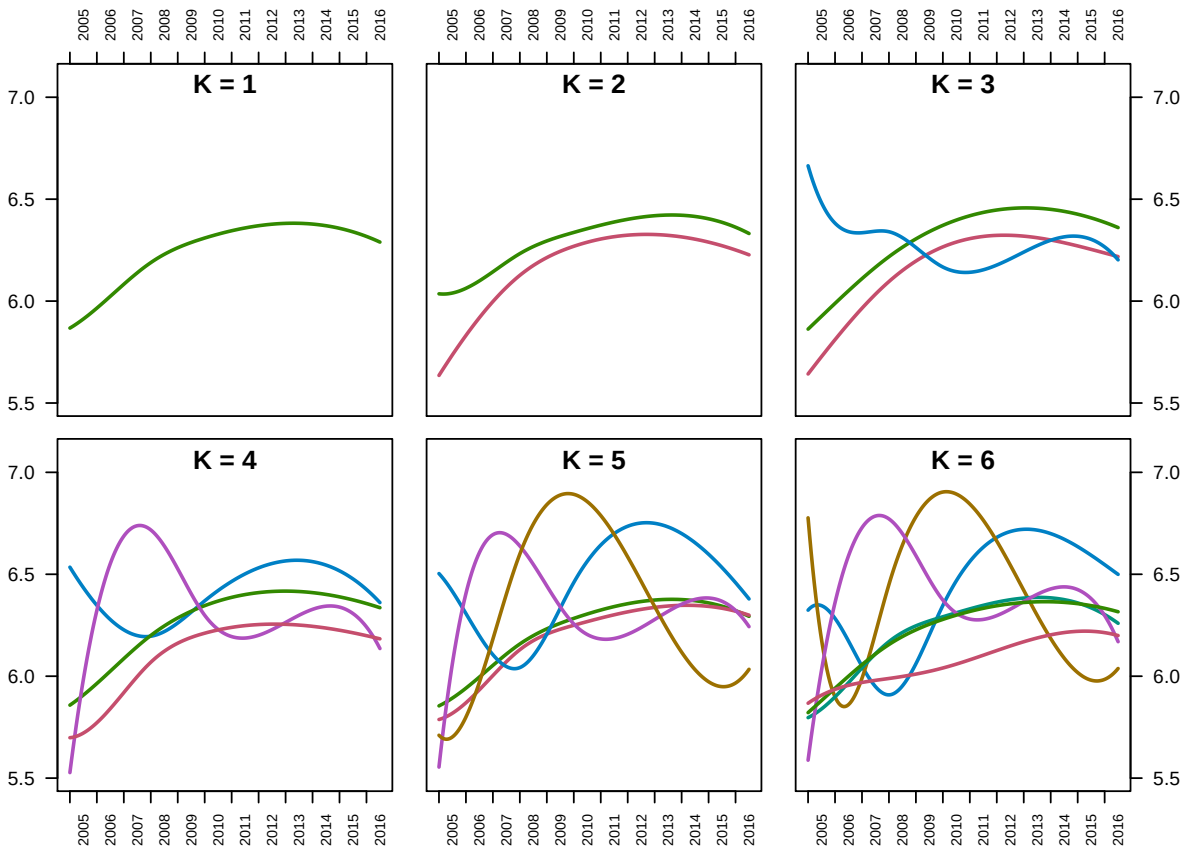

Figure 2: Spline curves for logarithm of the lowest income to make ends meet of households of unit weighted family size for different choice of the number of clusters K.

then the household is classified into the cluster that maximizes this probability. Otherwise, the household remains unclassified, which occurred in almost 23 % of cases.

5 Conclusions

We have developed a statistical model dealing with panel data of a mixed type. It was achieved by application of multivariate mixed effects model on numeric outcomes together with latent numeric outcomes which give rise to observed binary and ordinal outcomes. Mixture of such models was further used to discover different patterns in evolution of outcomes of interest. Using a fully Bayesian approach we were able to sort Czech households into three substantially different groups according to their ability to afford to pay for unexpected expenses, ability to make ends meet and the lowest needed income to do so.

Acknowledgements

This research was supported by the Czech Science Foundation (GAČR) grant 19-00015S.

References

[1] Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society, Series B*, **62**, 355 – 366.

[2] Fraley, C. and Raftery, A. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**, 611 – 631.

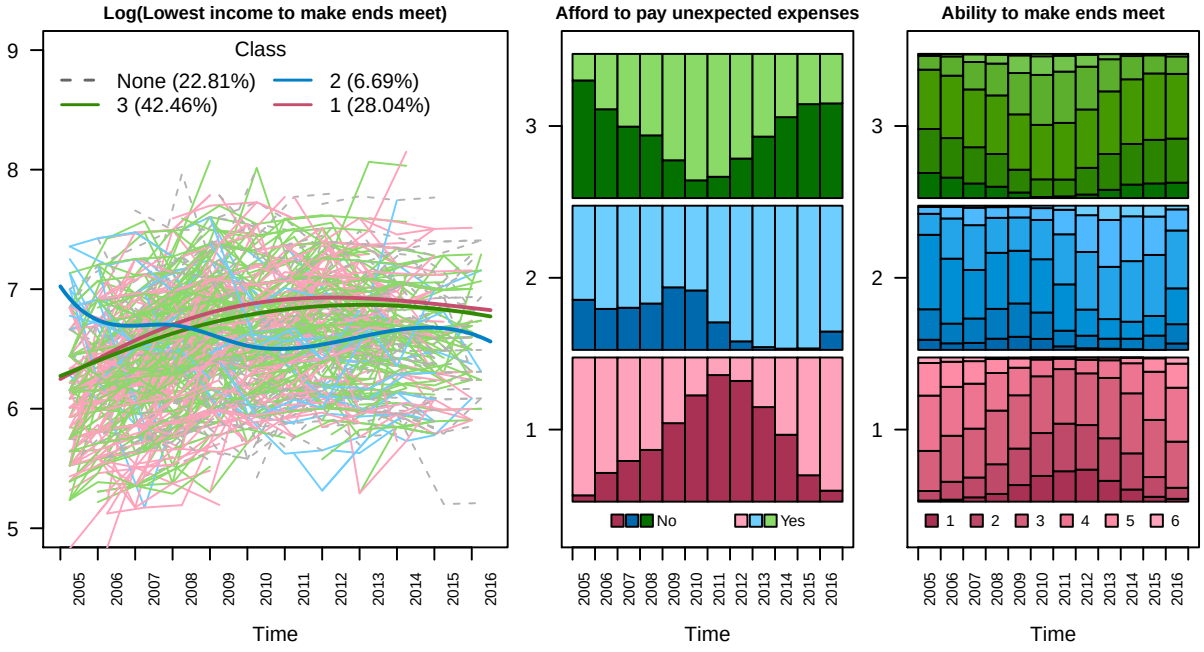

Figure 3: Longitudinal profiles of numeric, binary and ordinal outcomes of $n = 1\,000$ randomly selected Czech households. Bold curves on the left represent the estimated conditional expectation of response within $K = 3$ discovered groups for a household of weighted family size of 2. Categorical outcomes are presented by the proportions of categories in each year separately for the discovered groups. Some households remain unclassified.