Model based clustering using multivariate mixed type panel data

Arnošt Komárek[1], Jan Vávra[1], Vladislav Bína[2]

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

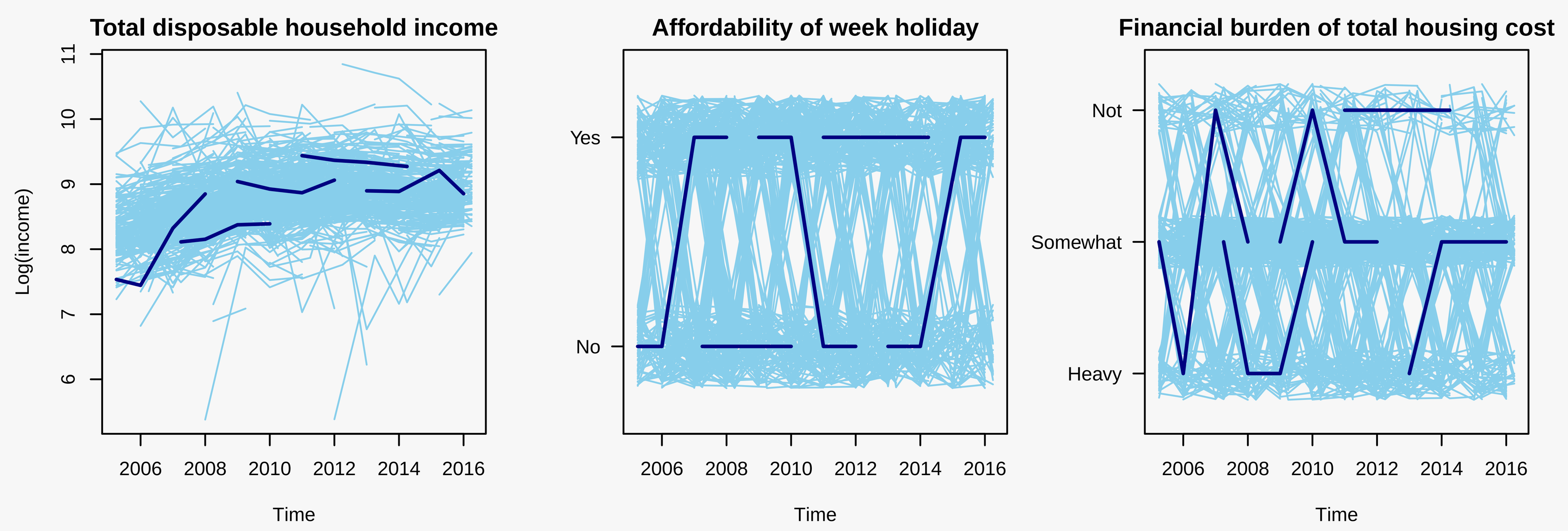
[2]Faculty of Management, University of Economics, Prague, Czech Republic

komarek@karlin.mff.cuni.cz, vavraj@karlin.mff.cuni.cz, vladislav.bina@vse.cz

EU-SILC dataset - mixed type data

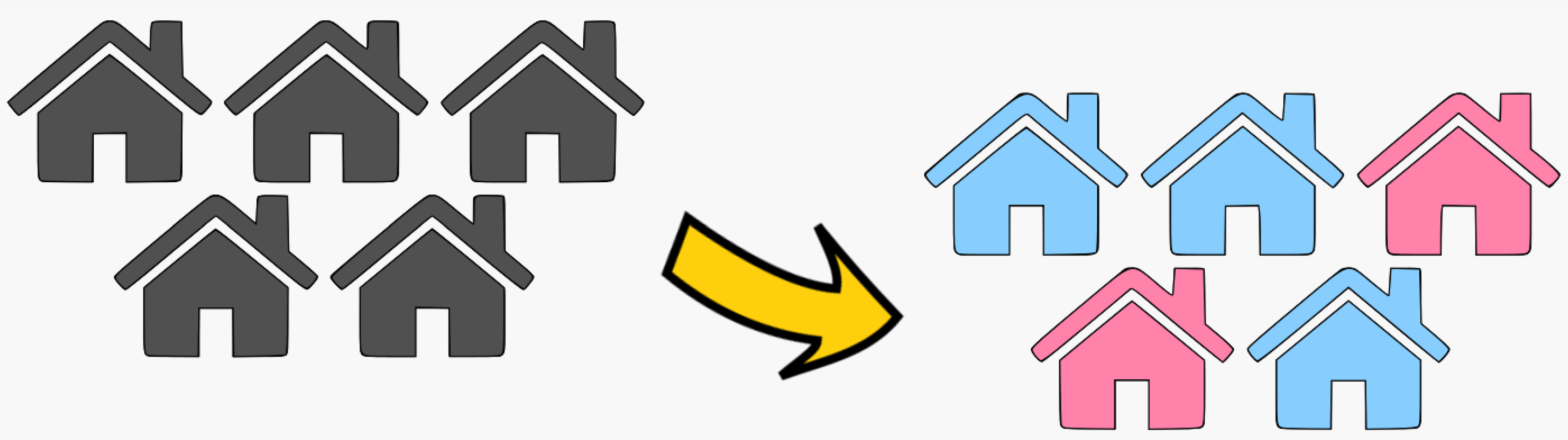
- EU-SILC = European Union - Statistics on Income and Living Conditions
- Longitudinal multidimensional data on income, poverty, social exclusion and living conditions measured on private households
- Annually gathered data via questionnaires targeted on both households and individuals living there
- Available data: $n = 29\,292$ households from the Czech Republic (years 2005 – 2016)
- Outcomes
  - Numeric - Total disposable household income, ...
  - Binary - Affordability of week holiday away from home, ...
  - Ordinal - Financial burden of total housing cost, ...

  } mixed type data
- Explanatory variables:
  - year, region, level of urbanization, dwelling type, weighted family size, ...



Research goals

- To discover unobserved heterogeneity in various socio-economic characteristics.
- To identify hidden groups of similar longitudinal evolution of these characteristics.
- To partition households into these groups to determine the level of social-economic status.
- To construct a set of general rules for classification of households.
- To uncover poverty and social exclusion temporal patterns.



Notation

- household $i \in \{1,\ldots,n\}$, visit number $j \in \{1,\ldots,n_i\}$, outcome $r \in \{1,\ldots,R\}$
- $Y^r_{i,j}$ - measured value of an outcome $\rightsquigarrow \boldsymbol{Y}^r_i, \boldsymbol{Y}_i$
- $t_{i,j}$ - time of the measurement $\rightsquigarrow \boldsymbol{t}_i$
- C_i - all explanatory information known to household i

Model based clustering (Banfield and Raftery, 1993)

- K: number of unobserved groups (initially assumed to be known)
- $U_i \in \{1,\ldots,K\}$: latent indicators of membership to one of the K groups for each household i
- $0 < w_k = \mathsf{P}[U_i = k]$: unknown probability of group $k \in \{1,\ldots,K\}$
- $f_k\left(\boldsymbol{y}_i;\ C_i, \boldsymbol{\psi}, \boldsymbol{\psi}^{(k)}\right)$: PDF of the probabilistic model for $\boldsymbol{Y}_i$ when household i belongs to group k
  - $\boldsymbol{\psi}$: parameters of the probabilistic model that are common to all groups
  - $\boldsymbol{\psi}^{(k)}$: group-specific parameters of model for group k
- $\boldsymbol{\theta} = \left(\boldsymbol{w}, \boldsymbol{\psi}, \boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(K)}\right)$: unknown parameters of interest
- Mixture likelihood: $$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} w_k f_k\left(\boldsymbol{Y}_i;\ C_i, \boldsymbol{\psi}, \boldsymbol{\psi}^{(k)}\right)$$
- By Bayes rule: $$p_{i,k}(\boldsymbol{\theta}) = \mathsf{P}\left[U_i = k \,|\, \boldsymbol{Y}_i = \boldsymbol{y}_i;\ C_i, \boldsymbol{\theta}\right] = \frac{w_k f_k\left(\boldsymbol{y}_i;\ C_i, \boldsymbol{\psi}, \boldsymbol{\psi}^{(k)}\right)}{\sum_{\ell=1}^{K} w_\ell f_\ell\left(\boldsymbol{y}_i;\ C_i, \boldsymbol{\psi}, \boldsymbol{\psi}^{(\ell)}\right)}$$
- Classification rule: $$\widehat{U}_i := k \iff k = \underset{\ell \in \{1,\ldots,K\}}{\arg\max}\ \widehat{p_{i,\ell}(\boldsymbol{\theta})}$$
- Estimation: Bayesian approach and MCMC methods (Gibbs sampling)
- Software: implemented in R using programming language C

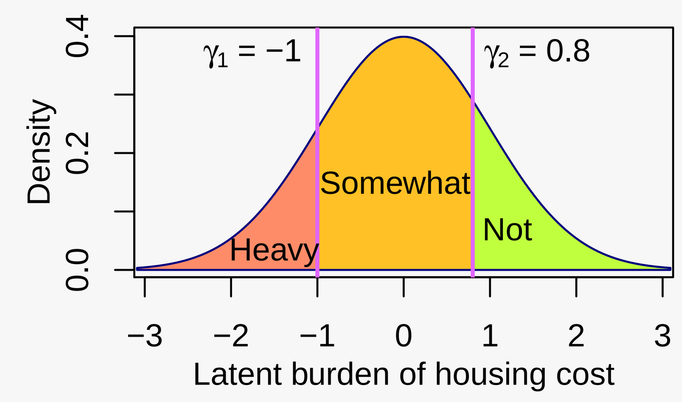
Models for single outcomes (in certain group)

- Numeric - Linear Mixed-effects Model (LMM)
  - Model formula: $Y^{\mathrm{N}}_{i,j} \sim \mathsf{N}\left(\left(\boldsymbol{X}^{\mathrm{N}}_{i,j}\right)^\top \boldsymbol{\beta}^{\mathrm{N}} + \left(\boldsymbol{Z}^{\mathrm{N}}_{i,j}\right)^\top \boldsymbol{b}^{\mathrm{N}}_i,\ \left(\tau^{\mathrm{N}}\right)^{-1}\right)$
  - Random effects: $\boldsymbol{b}^{\mathrm{N}}_i \overset{\text{iid}}{\sim} \mathsf{N}\left(\boldsymbol{\mu}^{\mathrm{N}}, \boldsymbol{\Sigma}^{\mathrm{NN}}\right)$
  - For example: $\log\left(Y^{\mathrm{N}}_{i,j}\right) = b^{\mathrm{N}}_i + \beta^{\mathrm{N}} \cdot t_{i,j} + \varepsilon^{\mathrm{N}}_{i,j}$
- Binary + Ordinal - thresholded latent numeric variable following LMM
  - $Y^{\mathrm{O}}_{i,j} = \ell \in \{1,\ldots,L\}$ (L ordered levels)
  - $Y^{\mathrm{O},\star}_{i,j} \sim$ LMM with $\boldsymbol{\beta}^{\mathrm{O}}, \boldsymbol{b}^{\mathrm{O}}_i$ and $\tau^{\mathrm{O}} = 1$
  - Observed $Y^{\mathrm{O}}_{i,j}$ determined by set of thresholds $\boldsymbol{\gamma}$

$$-\infty = \gamma_0 < -1 = \gamma_1 < \gamma_2 < \cdots < \gamma_{L-1} < \gamma_L = \infty$$

$$\begin{aligned} Y^{O}_{i,j} &= \text{Heavy} \iff Y^{\mathrm{O},\star}_{i,j} \le \gamma_1 \\ Y^{O}_{i,j} &= \text{Somewhat} \iff \gamma_1 < Y^{\mathrm{O},\star}_{i,j} \le \gamma_2 \\ Y^{O}_{i,j} &= \text{Not} \iff \gamma_2 < Y^{\mathrm{O},\star}_{i,j} \end{aligned}$$

  - In general $Y^{\mathrm{O}}_{i,j} = \ell \iff \gamma_{\ell-1} < Y^{\mathrm{O},\star}_{i,j} \le \gamma_\ell$



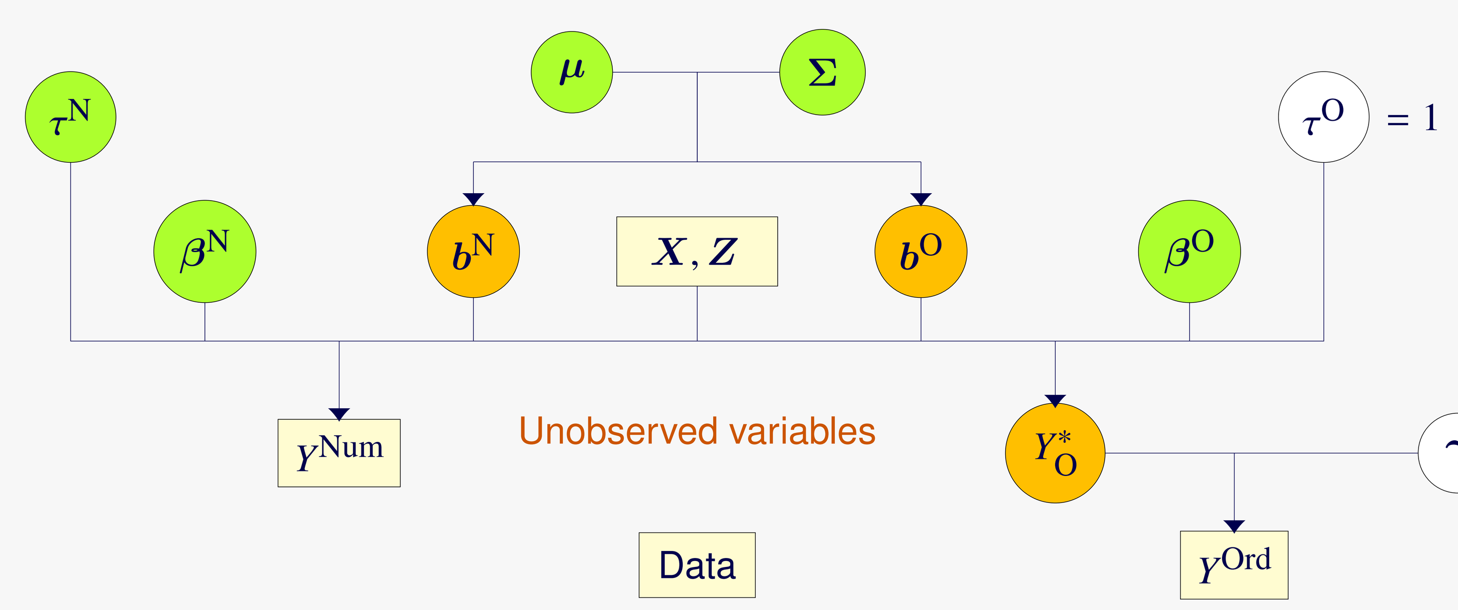
Joint modelling (in certain group)

- Outcomes cannot be considered to be independent of each other.
- Individual models are joined through joint distribution of random effects: $$\boldsymbol{b}_i = \begin{pmatrix} \boldsymbol{b}^{\mathrm{N}}_i \\ \boldsymbol{b}^{\mathrm{O}}_i \end{pmatrix} \overset{\text{iid}}{\sim} \mathsf{N}\left(\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}^{\mathrm{N}} \\ \boldsymbol{\mu}^{\mathrm{O}} \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}^{\mathrm{NN}} & \boldsymbol{\Sigma}^{\mathrm{NO}} \\ \boldsymbol{\Sigma}^{\mathrm{ON}} & \boldsymbol{\Sigma}^{\mathrm{OO}} \end{pmatrix}\right)$$
- Leading to a certain dependency structure:

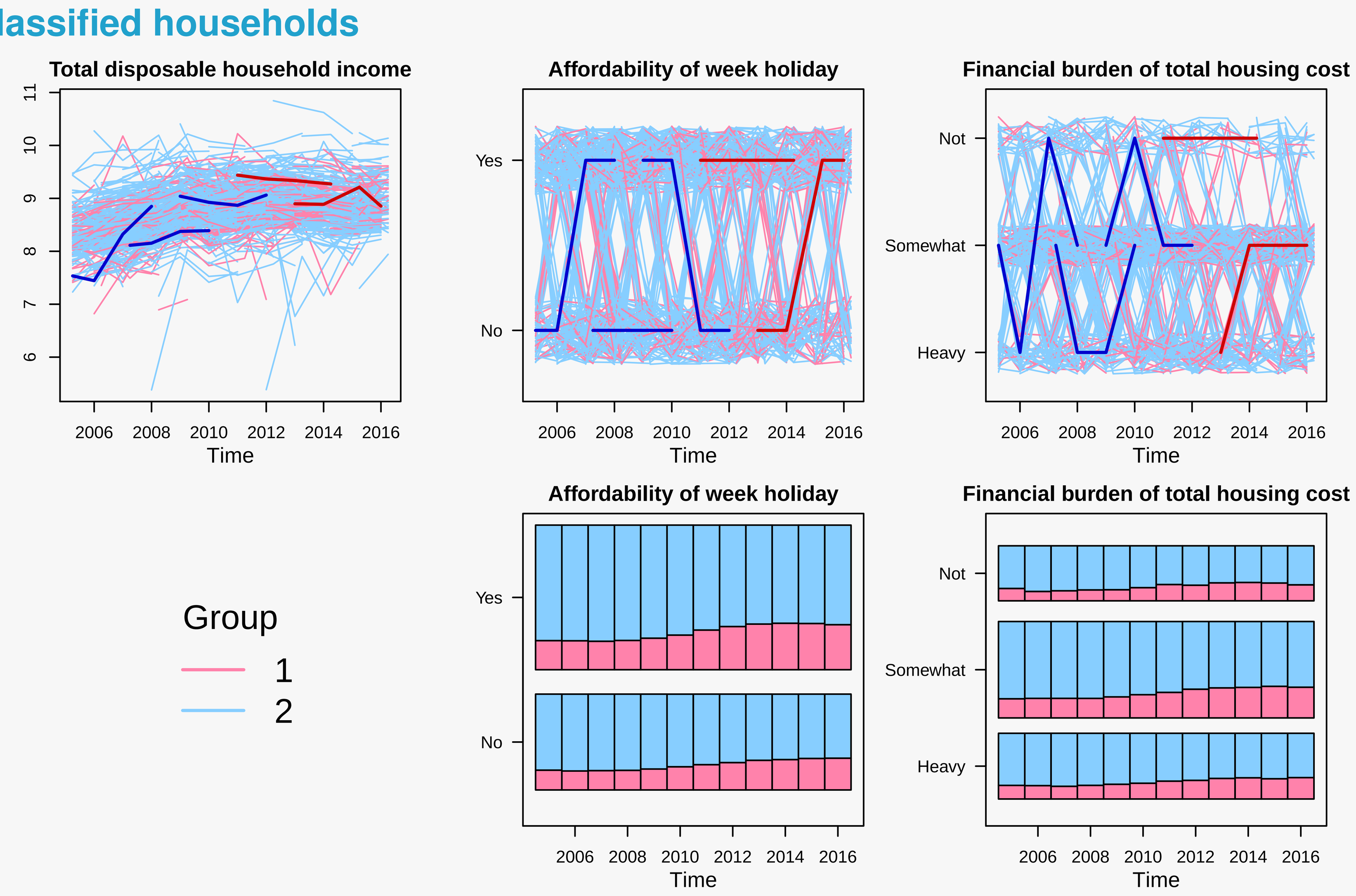
$$\operatorname{var}\left[\begin{pmatrix} Y^{\mathrm{N}}_{i,j} \\ Y^{\mathrm{O},\star}_{i,j} \end{pmatrix} \middle| C_{i,j}\right] = \begin{pmatrix} \left(\tau^{\mathrm{N}}\right)^{-1} + \left(\boldsymbol{Z}^{\mathrm{N}}_{i,j}\right)^\top \boldsymbol{\Sigma}^{\mathrm{NN}} \boldsymbol{Z}^{\mathrm{N}}_{i,j} & \left(\boldsymbol{Z}^{\mathrm{N}}_{i,j}\right)^\top \boldsymbol{\Sigma}^{\mathrm{NO}} \boldsymbol{Z}^{\mathrm{O}}_{i,j} \\ \left(\boldsymbol{Z}^{\mathrm{O}}_{i,j}\right)^\top \boldsymbol{\Sigma}^{\mathrm{ON}} \boldsymbol{Z}^{\mathrm{N}}_{i,j} & 1 + \left(\boldsymbol{Z}^{\mathrm{O}}_{i,j}\right)^\top \boldsymbol{\Sigma}^{\mathrm{OO}} \boldsymbol{Z}^{\mathrm{O}}_{i,j} \end{pmatrix}$$

- Group-specific parameters: $\boldsymbol{\psi}^{(k)} = \left(\boldsymbol{\beta}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}\right)$

Hierarchical Bayesian joint model for numeric and ordinal variable



Classified households



References

Banfield, J., D. and Raftery, A., E. (1993). Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, **49**(3), 803 – 821.

R Core Team (2019). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org.