FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

Mgr. Jan Vávra

Department of Probability and Mathematical Statistics

Classification Based on Longitudinal Data of a Mixed Type

6th June 2019

Outline

1. Motivation
2. Model based clustering
3. Modelling the mixed type longitudinal data
4. Application to EU-SILC data
5. Summary

EU-SILC longitudinal dataset

- $n = 29\,292$ households in the Czech Republic
- monitored periodically every year (2005 - 2016)
- each household at most for 4 years
- many outcomes (housing, living conditions, social status, ...) of different types (numeric, binary, ordinal, categorical)
- = mixed type data
- explanatory variables: time, location, level of urbanisation, type of dwelling, family size, other personal information, ...

Goal

Divide households into several groups of similar characteristics according to measured data.

Longitudinal data of mixed type

Household ID	Year	Weighted family size	HY020	HS040	HS050	HS060	HS090	HS110	HS140
1008400	2005	1.3	6228.79	1	1	1	2	1	2
	2006	1.3	7214.65	2	2	2	2	1	2
	2007	1.3	7566.56	1	1	1	1	1	2
	2008	1.5	7039.23	1	1	1	1	1	2
4329500	2014	1.5	5665.90	1	1	1	3	1	1
	2015	1.5	6362.58	1	1	1	3	1	2
	2016	1.5	6553.61	1	1	1	1	1	2

Classification - longitudinal data of a mixed type

Let’s apply classification in $\mathbb{R}^p$!

Wait!

- Different number of questionnaires per household?
- Different time periods?
- Distances between categorical variables?
- Can a suitable metric dealing with these problems be found?
- And if so, how can we interpret such results?

Model based clustering

- Origins: Banfield and Raftery (1993)
- Outcomes: $\boldsymbol{Y}_i, i = 1, \ldots, n$
- K models: $f_k\left(\boldsymbol{y}_i; \boldsymbol{x}_i, \psi, \psi^{(k)}\right), k = 1, \ldots, K$
- Group probabilities: $\boldsymbol{w} = (w_1, \ldots, w_K), 0 < w_k < 1, w_1 + \cdots + w_K = 1$
- Parameters of interest: $\boldsymbol{\theta} = \left(\boldsymbol{w}, \psi, \psi^{(1)}, \ldots, \psi^{(K)}\right)$
- Mixture likelihood:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} w_k f_k\left(\boldsymbol{y}_i; \boldsymbol{x}_i, \psi, \psi^{(k)}\right) \right)$$

Model based clustering - latent variable approach

- Conditional distribution point of view.
- $U_i \in \{1, \ldots, K\}$ latent (hidden, unobserved) variables
- $\boldsymbol{Y}_i$ generated from group $k \iff U_i = k$
- $\mathrm{P}\left[U_i = k\right] = w_k$
- $\boldsymbol{Y}_i | U_i = k \sim f_k$
- By Bayes Theorem:

$$p_{i,k}(\theta) = \mathrm{P}\left[U_i = k | \boldsymbol{Y}_i = \boldsymbol{y}_i; \boldsymbol{x}_i, \theta\right] = \frac{w_k f_k\left(\boldsymbol{y}_i; \boldsymbol{x}_i, \psi, \psi^{(k)}\right)}{\sum_{j=1}^{K} w_j f_j\left(\boldsymbol{y}_i; \boldsymbol{x}_i, \psi, \psi^{(j)}\right)}$$

- Estimation: MLE (EM-algorithm) or Bayesian approach

Model based clustering - Bayesian approach

- Choose models f_k for given outcome $\boldsymbol{Y}_i$.
- $$\left.\begin{array}{l}\text{Model parameters } \boldsymbol{\theta} = \left(\boldsymbol{w}, \psi, \psi^{(1)}, \ldots, \psi^{(K)}\right) \\ \text{Latent variables } U_i, i = 1, \ldots, n\end{array}\right\} \text{viewed as random}$$
- Choose suitable prior distributions.
- Construct an MCMC algorithm (Robert and Casella, 2004).
  - Gibbs sampling
- Generate ”a sample” from posterior distributions.
- Estimate parameters based on the obtained ”sample”.

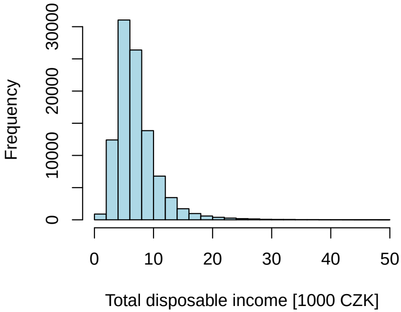
Details in NMST431 or NMTP539.

Numeric outcome

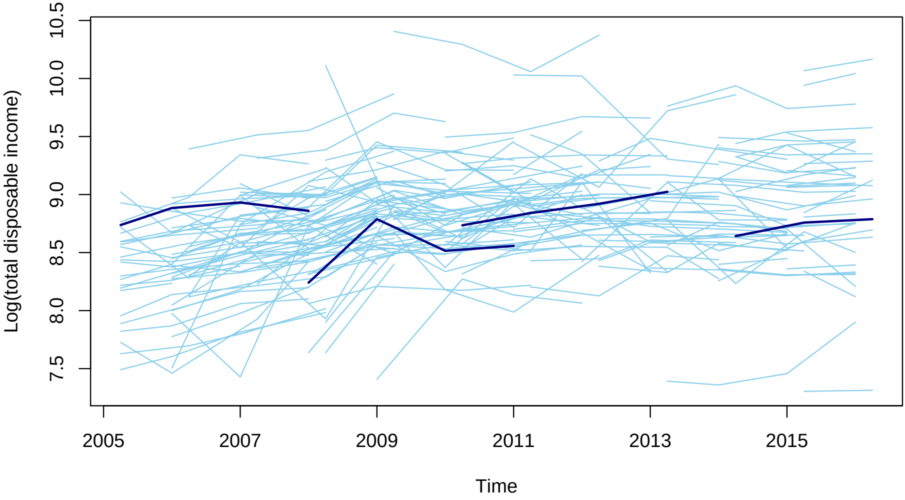
HY020 = Total disposable household income

Used model:

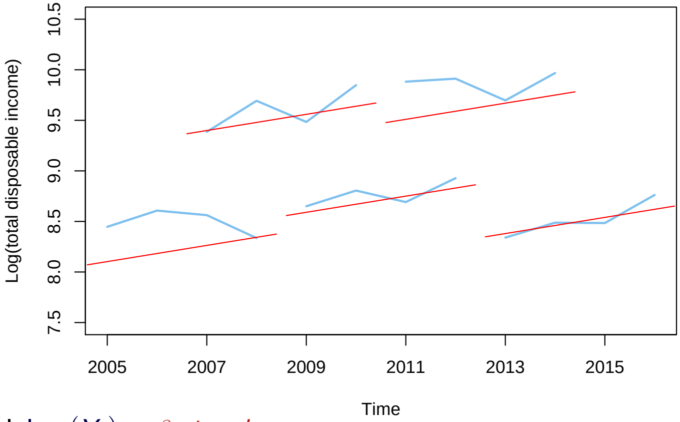
- Random effects models (LMM)
  - Laird and Ware (1983)
  - $\mathrm{N}\left(\boldsymbol{X}_i^\top \beta + \boldsymbol{Z}_i^\top \boldsymbol{b}_i, \sigma^2\right)$
  - $\boldsymbol{b}_i \sim \mathrm{N}\left(\boldsymbol{\mu}, \Sigma\right)$



Numeric outcome



Numeric outcome - random effects model



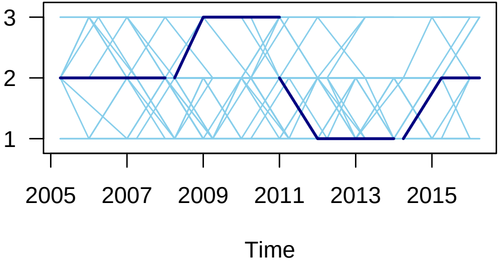
- Model: $\log\left(Y_{ij}\right) = \beta \cdot t_{ij} + b_i + \varepsilon_{ij}$
- Model error: $\varepsilon_{ij} \sim \mathrm{N}\left(0, \tau^{-1}\right)$
- Random effects (shift of y-axis): $b_i \sim \mathrm{N}\left(\mu, \Sigma\right)$
- Fixed effect (slope): β

Binary + Ordinal variable

HS140 = Financial burden of the total housing cost

$L = 3$ ordered categories

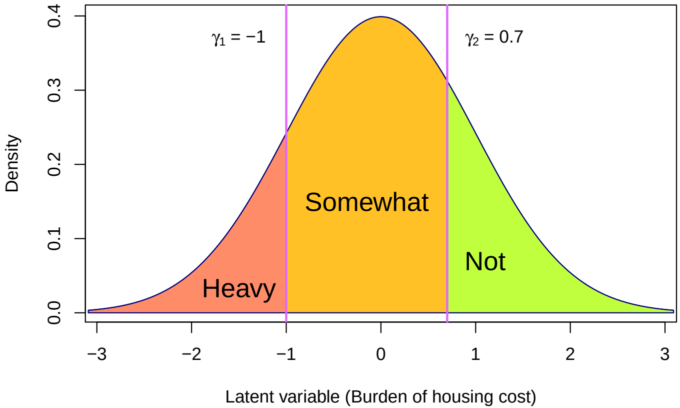
- 3 = Not a burden at all,
- 2 = Somewhat a burden,
- 1 = A heavy burden.



Used model

- Latent variable modelling: $Y|Y^*$
  - Y^* latent numeric outcome
  - Thresholding by $-\infty = \gamma_0 < \gamma_1 < \gamma_2 < \cdots < \gamma_{L-1} < \gamma_L = \infty$

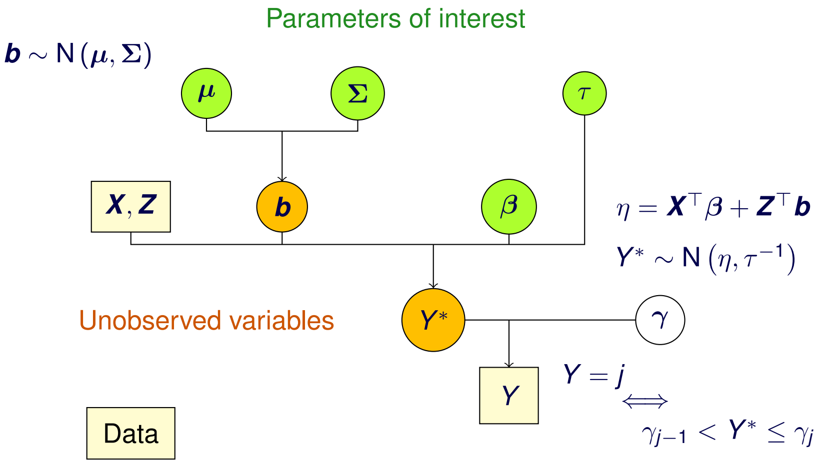
Binary + Ordinal variable - latent variable modelling



- Fixed threshold: $\gamma_1 = -1$
- Estimate other thresholds: $\gamma_2, \ldots$
- $Y_{ij}^{*} | \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}; \boldsymbol{b}_i \sim \mathrm{N}\left(\boldsymbol{X}_{ij}^{\top}\beta + \boldsymbol{Z}_{ij}^{\top}\boldsymbol{b}_i, 1\right)$

$$
\begin{aligned}
Y_{ij} = 1 &\iff & Y_{ij}^{*} &\leq \gamma_1 \\
Y_{ij} = 2 &\iff \gamma_1 < & Y_{ij}^{*} &\leq \gamma_2 \\
Y_{ij} = 3 &\iff \gamma_2 < & Y_{ij}^{*} &
\end{aligned}
$$

Diagram for modelling single outcome



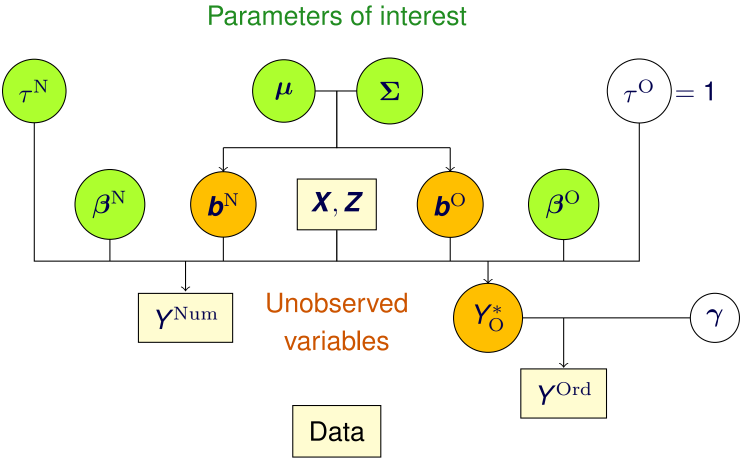
Joint modelling

How can we combine these models for several outcomes of different type?

- Independent models?
- Dependent outcomes!
- Several LMM can be combined into Multivariate LMM (Hendersen, 1984).
- Numeric variables + latent numeric variables
- Random effects from joint multivariate normal distribution.

$$\boldsymbol{b}_i = \begin{pmatrix} \boldsymbol{b}_i^N \\ \boldsymbol{b}_i^B \\ \boldsymbol{b}_i^O \end{pmatrix} \sim \mathrm{N}_p \left(\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}^N \\ \boldsymbol{\mu}^B \\ \boldsymbol{\mu}^O \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}^{NN} & \boldsymbol{\Sigma}^{NB} & \boldsymbol{\Sigma}^{NO} \\ \boldsymbol{\Sigma}^{BN} & \boldsymbol{\Sigma}^{BB} & \boldsymbol{\Sigma}^{BO} \\ \boldsymbol{\Sigma}^{ON} & \boldsymbol{\Sigma}^{OB} & \boldsymbol{\Sigma}^{OO} \end{pmatrix} \right)$$

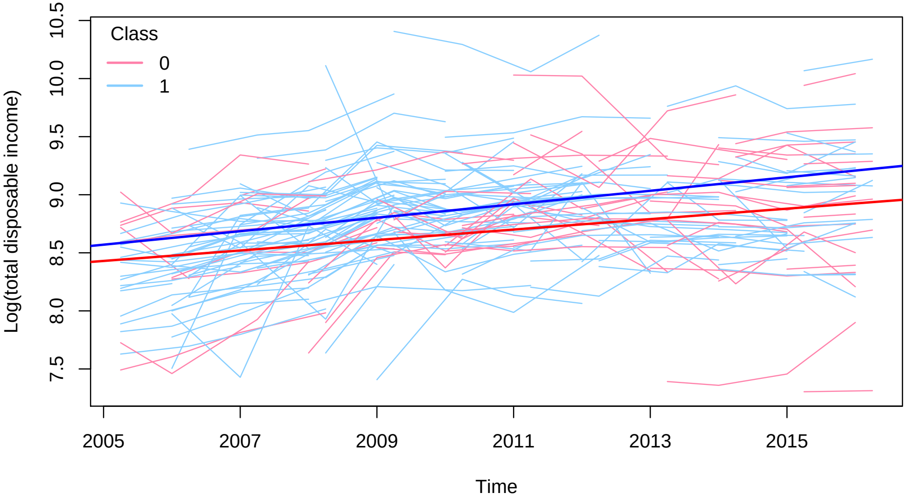
Diagram - numeric and ordinal variable jointly



Classification in joint model

- Use model based clustering.
- Each model $f_k, k = 1, \ldots, K$ of the same form.
- Choose class-specific parameters, e.g.
  - $\boldsymbol{\psi}^{(k)} = \left(\boldsymbol{\beta}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}\right),$
  - $\boldsymbol{\psi} = (\boldsymbol{\gamma}, \boldsymbol{\tau}).$
- Use Bayesian methodology to get estimates.

Log total disposable income - classified



Summary + Related contributions

- Mixed type data
- Model based clustering
- Combination of known methods
- Clustering in joint models in the world:
  - 1996 - Verbeke and Lesaffre - mixtures of LMM,
  - 2008 - Grün and Leisch - multivariate mixed type data (`flexmix`, EM-algorithm),
  - 2009 - Villarroel et al. - several numeric outcomes,
  - 2014 - Komárek, Komárková - MMGLMM (`mixAK`, no ordinal, Bayesian),
  - 2017 - Proust-Lima et al. - several outcomes of the same type (`lcmm`, ML).

Further development

- Add general categorical variables.
- Robustness to violation of normality assumption.
- Selection of important regressors.
- GLMM as an alternative to latent variable modelling.

Thank you for your attention.

Literature references

[1] A. Agresti and B. Lang, J. A proportional odds model with subject-specific effects for repeated ordered categorical responses. *Biometrika*, 80(3):527–534, 09 1993.

[2] D. Banfield, J. and E. Raftery, A. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3):803–821, 1993.

[3] C. Fraley and E. Raftery, A. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

[4] B. Grün and F. Leisch. FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, 28(4):1–35, 2008.

[5] F. Hedeker and J. Mermelstein, R. A multilevel thresholds of change model for analysis of stages of change data. *Multivariate Behavioral Research*, 33(4):427–455, 1998.

[6] R. Hendersen, C. *Applications of Linear Models in Animal Breeding*. University of Guelph, 1984.

[7] A. Komárek and L. Komárková. Capabilities of R package mixAK for clustering based on multivariate continuous and discrete longitudinal data. *Journal of Statistical Software*, 59(12):1–38, 2014.

[8] M. Laird, N. and H. Ware, J. Random-effects models for longitudinal data. *Biometrics*, 38:963–74, 01 1983.

[9] P. McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):109–142, 1980.

[10] C. Proust-Lima, V. Philipps, and B. Liquet. Estimation of extended mixed models using latent classes and latent processes: The R package lcmm. *Journal of Statistical Software*, 78(2):1–56, 2017.

[11] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Second edition. Springer-Verlag, New York, 2004.

[12] G. Verbeke and E. Lesaffre. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91(433):217–221, 1996.

[13] L. Villarroel, G. Marshall, and A. Baron. Cluster analysis using multivariate mixed effects models. *Statistics in Medicine*, 28:2552–65, 09 2009.