FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

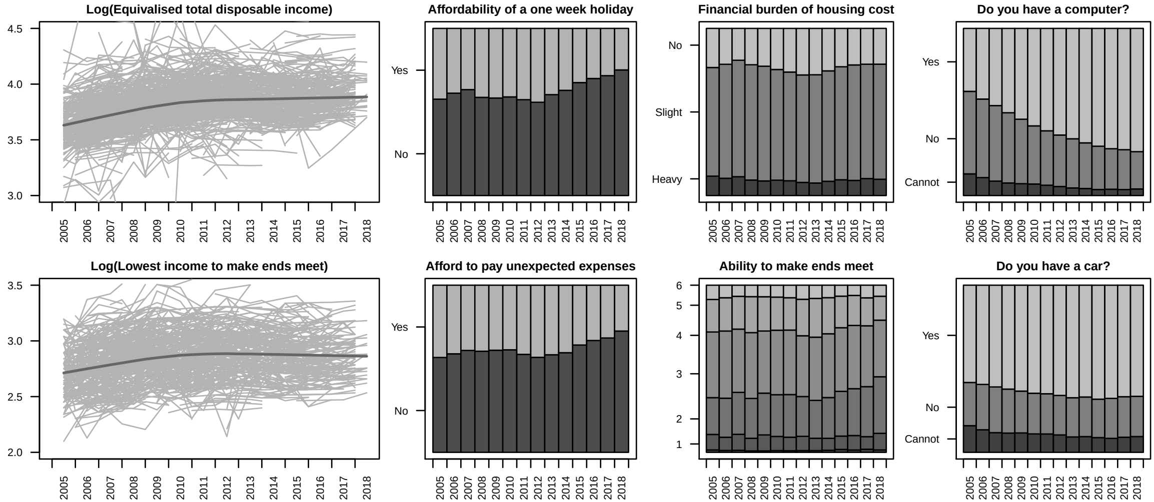
Jan Vávra

Department of Probability and Mathematical Statistics

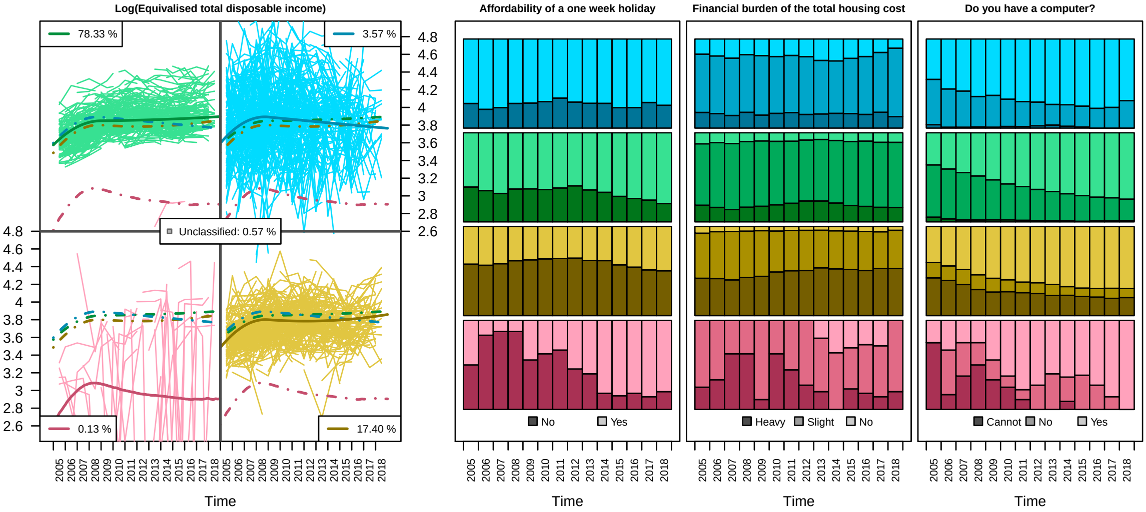
GLMM Based Clustering of Multivariate Mixed Type Longitudinal Data

15th June 2022 ROBUST 2022

EU-SILC - Outcomes of interest



Clustering into $G = 4$ groups



Random effects models

Each outcome is modelled by predictor η consisting of

$$\left.\begin{array}{l}\text{Fixed part - } \boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta} \\ \text{Random part - } \boldsymbol{Z}_{ij}^{\top}\boldsymbol{b}_i\end{array}\right\} \quad \eta = \boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}^{\top}\boldsymbol{b}_i$$

In case of EU-SILC application:

- Fixed:
  - intercept
  - time - quadratic spline with 3 equidistant knots
  - level of urbanisation
  - highest education level achieved in the household
  - presence of a student
  - presence of a baby
  - no interaction terms
- Random: random intercept $\boldsymbol{Z}_{ij}^{\top}\boldsymbol{b}_i = b_{i,0} \stackrel{\text{iid}}{\sim} \mathrm{N}\left(0, \sigma_b^2\right), i = 1, \ldots, n$

GLMM for different types of outcomes

Numeric outcome

- Linear Mixed-effects Model (LMM)
  - $Y_{ij}^N \mid \eta^N \sim \mathsf{N}\left(\eta^N, \tau^{-1}\right)$

Binary outcome

- Logistic Regression with random effects
  - $\mathsf{P}\left[Y^B = 1 \left| \eta^B \right.\right] = \mathrm{logit}^{-1}\left(\eta^B\right) = \dfrac{e^{\eta^B}}{1 + e^{\eta^B}}$

Ordinal outcome of K levels

- Ordinal Logistic Regression (OLR)
  - $p_k = \mathsf{P}\left[Y^O > k | \eta^O, \mathbf{c}\right] = \mathrm{logit}^{-1}\left(\eta^O - c_k\right)$
  - $q_k = \mathsf{P}\left[Y^O = k | \eta^O, \mathbf{c}\right] = p_{k-1} - p_k$
  - ordered intercepts $-\infty = c_0 < c_1 < \cdots < c_{K-1} < c_K = \infty$

General categorical outcome of K levels

- Multinomial Logistic Regression (MLR)
  - η_k^C specific for each level k

$$\mathsf{P}\left[Y^C = k | \eta_1^C, \ldots, \eta_{K-1}^C\right] = \\ = \mathrm{softmax}_k(\boldsymbol{\eta}^C) = \frac{e^{\eta_k^C}}{1 + \sum_{k'=1}^{K-1} e^{\eta_{k'}^C}}$$

Joint modelling

Models for individual outcomes

- Numeric: $Y_i^N \sim \text{LMM}\left(\beta_N, \boldsymbol{b}_i^N, \tau\right)$
- Ordinal: $Y_i^O \sim \text{ORL}\left(\beta_O, \boldsymbol{b}_i^O, \boldsymbol{c}\right)$
- Binary: $Y_i^B \sim \text{LR}\left(\beta_B, \boldsymbol{b}_i^B\right)$
- Categorical: $Y_i^C \sim \text{MLR}\left(\beta_{k,C}, \boldsymbol{b}_i^C\right)$

Join individual models through joint distribution of random effects

$$\boldsymbol{b}_i = \begin{pmatrix} \boldsymbol{b}_i^N \\ \boldsymbol{b}_i^B \\ \boldsymbol{b}_i^O \\ \boldsymbol{b}_i^C \end{pmatrix} \overset{\text{iid}}{\sim} \text{N}\left(\mathbf{0} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}^{NN} & \boldsymbol{\Sigma}^{NB} & \boldsymbol{\Sigma}^{NO} & \boldsymbol{\Sigma}^{NC} \\ \boldsymbol{\Sigma}^{BN} & \boldsymbol{\Sigma}^{BB} & \boldsymbol{\Sigma}^{BO} & \boldsymbol{\Sigma}^{BC} \\ \boldsymbol{\Sigma}^{ON} & \boldsymbol{\Sigma}^{OB} & \boldsymbol{\Sigma}^{OO} & \boldsymbol{\Sigma}^{OC} \\ \boldsymbol{\Sigma}^{CN} & \boldsymbol{\Sigma}^{CB} & \boldsymbol{\Sigma}^{CO} & \boldsymbol{\Sigma}^{CC} \end{pmatrix} \right), \quad i = 1, \ldots, n$$

Joint probability density function

$$h\left(\mathbf{y}_i;\,\mathcal{C}_i,\zeta\right) = p\left(\mathbb{Y}_i = \mathbf{y}_i \,|\, \boldsymbol{\beta}_N, \tau, \boldsymbol{\beta}_B, \boldsymbol{\beta}_O, \boldsymbol{c}, \boldsymbol{\beta}_C, \boldsymbol{\Sigma};\, \mathcal{C}_i\right) =$$

$$= \int \prod_{r=1}^{R} \prod_{j=1}^{n_i} \exp\left\{ \ell^{\text{type}(r)}\left(Y_{i,j}^r \,|\, \boldsymbol{b}_i^r, \zeta_r;\, \mathcal{C}_{i,j}\right)\right\} \cdot (2\pi)^{-\frac{\dim b_i}{2}} \, |\boldsymbol{\Sigma}|^{-1} \exp\left\{-\frac{1}{2} b_i^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{b}_i\right\} \mathrm{d}\boldsymbol{b}_i$$

- Requires methods for numerical evaluation of the integral

→ solved by *Laplacian approximation* or generally *Adaptive Gaussian Quadrature*

  - Pinheiro and Chao (2006)

Model based clustering

To identify different patterns:

- suppose G latent groups each following model given by $h\left(\boldsymbol{y}_i; \mathcal{C}_i, \zeta^{(g)}\right)$
- $\zeta^{(g)}$ consists of
  - $\boldsymbol{\psi}$ - parameters common to all latent groups
  - $\boldsymbol{\psi}^{(g)}$ - group-specific parameters (fixed effects for evolution in time, ...)
- group allocation indicators $U_i \in \{1, \ldots, G\}$
- marginal clustering probabilities $0 < w_g := \mathsf{P}[U_i = g] < 1, \quad w_1 + \cdots + w_G = 1$
- complete set of unknown parameters $\boldsymbol{\theta} = \left\{\boldsymbol{w}, \boldsymbol{\psi}, \boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(G)}\right\}$
- mixture distribution for $\boldsymbol{Y}_i$:

$$f(\boldsymbol{y}_i | \mathcal{C}_i; \boldsymbol{\theta}) = \sum_{g=1}^{G} w_g h\left(\boldsymbol{y}_i; \mathcal{C}_i, \zeta^{(g)}\right)$$

Clustering probabilities

- Estimation: Bayesian approach and MCMC methods (Gibbs sampling)
- Metropolis proposal steps for fixed and random effects
- By Bayes Theorem:

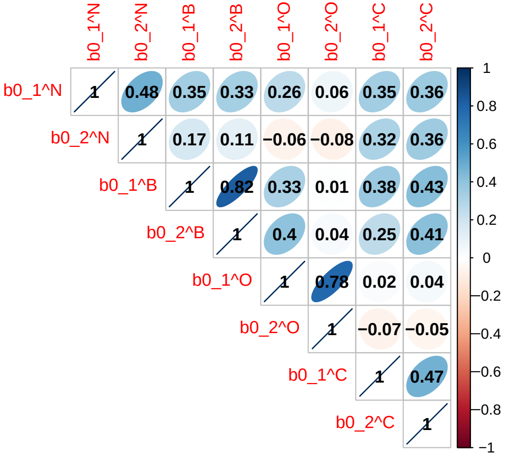
$$p_{i,g}(\boldsymbol{\theta}) = \mathsf{P}\left[U_i = g \middle| \boldsymbol{Y}_i = \boldsymbol{y}_i, \mathcal{C}_i; \boldsymbol{\theta}\right] = \frac{w_g h\left(\boldsymbol{y}_i; \mathcal{C}_i, \boldsymbol{\zeta}^{(g)}\right)}{\sum_{\ell=1}^{G} w_\ell h\left(\boldsymbol{y}_i; \mathcal{C}_i, \boldsymbol{\zeta}^{(\ell)}\right)}$$

  $\rightarrow$ integral approximation required
- Simple clustering rule:

$$\widehat{U}_i := g \iff g = \operatorname*{arg\,max}_{\ell \in \{1,\dots,G\}} \widehat{p_{i,\ell}}$$

- Alternatively use sampled cluster indicators U_i
- Software: implemented in R using the C programming language

Correlation matrix of random intercepts



Random intercepts for

Log(Total disposable income)

Log(Lowest income to make ends meet)

Affordability of a one week holiday

Afford to pay unexpected expenses

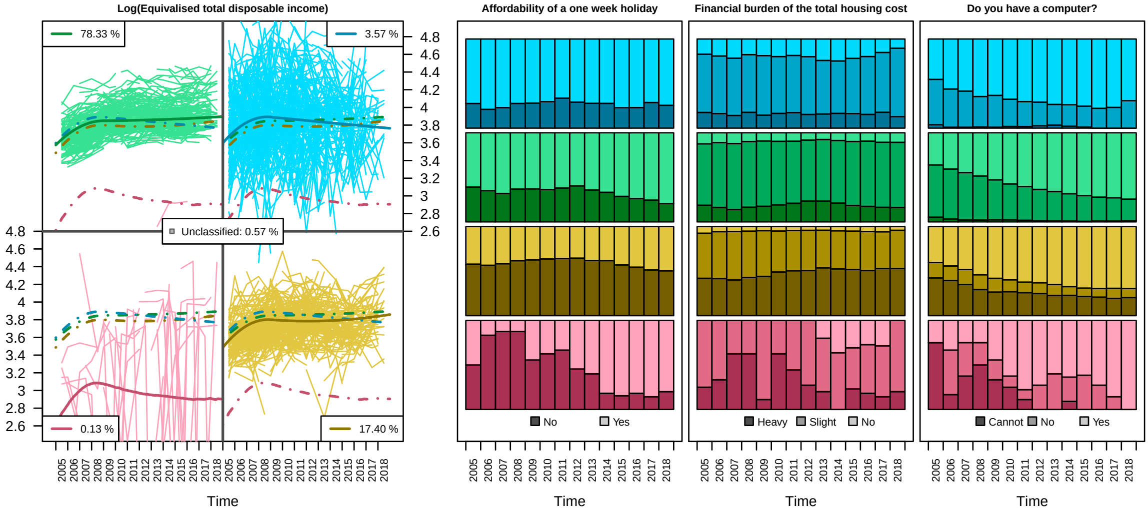
Ability to make ends meet

Financial burden of the total housing cost

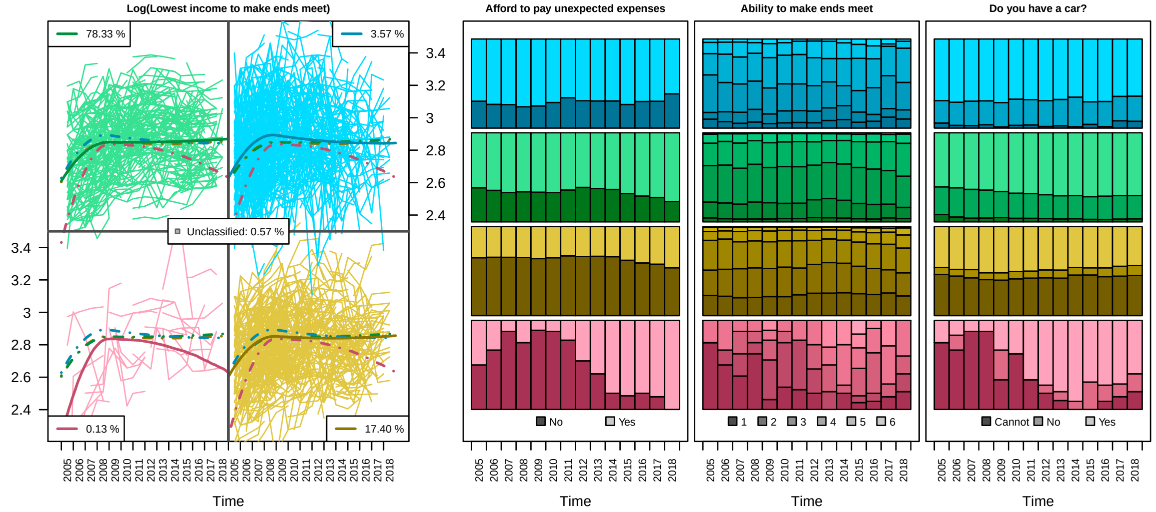
Do you have a computer?

Do you have a car?

Clustering into $\widehat{G}_+ = 4$ groups with respect to time



Clustering into $\widehat{G}_+ = 4$ groups with respect to time



Conclusion

See you at my poster (doors to *knihovna*) with application to `PBC` data

Conclusion

See you at my poster (doors to *knihovna*) with application to `PBC` data

Thank you for your attention.

Literature references

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

[2] E. Breslow, N. and G. Clayton, D. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25, 1983.

[3] Gelman A. Jones G. Brooks, S. and X. Meng. *Handbook for Markov chain Monte Carlo*. First edition. Taylor & Francis, New York, 05 2011.

[4] S. Frühwirth-Schnatter and G. Malsiner-Walli. From here to infinity - sparse finite versus dirichlet process mixtures in model-based clustering. *Advances in Data Analysis and Classification*, 13:33–64, 03 2019.

[5] J. Jiang. *Linear and Generalized Linear Mixed Models and Their Applications*. First edition. Springer-Verlag, New York, 05 2007.

[6] 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.

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

[8] J. C. Pinheiro and E. C. Chao. Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 15(1):58–81, 2006.

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