



FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

Jan Vávra, Arnošt Komárek

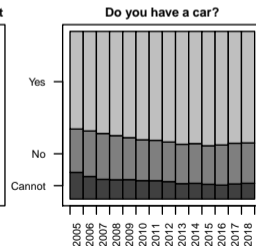
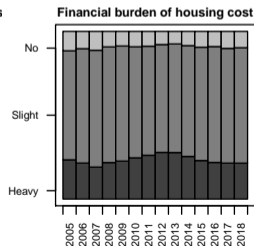
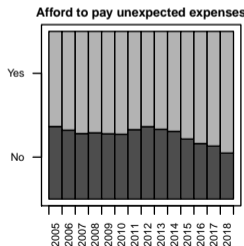
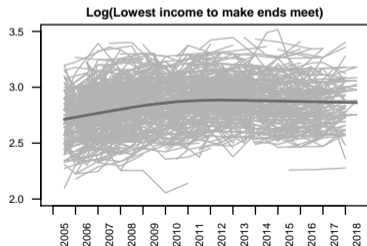
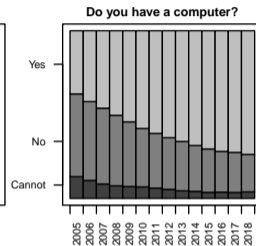
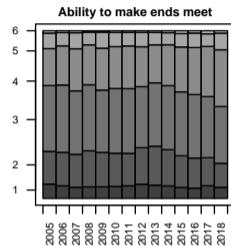
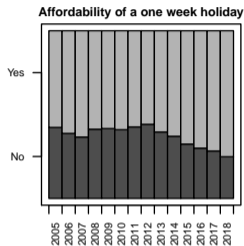
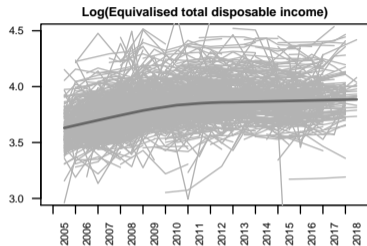
Department of Probability and Mathematical Statistics

**GLMM Based Clustering of Multivariate Mixed Type
Longitudinal Data**

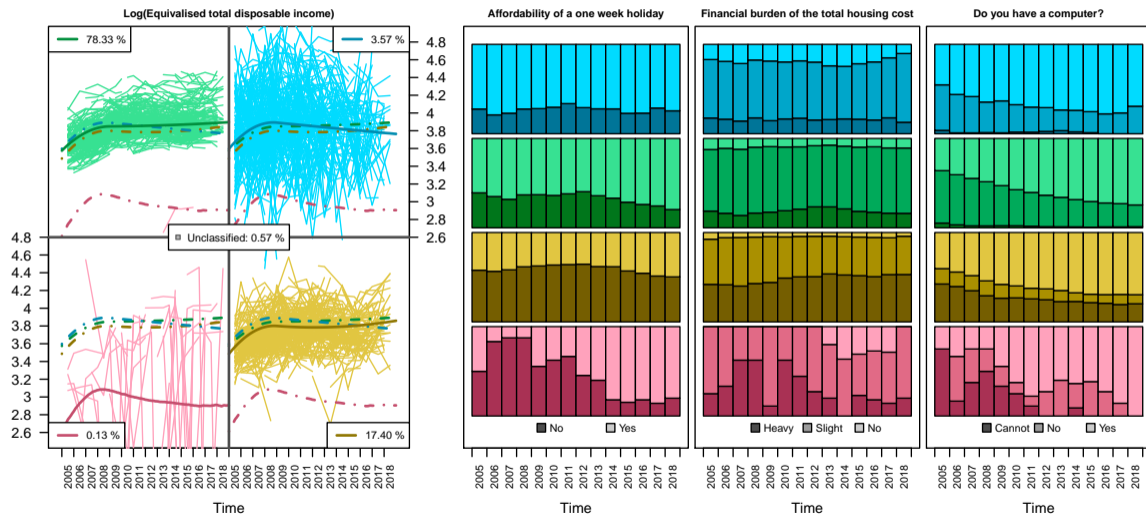
18th July 2022

IWSM 2022, Trieste

EU-SILC - Outcomes of interest



Clustering into $G = 4$ groups



Random effects models

Each outcome is modelled by predictor η consisting of

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

In case of EU-SILC application:

- Fixed:
 - intercept
 - time - quadratic spline with 3 equidistant knots
 - equalised household size
 - level of urbanisation
 - highest education level achieved in the household
 - presence of a student
 - presence of a baby
 - **no** interaction terms

- Random: random intercept $\mathbf{Z}_{ij}^{\top} \mathbf{b}_i = b_{i,0} \stackrel{\text{iid}}{\sim} \text{N}(0, \sigma_b^2), i = 1, \dots, n$

GLMM for different types of outcomes

Numeric outcome

- Linear Mixed-effects Model (LMM)
 - $Y_{ij}^N \mid \eta^N \sim N(\eta^N, \tau^{-1})$
 - τ - precision parameter

Ordinal outcome of K levels

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

Binary outcome

- Logistic Regression with random effects
 - $P[Y^B = 1 \mid \eta^B] = \text{logit}^{-1}(\eta^B) = \frac{e^{\eta^B}}{1 + e^{\eta^B}}$

General categorical outcome of K levels

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

$$\begin{aligned}
 P[Y^C = k \mid \eta_1^C, \dots, \eta_{K-1}^C] &= \\
 &= \text{softmax}_k(\boldsymbol{\eta}^C) = \frac{e^{\eta_k^C}}{1 + \sum_{k'=1}^{K-1} e^{\eta_{k'}^C}}
 \end{aligned}$$

Joint modelling

Models for individual outcomes

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

Join individual models through **joint distribution** of random effects

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

Association among outcomes through random effects

Simulated data: $n = 10\,000$, $n_i = 4$, $\Sigma = D \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} D$, $D = \text{diag}(2, 1.6)$, $t \sim \text{Unif}(0, 1)$, $\beta = 0.5$

Joint probability density function

$$\begin{aligned}
 h(\mathbf{y}_i; C_i, \zeta) &= p(\mathbb{Y}_i = \mathbf{y}_i \mid \beta_N, \tau, \beta_B, \beta_O, \mathbf{c}, \beta_C, \Sigma; C_i) = \\
 &= \int \prod_{r=1}^R \prod_{j=1}^{n_i} \exp \left\{ \ell^{\text{type}(r)} (Y_{i,j}^r \mid \mathbf{b}_i^r, \zeta_r; C_{i,j}) \right\} \cdot (2\pi)^{-\frac{\dim \mathbf{b}_i}{2}} |\Sigma|^{-1} \exp \left\{ -\frac{1}{2} \mathbf{b}_i^\top \Sigma^{-1} \mathbf{b}_i \right\} d\mathbf{b}_i
 \end{aligned}$$

- Requires methods for numerical evaluation of the integral
- solved by *Laplacian approximation* or generally *Adaptive Gaussian Quadrature*

Model based clustering

To identify different patterns:

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

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

Sparse finite mixture to estimate appropriate number of clusters

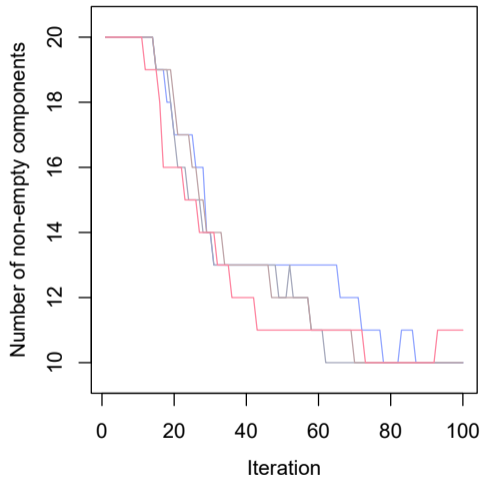
Method for G **not** known in advance

Number of **non-empty** clusters $G_+ = G - \sum_{g=1}^G \mathbb{I}_{(n_g=0)}$

prior $\mathbf{w} \sim \text{Dir}_G(\mathbf{e}_0)$

Idea [4]:

- choose initial large G
- force sparsity by $\mathbf{e}_0 \sim \Gamma(1, 0.01)$
- run MCMC
- converge to a solution of \hat{G}_+ clusters



Clustering probabilities

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

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

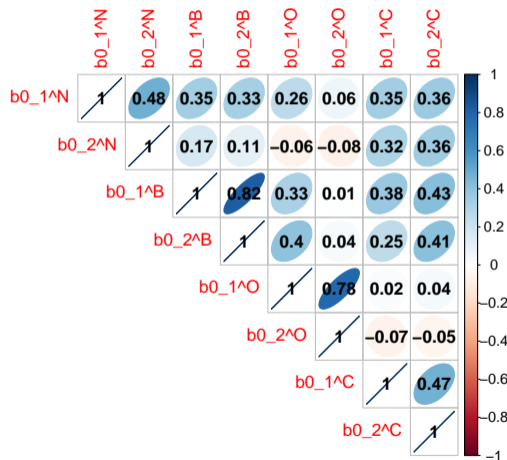
→ integral approximation required

- Simple clustering rule:

$$\hat{U}_i := g \iff g = \arg \max_{\ell \in \{1, \dots, G\}} \hat{p}_{i,\ell}$$

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

Correlation matrix of random intercepts b_i



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

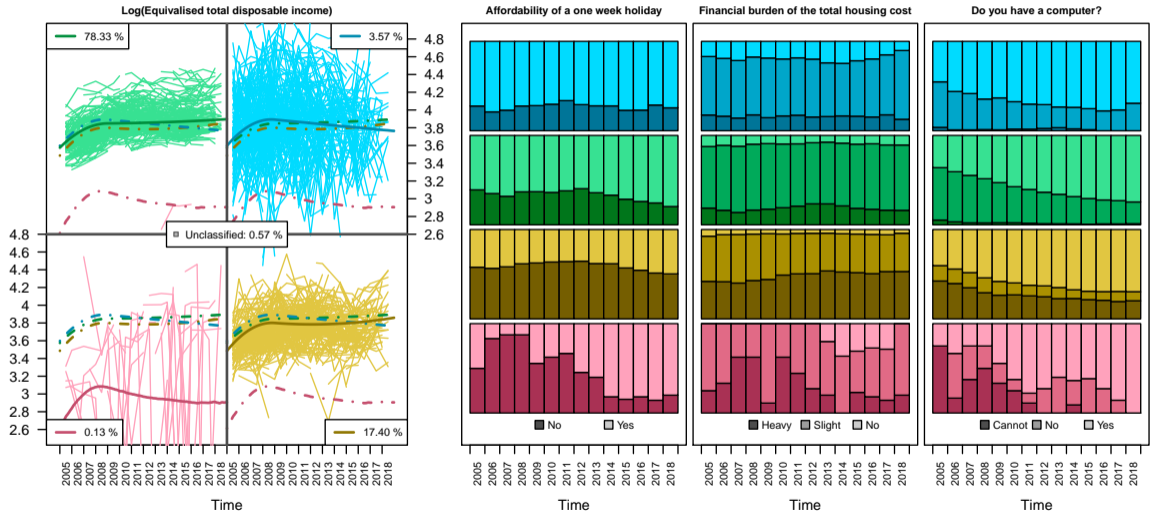
Ability to make ends meet

Financial burden of the total housing cost

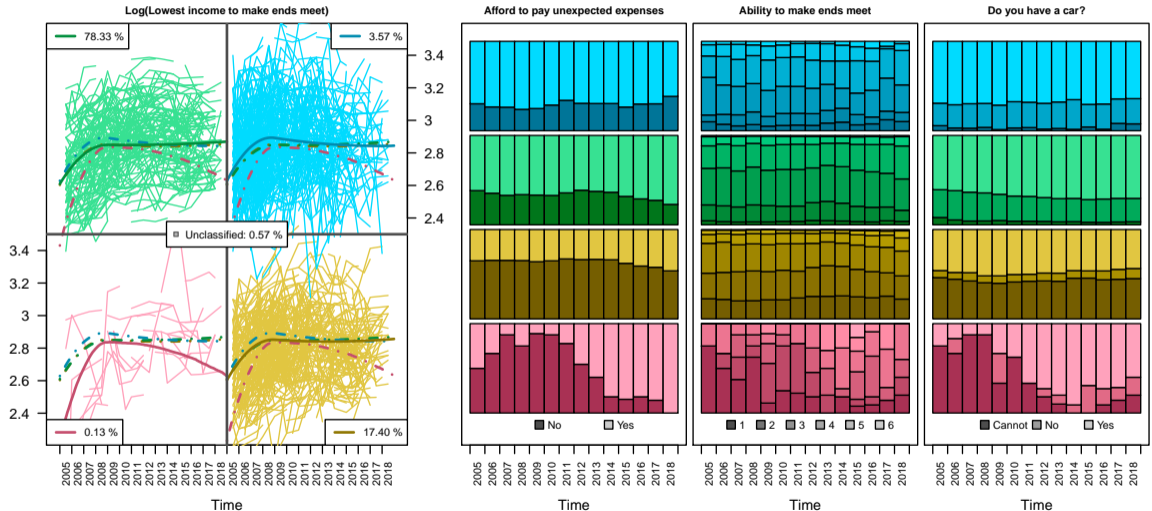
Do you have a computer?

Do you have a car?

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



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



Further development

- generalization to an arbitrary GLM family and link function
- structured parametrization of covariance matrix Σ
- robustness to violation of normality assumption
- selection of important regressors

Further development

- generalization to an arbitrary GLM family and link function
- structured parametrization of covariance matrix Σ
- robustness to violation of normality assumption
- selection of important regressors

Thank you for your attention!
You are welcomed to see our poster!

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.