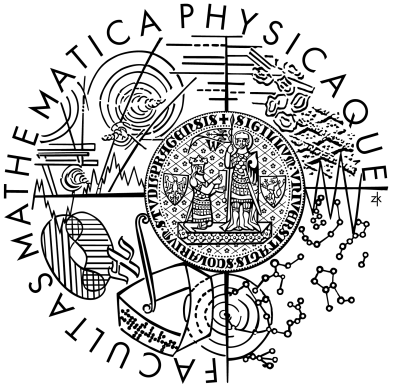
Dept. of Probability and Mathematical Statistics



FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

Arnošt Komárek & Jan Vávra

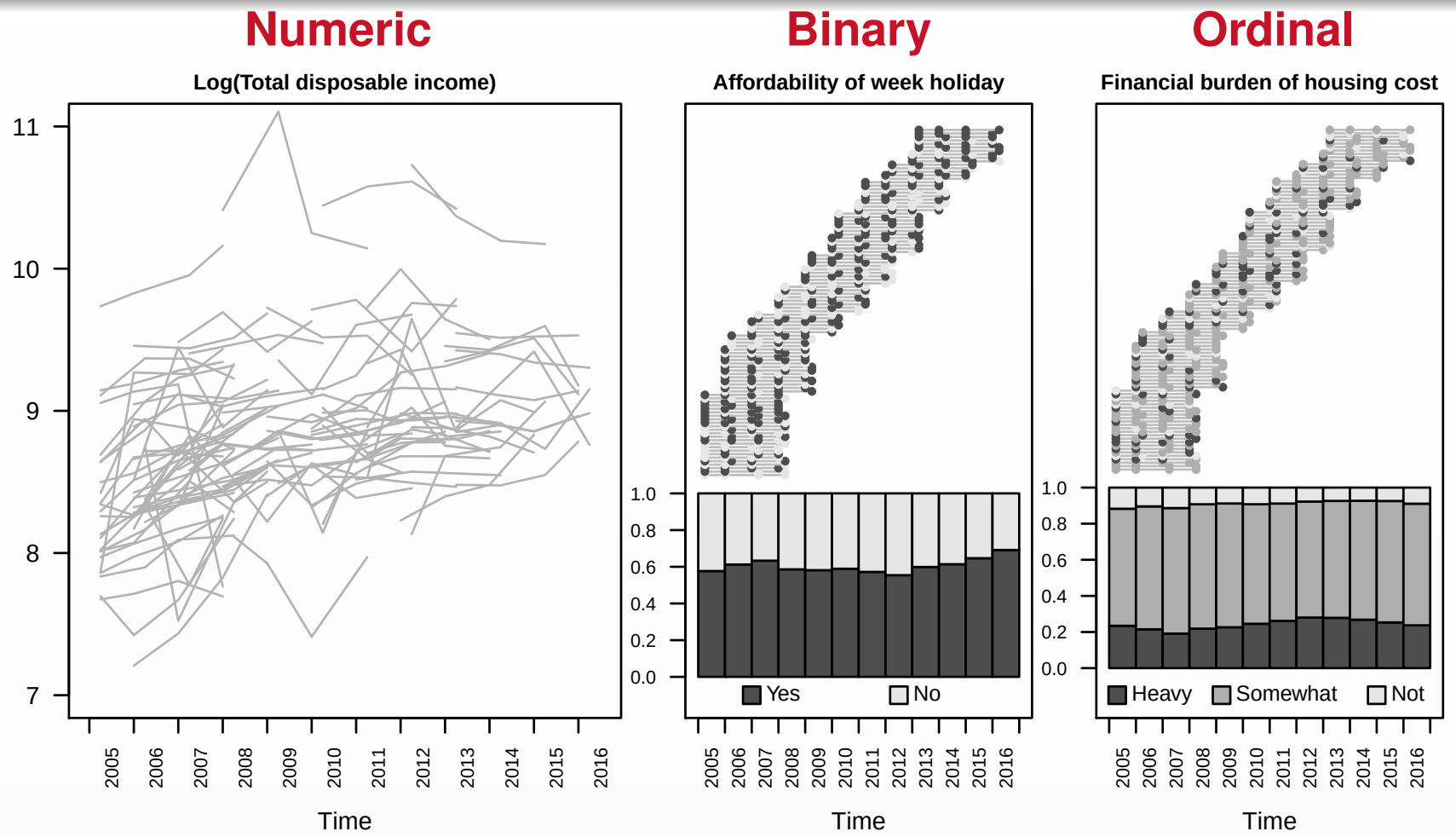
Classification Based on Mixed Type Numeric, Ordinal and Binary Longitudinal Data

Wirtschaftsuniversität Wien, 18. Juni 2021

I.

Introduction

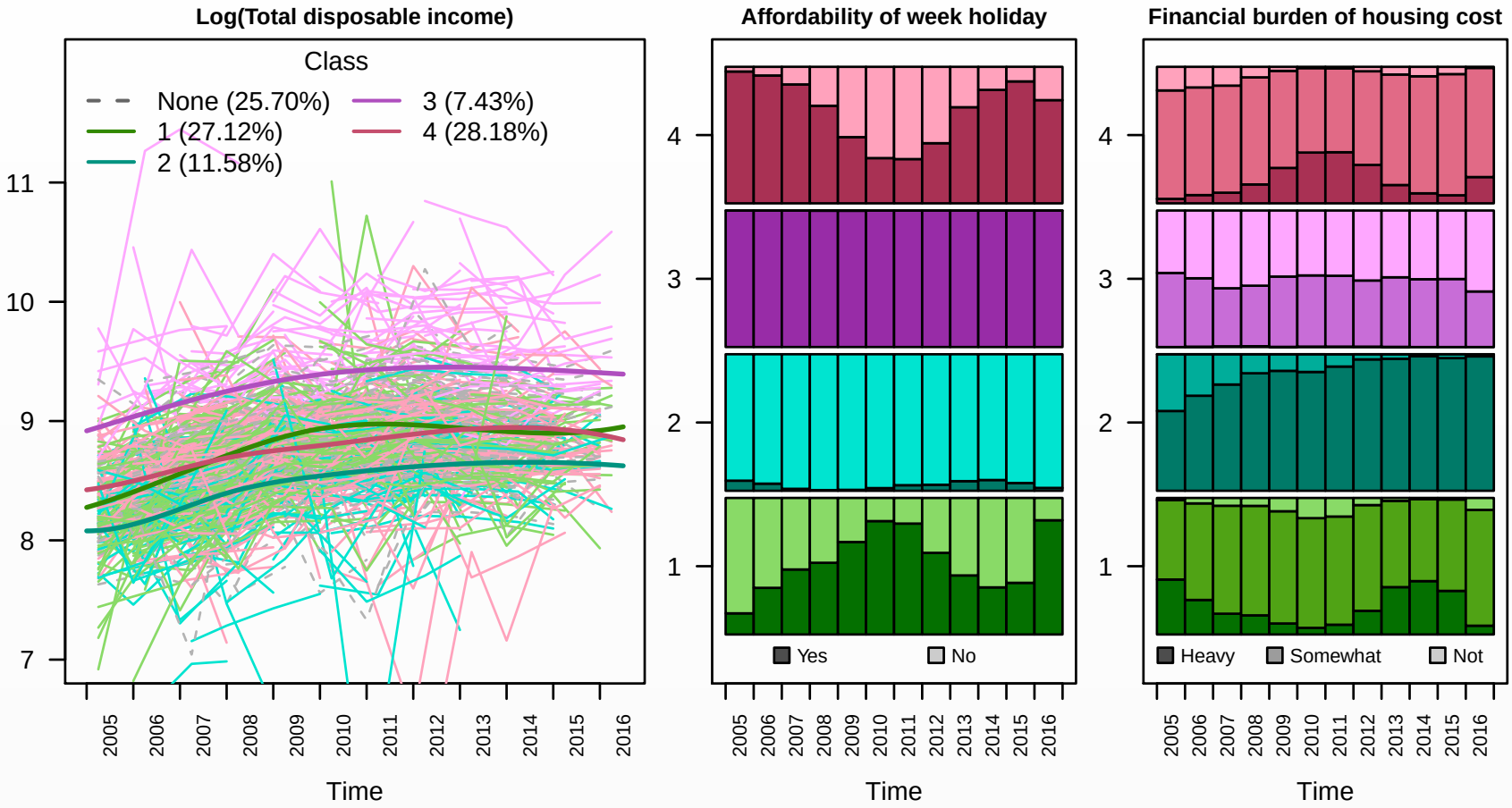
(Multivariate and mixed type) longitudinal data



$$\boldsymbol{Y}_i \equiv \big(\underbrace{Y^1_{i,1}, \ldots, Y^1_{i,n_i}}_{\boldsymbol{Y}^1_i}, \underbrace{Y^2_{i,1}, \ldots, Y^2_{i,n_i}}_{\boldsymbol{Y}^2_i}, \underbrace{Y^3_{i,1}, \ldots, Y^3_{i,n_i}}_{\boldsymbol{Y}^3_i}\big), \quad i = 1, \ldots, n$$

$\boldsymbol{v}^r_{i,j} \equiv \big(t_{i,j}, \ldots\big)$: additional covariates, $\quad j = 1, \ldots, n_i,\ r = 1, \ldots, R(= 3)$

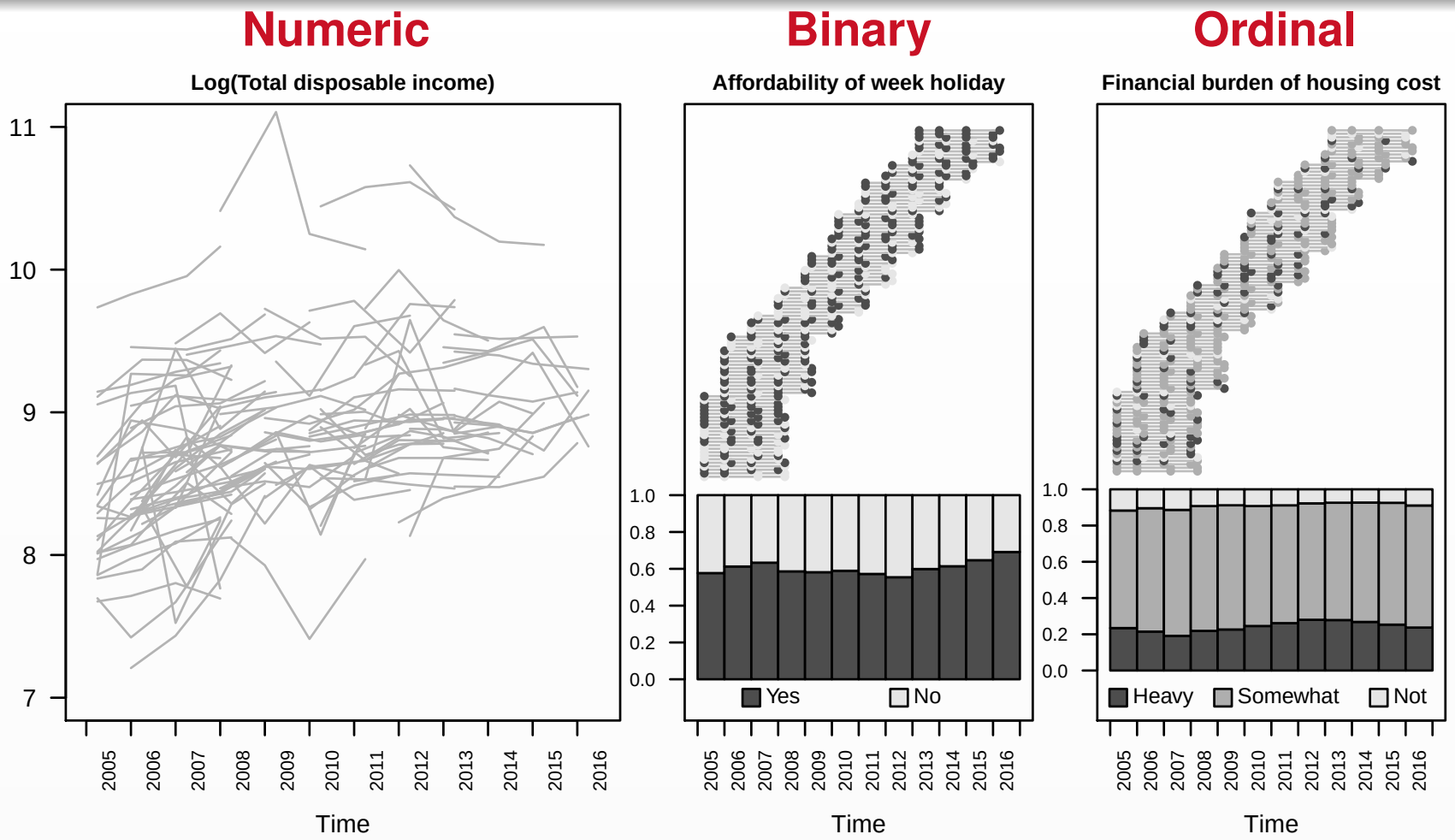
Clustering (unsupervised classification)



Classical clustering methods

- "Classical" data: $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \in \mathbb{R}^p$
- K-means, hierarchical methods, ...
  - based on a distance in p-dimensional Euclidean space
- (Mixed type) longitudinal data
  - $\boldsymbol{Y}_i \in \mathbb{R}^{R\, n_i}$ – different numbers of measurements per subject
  - irregulary spaced in time
  - some $Y_{i,j} \in \{0, 1\}, \in \{0, 1, 2\}, \ldots$
  - distance?

(Multivariate and mixed type) longitudinal data



$$\boldsymbol{Y}_i \equiv \big(\underbrace{Y^1_{i,1}, \ldots, Y^1_{i,n_i}}_{\boldsymbol{Y}^1_i}, \underbrace{Y^2_{i,1}, \ldots, Y^2_{i,n_i}}_{\boldsymbol{Y}^2_i}, \underbrace{Y^3_{i,1}, \ldots, Y^3_{i,n_i}}_{\boldsymbol{Y}^3_i}\big), \quad i = 1, \ldots, n$$

$\boldsymbol{v}^r_{i,j} \equiv \big(t_{i,j}, \ldots\big)$: additional covariates, $\quad j = 1, \ldots, n_i,\ r = 1, \ldots, R(= 3)$

II.

Model based clustering

Model based clustering

Data for subject i, $\mathcal{D}_i$:

$$\boldsymbol{Y}_i \equiv \big(\underbrace{Y^1_{i,1}, \ldots, Y^1_{i,n_i}}_{\boldsymbol{Y}^1_i}, \underbrace{Y^2_{i,1}, \ldots, Y^2_{i,n_i}}_{\boldsymbol{Y}^2_i}, \underbrace{Y^3_{i,1}, \ldots, Y^3_{i,n_i}}_{\boldsymbol{Y}^3_i}\big), \quad i = 1, \ldots, n$$

$\boldsymbol{v}^r_{i,j} \equiv \big(t_{i,j}, \ldots\big)$: additional covariates, $j = 1, \ldots, n_i$, $r = 1, \ldots, R(= 3)$

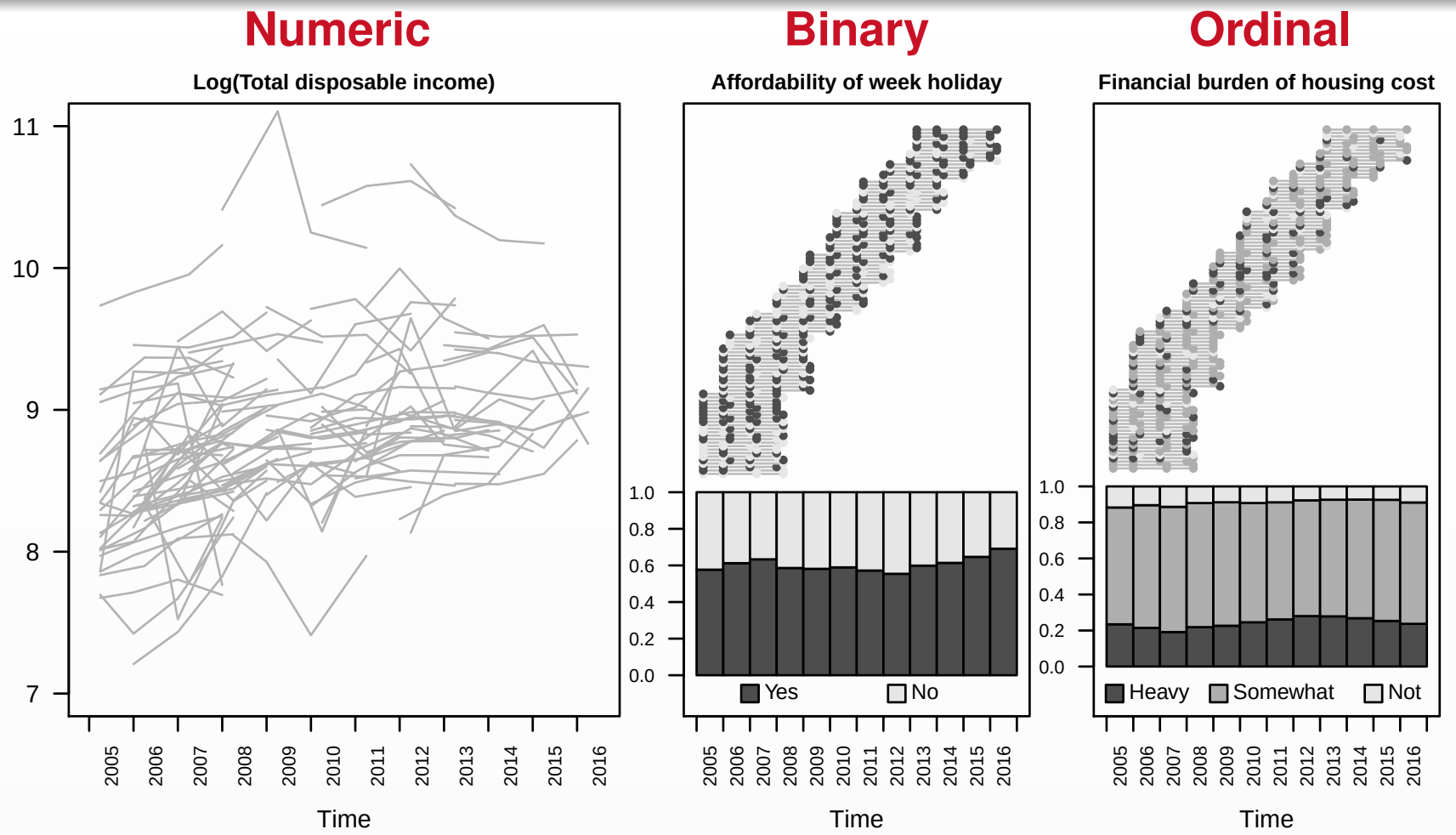
$\longrightarrow \mathcal{C}_i$

- (Whole) population consists of K subpopulations.
  In advance, we do not know who is who.
- Data in subpopulation $k \in \{1, \ldots, K\}$ follow certain statistical model

$\longrightarrow f_k\big(\boldsymbol{y}_i; \boldsymbol{\xi}^k, \boldsymbol{\xi}, \mathcal{C}_i\big)$

- $\boldsymbol{\xi}^k$: group-specific (unknown) model parameters
- $\boldsymbol{\xi}$: (unknown) model parameters shared by all groups

(Multivariate and mixed type) longitudinal data



$$\boldsymbol{Y}_i \equiv \big(\underbrace{Y^1_{i,1}, \ldots, Y^1_{i,n_i}}_{\boldsymbol{Y}^1_i}, \underbrace{Y^2_{i,1}, \ldots, Y^2_{i,n_i}}_{\boldsymbol{Y}^2_i}, \underbrace{Y^3_{i,1}, \ldots, Y^3_{i,n_i}}_{\boldsymbol{Y}^3_i}\big), \quad i = 1, \ldots, n$$

$\boldsymbol{v}^r_{i,j} \equiv \big(t_{i,j}, \ldots\big)$: additional covariates, $\quad j = 1, \ldots, n_i,\ r = 1, \ldots, R(= 3)$

Simple example

- Group-specific model $\equiv$ "simple" gaussian linear model

$$Y_{i,j} = \beta_0^k + \beta_1^k t_{i,j} + \varepsilon_{i,j}, \qquad \varepsilon_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

  - $\boldsymbol{\xi}^k \equiv \boldsymbol{\beta}^k \equiv (\beta_0^k, \beta_1^k)$
  - $\boldsymbol{\xi} \equiv \sigma^2$

- $f_k(\boldsymbol{y}_i; \boldsymbol{\xi}^k, \boldsymbol{\xi}, \mathcal{C}_i) = \prod_{j=1}^{n_i} \varphi(y_{i,j}; \beta_0^k + \beta_1^k t_{i,j}, \sigma^2)$

Longitudinal numeric outcome

- Group-specific model $\equiv$ linear mixed model

$$Y_{i,j} = \beta_0^k + \beta_1^k t_{i,j} + b_{i,0} + b_{i,1}\, t_{i,j} + \varepsilon_{i,j}, \qquad \varepsilon_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\boldsymbol{b}_i = \left(b_{i,0},\, b_{i,1}\right)^\top \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0},\, \mathbb{D})$$

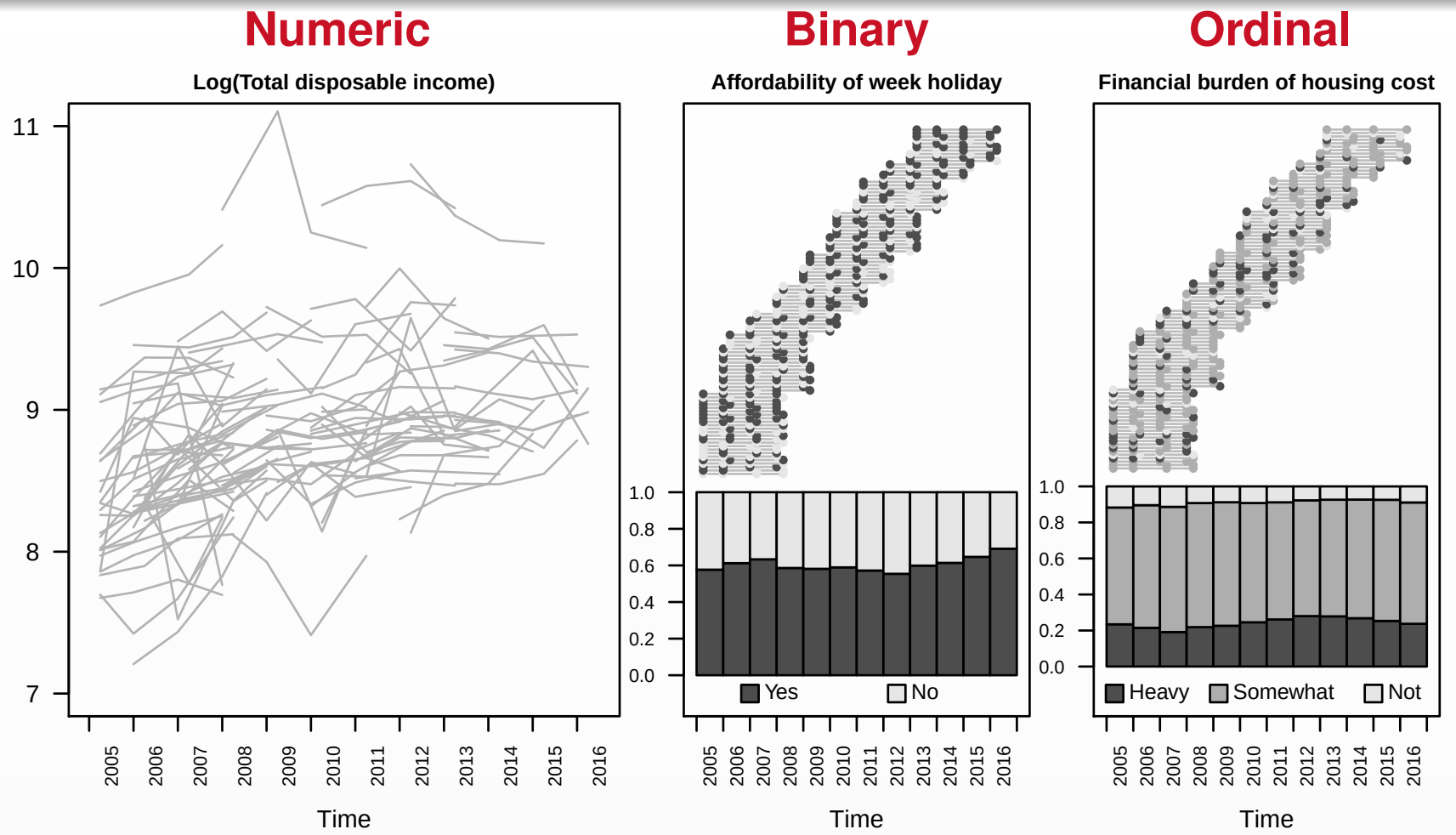
  - $\boldsymbol{\xi}^k \equiv \boldsymbol{\beta}^k \equiv (\beta_0^k,\, \beta_1^k)$
  - $\boldsymbol{\xi} \equiv (\sigma^2,\, \boldsymbol{\Sigma})$

- $f_k(\boldsymbol{y}_i;\, \boldsymbol{\xi}^k,\, \boldsymbol{\xi},\, \mathcal{C}_i) = \varphi\left(\boldsymbol{y}_i;\, \mathbb{X}_i\boldsymbol{\beta}^k,\, \mathbb{V}_i\right),$

$$\mathbb{X}_i = \begin{pmatrix} 1 & t_{i,1} \\ \vdots & \vdots \\ 1 & t_{i,n_i} \end{pmatrix}$$

$$\mathbb{V}_i = \mathbb{X}_i\, \boldsymbol{\Sigma}\, \mathbb{X}_i^\top + \sigma^2\, \boldsymbol{I}_{n_i}$$

(Multivariate and mixed type) longitudinal data



$$\boldsymbol{Y}_i \equiv \big(\underbrace{Y^1_{i,1}, \ldots, Y^1_{i,n_i}}_{\boldsymbol{Y}^1_i}, \underbrace{Y^2_{i,1}, \ldots, Y^2_{i,n_i}}_{\boldsymbol{Y}^2_i}, \underbrace{Y^3_{i,1}, \ldots, Y^3_{i,n_i}}_{\boldsymbol{Y}^3_i}\big), \quad i = 1, \ldots, n$$

$\boldsymbol{v}^r_{i,j} \equiv (t_{i,j}, \ldots)$: additional covariates, $\quad j = 1, \ldots, n_i, \; r = 1, \ldots, R(= 3)$

Model based clustering

- Data in subpopulation $k \in \{1, \ldots, K\}$ follow certain statistical model

$\longrightarrow$ $f_k(\boldsymbol{y}_i; \boldsymbol{\xi}^k, \boldsymbol{\xi}, \mathcal{C}_i)$

- (Yet) unknown allocation of subject i in either of K groups: $U_i \in \{1, \ldots, K\}$

$$f_k(\boldsymbol{y}_i; \boldsymbol{\xi}^k, \boldsymbol{\xi}, \mathcal{C}_i) = p(\boldsymbol{y}_i \,|\, U_i = k; \boldsymbol{\xi}^k, \boldsymbol{\xi}, \mathcal{C}_i)$$

- (Unknown) proportions of K groups in the whole population:

$$w_k = \mathrm{P}(U_i = k; \boldsymbol{w}), \quad k = 1, \ldots, K,$$

$$\boldsymbol{w} = (w_1, \ldots, w_K)^\top \in (0, 1)^K$$

Model based clustering

- Above implies the (marginal) distribution of response

$$
\begin{aligned}
f(\boldsymbol{y}_i;\, \boldsymbol{\theta}, \mathcal{C}_i) &= p(\boldsymbol{y}_i;\, \boldsymbol{\theta}, \mathcal{C}_i) \\
&= \sum_{k=1}^{K} p(\boldsymbol{y}_i \,|\, U_i = k;\, \boldsymbol{\xi}^k, \boldsymbol{\xi}, \mathcal{C}_i)\, \mathrm{P}(U_i = k;\, \boldsymbol{w}) \\
&= \underbrace{\sum_{k=1}^{K} w_k\, f_k(\boldsymbol{y}_i;\, \boldsymbol{\xi}^k, \boldsymbol{\xi}, \mathcal{C}_i)}_{\text{mixture of models}}
\end{aligned}
$$

- Unknown parameters of (the whole) model:

$$\boldsymbol{\theta} \equiv (\boldsymbol{w}, \boldsymbol{\xi}, \boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^K)$$

and a mixture likelihood (if independence across subjects assumed)

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{y}_i;\, \boldsymbol{\theta}, \mathcal{C}_i)$$

Classification?

- (Not yet) really Bayesian analyzis but a simple use of the Bayes theorem ($k = 1, \ldots, K$):

$$u_{i,k}(\boldsymbol{\theta}) := \mathrm{P}\left(U_i = k \,\middle|\, \boldsymbol{Y}_i = \boldsymbol{y}_i; \boldsymbol{\theta}, \mathcal{C}_i\right)$$

$$= \frac{p\left(\boldsymbol{y}_i \,\middle|\, U_i = k; \boldsymbol{\xi}^k, \boldsymbol{\xi}, \mathcal{C}_i\right) \mathrm{P}\left(U_i = k; \boldsymbol{w}\right)}{\sum_{l=1}^{K} p\left(\boldsymbol{y}_i \,\middle|\, U_i = l; \boldsymbol{\xi}^l, \boldsymbol{\xi}, \mathcal{C}_i\right) \mathrm{P}\left(U_i = l; \boldsymbol{w}\right)}$$

$$= \frac{w_k \, f_k\left(\boldsymbol{y}_i; \boldsymbol{\xi}^k, \boldsymbol{\xi}, \mathcal{C}_i\right)}{\sum_{l=1}^{K} w_l \, f_l\left(\boldsymbol{y}_i; \boldsymbol{\xi}^l, \boldsymbol{\xi}, \mathcal{C}_i\right)}$$

Classification?

- If we knew all model parameters $\theta \equiv \left(\boldsymbol{w}, \boldsymbol{\xi}, \boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^K\right)$, possible classification rule ("prediction" of the allocation variable U_i):

$$\widehat{U}_i = \underset{k=1,\ldots,K}{\operatorname{argmax}}\, u_{i,k}(\theta)$$

- In case we do not know θ, just estimate it ($\longrightarrow \widehat{\theta}$) and then

$$\widehat{U}_i = \underset{k=1,\ldots,K}{\operatorname{argmax}}\, u_{i,k}\left(\widehat{\theta}\right)$$

Maximum-likelihood estimation

- Likelihood

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

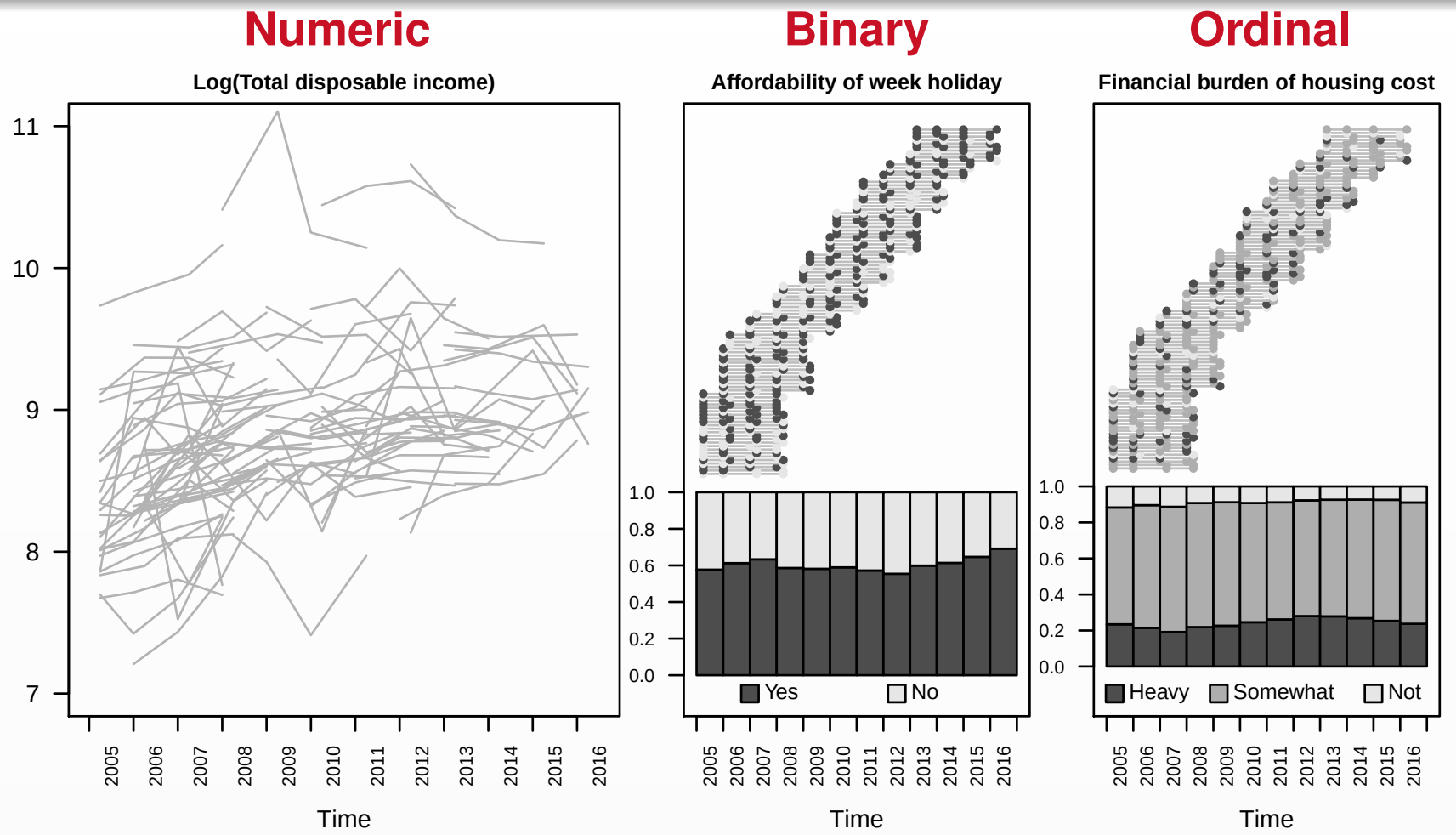
$$\overset{\text{LMM}}{=} \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} w_k \, \varphi\left(\boldsymbol{y}_i;\, \mathbb{X}_i \boldsymbol{\beta}^k,\, \mathbb{X}_i \, \boldsymbol{\Sigma} \, \mathbb{X}_i^\top + \sigma^2 \, \boldsymbol{I}_{n_i}\right) \right\}$$

- $\widehat{\theta} = \text{argmax } L(\theta)$
- EM algorithm?

III.

Model for mixed type longitudinal data

(Multivariate and mixed type) longitudinal data



$$\boldsymbol{Y}_i \equiv \big(\underbrace{Y^1_{i,1}, \ldots, Y^1_{i,n_i}}_{\boldsymbol{Y}^1_i}, \underbrace{Y^2_{i,1}, \ldots, Y^2_{i,n_i}}_{\boldsymbol{Y}^2_i}, \underbrace{Y^3_{i,1}, \ldots, Y^3_{i,n_i}}_{\boldsymbol{Y}^3_i}\big), \quad i = 1, \ldots, n$$

$\boldsymbol{v}^r_{i,j} \equiv (t_{i,j}, \ldots)$: additional covariates, $j = 1, \ldots, n_i$, $r = 1, \ldots, R(= 3) \rightarrow \mathcal{C}^r_i$

Model for mixed type longitudinal data

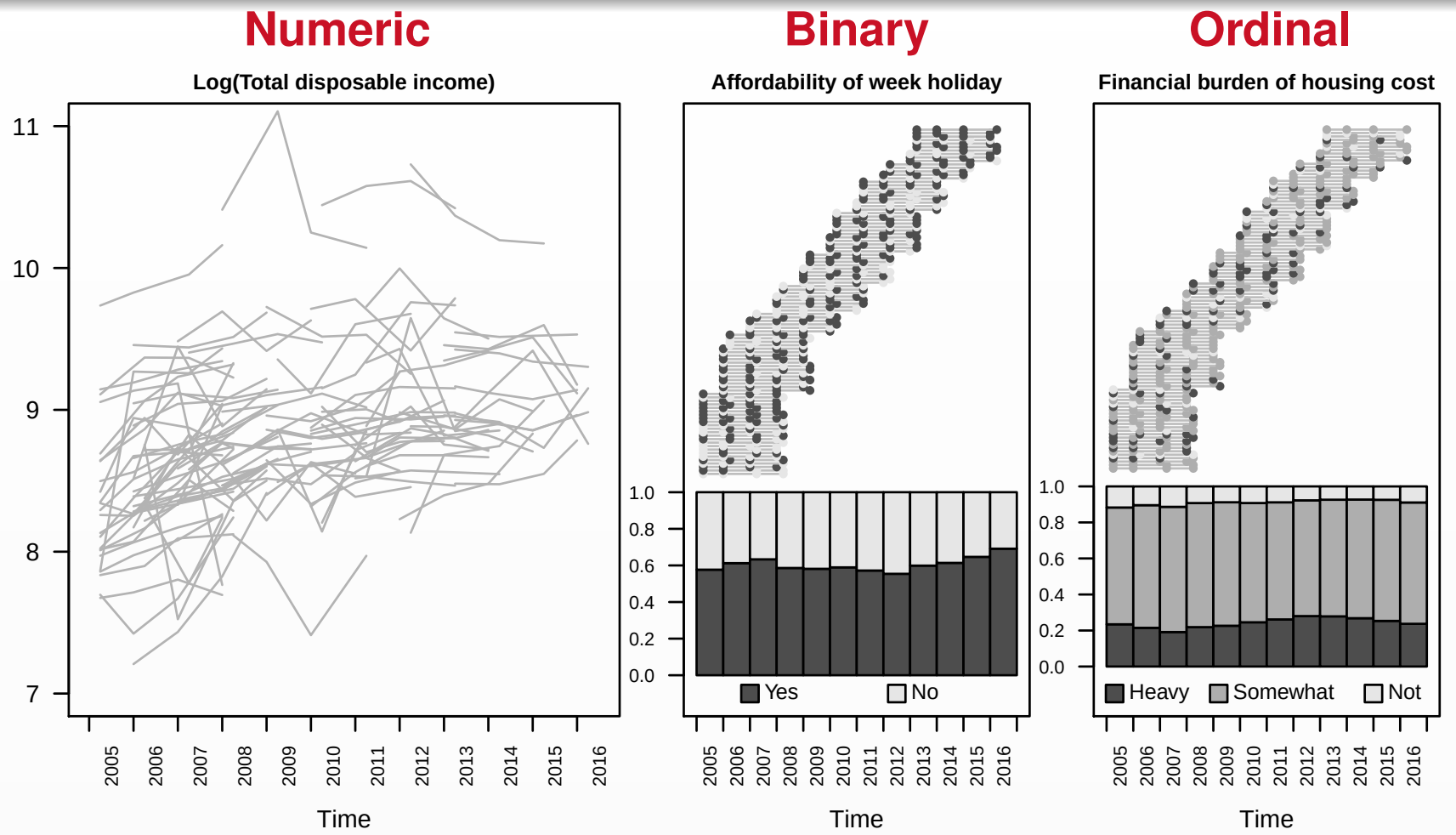
- Hierarchical model
- Numeric outcome: linear mixed model (LME), i.e.,

$$\boldsymbol{Y}_i^r \mid \boldsymbol{b}_i^r; \mathcal{C}_i^r \sim \mathcal{N}_{n_i}\left(\boldsymbol{\eta}_i^r, \tau_r^{-1}\mathbb{I}_{n_i}\right),$$

$$\boldsymbol{\eta}_i^r = \mathbb{X}_i^r \boldsymbol{\beta}^r + \mathbb{Z}_i^r \boldsymbol{b}_i^r$$

$$\boldsymbol{b}_i^r \overset{\text{i.i.d.}}{\sim} \text{wait a little bit}$$

(Multivariate and mixed type) longitudinal data



$$\boldsymbol{Y}_i \equiv \big(\underbrace{Y^1_{i,1}, \ldots, Y^1_{i,n_i}}_{\boldsymbol{Y}^1_i}, \underbrace{Y^2_{i,1}, \ldots, Y^2_{i,n_i}}_{\boldsymbol{Y}^2_i}, \underbrace{Y^3_{i,1}, \ldots, Y^3_{i,n_i}}_{\boldsymbol{Y}^3_i}\big), \quad i = 1, \ldots, n$$

$\boldsymbol{v}^r_{i,j} \equiv (t_{i,j}, \ldots)$: additional covariates, $j = 1, \ldots, n_i$, $r = 1, \ldots, R(= 3) \rightarrow \mathcal{C}^r_i$

Model for mixed type longitudinal data

- Binary and ordinal outcome: linear mixed model (LME) for a latent numeric variable
- Thresholding concept

$$\text{we observe } Y_{i,j}^r = l \quad \Longleftrightarrow \quad \gamma_l^r < Y_{i,j}^{\star,r} \leq \gamma_{l+1}^r$$

- $Y_{i,j}^{\star,r}$: latent (unobservable) numeric variable
- $-\infty = \gamma_0^r < \gamma_1^r < \cdots < \gamma_{L^r}^r = \infty$: unknown thresholds
- For identifiability purposes: $\gamma_1^r =$ fixed const
- Then as before:

$$\boldsymbol{Y}_i^{\star,r} \mid \boldsymbol{b}_i^r; \mathcal{C}_i^r \sim \mathcal{N}_{n_i}\left(\boldsymbol{\eta}_i^r, \tau_r^{-1}\mathbb{I}_{n_i}\right),$$

$$\tau_r = 1 \text{ for identifiability purposes}$$

$$\boldsymbol{\eta}_i^r = \mathbb{X}_i^r \boldsymbol{\beta}^r + \mathbb{Z}_i^r \boldsymbol{b}_i^r$$

Joint model

- How to model dependencies across different outcomes on one subject?
- Joint distribution for all random effects

$$\boldsymbol{b}_i = \begin{pmatrix} \boldsymbol{b}_i^{\text{N}} \\ \boldsymbol{b}_i^{\text{OB}} \end{pmatrix} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_{d^{\text{R}}} \left(\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}^{\text{N}} \\ \boldsymbol{\mu}^{\text{OB}} \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}^{\text{N}} & \boldsymbol{\Sigma}^{\text{NOB}} \\ \boldsymbol{\Sigma}^{\text{OBN}} & \boldsymbol{\Sigma}^{\text{OB}} \end{pmatrix} \right)$$

- $\boldsymbol{\mu}$: (unknown) mean of random effects
- $\boldsymbol{\Sigma}$: (unknown) covariance matrix of random effects

Likelihood

$$p\Big(\mathbb{Y}_i \,\Big|\, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\tau}, \boldsymbol{\gamma};\, \mathcal{C}_i\Big)$$

$$= \int\int p\Big(\mathbb{Y}_i^{\mathrm{N}}, \mathbb{Y}_i^{\mathrm{OB}}, \mathbb{Y}_i^{\star,\mathrm{OB}}, \boldsymbol{b}_i \,\Big|\, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\tau}, \boldsymbol{\gamma};\, \mathcal{C}_i\Big)\, \mathrm{d}\boldsymbol{b}_i\, \mathrm{d}\mathbb{Y}_i^{\star,\mathrm{OB}}$$

$$= \int\int \underbrace{p\big(\mathbb{Y}_i^{\mathrm{OB}} \,\big|\, \mathbb{Y}_i^{\star,\mathrm{OB}}, \boldsymbol{\gamma}\big)}_{\text{thresholding}} \cdot \underbrace{p\big(\mathbb{Y}_i^{\mathrm{N}}, \mathbb{Y}_i^{\star,\mathrm{OB}} \,\big|\, \boldsymbol{b}_i, \boldsymbol{\beta}, \boldsymbol{\tau};\, \mathcal{C}_i\big)}_{\text{MV LME}}$$

$$\cdot \underbrace{p\,(\boldsymbol{b}_i \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})}_{\text{normality}}\, \mathrm{d}\boldsymbol{b}_i\, \mathrm{d}\mathbb{Y}_i^{\star,\mathrm{OB}}$$

- In principle, the likelihood corresponds to the CDF of truncated multivariate normal distribution

$$p\left(\mathbb{Y}_i^{\mathrm{OB}} \,\middle|\, \mathbb{Y}_i^{\star,\mathrm{OB}}, \boldsymbol{\gamma}\right) = \prod_{r \in \mathcal{R}^{\mathrm{OB}}} \prod_{j=1}^{n_i} \left[\sum_{l=0}^{L^r - 1} \mathbb{I}_{\{l\}}\left(y_{i,j}^r\right) \mathbb{I}_{\left(\gamma_l^r\,,\,\gamma_{l+1}^r\right]}\left(y_i^{\star,\mathrm{OB}}\right) \right]$$

$$p\left(\mathbb{Y}_i^{\mathrm{N}}, \mathbb{Y}_i^{\star,\mathrm{OB}} \,\middle|\, \boldsymbol{b}_i, \boldsymbol{\beta}, \boldsymbol{\tau}; \mathcal{C}_i\right)$$

$$= \prod_{r \in \mathcal{R}^{\mathrm{Num}}} \prod_{j=1}^{n_i} \varphi\left(y_{i,j}^r;\ \eta_{i,j}^r,\ \tau_r^{-1}\right) \cdot \prod_{r \in \mathcal{R}^{\mathrm{OB}}} \prod_{j=1}^{n_i} \varphi\left(y_{i,j}^{\star,r};\ \eta_{i,j}^r,\ 1\right)$$

$$p\left(\boldsymbol{b}_i \,\middle|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \varphi\left(\boldsymbol{b}_i;\ \boldsymbol{\mu},\ \boldsymbol{\Sigma}\right)$$

Mixture likelihood

- Do not forget, we will mix the above model
- On top of random effects and latent variables for binary/ordinal outcomes, there will be also latent component allocations U_i...
- The mixture likelihood is then

$$L(\theta) = \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} w_k \, p\left(\mathbb{Y}_i \,\middle|\, \beta^{(k)}, \mu^{(k)}, \Sigma^{(k)}, \tau^{(k)}, \gamma; \mathcal{C}_i \right) \right\},$$

$$p\left(\mathbb{Y}_i \,\middle|\, \beta, \mu, \Sigma, \tau, \gamma; \mathcal{C}_i \right)$$

$$= \int \int \underbrace{p\left(\mathbb{Y}_i^{\text{OB}} \,\middle|\, \mathbb{Y}_i^{\star,\text{OB}}, \gamma \right)}_{\text{thresholding}} \cdot \underbrace{p\left(\mathbb{Y}_i^{\text{N}}, \mathbb{Y}_i^{\star,\text{OB}} \,\middle|\, b_i, \beta, \tau; \mathcal{C}_i \right)}_{\text{MV LME}}$$

$$\cdot \underbrace{p\left(b_i \middle| \mu, \Sigma \right)}_{\text{normality}} \, \mathrm{d}b_i \, \mathrm{d}\mathbb{Y}_i^{\star,\text{OB}}$$

Maximum likelihood?

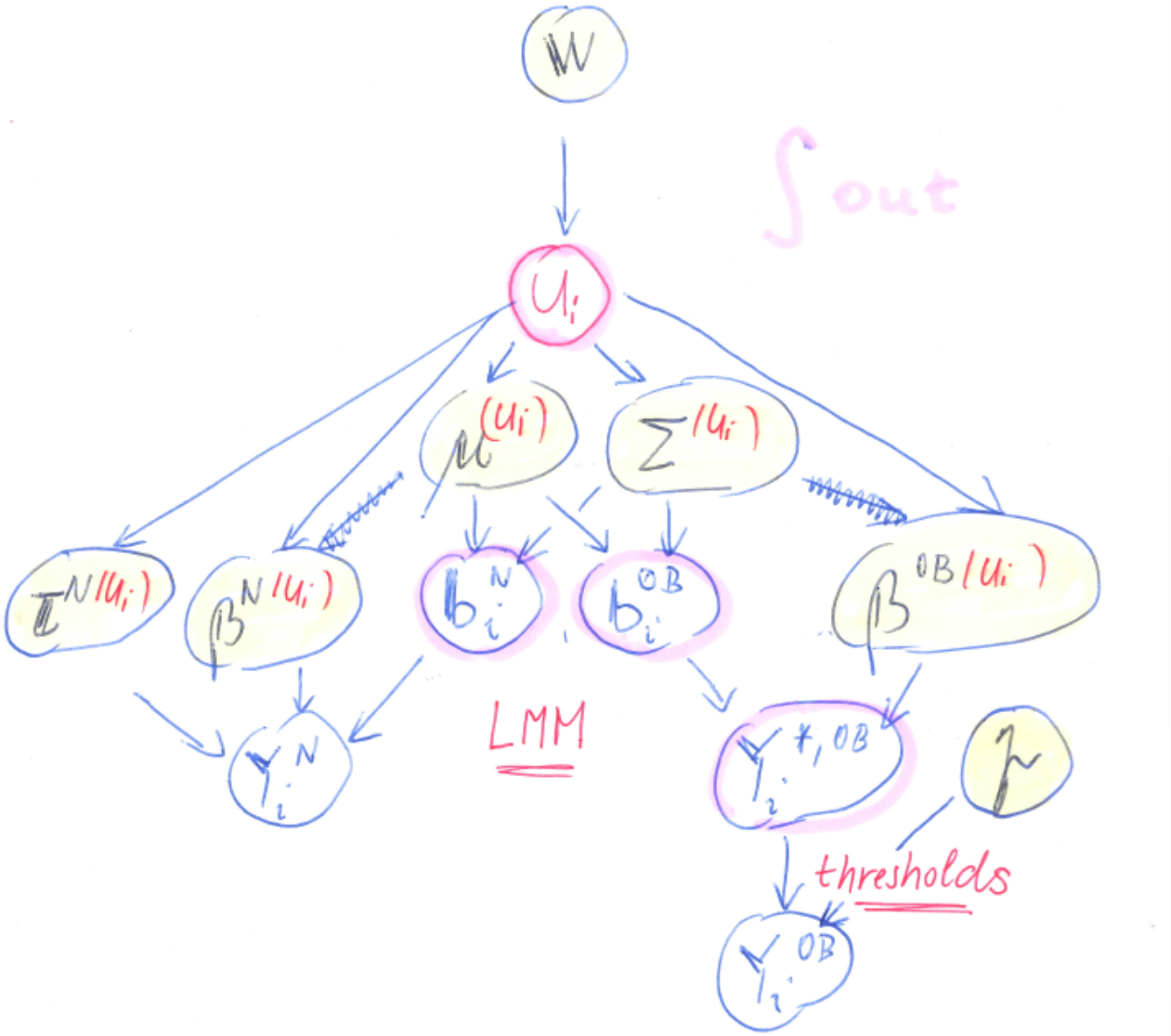
- Maximum-likelihood estimation?
- EM algorithm?
- Good luck. . .

IV.

Bayesian inference

Hierarchical model

The model is fully hierarchical:



Bayesian inference

- Specify a prior distribution for θ (in a *reasonable* way)
- Principles of Bayesian Data Augmentation (BDA) can be used towards Gibbs sampling and MCMC to obtain the posterior sample $\theta^{(m)}$, $m = 1, \ldots, M$ for inference
- Has been implemented (by JV) in  and 

Bayesian classification

Remember ($k = 1, \ldots, K$):

$$u_{i,k}(\boldsymbol{\theta}) = \mathrm{P}\left(U_i = k \,\middle|\, \boldsymbol{Y}_i = \boldsymbol{y}_i;\, \boldsymbol{\theta}, \mathcal{C}_i\right)$$

$$= \frac{w_k\, f_k\left(\boldsymbol{y}_i;\, \boldsymbol{\xi}^k, \boldsymbol{\xi}, \mathcal{C}_i\right)}{\sum_{l=1}^{K} w_l\, f_l\left(\boldsymbol{y}_i;\, \boldsymbol{\xi}^l, \boldsymbol{\xi}, \mathcal{C}_i\right)}$$

Here:

$$f_k\left(\boldsymbol{y}_i;\, \boldsymbol{\xi}^k, \boldsymbol{\xi}, \mathcal{C}_i\right) \equiv p\left(\mathbb{Y}_i \,\middle|\, \boldsymbol{\beta}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \boldsymbol{\tau}^{(k)}, \boldsymbol{\gamma};\, \mathcal{C}_i\right)$$

$$= \int\int \underbrace{p\left(\mathbb{Y}_i^{\mathrm{OB}} \,\middle|\, \mathbb{Y}_i^{\star,\mathrm{OB}}, \boldsymbol{\gamma}\right)}_{\text{thresholding}} \cdot \underbrace{p\left(\mathbb{Y}_i^{\mathrm{N}}, \mathbb{Y}_i^{\star,\mathrm{OB}} \,\middle|\, \boldsymbol{b}_i, \boldsymbol{\beta}^{(k)}, \boldsymbol{\tau}^{(k)};\, \mathcal{C}_i\right)}_{\text{MV LME}}$$

$$\cdot \underbrace{p\left(\boldsymbol{b}_i \middle|\, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}\right)}_{\text{normality}} \mathrm{d}\boldsymbol{b}_i\, \mathrm{d}\mathbb{Y}_i^{\star,\mathrm{OB}}$$

Bayesian classification

- For each i and k, quantity $u_{i,k}(\boldsymbol{\theta})$ is just some function of unknown parameters
- In a Bayesian classification procedure, rather than $u_{i,k}\left(\widehat{\boldsymbol{\theta}}\right)$, one would use some estimate of $u_{i,k}(\boldsymbol{\theta})$ directly, e.g.,
- The posterior mean

$$\widehat{U}_{i,k} = \int u_{i,k}(\boldsymbol{\theta})\, p\left(\boldsymbol{\theta} \,\middle|\, \mathcal{D}\right) \mathrm{d}\boldsymbol{\theta} \approx \frac{1}{M} \sum_{m=1}^{M} u_{i,k}\left(\boldsymbol{\theta}^{(m)}\right)$$

- "Small" complication: integrals over the multivariate normal density must be calculated
- Next to the point estimate, also the (95% HPD) credible interval for $u_{i,k}(\boldsymbol{\theta})$ can be calculated to quantify uncertainty

$$\longrightarrow \qquad \left(\widehat{U}_{i,k}^{LOW},\ \widehat{U}_{i,k}^{UPP}\right)$$

Bayesian classification

- Traditional rule: $\widehat{U}_i = \underset{k=1,\ldots,K}{\text{argmax}}\, \widehat{U}_{i,k}$

$\longrightarrow$ may lead to high misclassification rates if two (or more) clusters are not really apart

- Uncertainty in classification may, e.g., be taken into account as follows:

  Subject i is classified in group $\widehat{U}_i = k$ if and only if the lower bound $\widehat{U}^{LOW}_{i,k}$ of the credible interval is higher than any other upper bound $\widehat{U}^{UPP}_{i,l}$, $l \neq k$. If this does not happen, subject i remains unclassified.

$\longrightarrow$ fills clusters with their most typical representatives

$\longrightarrow$ keep indecisive subjects aside

Classification of a "new" subject

- The whole procedure can also be used to classify a "new" subject without re-fitting the model (running the MCMC procedure
- Allocation probabilities are just a function of model parameters

$$u_{new,k}(\boldsymbol{\theta}) = \mathrm{P}\left(U_{new} = k \,\middle|\, \boldsymbol{Y}_{new} = \boldsymbol{y}_{new};\, \boldsymbol{\theta},\, \mathcal{C}_{new}\right)$$

$$= \frac{w_k\, f_k\left(\boldsymbol{y}_{new};\, \boldsymbol{\xi}^k,\, \boldsymbol{\xi},\, \mathcal{C}_{new}\right)}{\sum_{l=1}^{K} w_l\, f_l\left(\boldsymbol{y}_{new};\, \boldsymbol{\xi}^l,\, \boldsymbol{\xi},\, \mathcal{C}_{new}\right)}$$

- The posterior mean (and the credible interval) does not need to include the "new" observation in the posterior distribution

$$\widehat{U}_{new,k} = \int u_{new,k}(\boldsymbol{\theta})\, p\left(\boldsymbol{\theta} \,\middle|\, \mathcal{D}(old)\right) \mathrm{d}\boldsymbol{\theta} \approx \frac{1}{M} \sum_{m=1}^{M} u_{new,k}\left(\boldsymbol{\theta}^{(m)}\right)$$

Classification of a "new" subject

- Allocation probabilities are just a function of model parameters

$$u_{new,k}(\boldsymbol{\theta}) = \mathrm{P}\left(U_{new} = k \,\middle|\, \boldsymbol{Y}_{new} = \boldsymbol{y}_{new};\, \boldsymbol{\theta},\, \mathcal{C}_{new}\right)$$

$$= \frac{w_k\, f_k\left(\boldsymbol{y}_{new};\, \boldsymbol{\xi}^k,\, \boldsymbol{\xi},\, \mathcal{C}_{new}\right)}{\sum_{l=1}^{K} w_l\, f_l\left(\boldsymbol{y}_{new};\, \boldsymbol{\xi}^l,\, \boldsymbol{\xi},\, \mathcal{C}_{new}\right)}$$

- With longitudinal data $\boldsymbol{Y}_{new} = \left(Y_{new,1},\, Y_{new,2},\, \ldots,\, Y_{new,n_{new}}\right)^\top$, classification can proceed dynamically. We will follow the "new" subject only until the moment when it is clear to which cluster it belongs to

V.

Simulation

Simulation setting

- Three longitudinal outcomes: numeric, binary, ordinal (3 levels)
- Four measurement occasions: $0 < t_{i,1} < t_{i,2} < t_{i,3} < t_{i,4} < 1$
- Linear predictor (for all outcomes)

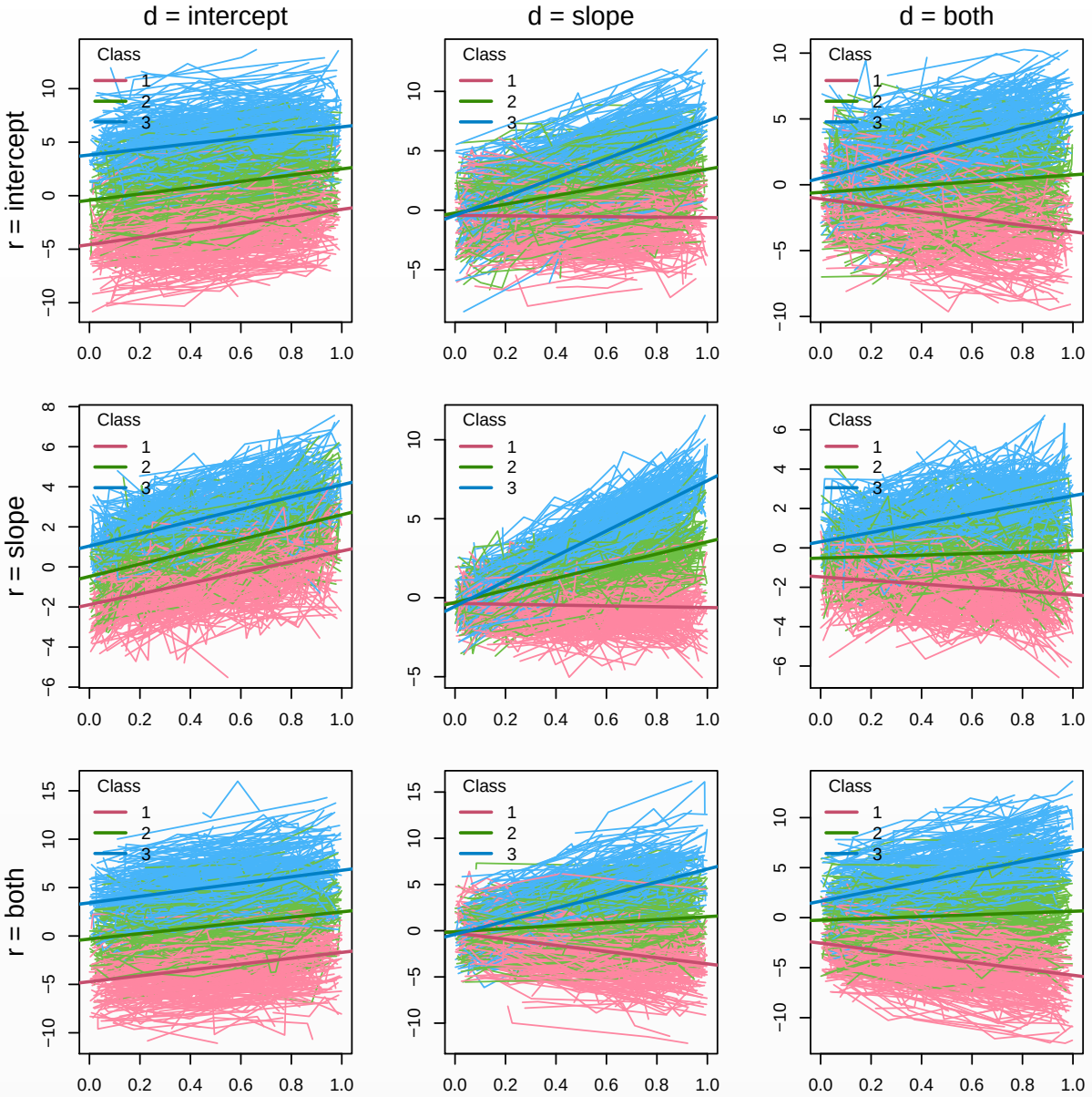
$$1\, X^1_{i,j} \; - \; 2\, X^2_{i,j} \; + \; \cdots \, ,$$

$$X^1_{i,j} \overset{\text{i.i.d.}}{\sim} \text{Alt}(0.5), \qquad X^2_{i,j} \overset{\text{i.i.d.}}{\sim} \text{Unif}(0,\, 1)$$

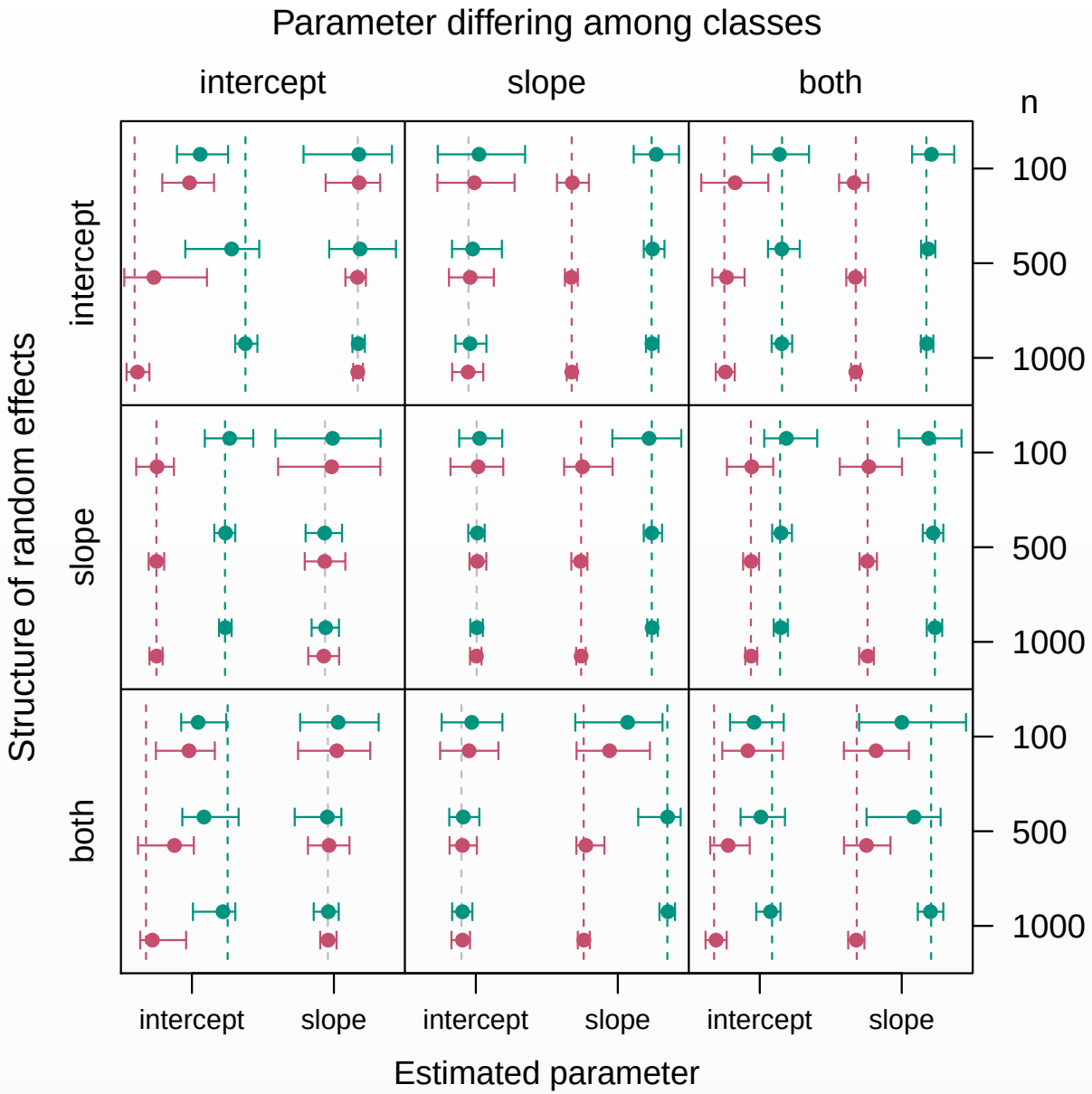
- $\cdots$
  1. (r = intercept): $b_{0,i} + \beta_1 t_{i,j}$, $\quad \text{E}(b_{0,i}) = \beta_0$
  2. (r = slope): $\beta_0 + b_{1,i} t_{i,j}$, $\quad \text{E}(b_{1,i}) = \beta_1$
  3. (r = both): $b_{0,i} + b_{1,i} t_{i,j}$, $\quad \text{E}(b_{0,i},\, b_{1,i}) = (\beta_0,\, \beta_1)$
- Class specific?
  1. (d = intercept): $\beta_0 \equiv \beta_0^{(k)}$
  2. (d = slope): $\beta_1 \equiv \beta_1^{(k)}$
  3. (d = both): $\beta_0 \equiv \beta_0^{(k)}$, $\beta_1 \equiv \beta_1^{(k)}$

$\longrightarrow$ $3 \times 3 = 9$ scenarios with $K = 2, 3$ and $n = 100,\ 500,\ 1\,000$

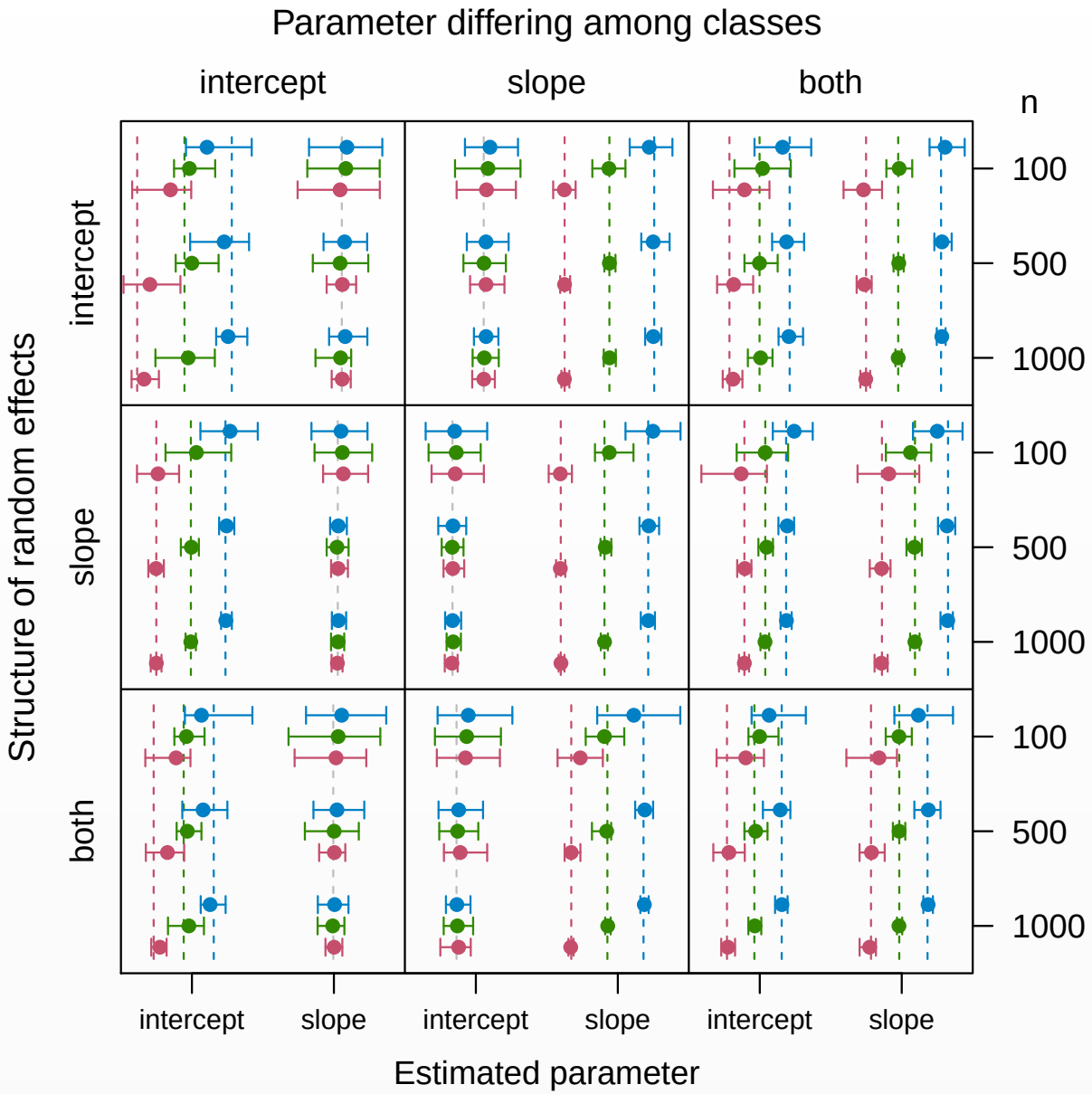
Simulation setting, $K = 3$



Statistical properties of posterior means, $K = 2$



Statistical properties of posterior means, $K = 3$



Classification ability

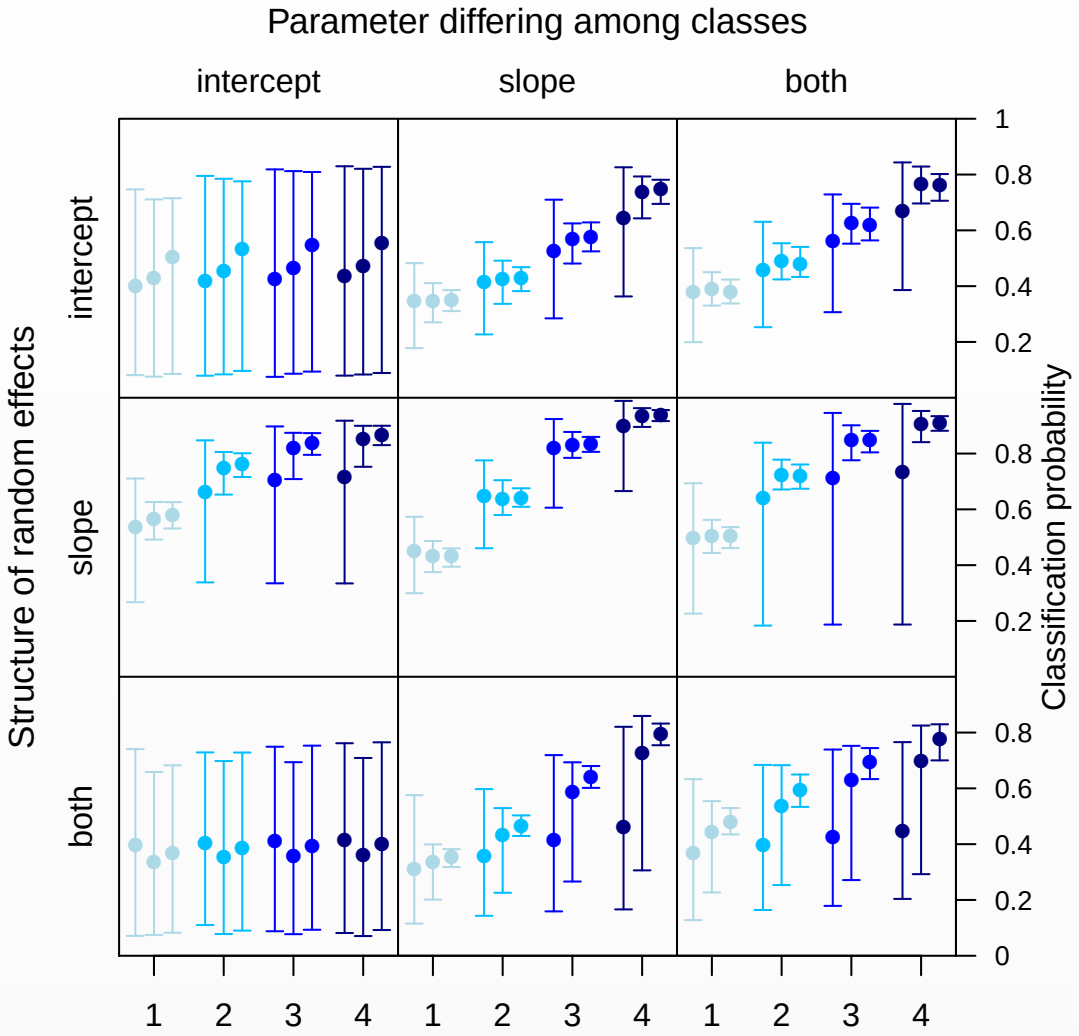
r	d	n	$K = 2$ Correct [%]	$K = 2$ Uncl. [%]	$K = 2$ Miscl. [%]	$K = 3$ Correct [%]	$K = 3$ Uncl. [%]	$K = 3$ Miscl. [%]
intercept	intercept	100	27.0 (17.2)	63.2 (25.4)	9.8 (13.7)	23.0 (17.5)	70.2 (21.0)	6.8 (9.4)
		500	62.5 (27.2)	33.0 (27.3)	4.4 (3.8)	44.3 (20.6)	50.8 (22.1)	4.9 (4.4)
		1000	85.1 (6.7)	10.1 (7.1)	4.8 (0.9)	58.6 (16.9)	35.5 (17.2)	6.0 (3.1)
intercept	slope	100	76.8 (5.4)	20.3 (5.5)	2.9 (1.9)	56.0 (8.5)	40.4 (8.9)	3.6 (2.4)
		500	86.1 (1.8)	8.9 (1.8)	5.0 (1.0)	74.6 (2.0)	19.0 (2.1)	6.4 (1.2)
		1000	87.5 (1.1)	6.7 (0.9)	5.9 (0.7)	78.2 (1.5)	13.8 (1.5)	8.0 (0.8)
intercept	both	100	86.5 (4.4)	12.0 (4.4)	1.5 (1.1)	58.0 (9.4)	38.5 (10.2)	3.4 (2.2)
		500	92.9 (1.4)	4.5 (1.1)	2.6 (0.7)	76.9 (2.5)	16.5 (2.5)	6.7 (1.1)
		1000	93.8 (0.8)	3.3 (0.6)	2.9 (0.5)	79.4 (1.6)	12.8 (1.6)	7.8 (0.8)
slope	intercept	100	96.2 (2.6)	3.4 (2.5)	0.4 (0.6)	61.2 (15.5)	36.4 (15.7)	2.3 (1.8)
		500	97.9 (0.5)	1.5 (0.5)	0.6 (0.4)	87.6 (2.2)	9.2 (2.2)	3.2 (0.7)
		1000	98.3 (0.4)	0.9 (0.3)	0.8 (0.3)	90.2 (1.2)	6.2 (1.1)	3.6 (0.5)
slope	slope	100	80.1 (20.4)	16.3 (19.0)	3.6 (8.7)	85.7 (13.5)	13.3 (13.6)	1.0 (1.2)
		500	92.8 (1.5)	4.6 (1.4)	2.6 (0.7)	94.9 (1.2)	3.6 (1.0)	1.5 (0.5)
		1000	93.9 (0.9)	3.3 (0.7)	2.8 (0.5)	95.5 (0.7)	2.6 (0.5)	1.9 (0.4)
slope	both	100	85.3 (18.0)	13.8 (18.0)	0.9 (0.9)	62.2 (23.5)	35.8 (23.4)	2.0 (2.7)
		500	96.2 (1.0)	2.6 (0.9)	1.3 (0.6)	92.4 (1.7)	5.5 (1.5)	2.1 (0.8)
		1000	96.7 (0.6)	1.8 (0.4)	1.5 (0.4)	93.3 (0.9)	4.1 (0.9)	2.5 (0.5)

Classification ability

both	intercept	100	18.8	(13.7)	76.0	(16.6)	5.2	(7.2)	18.7	(15.2)	78.1	(16.7)	3.2	(4.1)
		500	35.4	(25.2)	58.7	(27.2)	6.0	(8.5)	30.6	(18.5)	65.1	(20.5)	4.3	(3.9)
		1000	70.5	(22.4)	24.3	(23.4)	5.2	(1.9)	46.4	(12.1)	48.2	(13.9)	5.4	(2.4)
both	slope	100	16.2	(13.2)	79.2	(16.8)	4.5	(6.0)	23.4	(22.3)	74.9	(23.6)	1.6	(2.3)
		500	69.7	(18.1)	24.7	(19.4)	5.6	(2.0)	69.8	(13.4)	25.2	(14.4)	5.0	(1.4)
		1000	80.5	(3.0)	12.0	(3.3)	7.4	(1.2)	81.1	(2.2)	11.9	(2.1)	7.0	(0.8)
both	both	100	16.7	(14.5)	80.3	(17.3)	3.0	(5.5)	19.4	(19.8)	79.7	(20.7)	0.9	(1.4)
		500	43.6	(30.5)	53.3	(32.3)	3.0	(2.8)	66.3	(19.6)	29.1	(21.1)	4.5	(1.9)
		1000	80.3	(10.5)	13.5	(11.1)	6.2	(1.2)	80.9	(3.3)	12.1	(3.5)	7.0	(1.0)

Dynamic classification

- Remember: $u_{i,k}(\boldsymbol{\theta}) = \mathrm{P}\left(U_i = k \,\middle|\, \boldsymbol{Y}_i = \boldsymbol{y}_i; \boldsymbol{\theta}, \mathcal{C}_i\right)$
- $\boldsymbol{Y}_i = Y_{i,1}$, then $\boldsymbol{Y}_i = (Y_{i,1}, Y_{i,2})$, then $\boldsymbol{Y}_i = (Y_{i,1}, Y_{i,2}, Y_{i,3}), \ldots$

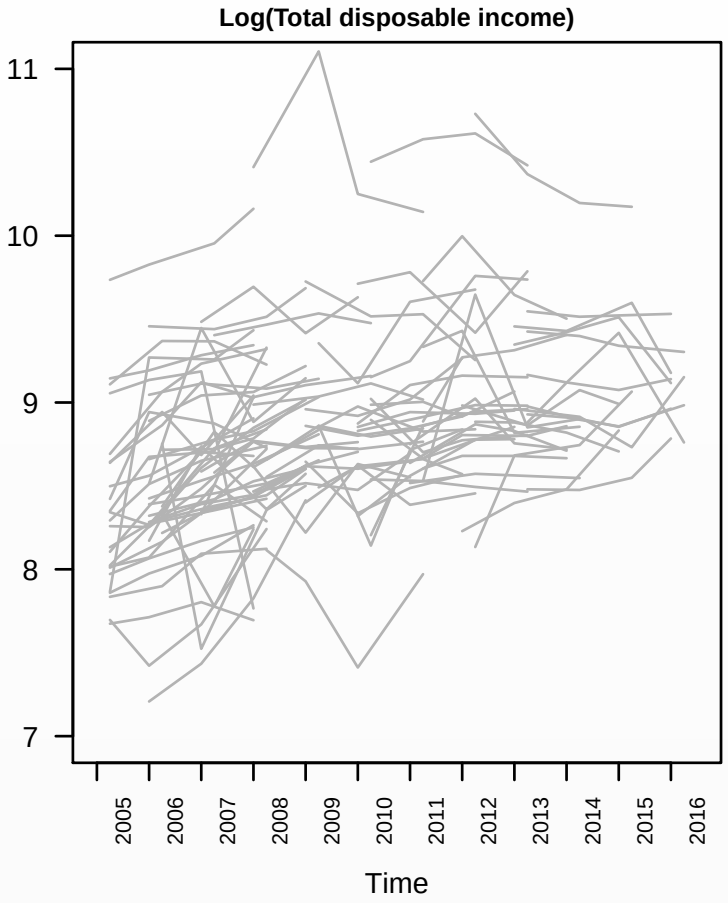


VI.

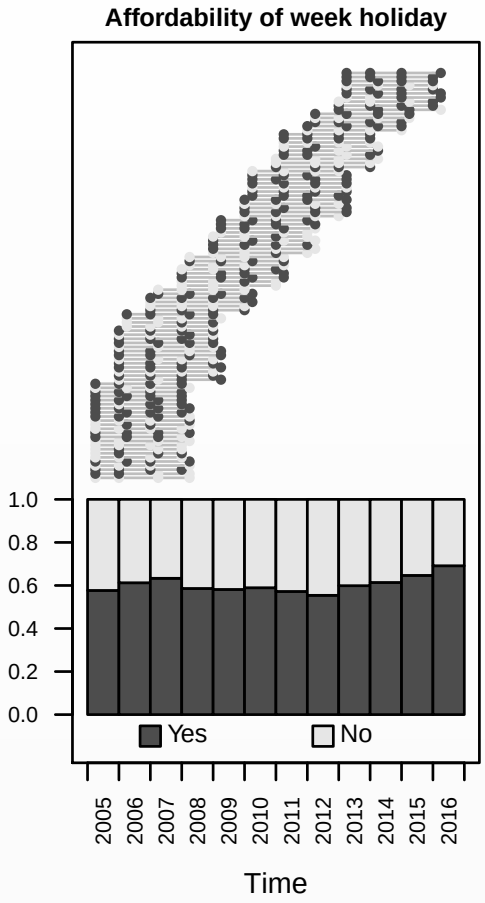
EU-SILC (Czech)

EU-SILC (Czech)

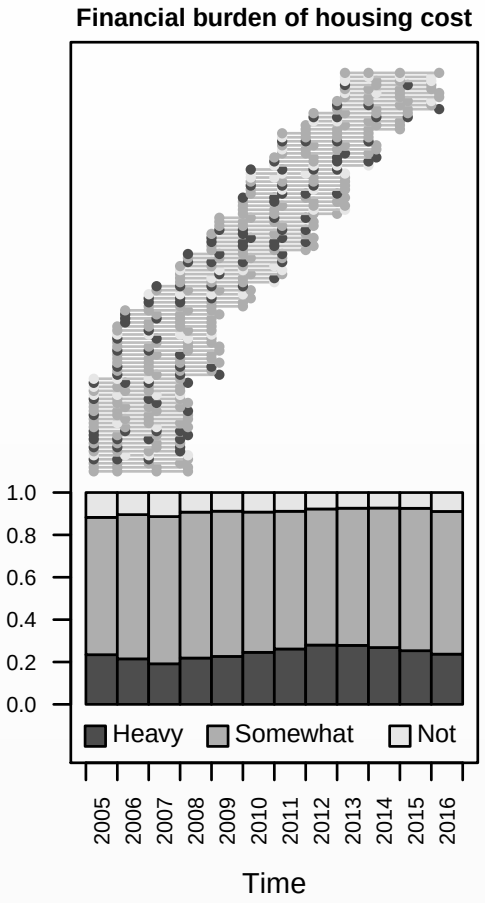
Numeric



Binary



Ordinal



Model

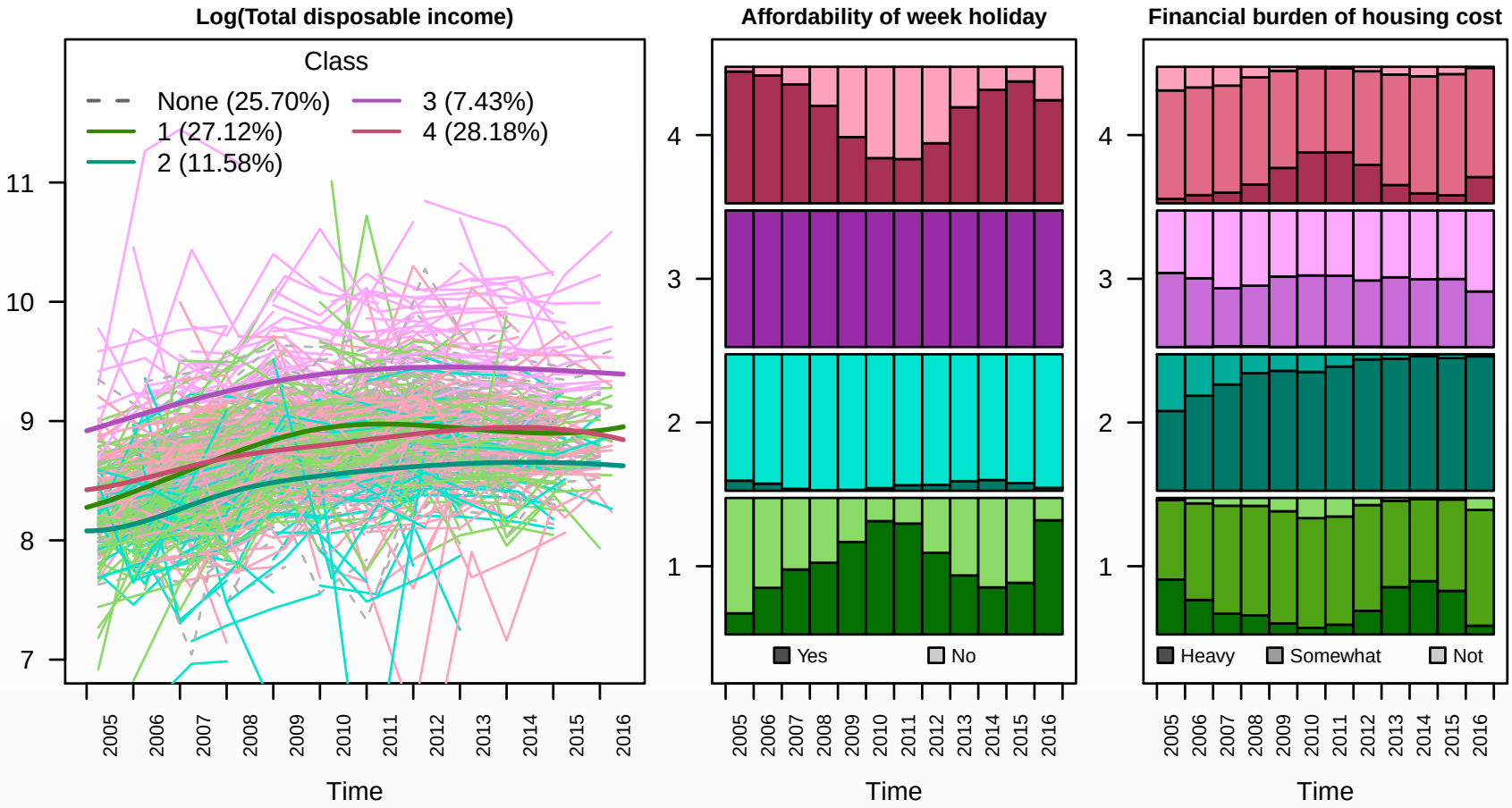
- Linear predictor (for outcome $r \in \{1, 2, 3\}$):

$$b^r_{0,i} + \underbrace{\beta^r_0 + \beta^r_1 B_1(t_{i,j}) + \cdots + \beta^r_5 B_5(t_{i,j})}_{\text{spline in time}} + \beta^r_6 \underbrace{S_{i,j}}_{\text{weighted family size}}$$

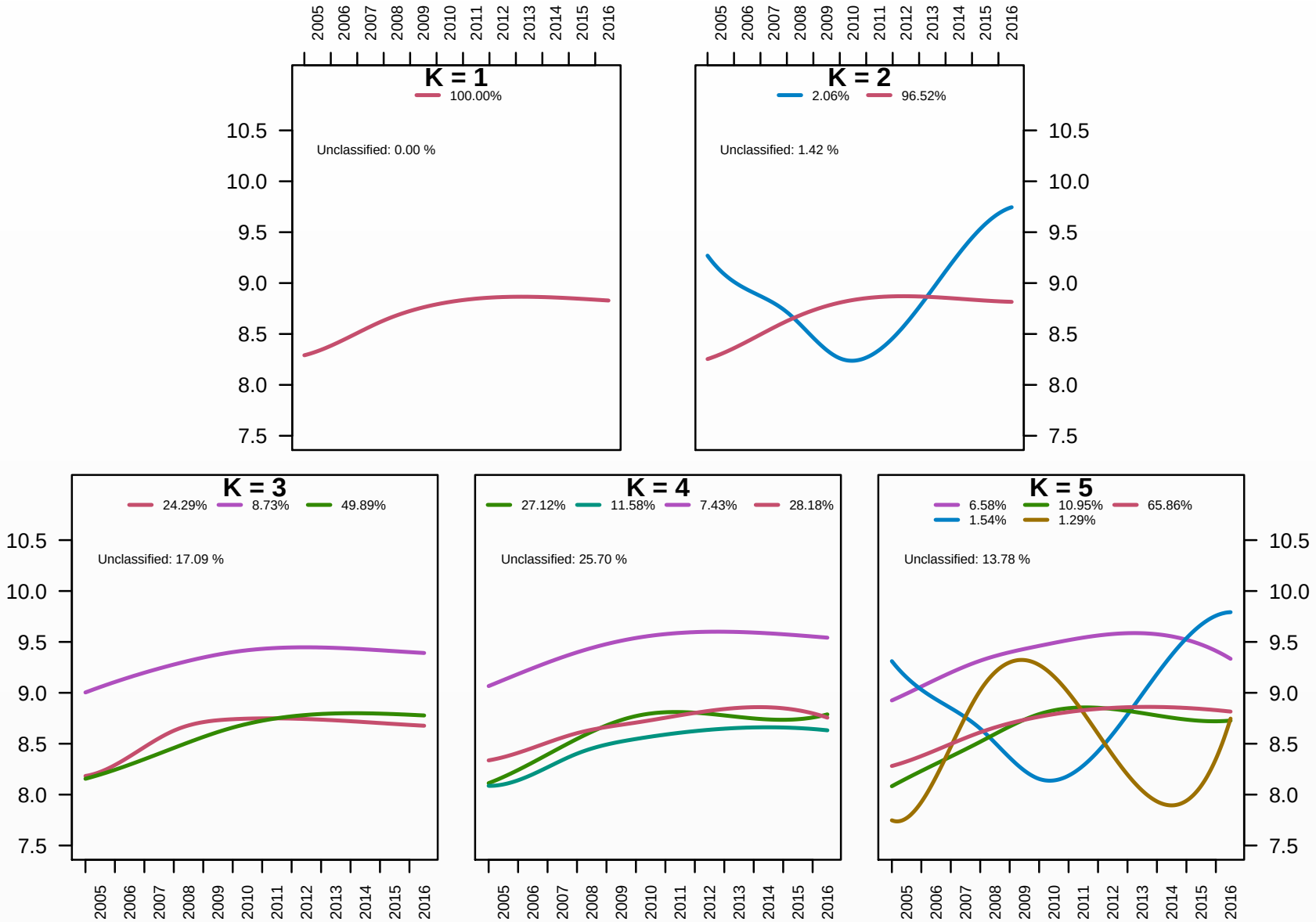
$$\Longrightarrow \text{all } \beta\text{'s possibly class specific (depend on } k)$$

- Random effect vector: $\boldsymbol{b}_{0,i} = \left(b^1_{0,i}, b^2_{0,i}, b^3_{0,i}\right)^\top$

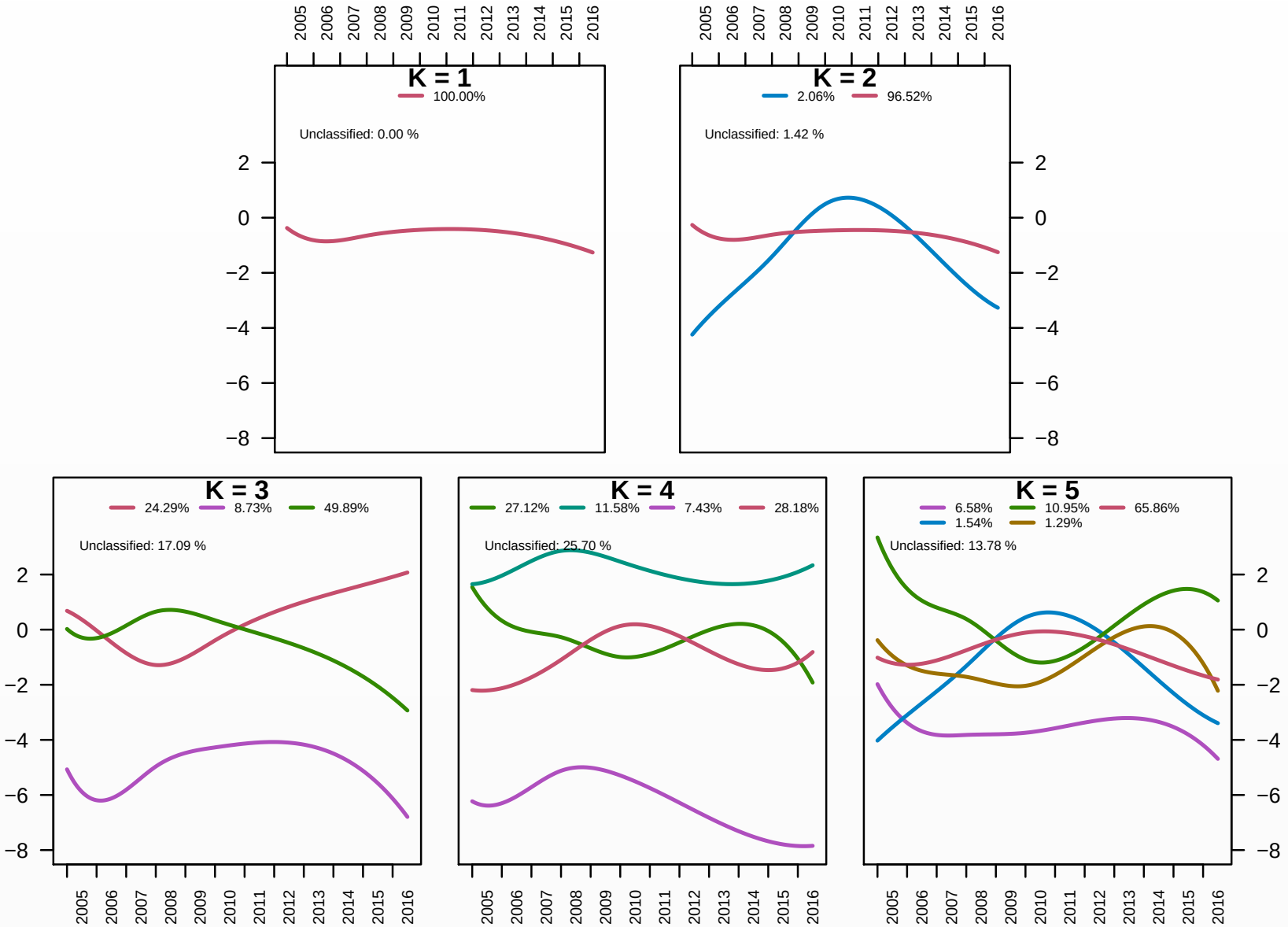
EU-SILC (Czech), $K = 4$



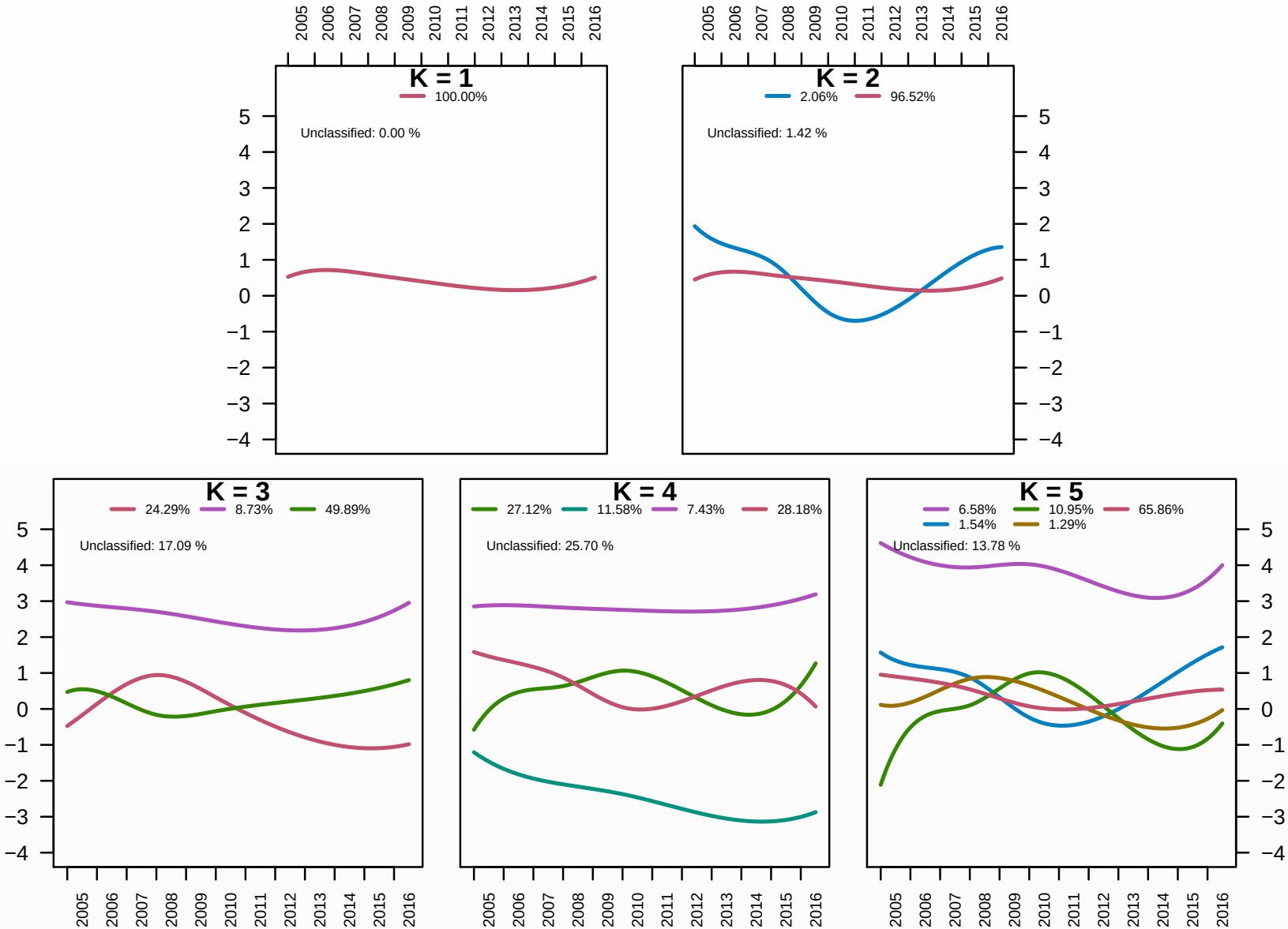
EU-SILC (Czech), log(total disposable income)



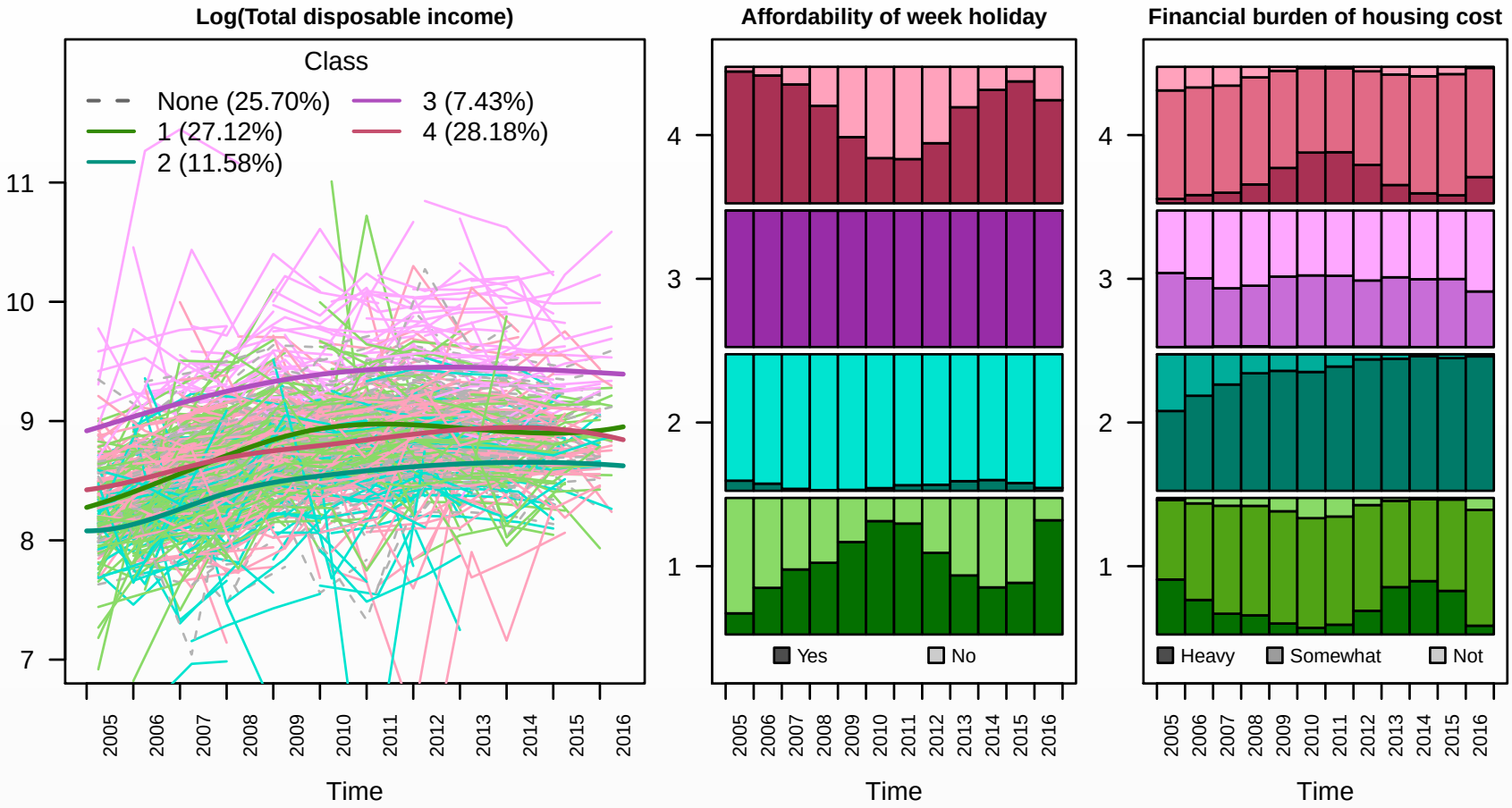
EU-SILC (Czech), affordability of week holiday)



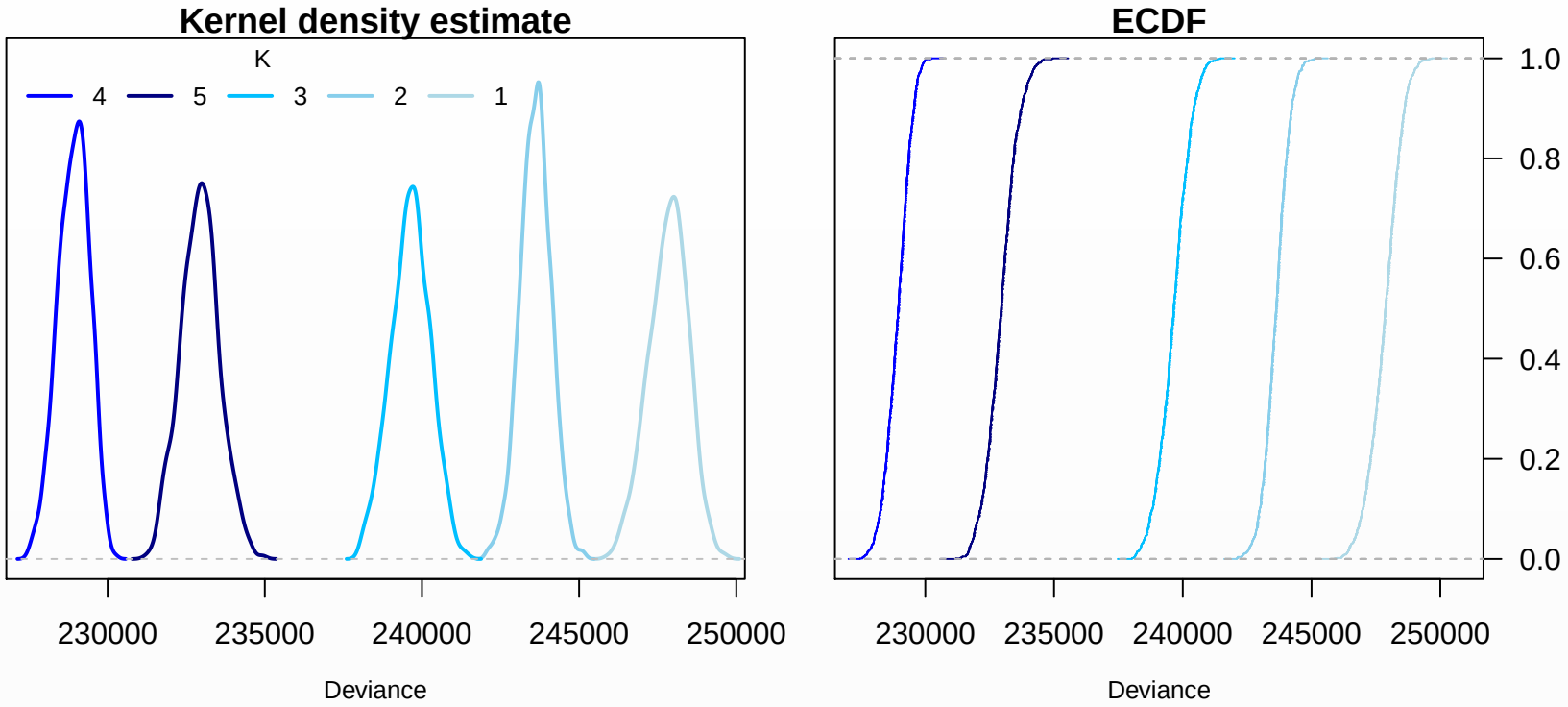
EU-SILC (Czech), financial burden of housing cost)



EU-SILC (Czech), $K = 4$



How many groups?



$$D^K(\boldsymbol{\theta}; \mathbb{Y}, \mathcal{C}) = -2 \log p(\mathbb{Y}|\boldsymbol{\theta}; \mathcal{C}) = -2 \sum_{i=1}^{n} \log p(\mathbb{Y}_i|\boldsymbol{\theta}; \mathcal{C}_i)$$

VII.

Conclusions

Conclusions

- Model based clustering is a general clustering method applicable with almost any data structures
- The only thing we need to do is to specify a model for data at hand which also expresses presumable differences between clusters
- Bayesian MCMC calculation solves (relatively easily) problems with a mixture likelihood and other possibly ugly integrals as soon as the model can be written hierarchically
- With MCMC based inference, also uncertainty in classification can (realively easily) be taken into account
- MBC also naturally allows for development of, in fact, a discriminant procedure allowing to classify "new" subjects (diagnosis in medicine, . . .)

Work in progress

- Also nominal categorical data are quite frequent (especially in social sciences. . .)
- Thresholding approach could in principle be used as well to link categorical outcome to the numeric one

$\longrightarrow$ Interpretation of the model parameters (trends) would, however, be somehow strange

- Alternative: sort of generalized linear mixed model (GLMM) to be used for categorical outcomes

Vielen Dank für Ihre Aufmerksamkeit!