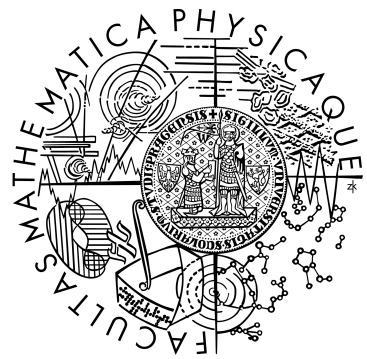


FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

Jan Vávra, Arnošt Komárek

Department of Probability and Mathematical Statistics

Classification Based on Longitudinal Data of a Mixed Type

16th July 2019

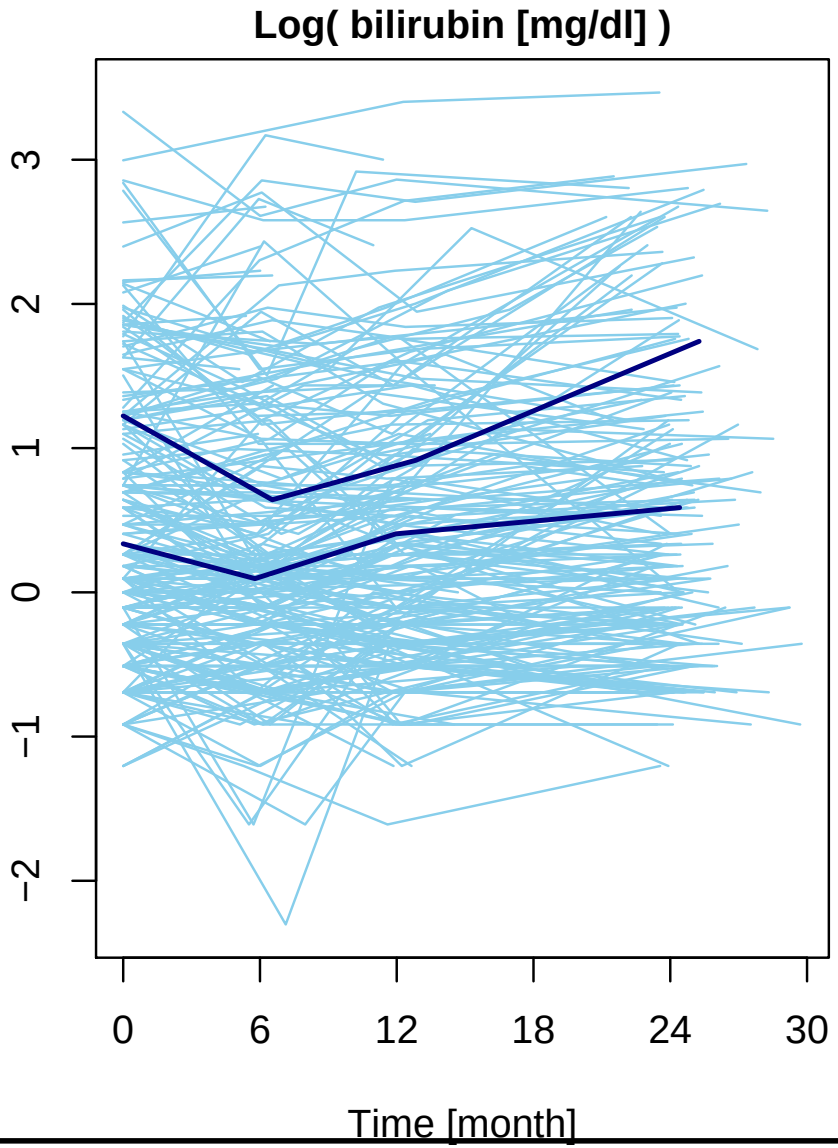
Mayo Clinic PBC Data (1974 – 1984)

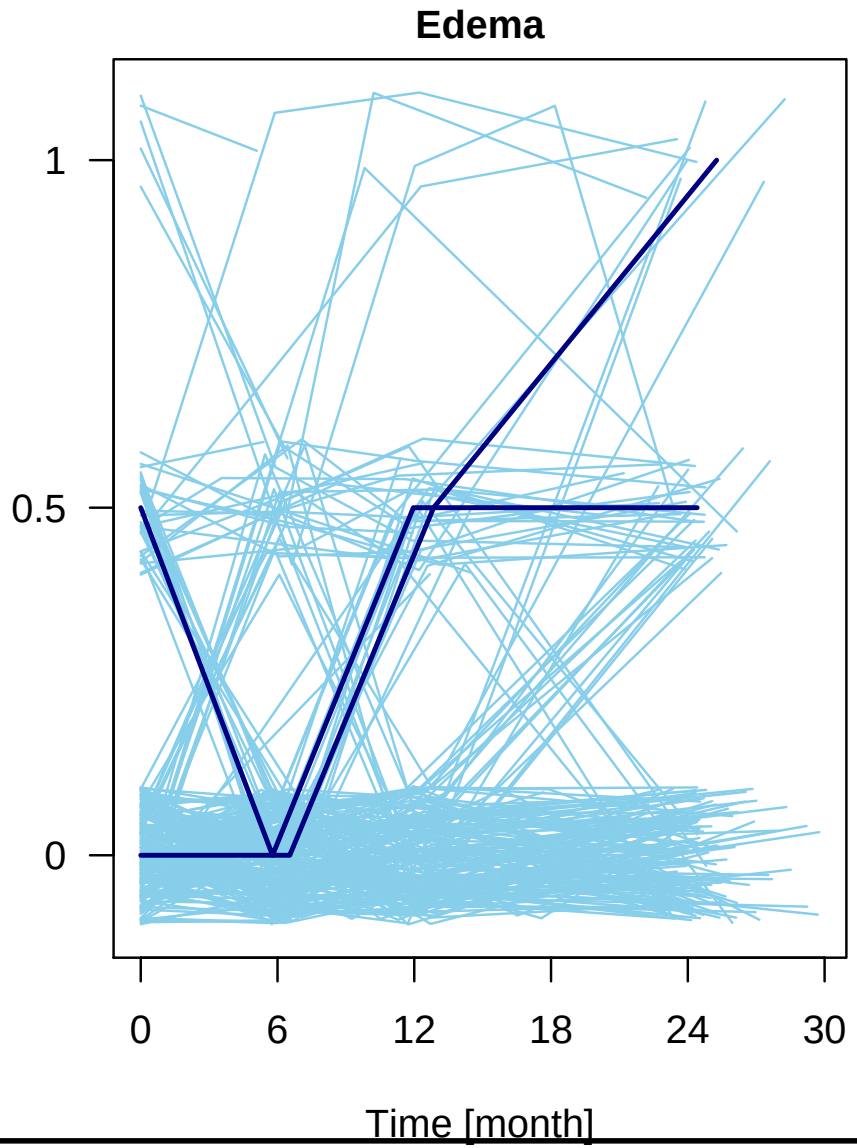
Flemming and Harrington (1991)

Subset PBC910 (see `mixAK` in )

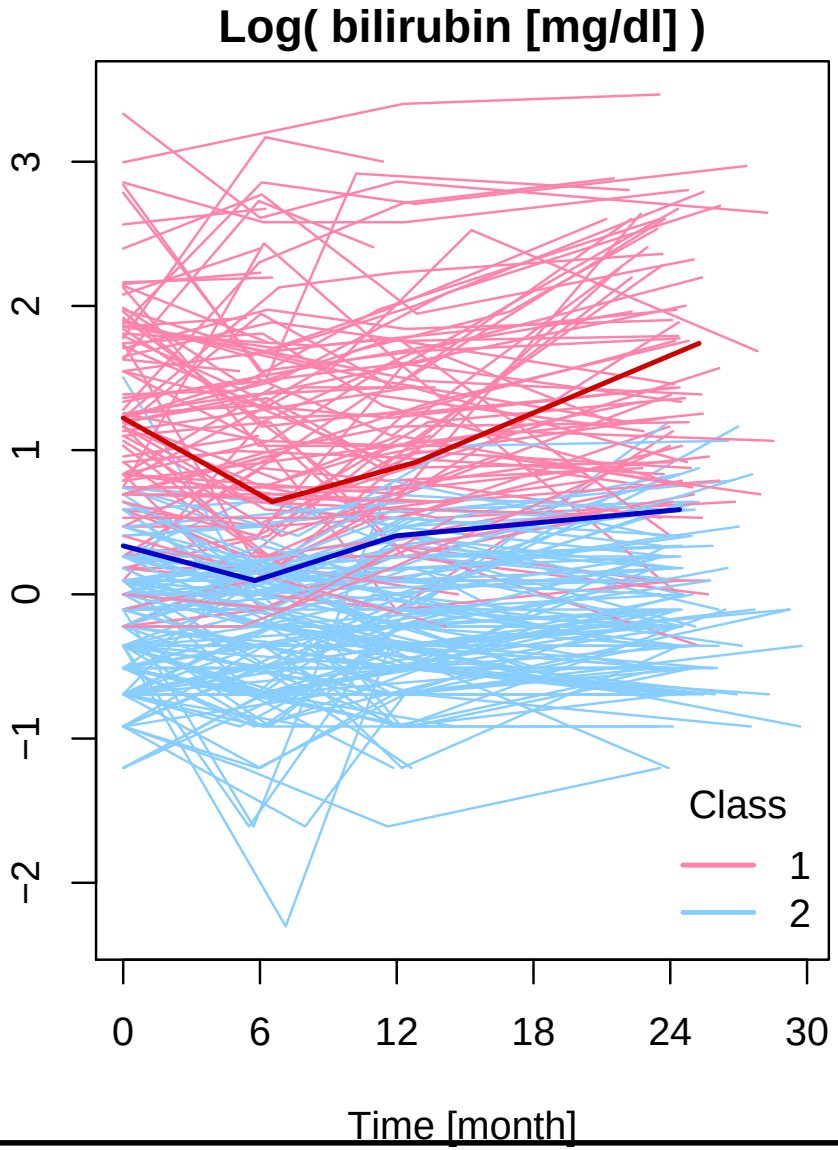
Patient ID	Month	Gender	Drug	Log bilirubin	Spiders	Edema	⋯
3	0.00	male	D-penicillamine	0.34	0	0.5	⋯
	5.78	male	D-penicillamine	0.10	1	0	⋯
	11.96	male	D-penicillamine	0.41	0	0.5	⋯
	24.41	male	D-penicillamine	0.59	1	0.5	⋯
5	0.00	female	placebo	1.22	1	0	⋯
	6.54	female	placebo	0.64	0	0	⋯
	12.85	female	placebo	0.92	0	0.5	⋯
	25.26	female	placebo	1.74	1	1	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

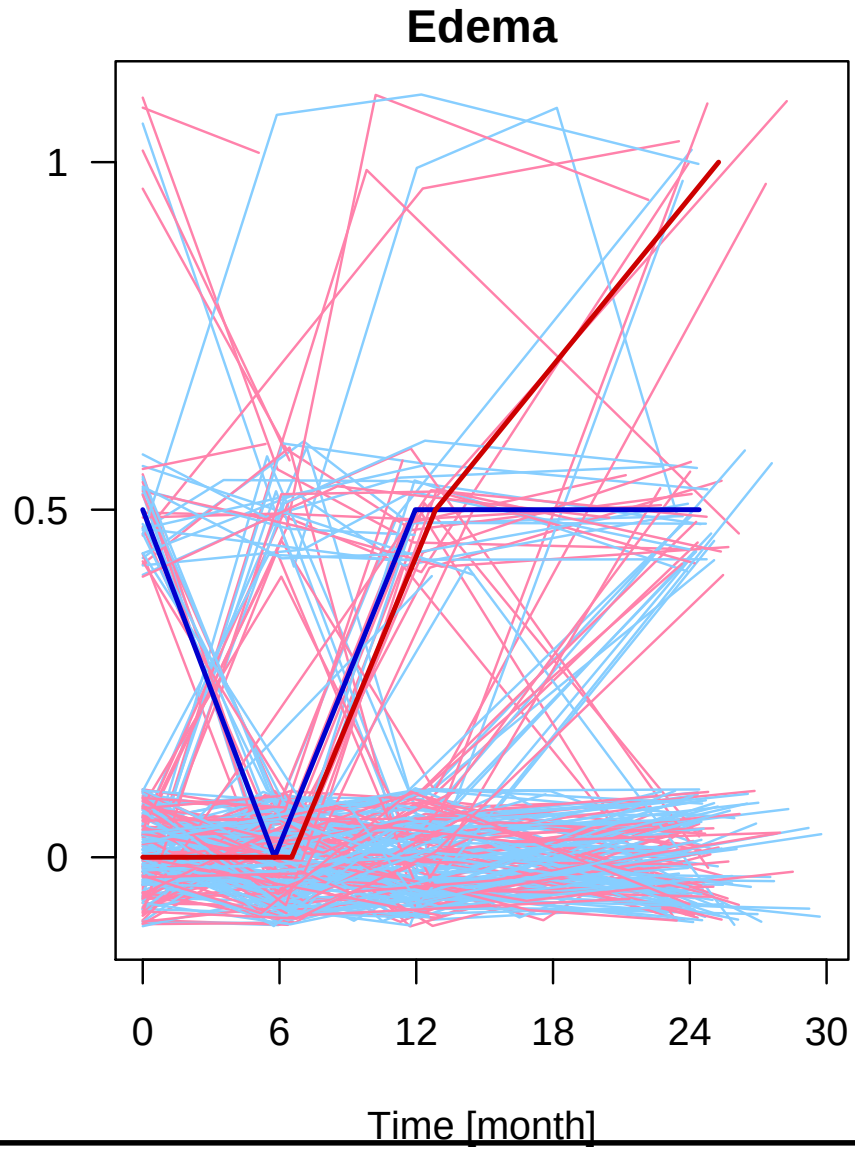
Mayo Clinic PBC Data - longitudinal profiles



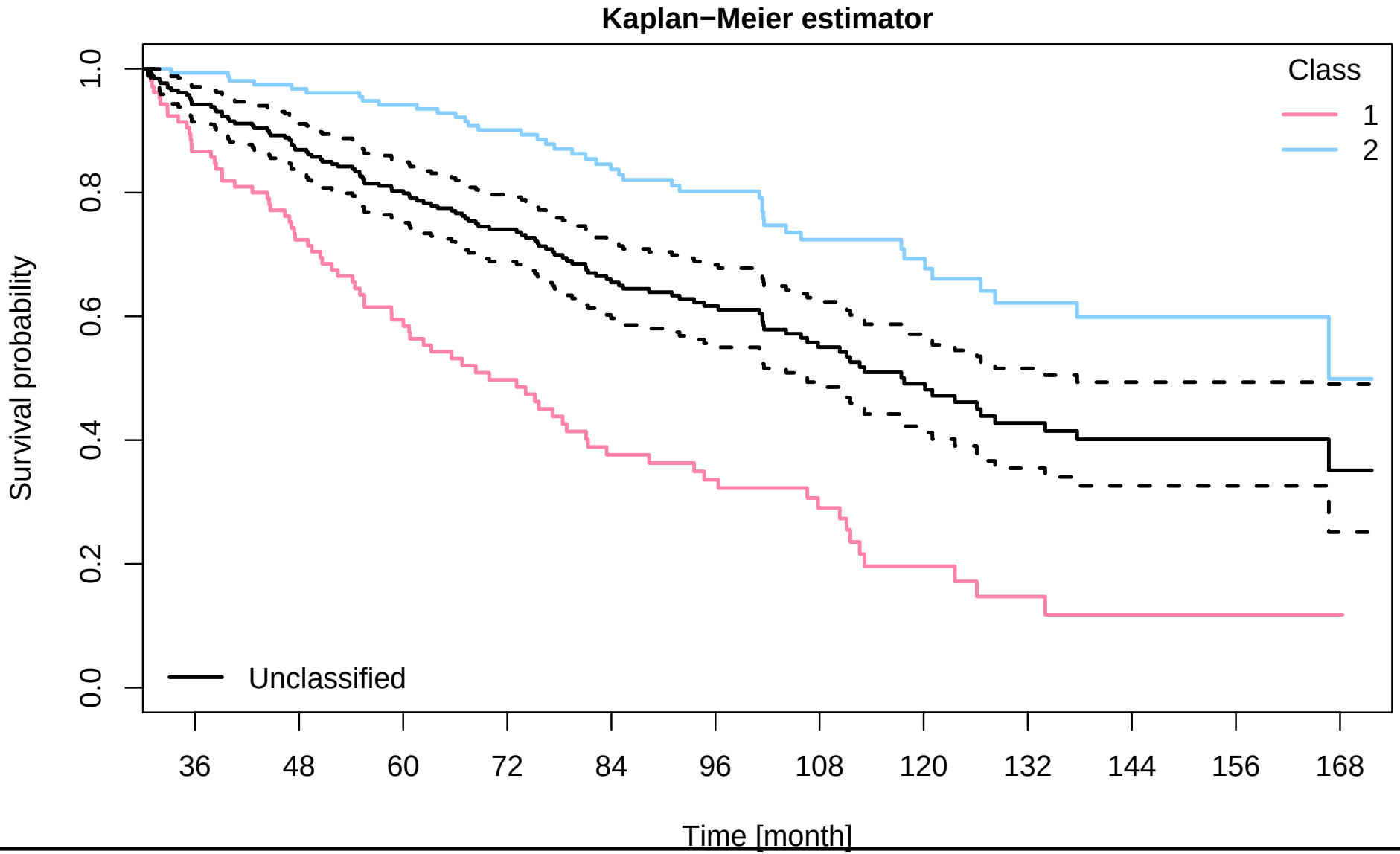


Mayo Clinic PBC Data - after classification





Mayo Clinic PBC Data - usage of classification



Model based clustering

- Origins: Banfield and Raftery (1993)
- Outcomes: $\boldsymbol{Y}_i, i = 1, \ldots, n$
- K models: $f_k\left(\boldsymbol{y}_i; \boldsymbol{x}_i, \psi, \psi^{(k)}\right), k \in \{1, \ldots, K\}$
- Latent group indicators: $U_i \in \{1, \ldots, K\}, \quad \boldsymbol{Y}_i | U_i = k \sim f_k$
- Group probabilities: $0 < w_k = \mathrm{P}\left[U_i = k\right]$
- 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 - classification rule

- By Bayes Theorem:

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

- Classification rule:

$$\widehat{U}_i := k \iff k = \underset{\ell \in \{1,\ldots,K\}}{\arg\max}\ \widehat{p_{i,\ell}}$$

- Estimation: Bayesian approach and MCMC methods (Gibbs sampling)
- Software: implemented in R using the C programming language

Numeric outcome

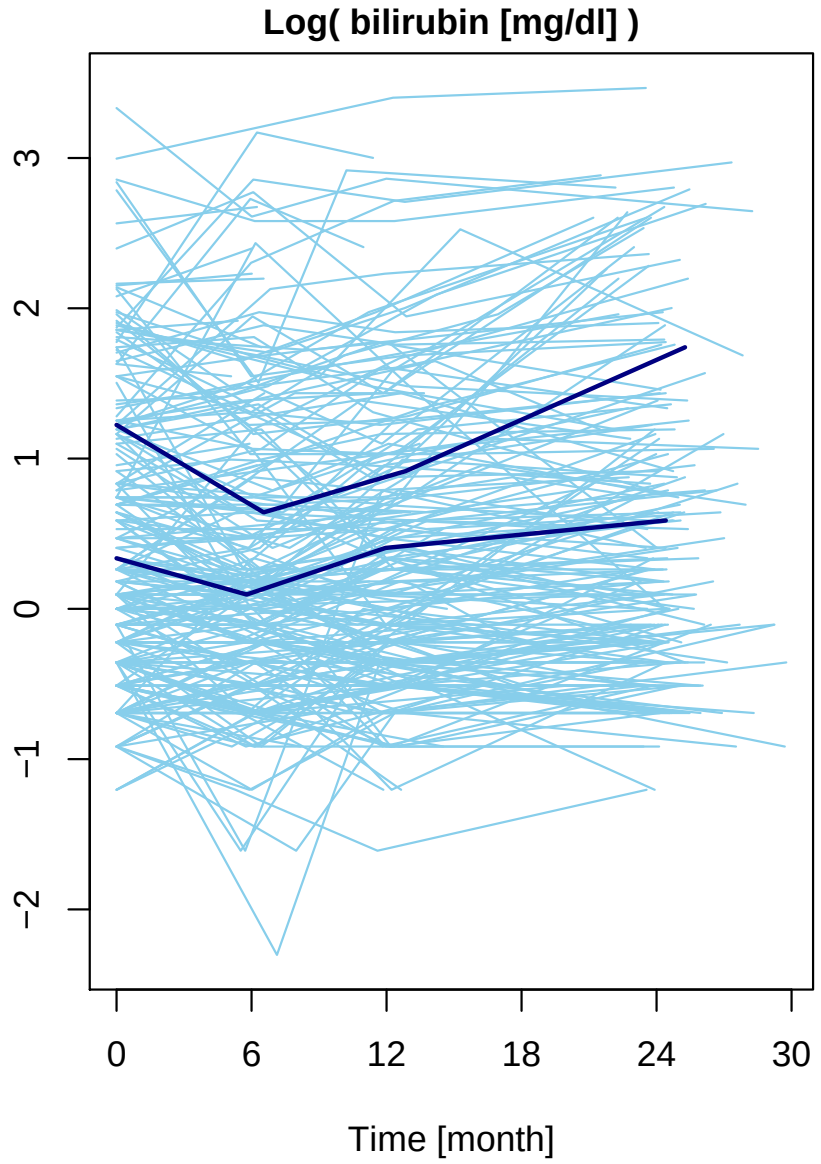
Logarithm of serum bilirubin [mg/dl]

Used model:

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

For example:

- Fixed: $\beta_0 + \beta_1 \cdot \text{Month} + \cdots$
  $\cdots +$ other regressors
- Random: $b_0 + b_1 \cdot \text{Month}$



Binary + Ordinal outcome

Presence and status of edema

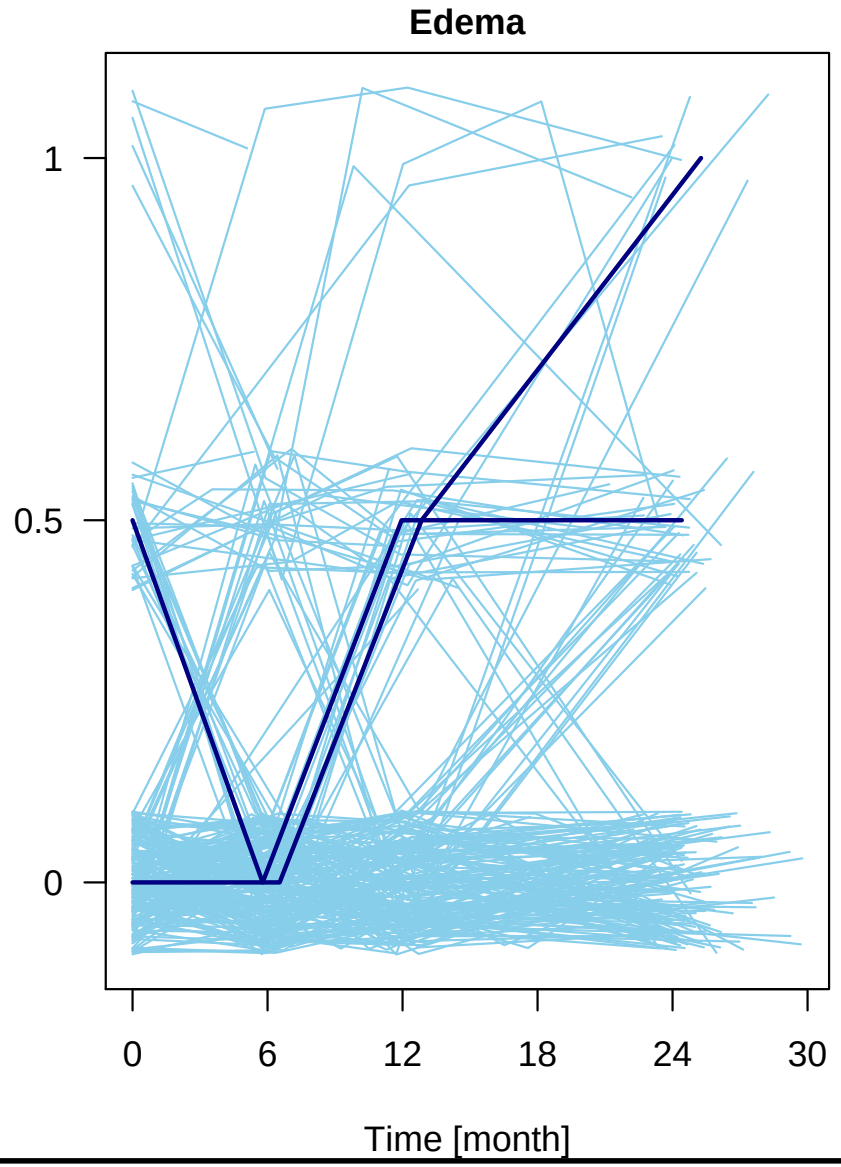
$L = 3$ ordered categories:

- 1 = Edema despite diuretic therapy,
- 0.5 = Untreated/Succesfully treated,
- 0 = No edema.

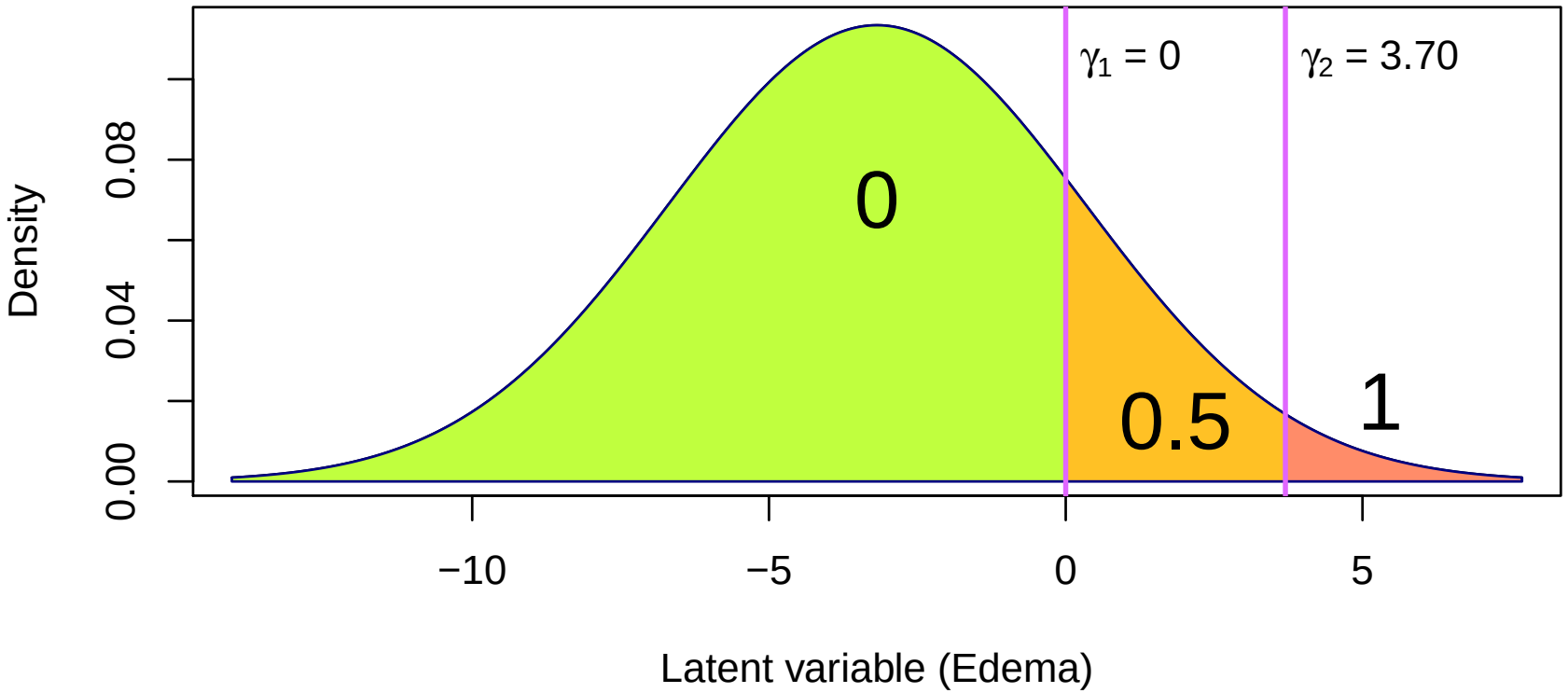
Used model

- Latent variable modelling: $Y | Y^\star$
  - $Y^\star$ latent numeric outcome
  - Thresholding by

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



Binary + Ordinal outcome - latent variable modelling



- Fixed threshold: $\gamma_1 = 0$
- Estimate other thresholds: $\gamma_2, \ldots$
- $Y_{ij}^{\star} | \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}; \boldsymbol{b}_i \sim \mathrm{N}\left(\boldsymbol{X}_{ij}^{\top}\beta + \boldsymbol{Z}_{ij}^{\top}\boldsymbol{b}_i, 1\right)$

$$
\begin{aligned}
Y_{ij} = 0 &\iff & Y_{ij}^{\star} &\leq \gamma_1 \\
Y_{ij} = 0.5 &\iff \gamma_1 < & Y_{ij}^{\star} &\leq \gamma_2 \\
Y_{ij} = 1 &\iff \gamma_2 < & Y_{ij}^{\star} &
\end{aligned}
$$

Joint modelling

Models for individual outcomes

- Numeric: $Y_{ij}^{\mathrm{N}} \left| \boldsymbol{X}_{ij}^{\mathrm{N}}, \boldsymbol{Z}_{ij}^{\mathrm{N}}; \boldsymbol{b}_i^{\mathrm{N}} \sim \mathrm{N}\left(\left(\boldsymbol{X}_{ij}^{\mathrm{N}}\right)^\top \beta^{\mathrm{N}} + \left(\boldsymbol{Z}_{ij}^{\mathrm{N}}\right)^\top \boldsymbol{b}_i^{\mathrm{N}}, \left(\tau^{\mathrm{N}}\right)^{-1} \right) \right.$
- Binary/Ordinal: $Y_{ij}^{\mathrm{O},\star} \left| \boldsymbol{X}_{ij}^{\mathrm{O}}, \boldsymbol{Z}_{ij}^{\mathrm{O}}; \boldsymbol{b}_i^{\mathrm{O}} \sim \mathrm{N}\left(\left(\boldsymbol{X}_{ij}^{\mathrm{O}}\right)^\top \beta^{\mathrm{O}} + \left(\boldsymbol{Z}_{ij}^{\mathrm{O}}\right)^\top \boldsymbol{b}_i^{\mathrm{O}}, 1 \right) \right.$
  - $Y_{ij}^{\mathrm{O},\star} \overset{\text{thresholding}}{\longrightarrow} Y_{ij}^{\mathrm{O}}$

Joining individual models through joint distribution of random effects:

$$\boldsymbol{b}_i = \begin{pmatrix} \boldsymbol{b}_i^N \\ \boldsymbol{b}_i^O \end{pmatrix} \sim \mathrm{N}\left(\boldsymbol{\mu} = \begin{pmatrix} \mu^N \\ \mu^O \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma^{NN} & \Sigma^{NO} \\ \Sigma^{ON} & \Sigma^{OO} \end{pmatrix} \right) \quad \text{leading to}$$

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

Final model

$$\boldsymbol{Y}_i = \begin{pmatrix} \boldsymbol{Y}_i^{\mathrm{N}} \\ \boldsymbol{Y}_i^{\mathrm{O}} \end{pmatrix}, \quad \boldsymbol{Y}_i^{\star} = \begin{pmatrix} \boldsymbol{Y}_i^{\mathrm{N}} \\ \boldsymbol{Y}_i^{\mathrm{O},\star} \end{pmatrix}, \quad \mathcal{C}_i = \begin{Bmatrix} \boldsymbol{X}_{i1}^{\mathrm{N}} & \cdots & \boldsymbol{X}_{in_i}^{\mathrm{N}} & \boldsymbol{Z}_{i1}^{\mathrm{N}} & \cdots & \boldsymbol{Z}_{in_i}^{\mathrm{N}} \\ \boldsymbol{X}_{i1}^{\mathrm{O}} & \cdots & \boldsymbol{X}_{in_i}^{\mathrm{O}} & \boldsymbol{Z}_{i1}^{\mathrm{O}} & \cdots & \boldsymbol{Z}_{in_i}^{\mathrm{O}} \end{Bmatrix}$$

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

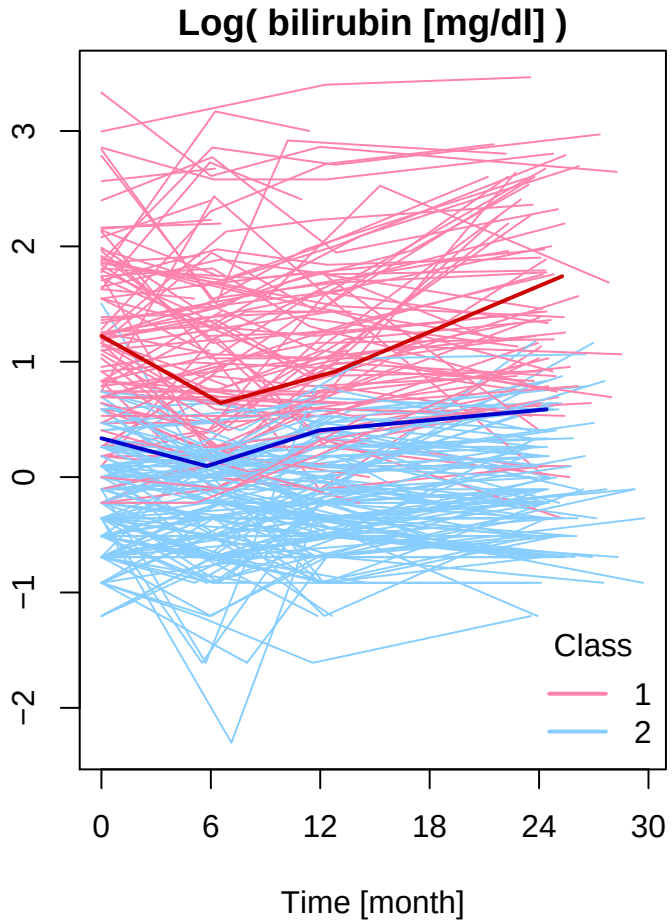
PDF of k-th model:

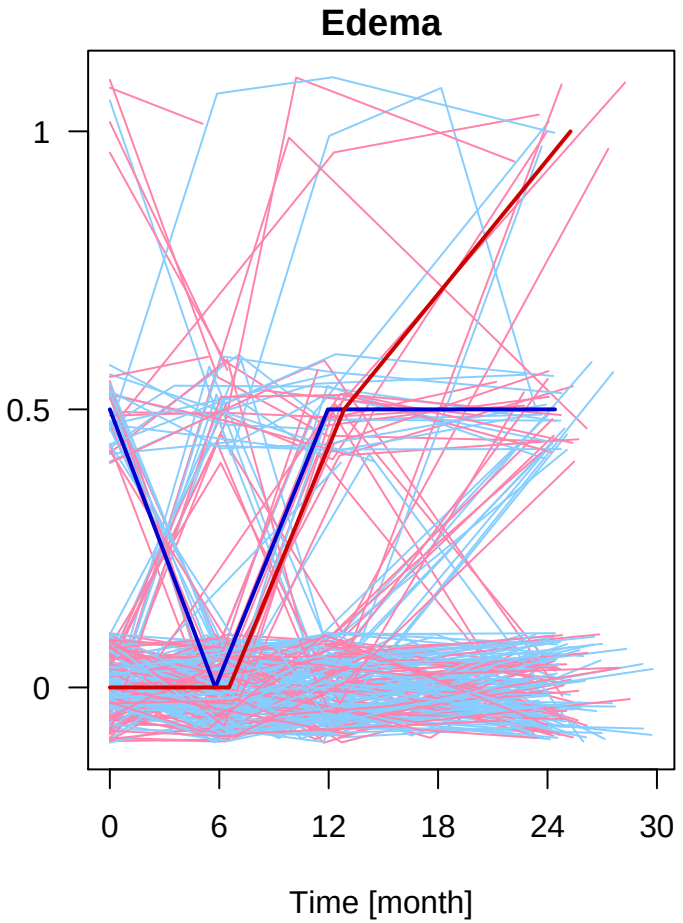
$$f_k\left(\boldsymbol{y}_i \left| \mathcal{C}_i; \gamma, \boldsymbol{\tau}, \beta^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}\right.\right) = \int\int p\left(\boldsymbol{y}_i^{\mathrm{O}} \left| \boldsymbol{y}_i^{\mathrm{O},\star}; \gamma\right.\right) \cdot$$
$$\cdot\, p\left(\boldsymbol{y}_i^{\star} \left| \mathcal{C}_i, \boldsymbol{b}_i; \beta^{(k)}, \boldsymbol{\tau}\right.\right) \cdot p\left(\boldsymbol{b}_i \left| \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}\right.\right) \mathrm{d}\boldsymbol{b}_i \mathrm{d}\boldsymbol{y}_i^{\mathrm{O},\star}$$

MLE complicated $\longrightarrow$ MCMC estimation

Classification probability

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





Classification rule of new patient

New patient: $\boldsymbol{Y}_{\text{new}}, \boldsymbol{X}_{\text{new}}, \boldsymbol{Z}_{\text{new}}$ and unobserved $U_{\text{new}}, \boldsymbol{b}_{\text{new}}, \boldsymbol{Y}_{\text{new}}^{\text{O},\star}$

Bayesian estimator: $\widehat{p}_{\text{new},k} = \mathrm{E}\left[p_{\text{new},k}\left(\boldsymbol{\theta}\right) \mid \text{data } \mathbb{Y}, \mathcal{C}\right]$

Calculation: use generated sample $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^M$ from $p(\boldsymbol{\theta} \mid \text{data } \mathbb{Y}, \mathcal{C})$

Gradual classification:

$$
\begin{array}{llll}
\boldsymbol{y}_{\text{new},1} & & & \longrightarrow \; p^1_{\text{new},k} \\
\boldsymbol{y}_{\text{new},1} & \boldsymbol{y}_{\text{new},2} & & \longrightarrow \; p^2_{\text{new},k} \\
\vdots & \vdots & \ddots & \vdots \\
\boldsymbol{y}_{\text{new},1} & \boldsymbol{y}_{\text{new},2} & \cdots \quad \boldsymbol{y}_{\text{new},n_{\text{new}}} & \longrightarrow \; p^{n_{\text{new}}}_{\text{new},k}
\end{array}
$$

Further development

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

Further development

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

Any interesting suggestions?

Further development

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

Any interesting suggestions?

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] R. Fleming, T. and P. Harrington, D. *Counting Processes and Survival Analysis*. New York: John Wiley and Sons, New York, 1991.

[3] 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.

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

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

[6] S. Wilhelm and B. Manjunath. tmvtnorm: A package for the truncated multivariate normal distribution. *The R Journal*, 2, 06 2010.