Sdružené modelování spojitých i diskrétních longitudinálních dat s exkurzí do diskriminační a shlukové analýzy

Arnošt Komárek

Katedra pravděpodobnosti a matematické statistiky

Matematicko-fyzikální fakulta
Univerzity Karlovy v Praze

XVI. ROBUST 2010

Králíky, 31. ledna – 5. února 2010

Obsah

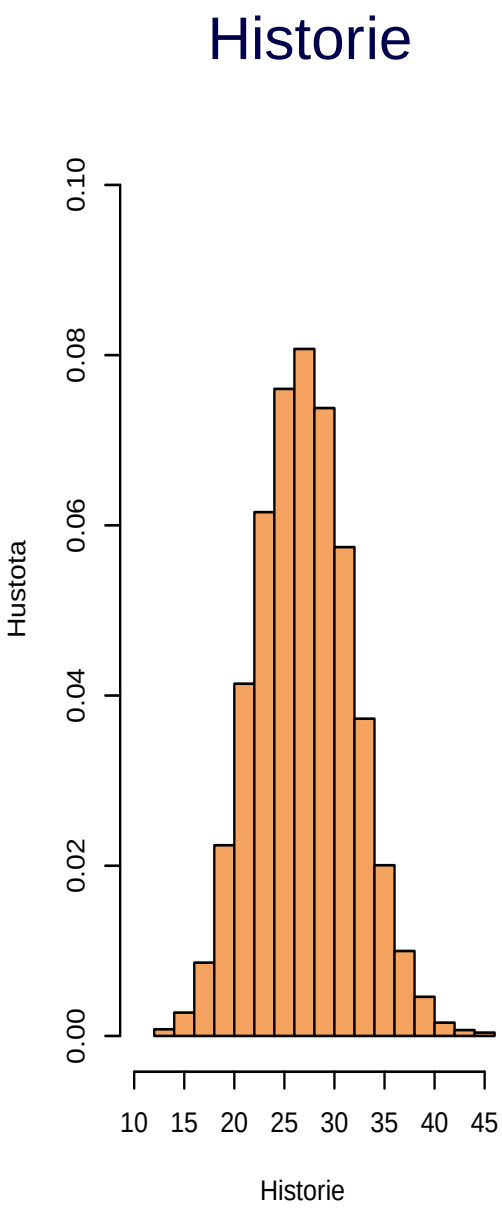
Část I

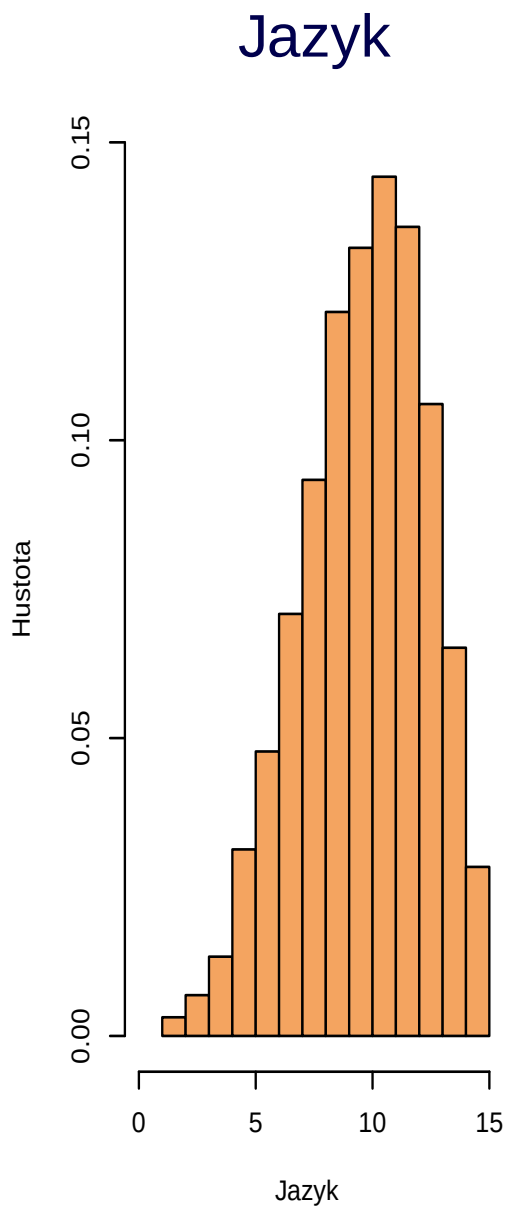
Zpět do Nečtin a taky trochu do Roháčů

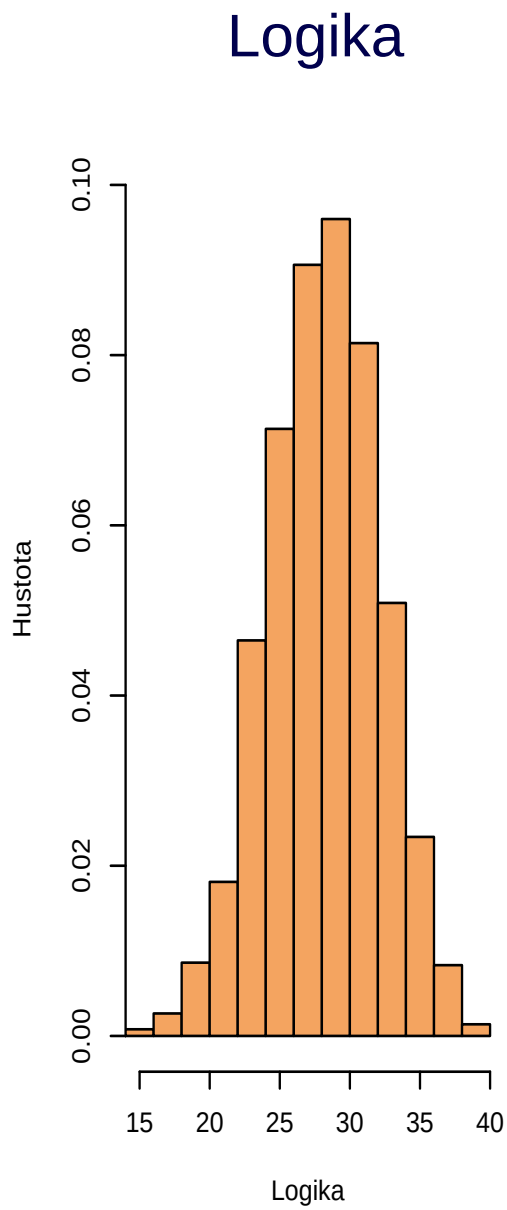
XI. ROBUST (září 2000, Nečtiny)

- *Matematika mezi §§§ aneb něco málo o diskriminaci*
- 12. termín přijímaček na Právnickou fakultu UK v Praze v roce 1999 s únikem (a prodejem) zadání testů
  (které byly dle tehdejšího rektora UK i děkana PF dílem gangsterské mafie stojící mimo fakultu)
- Lze vytipovat uchazeče napojené na tuto mafii?

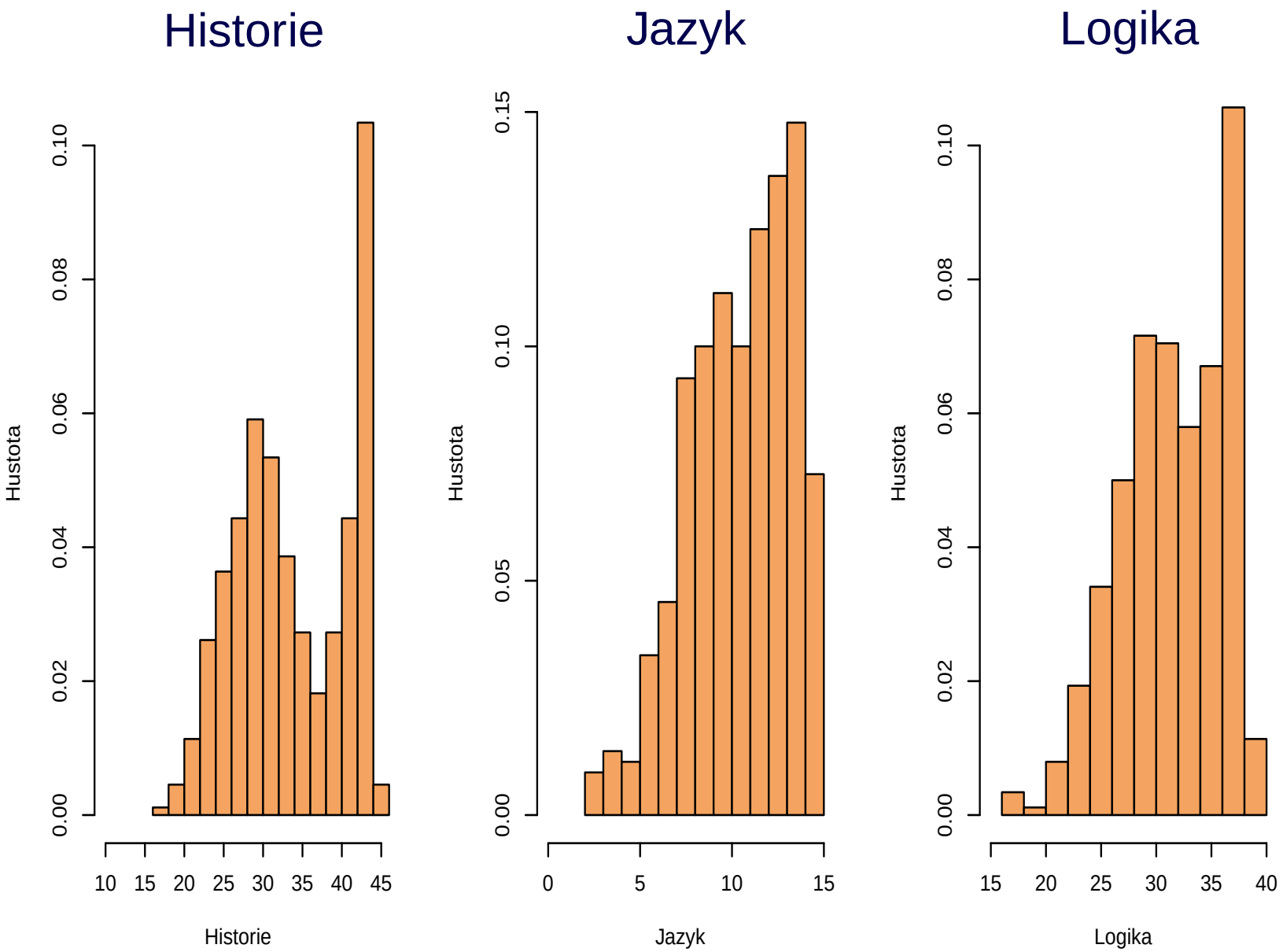
Přijímačky na PF UK v roce 1999 (Termíny č. 1–11)



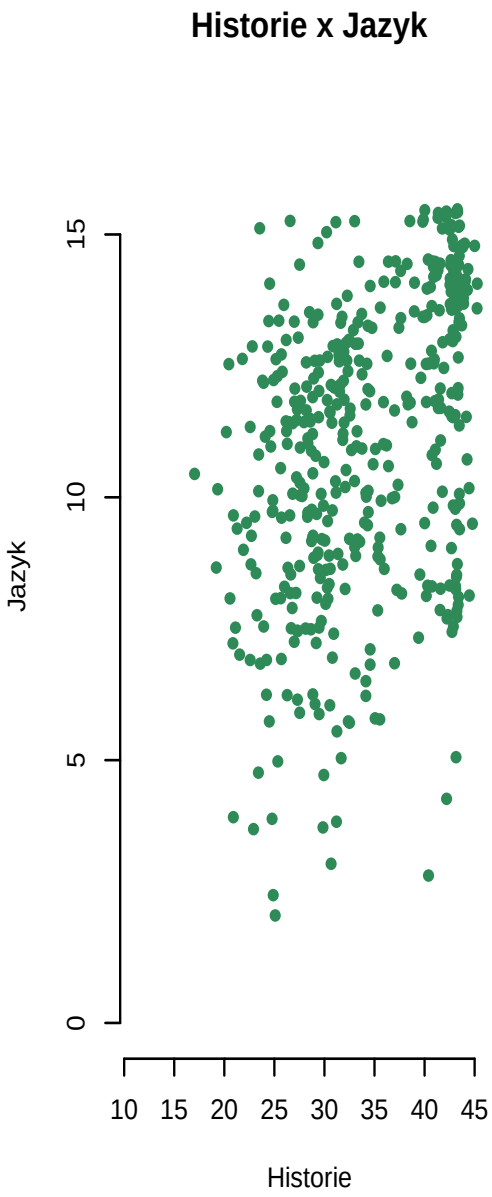


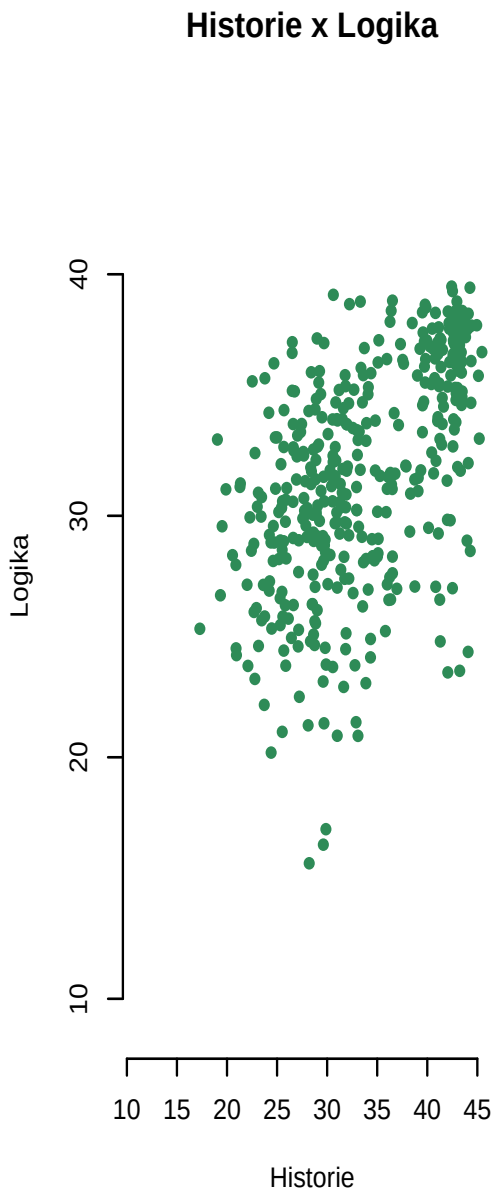


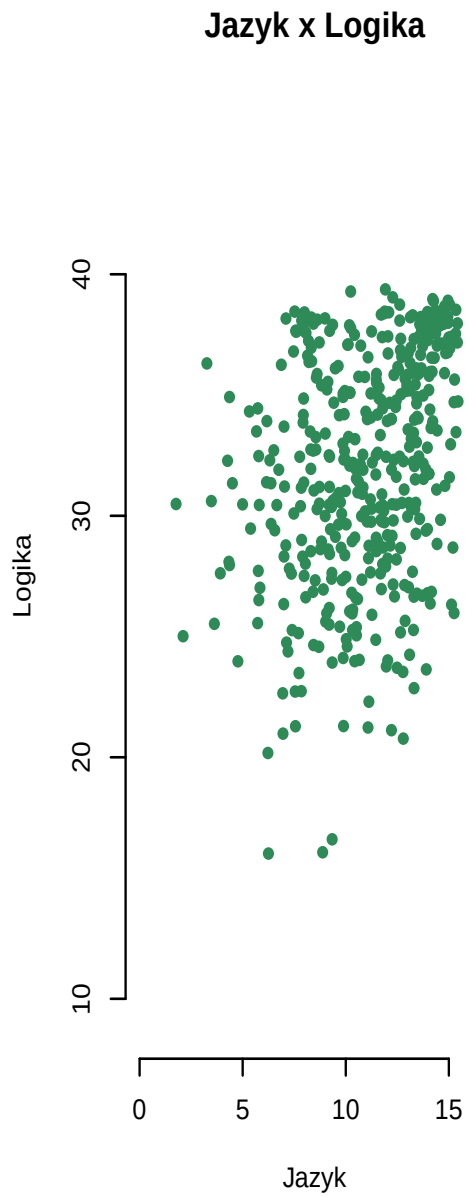
Přijímačky na PF UK v roce 1999 (Termín č. 12)



Přijímačky na PF UK v roce 1999 (Termín č. 12)





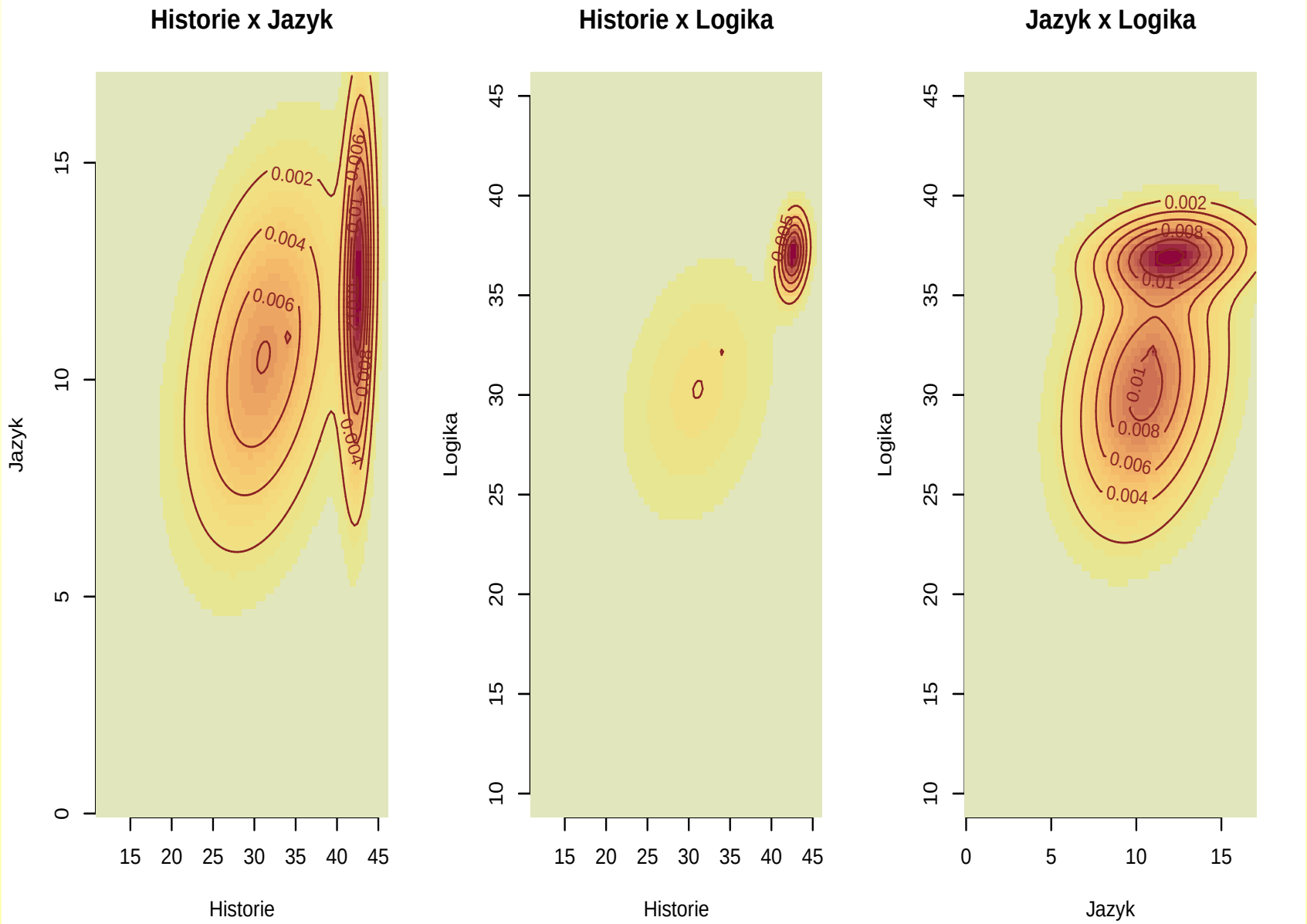


Modelově založená shluková analýza (Model based clustering)

- $\boldsymbol{Y} = (Y_1, \ldots, Y_p)'$: náhodný vektor pro odezvu (např. body z jednotlivých testů u náhodně vybraného uchazeče) s hustotou $p(\boldsymbol{y})$
- Směsový model pro $\boldsymbol{Y}$: $p(\boldsymbol{y}) = \sum_{k=1}^{K} w_k \, \phi(\boldsymbol{y}; \, \boldsymbol{\eta}_k)$
- K . . . počet shluků
- $\boldsymbol{w} = (w_1, \ldots, w_K)'$ . . . proporce (váhy) jednotlivých shluků
- $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K$ . . . parametry rozdělení odezvy v jednotlivých shlucích
- ϕ . . . vhodná hustota
- $\boldsymbol{\theta} = (w_1, \ldots, w_K, \, \boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_K')'$ . . . parametry modelu

- $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{y}) = \sum_{k=1}^{K} w_k \, \phi(\boldsymbol{y}; \boldsymbol{\eta}_k)$
- Nečtiny (XI. ROBUST): $\phi(\boldsymbol{y}; \boldsymbol{\eta}_k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$
  - Odhad metodou maximální věrohodnosti (EM algoritmus)
- Roháče (XV. ROBUST): $\phi(\boldsymbol{y}; \boldsymbol{\eta}_k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
  - Bayesovský odhad přes MCMC

Přijímačky na PF UK v roce 1999 (Termín č. 12)



- Pro shlukování: i.i.d. náhodné veličiny $U_1, \ldots, U_n$
- $U_i \in \{1, \ldots, K\}$
- $\mathrm{P}(U_i = k) = w_k, k = 1, \ldots, K, \; i = 1, \ldots, n$

Modelově založená shluková analýza

Směsový model zapsaný hierarchicky

- $\boldsymbol{Y}_i \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{y}) = \sum_{k=1}^{K} w_k \, \phi(\boldsymbol{y}; \boldsymbol{\eta}_k)$

- Hierarchicky:

$$\left.\begin{array}{rcl} \mathrm{P}(U_i = k) & = & w_k \\ p(\boldsymbol{y}_i \,|\, U_i = k) & = & \phi(\boldsymbol{y}_i; \boldsymbol{\eta}_k) \end{array}\right\} \quad k = 1, \ldots, K$$

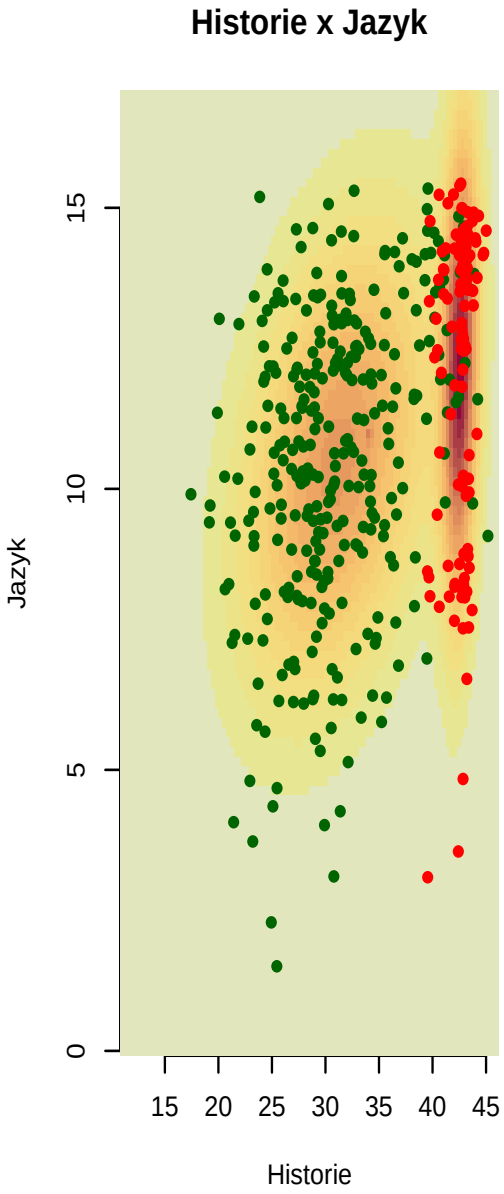
- Bayesova věta:

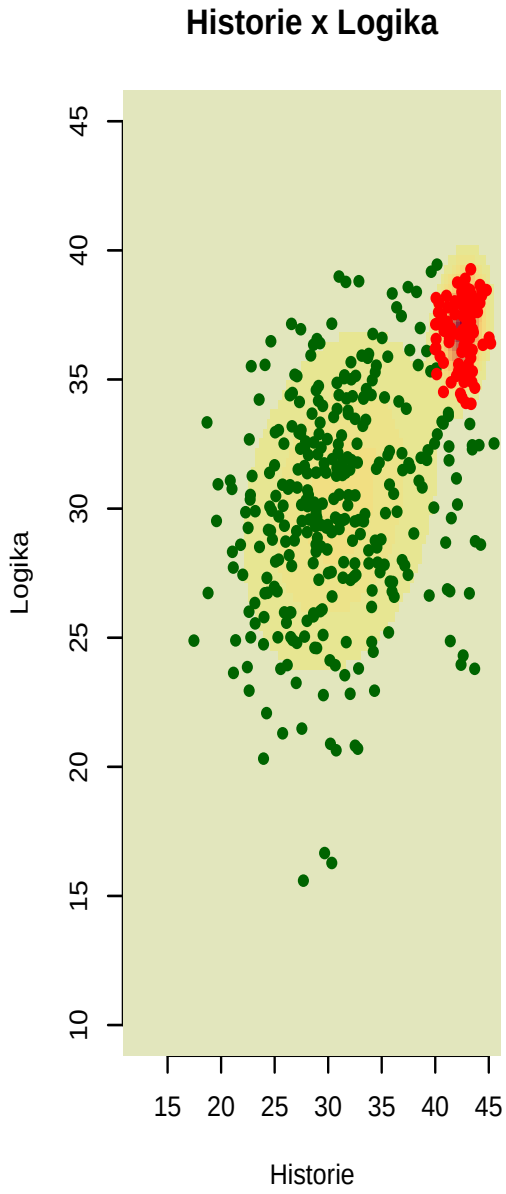
$$\pi_{i,k}(\boldsymbol{\theta}) \equiv \mathrm{P}(U_i = k \,|\, \boldsymbol{y}_i) = \frac{w_k \, \phi(\boldsymbol{y}_i; \boldsymbol{\eta}_k)}{\sum_{j=1}^{K} w_j \, \phi(\boldsymbol{y}_i; \boldsymbol{\eta}_j)}$$

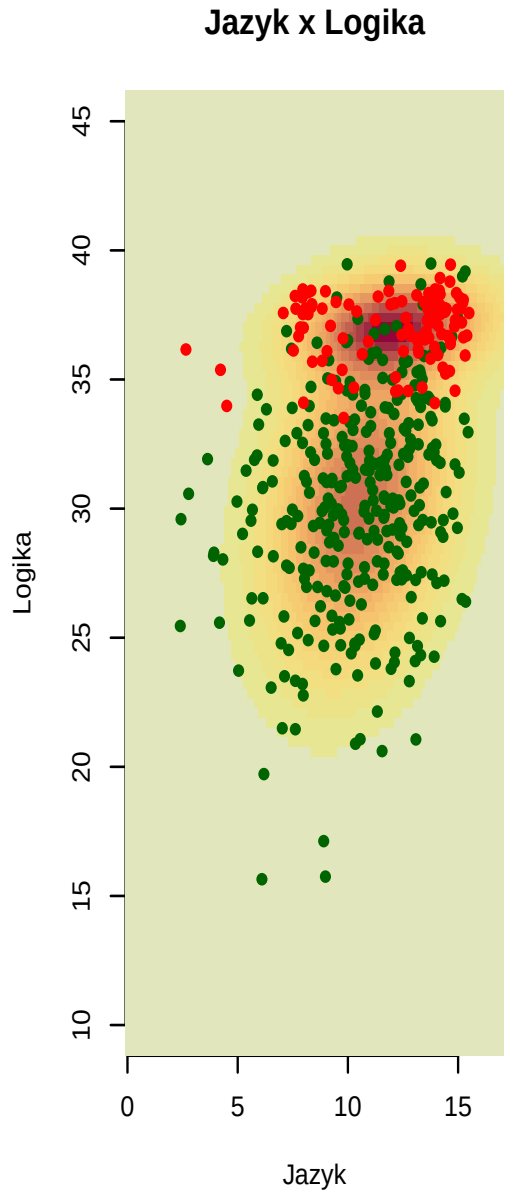
- Zařaď i-té pozorování do shluku s indexem j :

$$\pi_{i,j} = \max\{\pi_{i,1}, \ldots, \pi_{i,K}\}$$

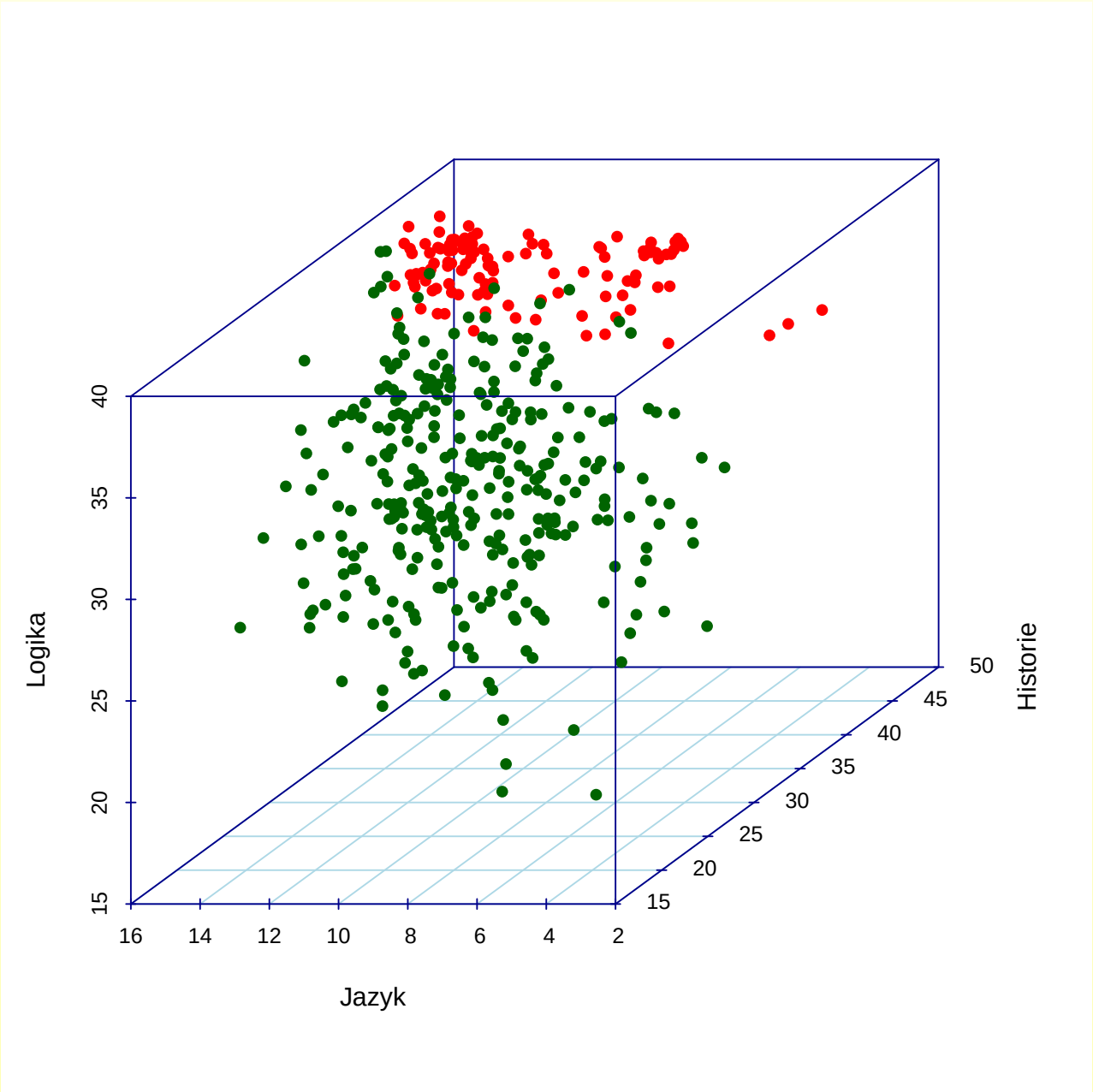
Přijímačky na PF UK v roce 1999 (Termín č. 12)







Přijímačky na PF UK v roce 1999 (Termín č. 12)



Část II

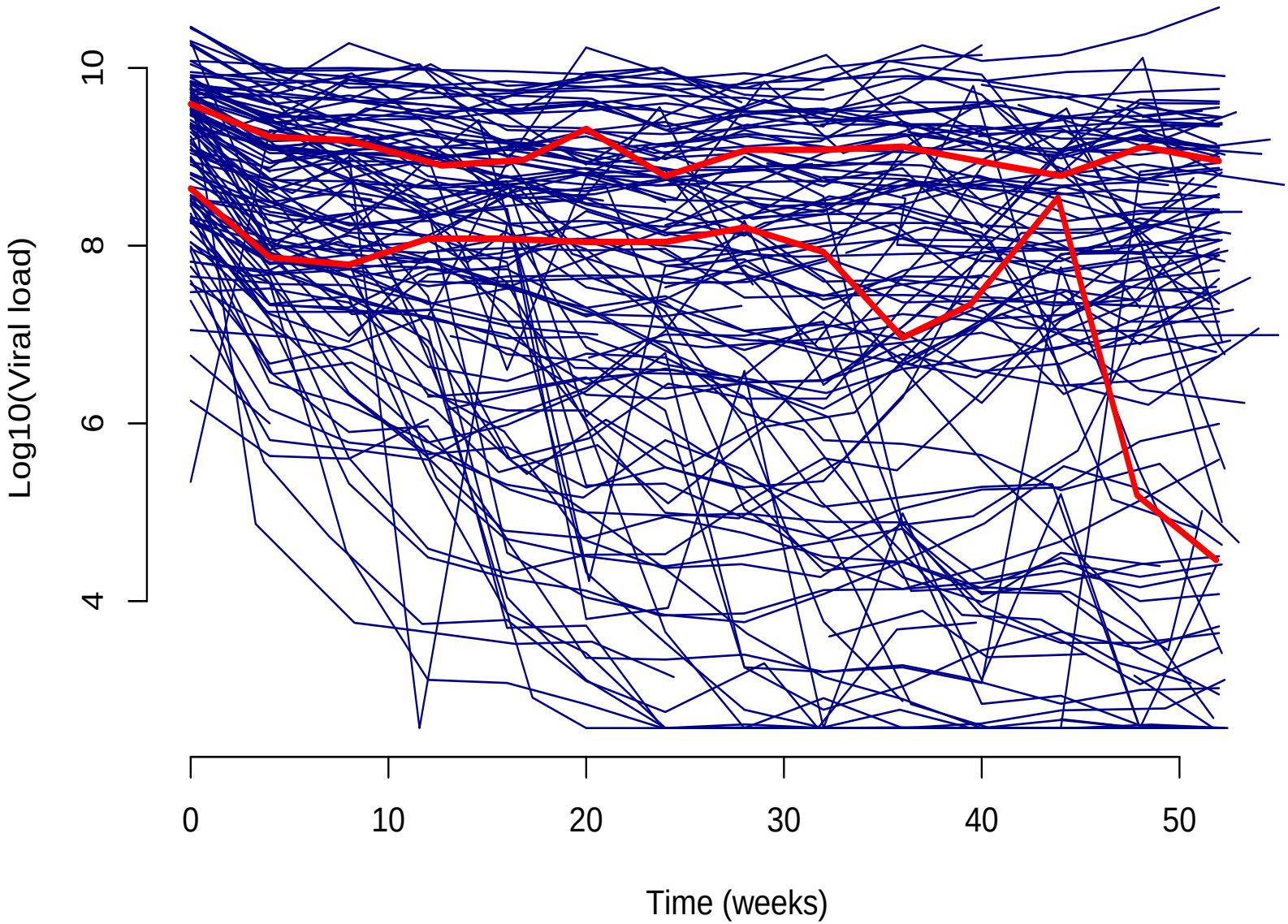
Na skok do Hejnic

Chronická hepatitida B

- Mezinárodní klinická studie koordinovaná na Erasmus Medisch Centrum v Rotterdamu
- Léčba pomocí PEG-INF (na ROBUSTu nekombinováno s lamivudinem) po dobu 52 týdnů
- Má to nehezké vedlejší účinky a často je stejně k ničemu
- Podrobnosti:
  Janssen, van Zonneveld, Senturk, Akarca, Cakaloglu, Simon, So, Gerken, de Man, Niesters, Zondervan, Hansen, Schalm (2005).
  Pegylated interferon alfa-2b alone or in combination with lamivudine for HBeAg-positive chronic hepatitis B: a randomised trial. *Lancet,* **365**, 123–129.

Chronická hepatitida B

Log10(Viral load)



Chronická hepatitida B

- Snaha roztřídit pacienty do skupin
  - může nějak souviset s diagnózou či prognózou
  - ...

- Jeden pacient: $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,n_i})'$
- $\boldsymbol{Y}_i \sim p(\boldsymbol{y})$

- Liší se to nějak od právnické mafie?

Model pro shlukování?

- ✧ Jeden pacient: $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,n_i})'$
- ✏ n_i (počet návštěv) není nutně stejné pro všechny pacienty

---

- ✧ Ve skutečnosti je $Y_{i,j} = Y_{i,j}(t_{i,j})$
  - ✧ $t_{i,j}$ . . . doba, po kterou je pacient léčen
- ✏ Posloupnost časů návštěv $(t_{i,1}, \ldots, t_{i,n_i})$ není stejná pro všechny pacienty

---

- ➠ Těžko bude $\boldsymbol{Y}_i \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{y})$

Model pro shlukování?

- Co takhle reprezentovat každého pacienta nějakým (nepřímo pozorovatelným) náhodným vektorem $\boldsymbol{b}_i$ tak, aby $\boldsymbol{b}_i \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{b})$

XII. ROBUST (únor 2002, Hejnice)

- *On a fitting of a linear mixed model with a finite normal mixture as random-effects distribution*
- Lineární smíšený model (LMM) pro $\boldsymbol{Y}_i$:

$$\boldsymbol{Y}_i = \mathbb{X}_i \boldsymbol{\alpha} + \mathbb{Z}_i \boldsymbol{b}_i + \boldsymbol{\varepsilon}_i$$

  - $\mathbb{X}_i, \mathbb{Z}_i$ . . . matice regresorů
  - $\boldsymbol{\alpha}$ . . . (neznámé) regresní parametry
  - $\boldsymbol{b}_i \overset{\text{i.i.d.}}{\sim} \sum_{k=1}^{K} w_k \mathcal{N}(\boldsymbol{\mu}_k, \mathbb{D}_k)$
  - $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n_i})$, nezávislé pro $i = 1, \ldots, n$

- Hejnice (XII. ROBUST):
  - Pouze homoskedastická směs
  - Maximální věrohodnost přes EM algoritmus s využitím software pro standardní (nesměsový) LMM
- Králíky (XVI. ROBUST):
  - Směs může být i heteroskedastická
  - Bayesovský odhad přes MCMC
  - Nestandardní podsoftware (balíček) standardního software (R)

- Pro shlukování: i.i.d. náhodné veličiny $U_1, \ldots, U_n$
- $U_i \in \{1, \ldots, K\}$
- $\mathrm{P}(U_i = k) = w_k, k = 1, \ldots, K, \; i = 1, \ldots, n$

✧ Hierarchicky zapsáno:

$$[\boldsymbol{Y}_i \,|\boldsymbol{b}_i] \sim \mathcal{N}\left(\mathbb{X}_i\boldsymbol{\alpha} + \mathbb{Z}_i\boldsymbol{b}_i,\, \sigma^2 I_{n_i}\right)$$

$$\left.\begin{array}{rcl} \mathrm{P}(U_i = k) & = & w_k \\ [\boldsymbol{b}_i \,|\, U_i = k] & \sim & \mathcal{N}(\boldsymbol{\mu}_k,\, \mathbb{D}_k) \end{array}\right\} \quad k = 1, \ldots, K$$

➠ $$\boldsymbol{Y}_i \sim \sum_{k=1}^{K} w_k \mathcal{N}\left(\mathbb{X}_i\boldsymbol{\alpha} + \mathbb{Z}_i\boldsymbol{\mu}_k,\; \mathbb{Z}_i\mathbb{D}_k\mathbb{Z}_i' + \sigma^2 I_{n_i}\right)$$

➠ $$[\boldsymbol{Y}_i \,|\, U_i = k] \sim \mathcal{N}\left(\mathbb{X}_i\boldsymbol{\alpha} + \mathbb{Z}_i\boldsymbol{\mu}_k,\; \mathbb{Z}_i\mathbb{D}_k\mathbb{Z}_i' + \sigma^2 I_{n_i}\right)$$

Chronická hepatitida B

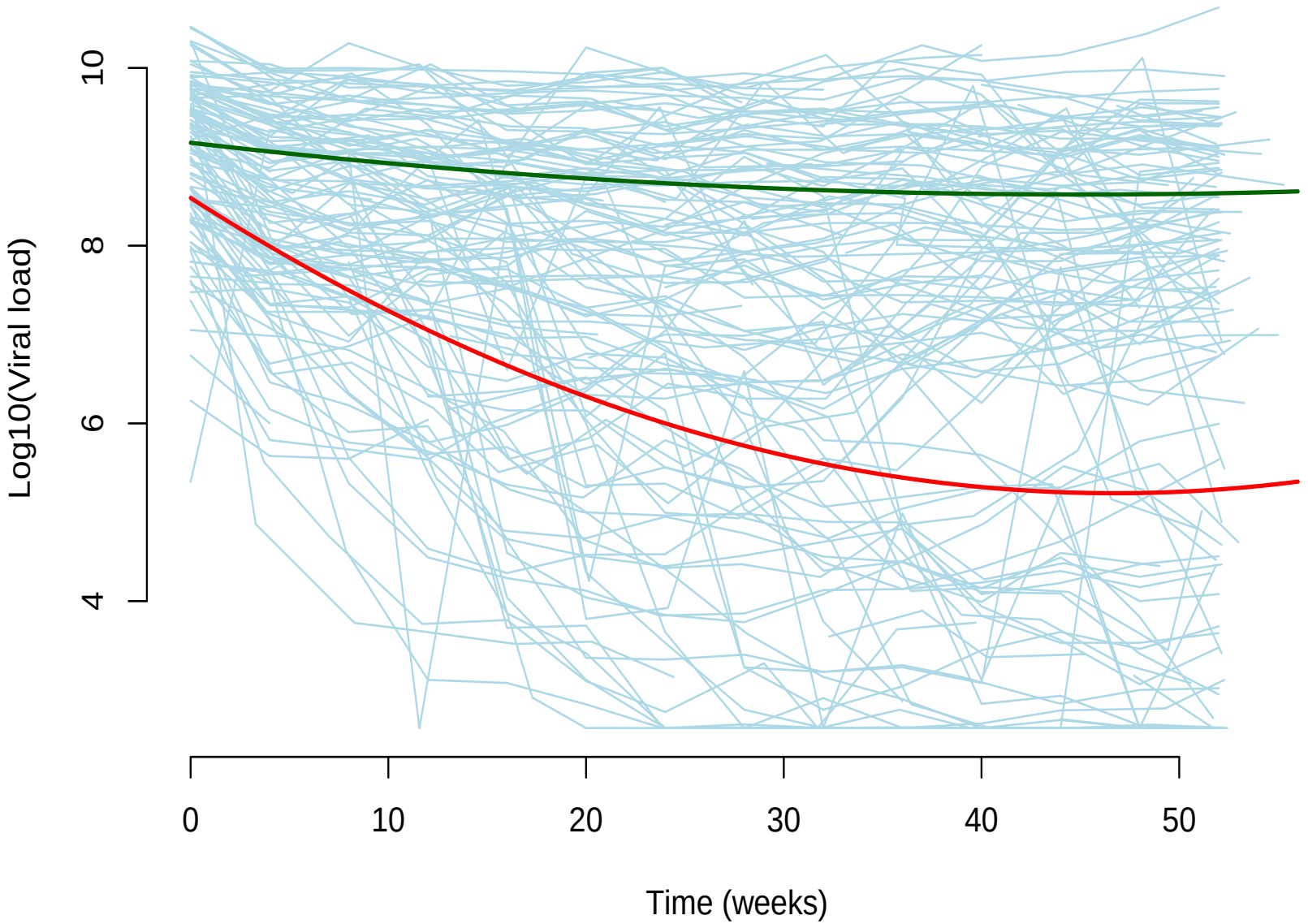
- $Y_{i,j} = b_{i,1} + b_{i,2}B_2(t_{i,j}) + b_{i,3}B_3(t_{i,j}) + \varepsilon_{i,j}$
- $1, B_2(t), B_3(t)$

  ... kvadratická B-splinová báze s uzly v $t = 0$ a 56
- $\boldsymbol{b}_i = (b_{i,1}, b_{i,2}, b_{i,3})' \overset{\text{i.i.d.}}{\sim} \sum_{k=1}^{2} w_k \mathcal{N}(\boldsymbol{\mu}_k, \mathbb{D}_k)$

$$\mathrm{E}\left(Y_{i,j} \,\middle|\, U_i = k\right) = \mu_{k,1} + \mu_{k,2}B_2(t_{i,j}) + \mu_{k,3}B_3(t_{i,j})$$

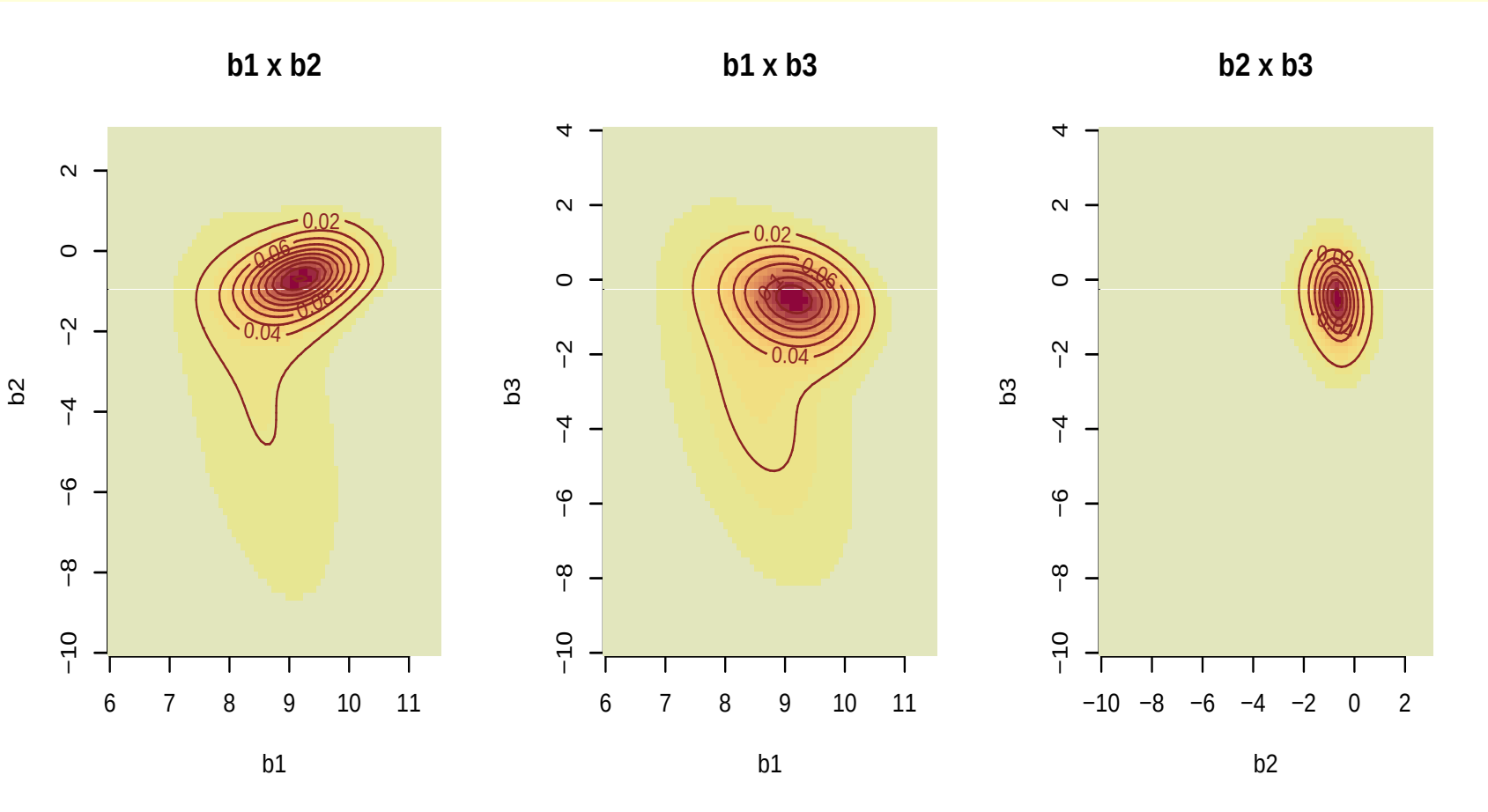
Chronická hepatitida B

Odhadnutý průměrný vývoj v čase ve dvou skupinách



Chronická hepatitida B

Rozdělení náhodných efektů



Modelově založená shluková analýza

- Model:

$$[\boldsymbol{Y}_i \,|\boldsymbol{b}_i] \sim \mathcal{N}\left(\mathbb{X}_i\boldsymbol{\alpha} + \mathbb{Z}_i\boldsymbol{b}_i,\, \sigma^2 I_{n_i}\right)$$

$$\left.\begin{array}{rcl} \mathrm{P}(U_i = k) & = & w_k \\ [\boldsymbol{b}_i \,|\, U_i = k] & \sim & \mathcal{N}(\boldsymbol{\mu}_k,\, \mathbb{D}_k) \end{array}\right\} \quad k = 1, \ldots, K$$

- Parametry modelu:

$$\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{w}', \boldsymbol{\mu}_1', \ldots, \boldsymbol{\mu}_K', \mathrm{vec}(\mathbb{D}_1)', \ldots, \mathrm{vec}(\mathbb{D}_K)', \sigma^2)' \in \Theta$$

- Nepřímo pozorovatelné náhodné veličiny/vektory:

$$\boldsymbol{\eta} = (\boldsymbol{b}_1', \ldots, \boldsymbol{b}_n', U_1, \ldots, U_n)' \in \mathbb{R}^q \times \cdots \times \mathbb{R}^q \times \{1, \ldots, K\}^n$$

Modelově založená shluková analýza

- Bayesova věta:

$$\pi_{i,k}(\boldsymbol{\theta}) \equiv \mathrm{P}\left(U_i = k \,\middle|\, \boldsymbol{y}_i, \boldsymbol{\theta}\right)$$

$$\propto w_k \, p(\boldsymbol{y}_i \,|\, U_i = k, \boldsymbol{\theta})$$

$$= w_k \, \varphi\left(\boldsymbol{y}_i;\ \mathbb{X}_i\boldsymbol{\alpha} + \mathbb{Z}_i\boldsymbol{\mu}_k,\ \mathbb{Z}_i\mathbb{D}_k\mathbb{Z}_i' + \sigma^2 I_{n_i}\right)$$

- Po chvíli počítání:

$$\pi_{i,k}(\boldsymbol{\theta}) \propto$$

$$w_k \, |\mathbb{D}_k|^{-1/2} \, |\mathbb{A}_{i,k}|^{-1/2} \, \exp\left\{-\frac{1}{2}\left(\boldsymbol{\mu}_k'\mathbb{D}_k^{-1}\boldsymbol{\mu}_k - \boldsymbol{c}_{i,k}'\mathbb{A}_{i,k}^{-1}\boldsymbol{c}_{i,k}\right)\right\}$$

$$\mathbb{A}_{i,k} = \sigma^{-2}\,\mathbb{Z}_i'\mathbb{Z}_i + \mathbb{D}_k^{-1}$$

$$\boldsymbol{c}_{i,k} = \sigma^{-2}\mathbb{Z}_i'\left(\boldsymbol{y}_i - \mathbb{X}_i\boldsymbol{\alpha}\right) + \mathbb{D}_k^{-1}\boldsymbol{\mu}_k$$

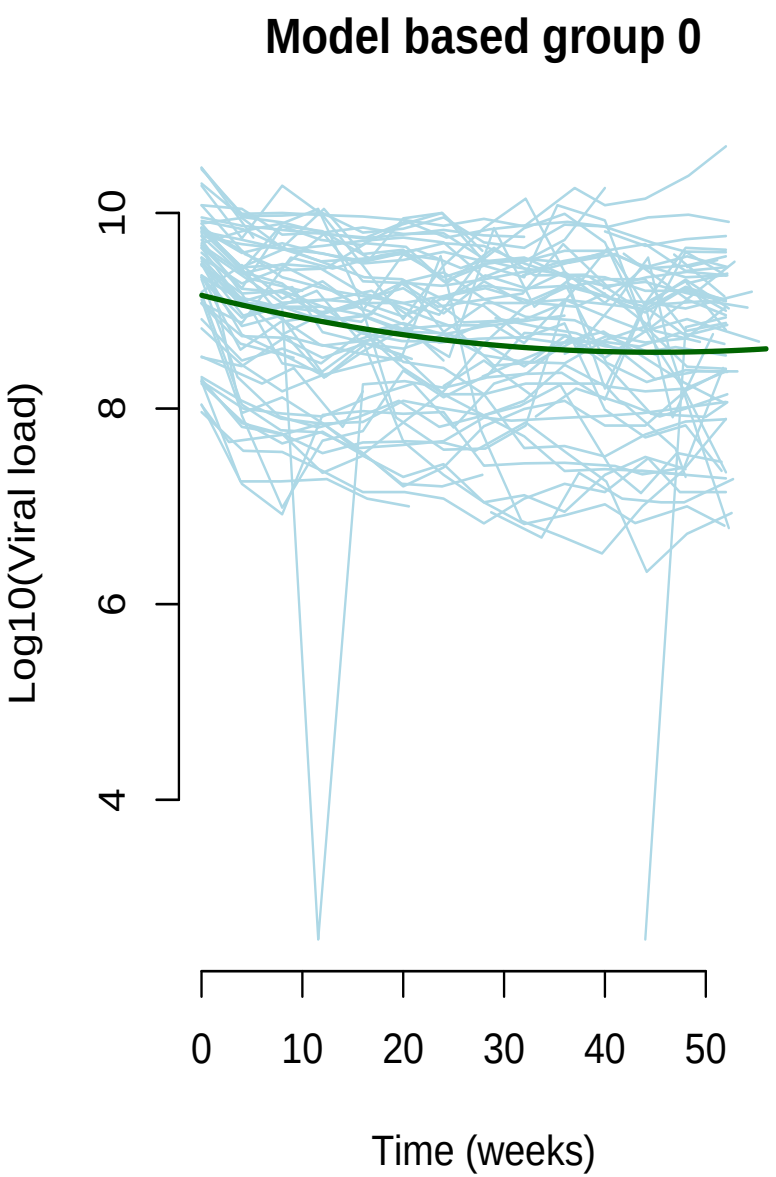
- Je-li $\hat{\boldsymbol{\theta}}$ ML odhad $\boldsymbol{\theta}$:
  - Shlukuj na základě $\pi_{i,k}(\hat{\boldsymbol{\theta}})$
- Je-li při Bayesovském odhadu, $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}$ náhodný výběr z aposteriorního rozdělení $p(\boldsymbol{\theta} \,|\, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$:
  - Shlukuj na základě

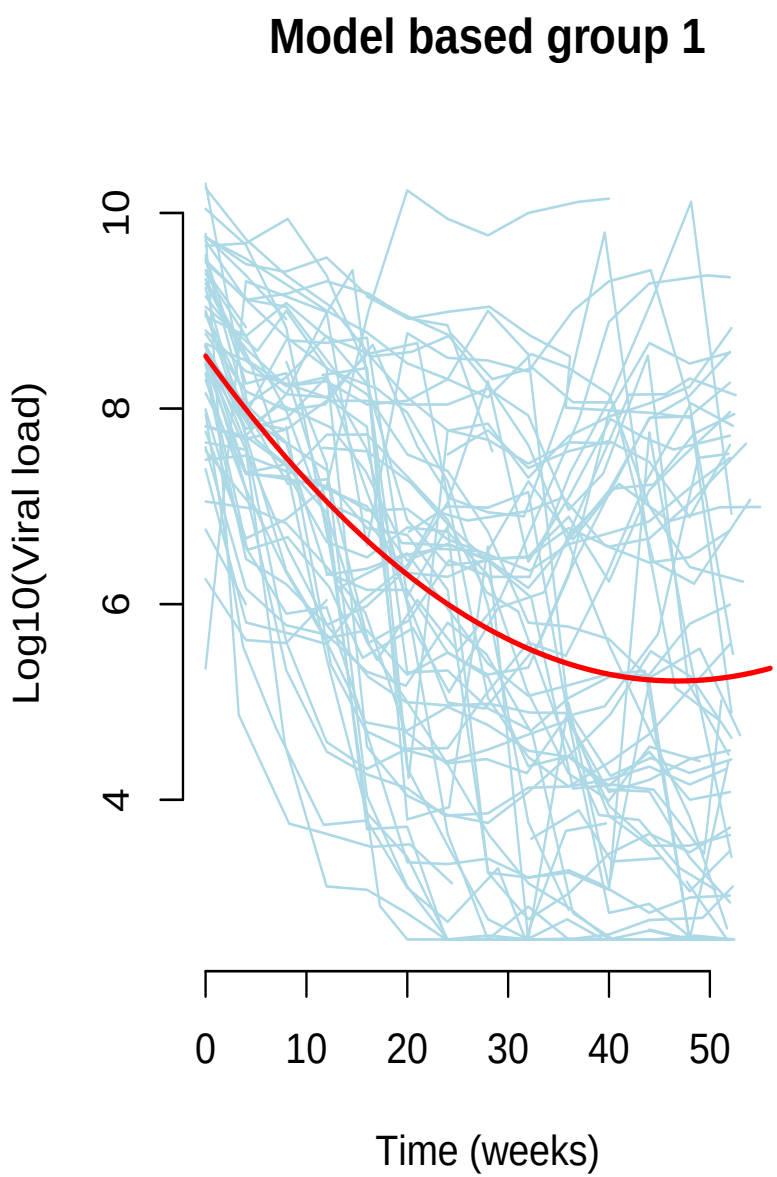
$$\widehat{\mathrm{P}}\big(U_i = k \,\big|\, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\big) = \widehat{\mathrm{E}}_{\boldsymbol{\theta}}\big\{\pi_{i,k}(\boldsymbol{\theta}) \,\big|\, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\big\}$$

$$= M^{-1} \sum_{m=1}^{M} \pi_{i,k}\big(\boldsymbol{\theta}^{(m)}\big)$$

Chronická hepatitida B

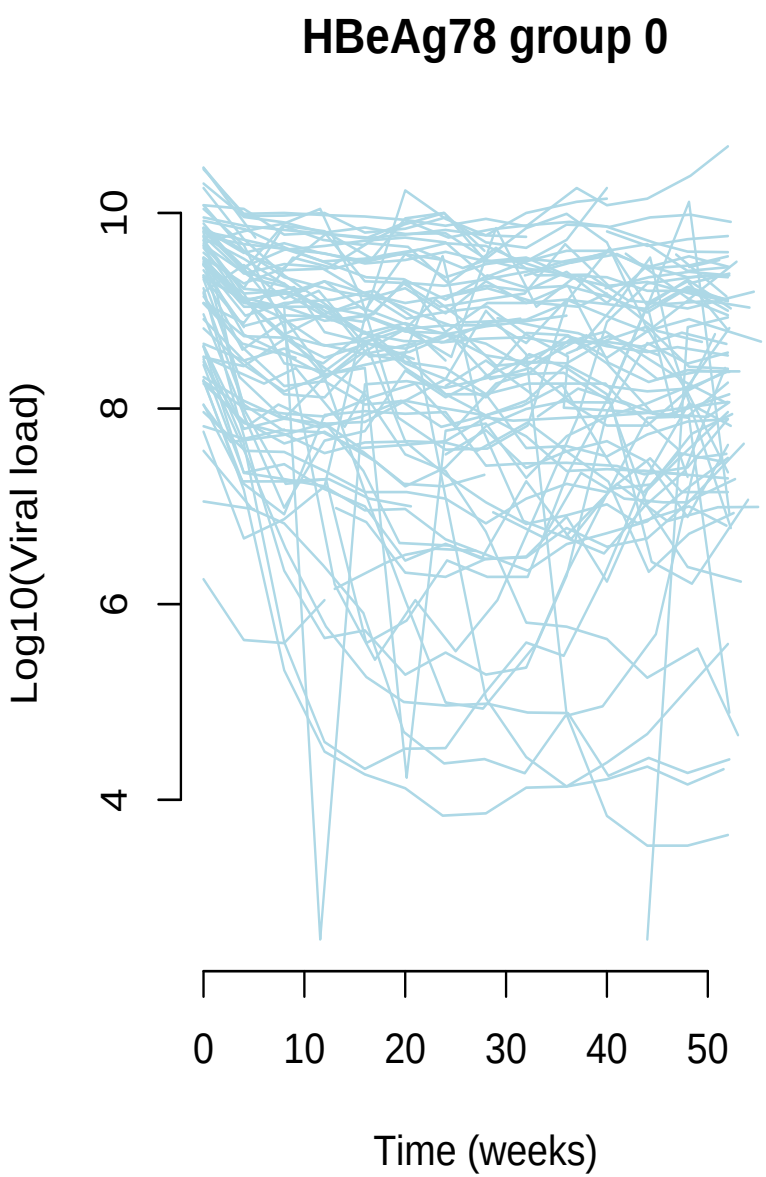
Shluky dle modelu pro Log10(Viral load)

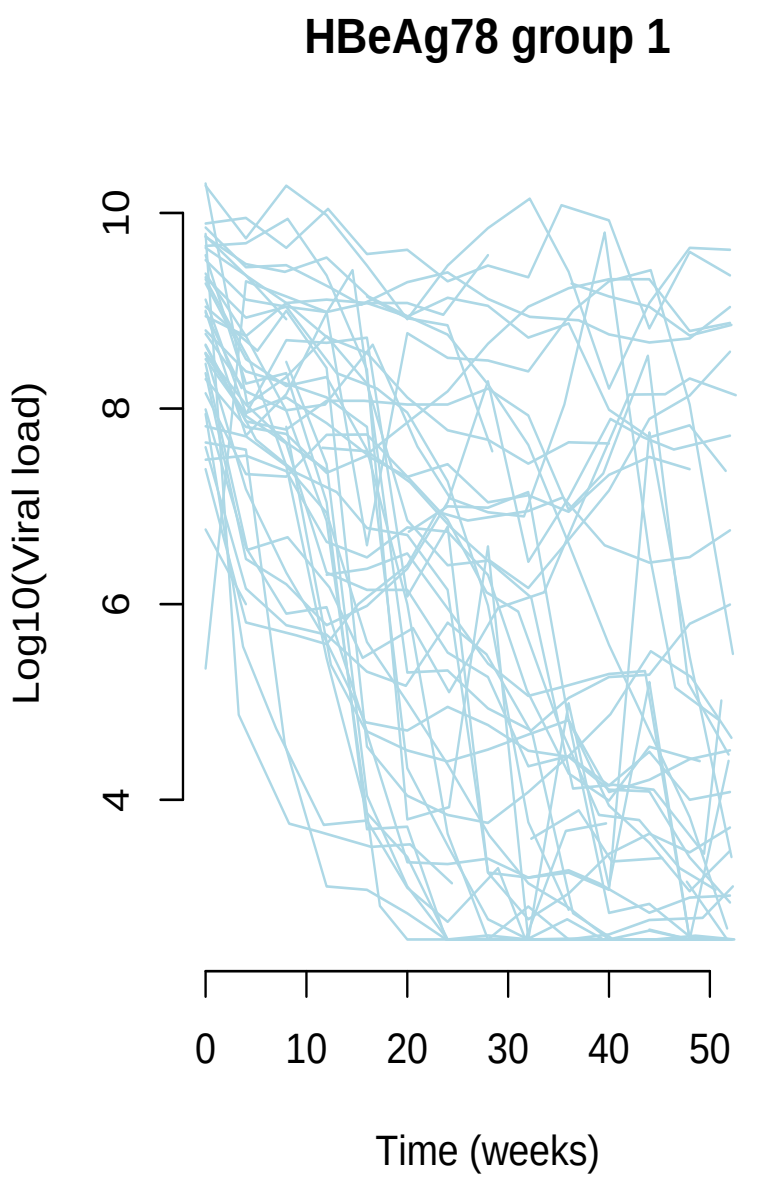




Chronická hepatitida B

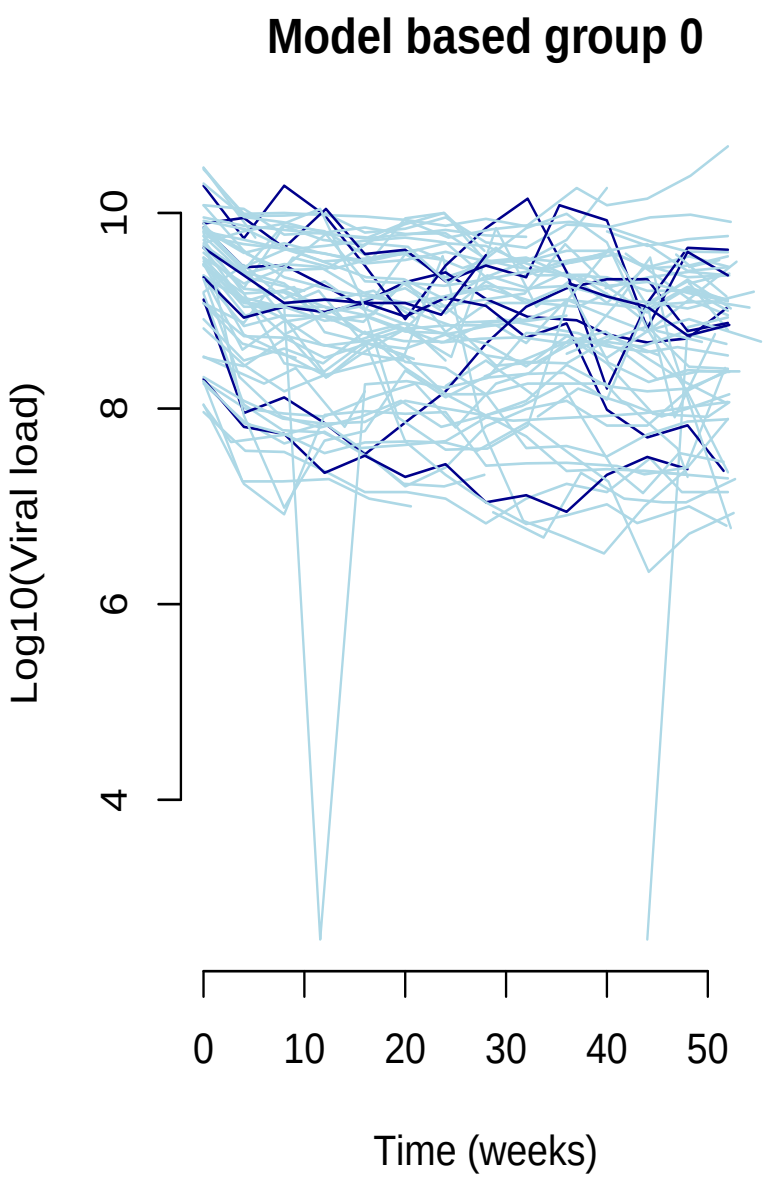
Shluky dle HBeAg v 78. týdnu

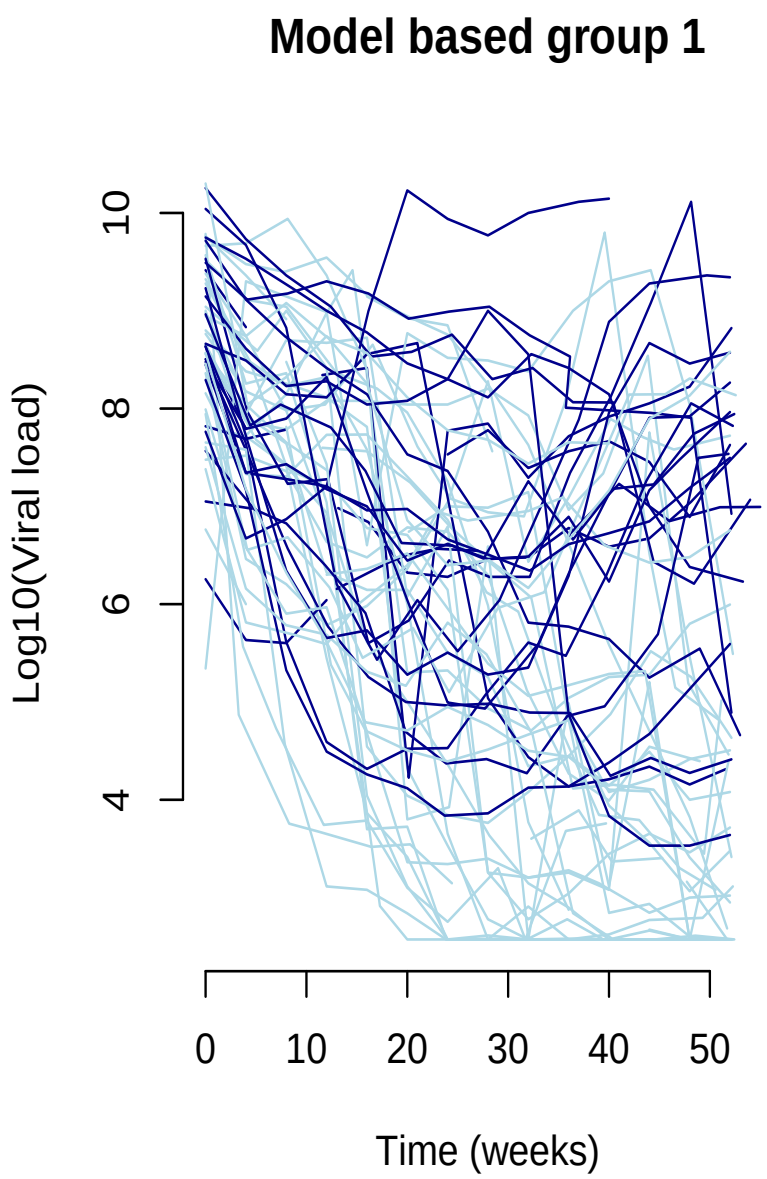




Chronická hepatitida B

Shluky dle modelu pro Log10(Viral load), rozdíl oproti HBeAg78 shlukům





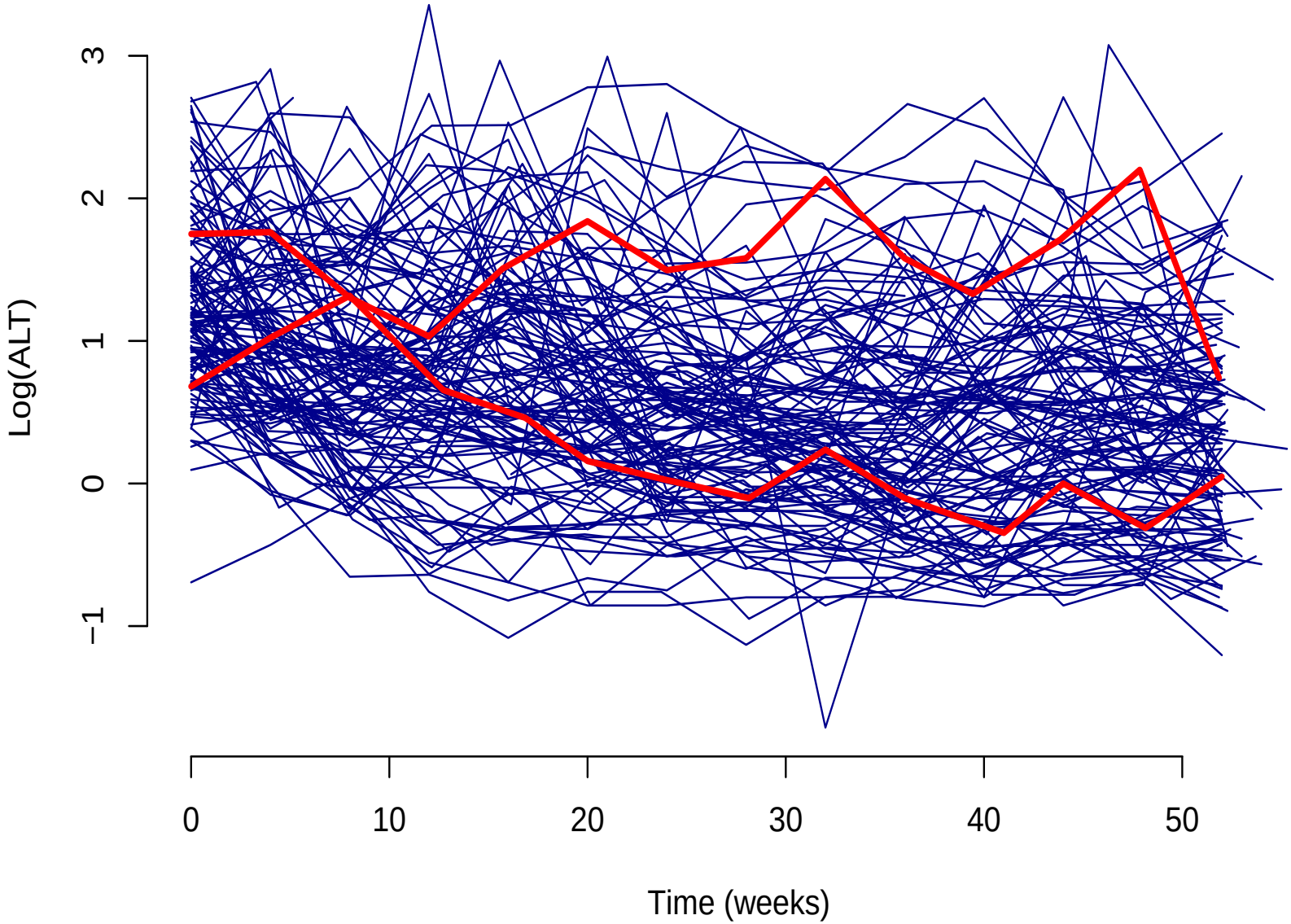
Chronická hepatitida B

- Nešlo by zlepšit odhalení skutečné HBeAg78 skupiny?
- Kromě množství viru v krvi se měřilo ještě ALT (míra aktivity choroby)

- Co takhle shlukovat na základě obou ukazatelů?

Chronická hepatitida B

Log(ALT)



Část III

Zpátky do Králík

LMM pro dvě různé odezvy

- Jeden pacient:

$$\boldsymbol{Y}_{i,1} = (Y_{i,1,1}, \ldots, Y_{i,1,n_{i,1}})' \qquad \text{Log10(Viral load)}$$

$$\boldsymbol{Y}_{i,2} = (Y_{i,2,1}, \ldots, Y_{i,2,n_{i,2}})' \qquad \text{Log(ALT)}$$

- LMM pro odezvu 1 (Log10(Viral load)):

$$\boldsymbol{Y}_{i,1} = \mathbb{X}_{i,1}\boldsymbol{\alpha}_1 + \mathbb{Z}_{i,1}\boldsymbol{b}_{i,1} + \boldsymbol{\varepsilon}_{i,1}$$

- LMM pro odezvu 2 (Log(ALT)):

$$\boldsymbol{Y}_{i,2} = \mathbb{X}_{i,2}\boldsymbol{\alpha}_2 + \mathbb{Z}_{i,2}\boldsymbol{b}_{i,2} + \boldsymbol{\varepsilon}_{i,2}$$

LMM pro dvě různé odezvy

- Lineární smíšené modely:

$$\boldsymbol{Y}_{i,1} = \mathbb{X}_{i,1}\boldsymbol{\alpha}_1 + \mathbb{Z}_{i,1}\boldsymbol{b}_{i,1} + \boldsymbol{\varepsilon}_{i,1}$$

$$\boldsymbol{Y}_{i,2} = \mathbb{X}_{i,2}\boldsymbol{\alpha}_2 + \mathbb{Z}_{i,2}\boldsymbol{b}_{i,2} + \boldsymbol{\varepsilon}_{i,2}$$

- $\boldsymbol{\varepsilon}_{i,1} \sim \mathcal{N}(\mathbf{0}, \sigma_1^2 I_{n_{i,1}})$, nezávislé pro $i = 1, \ldots, n$
- $\boldsymbol{\varepsilon}_{i,2} \sim \mathcal{N}(\mathbf{0}, \sigma_2^2 I_{n_{i,2}})$, nezávislé pro $i = 1, \ldots, n$
- $\boldsymbol{b}_i = (b_{i,1,1}, \ldots, b_{i,1,q_1}, b_{i,2,1}, \ldots, b_{i,2,q_2})'$
- $\boldsymbol{b}_i \overset{\text{i.i.d.}}{\sim} \sum_{k=1}^{K} w_k \mathcal{N}(\boldsymbol{\mu}_k, \mathbb{D}_k)$

LMM pro dvě různé odezvy napsaný jako jeden LMM

- Lze to napsat najednou:

$$\begin{pmatrix} \boldsymbol{Y}_{i,1} \\ \boldsymbol{Y}_{i,2} \end{pmatrix} = \begin{pmatrix} \mathbb{X}_{i,1} & \mathbb{0} \\ \mathbb{0} & \mathbb{X}_{i,2} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix} + \begin{pmatrix} \mathbb{Z}_{i,1} & \mathbb{0} \\ \mathbb{0} & \mathbb{Z}_{i,2} \end{pmatrix} \begin{pmatrix} \boldsymbol{b}_{i,1} \\ \boldsymbol{b}_{i,2} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \end{pmatrix}$$

- $\begin{pmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \sigma_1^2 I_{n_{i,1}} & \mathbb{0} \\ \mathbb{0} & \sigma_2^2 I_{n_{i,2}} \end{pmatrix}\right),$

  nezávislé pro $i = 1, \ldots, n$

- $\boldsymbol{b}_i = (b_{i,1,1}, \ldots, b_{i,1,q_1}, b_{i,2,1}, \ldots, b_{i,2,q_2})'$

- $\boldsymbol{b}_i \overset{\text{i.i.d.}}{\sim} \sum_{k=1}^{K} w_k \mathcal{N}(\boldsymbol{\mu}_k, \mathbb{D}_k)$

LMM pro dvě různé odezvy

- Při $K = 1$ lze ML odhadnout standardním softwarem (R balíčky `nlme`, `lme4`, SAS `PROC MIXED`)
- Při $K > 1$ a homoskedastické směsi v rozdělení náhodných efektů lze teoreticky použít přístup prezentovaný v Hejnicích na XII. ROBUSTu

  (EM algoritmus s M-krokem provedeným standardním softwarem)

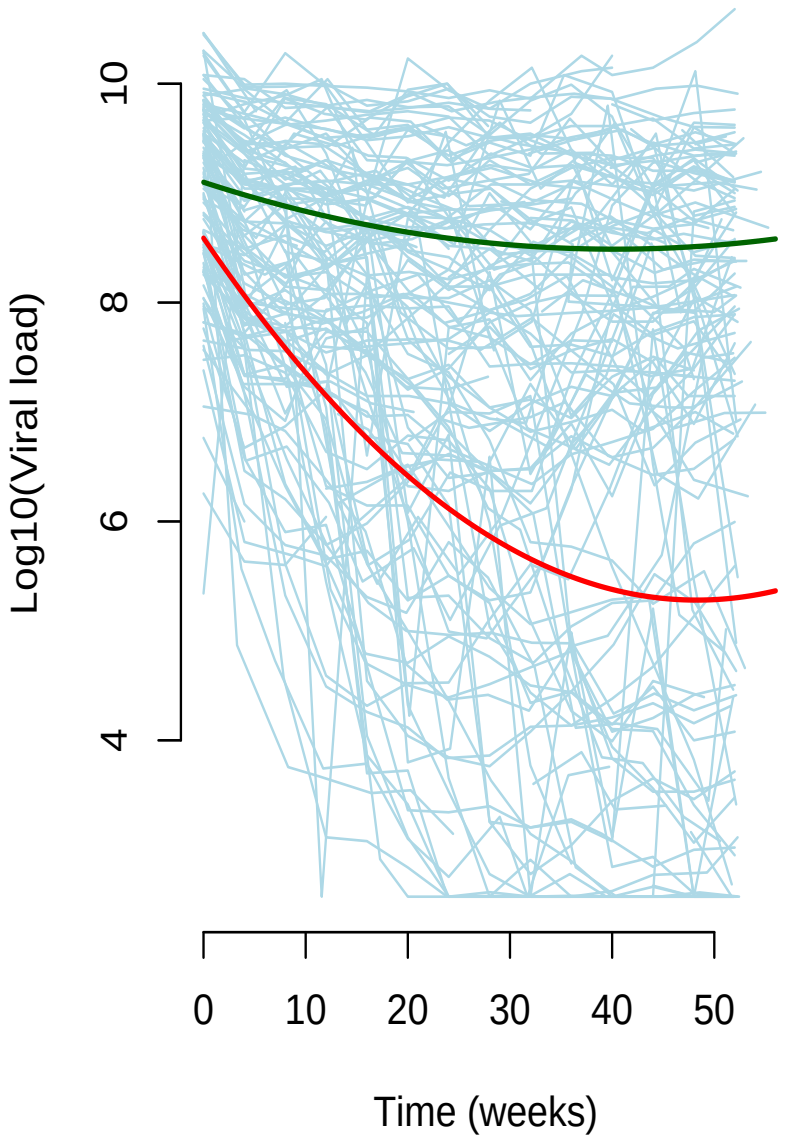
- Při použití standardního software se zbytečně počítá s mnoha nulami v maticích $\mathbb{X}$, $\mathbb{Z}$
- Dlouho to trvá
- Přináší numerické problémy (i při $K = 1$)

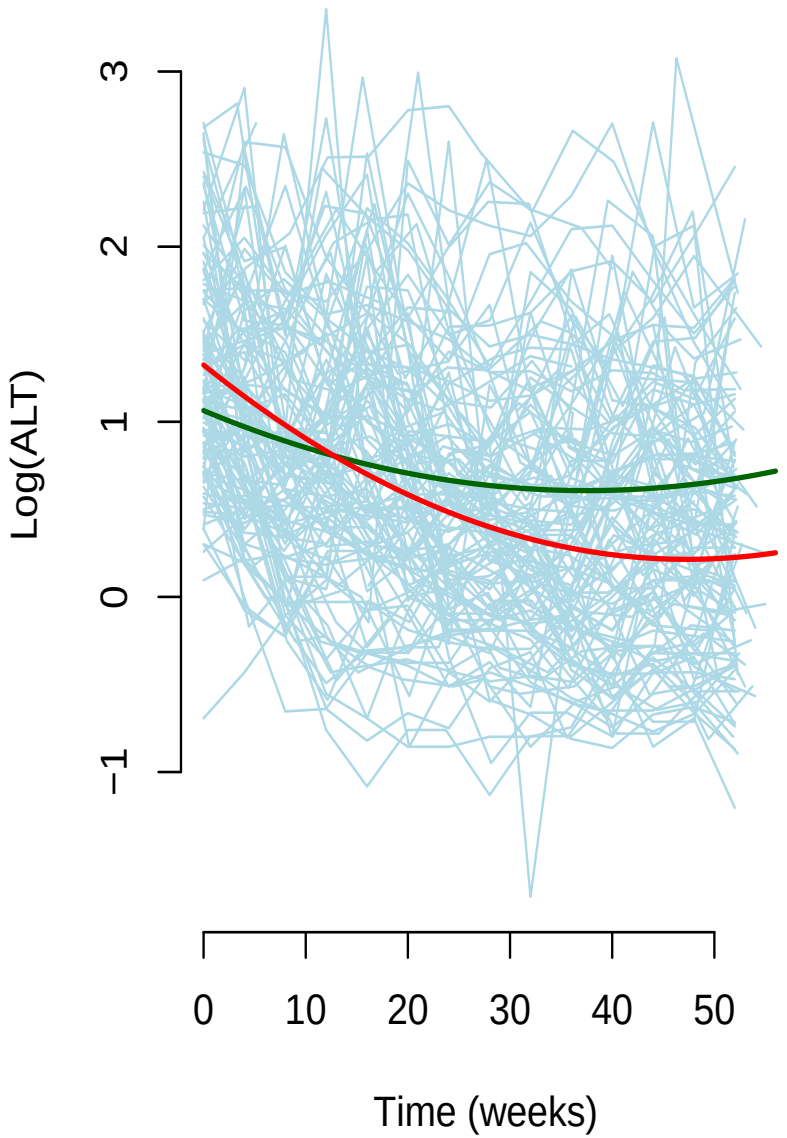
XVI. ROBUST (únor 2010, Králíky)

- Směs v rozdělení náhodných efektů může být i heteroskedastická
- Bayesovský odhad přes MCMC
- Nestandardní podsoftware (balíček) standardního software (R)
- Využití známé struktury matic $\mathbb{X}$, $\mathbb{Z}$

Chronická hepatitida B

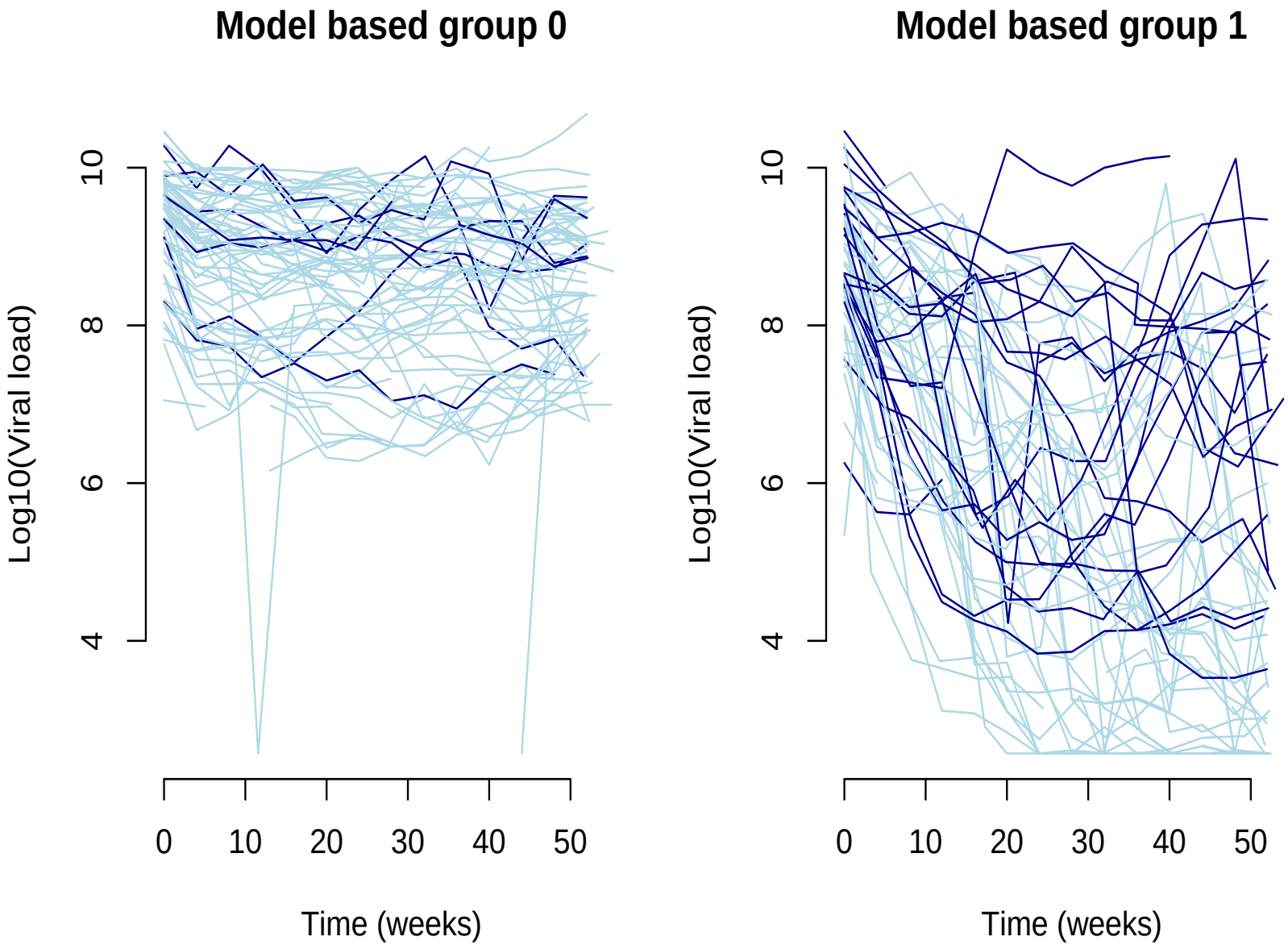
Odhadnutý průměrný vývoj v čase ve dvou skupinách





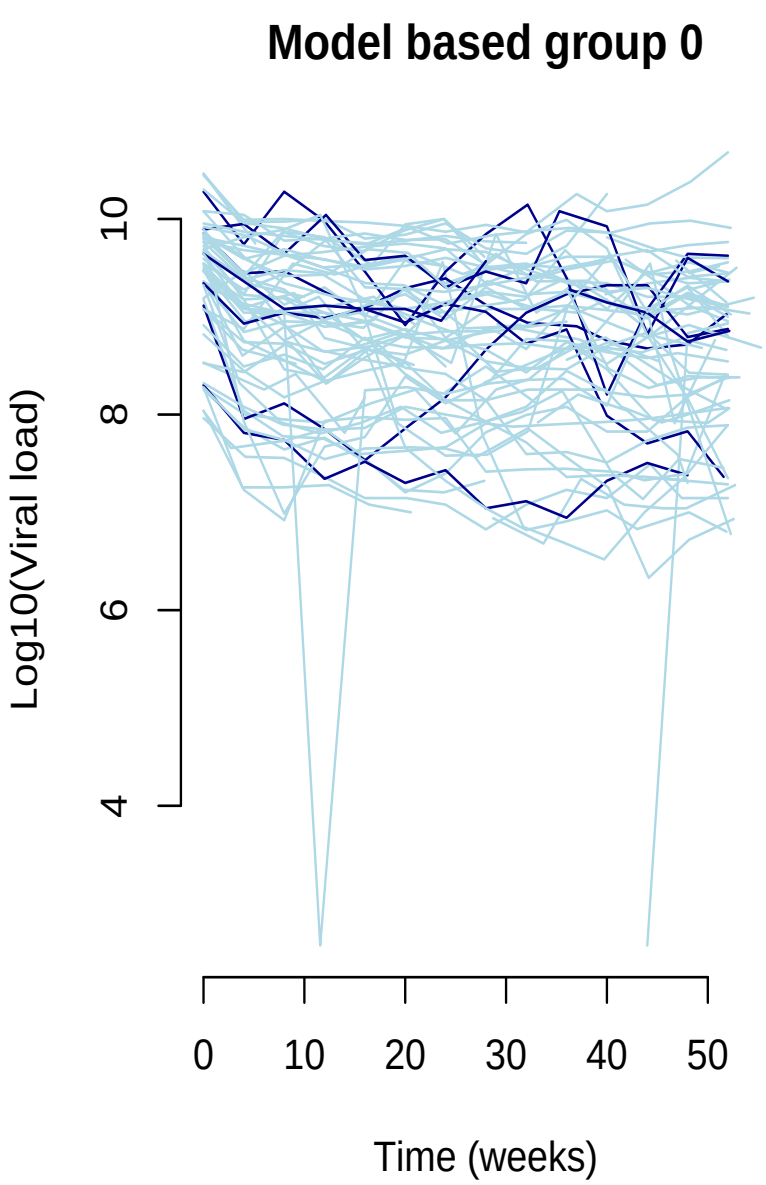
Chronická hepatitida B

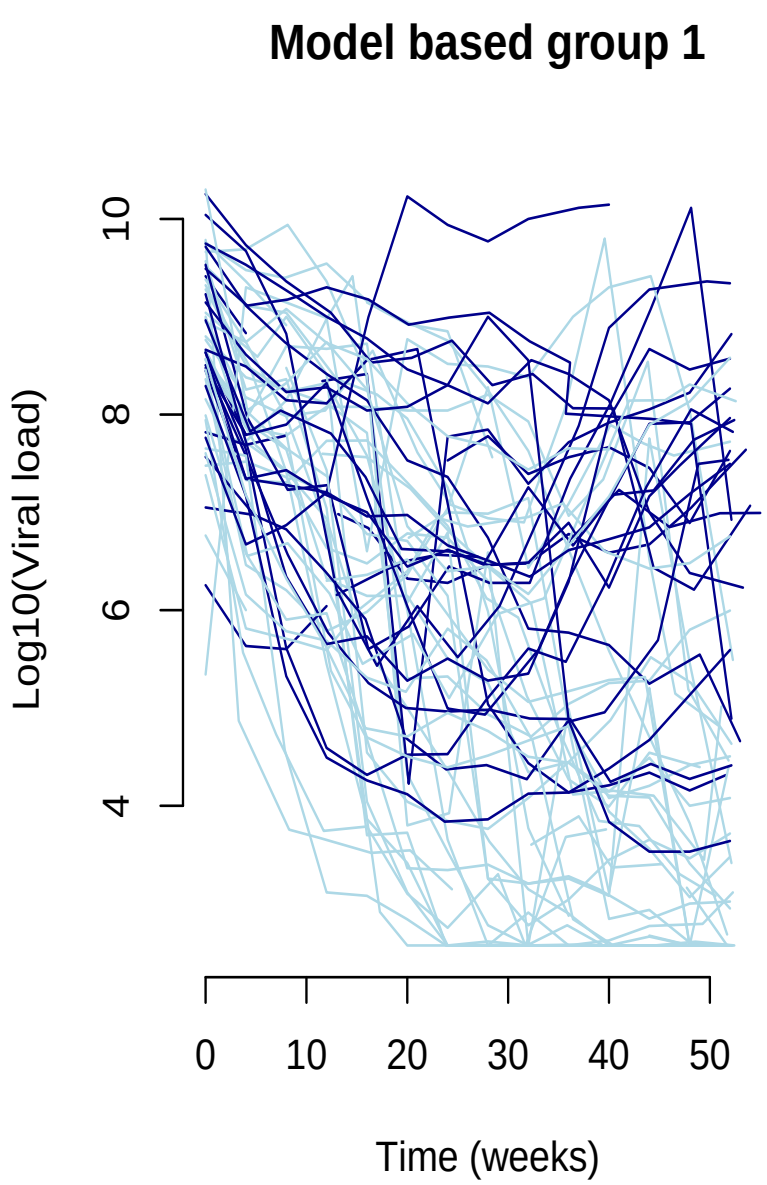
Shluky dle sdruženého modelu pro Log10(Viral load) a Log(ALT)



Chronická hepatitida B

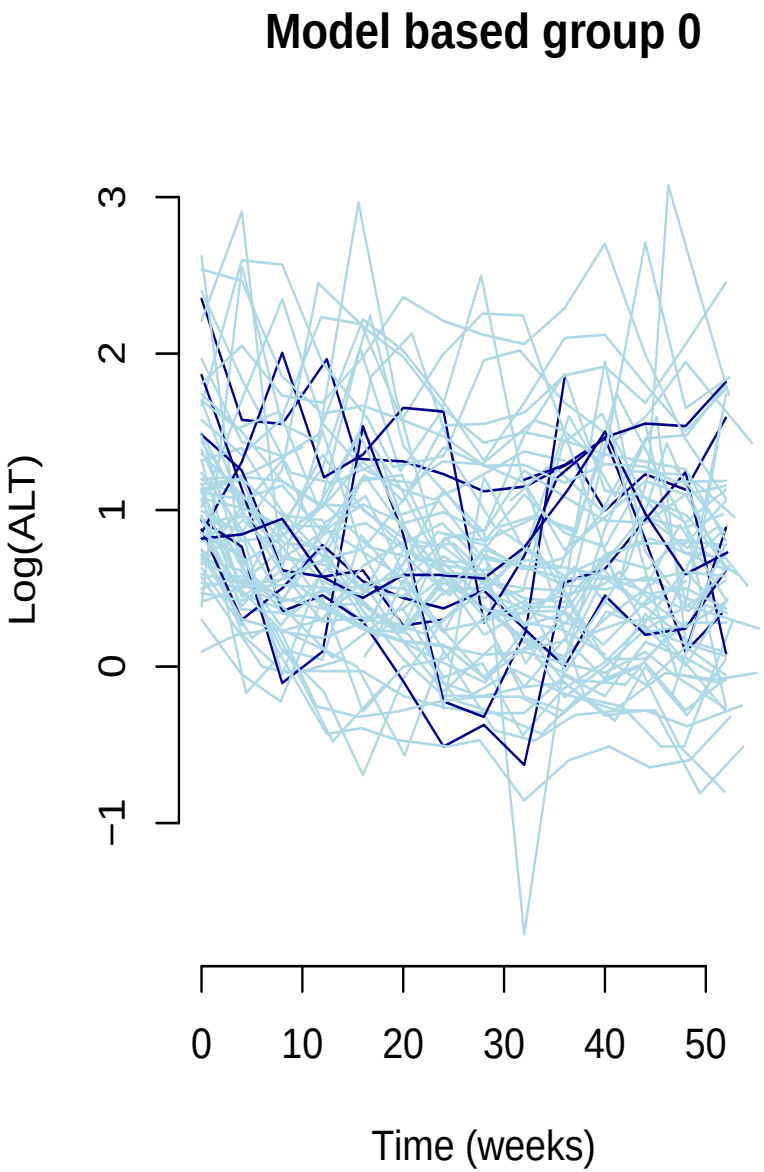
Shluky dle modelu pro Log10(Viral load)

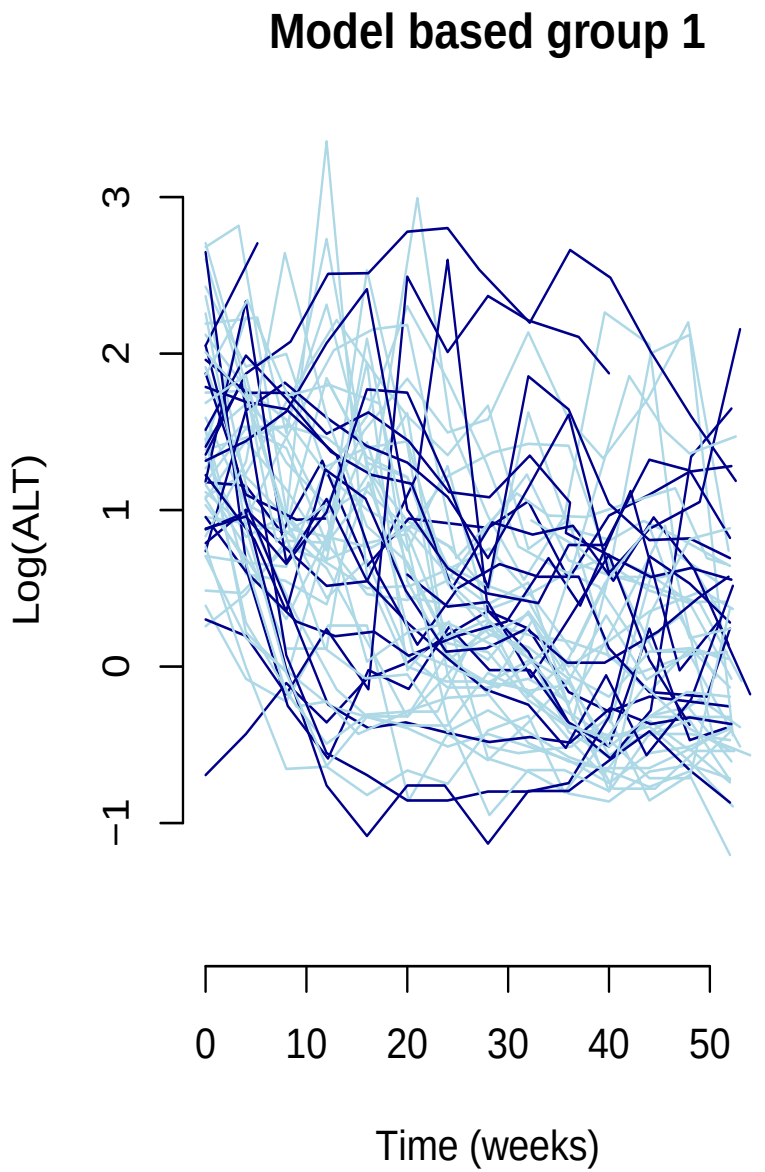




Chronická hepatitida B

Shluky dle sdruženého modelu pro Log10(Viral load) a Log(ALT)





Část IV

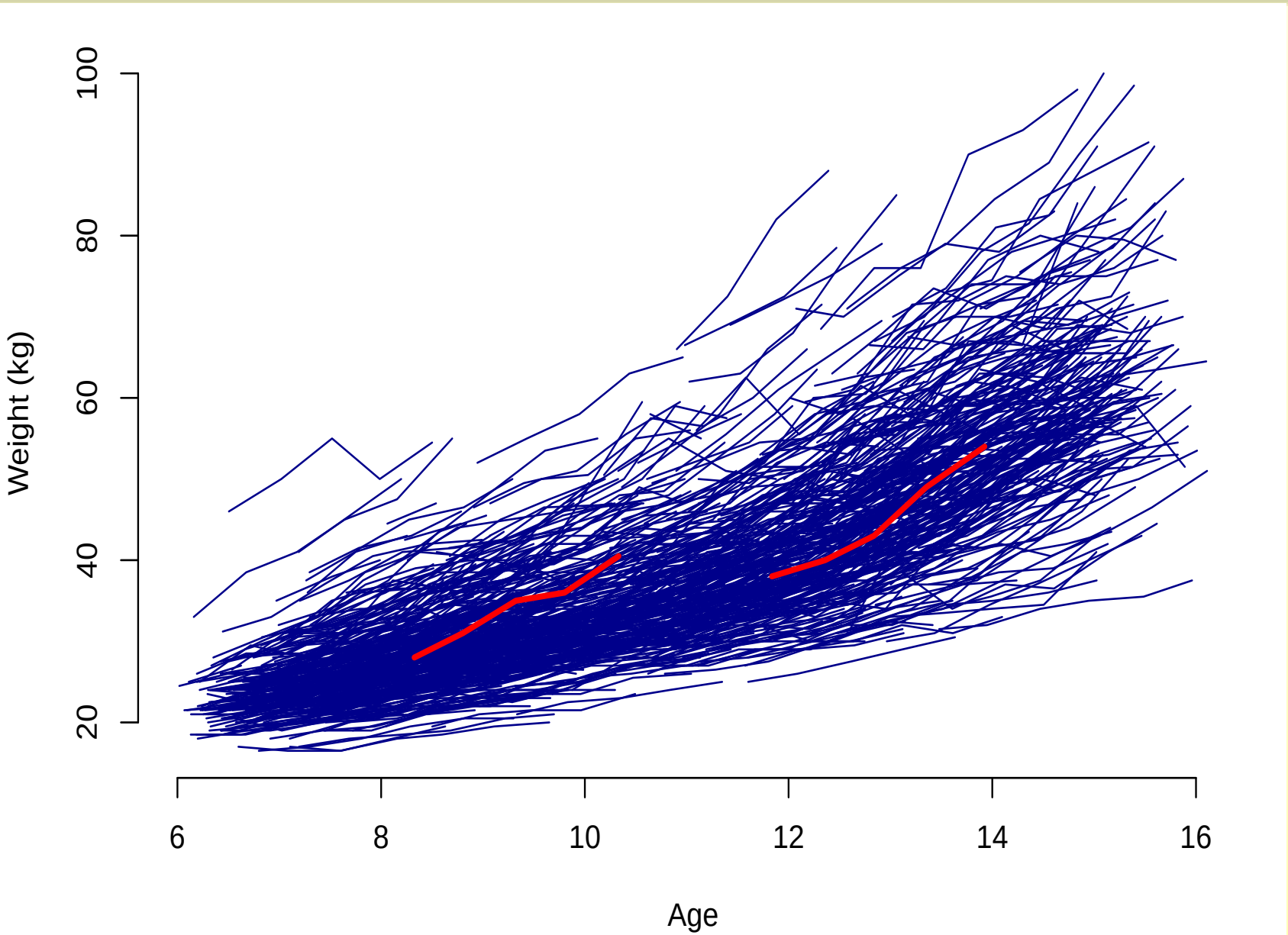
Stále v Králíkách

Růst českých dětí

- Semilongitudinální studie (1997-2000)
- České děti ve věku 6–16 let
- Každé dítě měřeno/váženo až 6x v intervalu cca 6 měsíců
- Podrobnosti:
  Bláha, Krejčovský, Jiroutová, Kobzová, Sedlák, Brabec, Riedlová, Vignerová (2006).
  Somatický vývoj současných českých dětí. Semilongitudinální studie.
  UK v Praze a SZÚ Praha

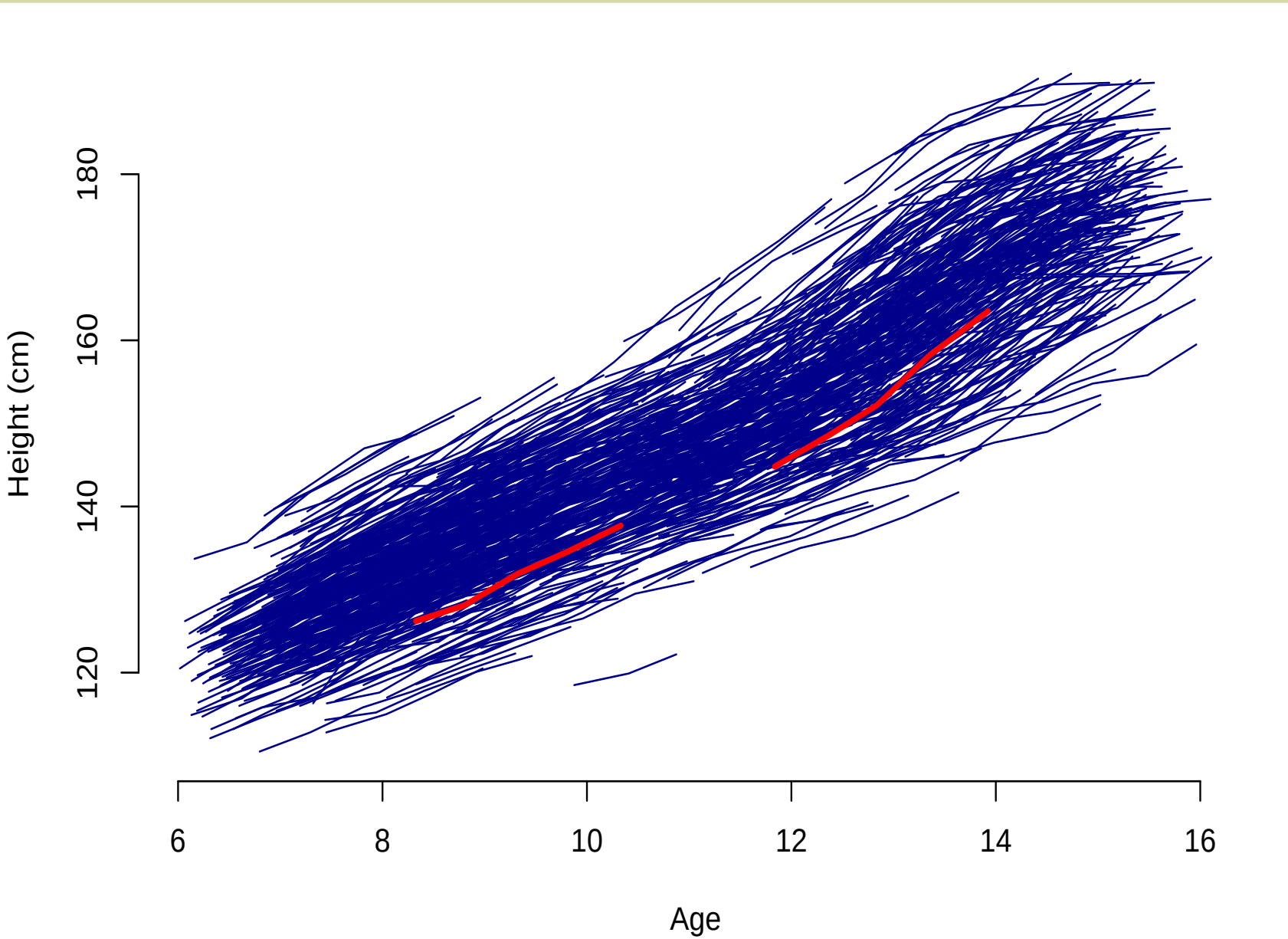
Růst českých chlapců

Hmotnost



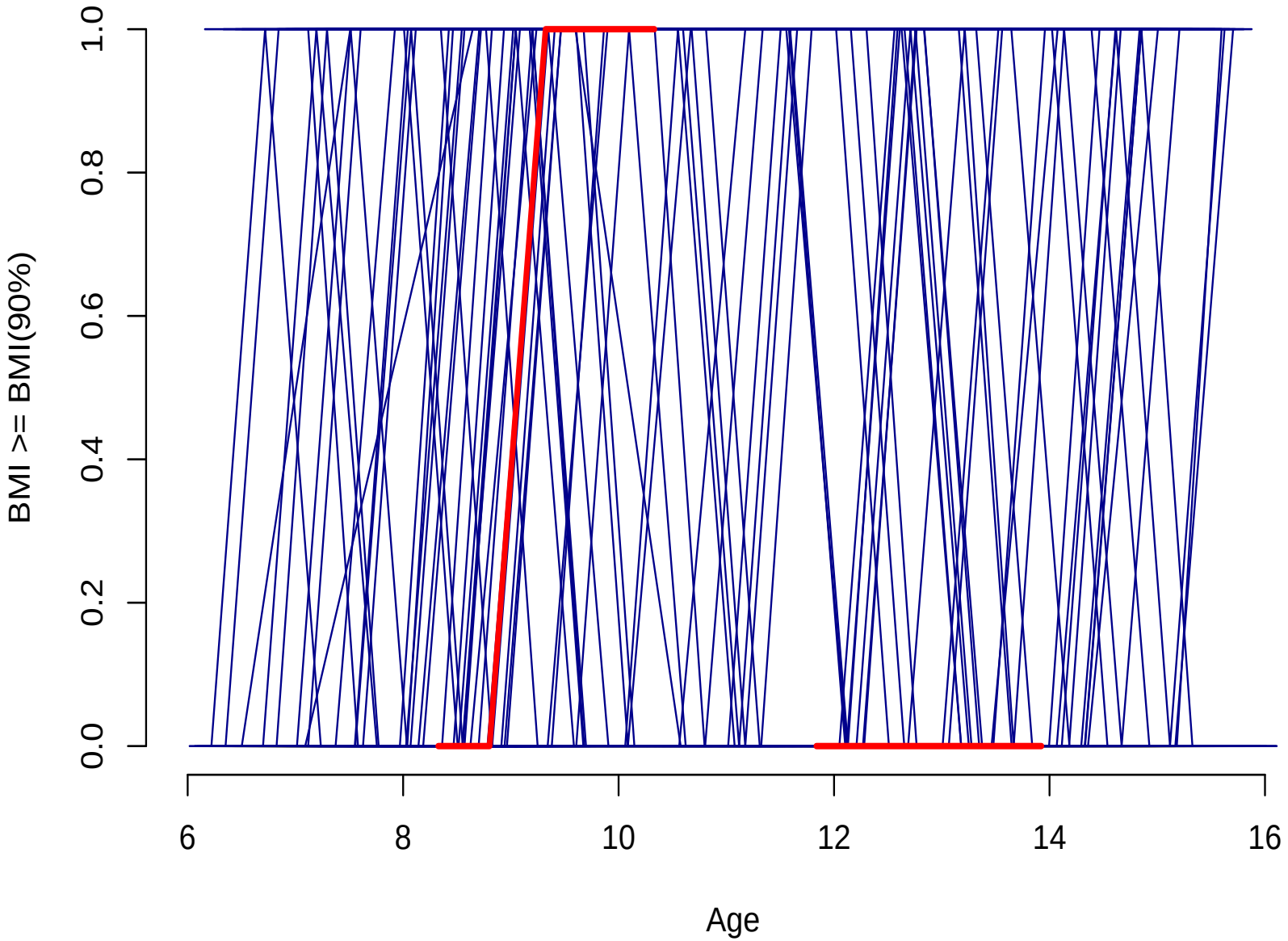
Růst českých chlapců

Výška



Růst českých chlapců

BMI nad populačním 90% percentilem



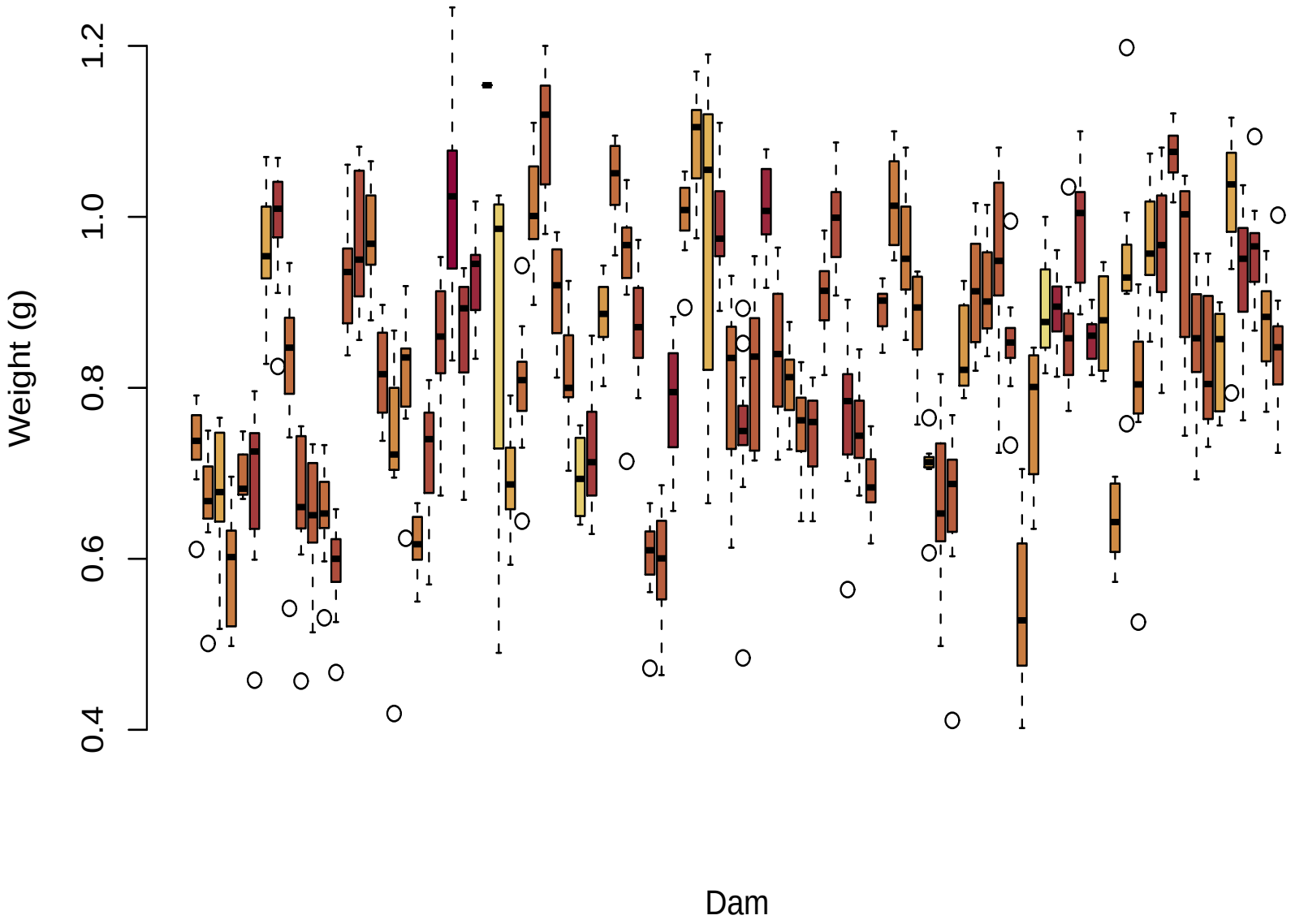
Myši s etylen-glykolem

- Klinická studie (National Toxicology Program, National Institute of Health)
- Těhotným myším byl podáván etylen-glykol
- V 17. dnu gravidity byla zjištěna hmotnost plodů a fakt, je-li plod poškozen
- Celkem 94 myší, každá 1–16 plodů

- Ve statistické literatuře poměrně profláknutá data
  - Catalano, Ryan (1992, JASA)
  - ⋮
  - Faes, Molenberghs, Aerts, Verbeke, Kenward (2009, Amer. Stat.)

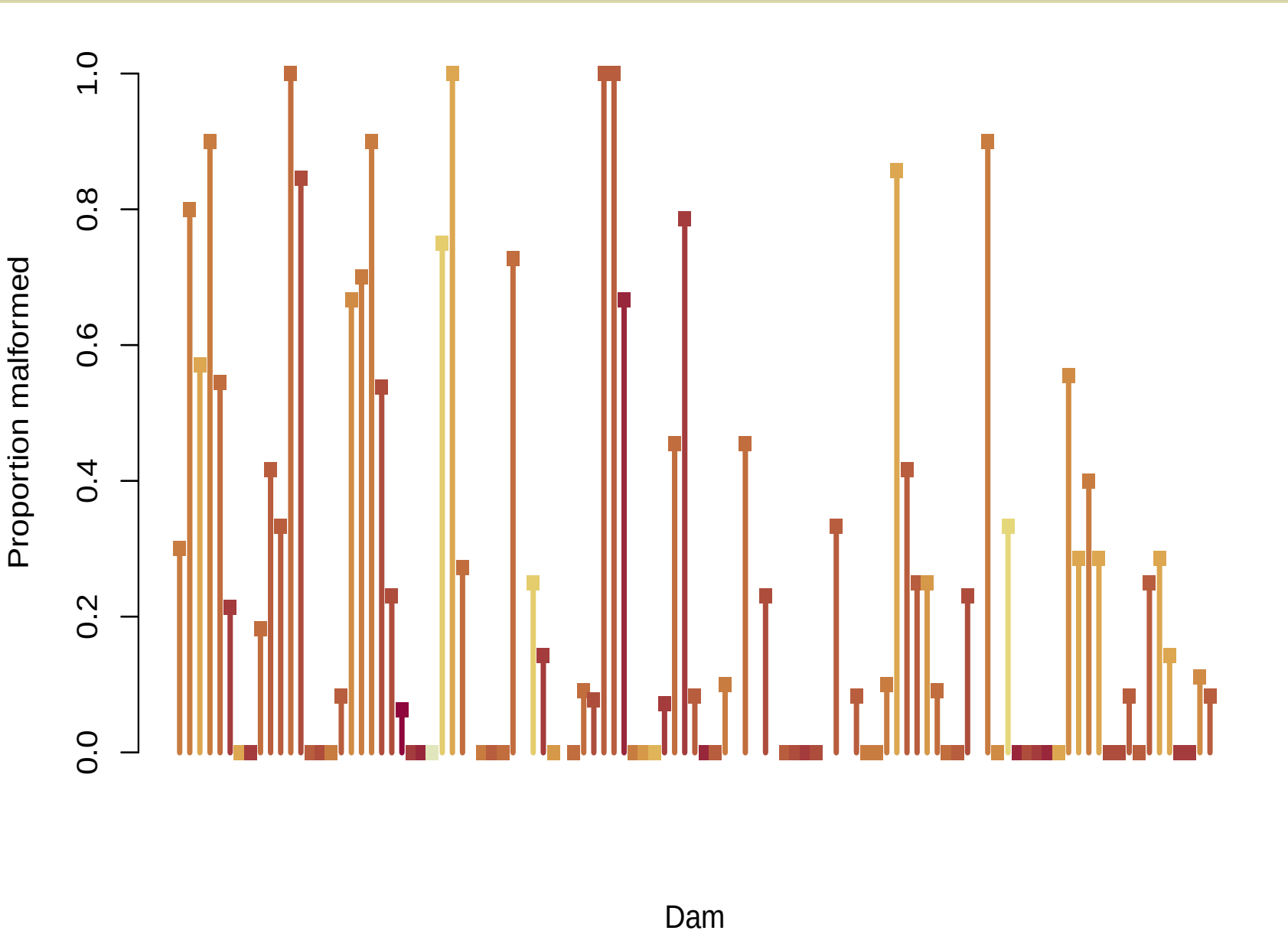
Myši s etylen-glykolem

Hmotnost plodů



Myši s etylen-glykolem

Proporce poškozených plodů



Myši s etylen-glykolem

- Nejsou to sice longitudinální data
- Ale jsou zde závislá pozorování (plody jedné samice)
- ,,čas“ $\equiv$ index plodu v rámci samice
- Takže metodologicky se to od longitudinálních dat moc neliší

Pravděpodobnostní model

- Jeden chlapec/jedna myš/jeden subjekt:

$$\boldsymbol{Y}_{i,1} = (Y_{i,1,1}, \ldots, Y_{i,1,n_{i,1}})' \qquad \text{(odezva č. 1)}$$

$$\boldsymbol{Y}_{i,2} = (Y_{i,2,1}, \ldots, Y_{i,2,n_{i,2}})' \qquad \text{(odezva č. 2)}$$

$$\vdots$$

$$\boldsymbol{Y}_{i,R} = (Y_{i,R,1}, \ldots, Y_{i,R,n_{i,R}})' \qquad \text{(odezva č. R)}$$

- $Y_{i,j,l}$ ne nutně spojité
- Budeme chtít shlukovat na základě

$$\boldsymbol{Y}_i = (\boldsymbol{Y}_{i,1}, \boldsymbol{Y}_{i,2}, \ldots, \boldsymbol{Y}_{i,R})'$$

- Je potřeba mít sdružený model pro $\boldsymbol{Y}_i$

GLMM pro několik různých odezev

- Zobecněný lineární smíšený model (GLMM) pro r-tou odezvu $(r = 1, \ldots, R)$:

$$h_r^{-1}\left\{\mathrm{E}\left(Y_{i,r,j} \middle| \boldsymbol{b}_{i,r}\right)\right\} = \boldsymbol{x}'_{i,r,j}\boldsymbol{\alpha}_r + \boldsymbol{z}_{i,r,j}\boldsymbol{b}_{i,r}$$

  - h_r^{-1} . . . linková funkce:
    - ✔ identita, je-li $\boldsymbol{Y}_{i,r} \mid \boldsymbol{b}_{i,r}$ normální
    - ✔ logit, je-li $\boldsymbol{Y}_{i,r} \mid \boldsymbol{b}_{i,r}$ alternativní (Bernoulli)
    - ✔ log, je-li $\boldsymbol{Y}_{i,r} \mid \boldsymbol{b}_{i,r}$ Poisson
  - $\boldsymbol{x}_{i,1,1}, \ldots, \boldsymbol{x}_{i,R,n_{i,R}}, \boldsymbol{z}_{i,1,1}, \ldots, \boldsymbol{z}_{i,R,n_{i,R}}$ . . . vektory regresorů
  - $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_R$ . . . (neznámé) regresní parametry
  - $\boldsymbol{b}_i = (\boldsymbol{b}'_{i,1}, \ldots, \boldsymbol{b}'_{i,R})' \overset{\text{i.i.d.}}{\sim} \sum_{k=1}^{K} w_k \mathcal{N}(\boldsymbol{\mu}_k, \mathbb{D}_k)$

- Pro shlukování: i.i.d. náhodné veličiny $U_1, \ldots, U_n$
- $U_i \in \{1, \ldots, K\}$
- $\mathrm{P}(U_i = k) = w_k, k = 1, \ldots, K, \; i = 1, \ldots, n$

Modelově založená shluková analýza

Vícerozměrný GLMM s normální směsí v rozdělení náhodných efektů zapsaný hierarchicky

- Hierarchicky zapsáno:

$$[\underbrace{\boldsymbol{Y}_{i,1},\ldots,\boldsymbol{Y}_{i,R}}_{\boldsymbol{Y}_i} \mid \underbrace{\boldsymbol{b}_{i,1},\ldots,\boldsymbol{b}_{i,R}}_{\boldsymbol{b}_i}] \sim \prod_{r=1}^{R}\prod_{j=1}^{n_{i,r}} \underbrace{[y_{i,r,j} \mid \boldsymbol{b}_{i,r}]}_{\text{normální, Bernoulli, Poisson, } \ldots}$$

se střední hodnotou $h_r(\boldsymbol{x}_{i,r,j}\boldsymbol{\alpha}_r + \boldsymbol{z}_{i,r,j}\boldsymbol{b}_{i,r})$

$$\left.\begin{aligned} \mathrm{P}(U_i = k) &= w_k \\ [\boldsymbol{b}_i \mid U_i = k] &\sim \mathcal{N}(\boldsymbol{\mu}_k, \mathbb{D}_k) \end{aligned}\right\} \quad k = 1,\ldots,K$$

Modelově založená shluková analýza

Vícerozměrný GLMM s normální směsí v rozdělení náhodných efektů

$$p(\boldsymbol{y}_i) = \int_{\mathbb{R}^q} \Big\{ \prod_{r=1}^{R} \prod_{j=1}^{n_{i,r}} \underbrace{p(y_{i,r,j} \mid \boldsymbol{b}_{i,r})}_{\text{normální, Bernoulli, Poisson, \ldots}} \Big\} \times \underbrace{\Big\{ \sum_{k=1}^{K} w_k \, \varphi\big(\boldsymbol{b}_i;\ \boldsymbol{\mu}_k, \mathbb{D}_k\big) \Big\}}_{\text{normální směs}} d\boldsymbol{b}_i$$

$$p(\boldsymbol{y}_i \mid U_i = k) = \int_{\mathbb{R}^q} \Big\{ \prod_{r=1}^{R} \prod_{j=1}^{n_{i,r}} \underbrace{p(y_{i,r,j} \mid \boldsymbol{b}_{i,r})}_{\text{normální, Bernoulli, Poisson, \ldots}} \Big\} \times \underbrace{\varphi\big(\boldsymbol{b}_i;\ \boldsymbol{\mu}_k, \mathbb{D}_k\big)}_{\text{normální}} d\boldsymbol{b}_i$$

Modelově založená shluková analýza

Vícerozměrný GLMM s normální směsí v rozdělení náhodných efektů

- Model:

$$[\boldsymbol{Y}_i \,|\boldsymbol{b}_i] \sim \prod_{r=1}^{R} \prod_{j=1}^{n_{i,r}} \text{GLMM}(Y_{i,r,j} \,|\, \boldsymbol{x}'_{i,r,j}\boldsymbol{\alpha}_r + \boldsymbol{z}'_{i,r,j}\boldsymbol{b}_{i,r})$$

$$\left.\begin{array}{rcl} \mathrm{P}(U_i = k) & = & w_k \\ [\boldsymbol{b}_i \,|\, U_i = k] & \sim & \mathcal{N}(\boldsymbol{\mu}_k, \mathbb{D}_k) \end{array}\right\} \quad k = 1, \ldots, K$$

- Parametry modelu:

$$\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{w}', \boldsymbol{\mu}'_1, \ldots, \boldsymbol{\mu}'_K, \text{vec}(\mathbb{D}_1)', \ldots, \text{vec}(\mathbb{D}_K)', \text{rozptyl GLMM})' \in \Theta$$

- Nepřímo pozorovatelné náhodné veličiny/vektory:

$$\boldsymbol{\eta} = (\boldsymbol{b}'_1, \ldots, \boldsymbol{b}'_n, U_1, \ldots, U_n)' \in \mathbb{R}^q \times \cdots \times \mathbb{R}^q \times \{1, \ldots, K\}^n$$

- Bayesova věta:

$$\pi_{i,k}(\boldsymbol{\theta}) \equiv \mathrm{P}\left(U_i = k \,\middle|\, \boldsymbol{y}_i, \boldsymbol{\theta}\right)$$

$$\propto w_k \, p(\boldsymbol{y}_i \mid U_i = k, \boldsymbol{\theta})$$

$$= w_k \int_{\mathbb{R}^q} \left\{ \prod_{r=1}^{R} \prod_{j=1}^{n_{i,r}} \underbrace{p(y_{i,r,j} \mid \boldsymbol{b}_{i,r})}_{\text{normální, Bernoulli, Poisson, } \ldots} \right\} \times$$

$$\underbrace{\varphi\left(\boldsymbol{b}_i;\ \boldsymbol{\mu}_k, \mathbb{D}_k\right)}_{\text{normální}} d\boldsymbol{b}_i$$

- I při známé hodnotě $\boldsymbol{\theta}$ analyticky obecně nevyjádřitelné
- Navíc, $\boldsymbol{\theta}$ je potřeba nejprve odhadnout...

Část V

Co v Králíkách bude možná až po 22. hodině

- Parametry modelu:

$$\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{w}', \boldsymbol{\mu}_1', \ldots, \boldsymbol{\mu}_K', \text{vec}(\mathbb{D}_1)', \ldots, \text{vec}(\mathbb{D}_K)', \text{rozptyl GLMM})' \in \Theta$$

- Věrohodnost:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{y}_i \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \int_{\mathbb{R}^q} \left\{ \prod_{r=1}^{R} \prod_{j=1}^{n_{i,r}} \underbrace{p(y_{i,r,j} \mid \boldsymbol{\alpha}, \boldsymbol{b}_{i,r})}_{\text{normální, Bernoulli, Poisson, } \ldots} \right\} \times$$

$$\underbrace{\left\{ \sum_{k=1}^{K} w_k \, \varphi(\boldsymbol{b}_i;\ \boldsymbol{\mu}_k, \mathbb{D}_k) \right\}}_{\text{normální směs}} d\boldsymbol{b}_i$$

- $\max_{\boldsymbol{\theta}\in\Theta} \log\{L(\boldsymbol{\theta})\}$ by šlo hledat EM algoritmem

- $\boldsymbol{\theta}^{(m)}$ . . . hodnota $\boldsymbol{\theta}$ v m-té iteraci EM algoritmu

  E-krok: spočítej (pro $i = 1, \ldots, n, k = 1, \ldots, K$)

$$\pi_{i,k}(\boldsymbol{\theta}^{(m)}) \equiv \mathrm{P}\left(U_i = k \,\middle|\, \boldsymbol{y}_i, \boldsymbol{\theta}^{(m)}\right)$$

$$\propto w_k^{(m)} \int_{\mathbb{R}^q} \left\{ \prod_{r=1}^{R} \prod_{j=1}^{n_{i,r}} \underbrace{p(y_{i,r,j} \mid \boldsymbol{\alpha}^{(m)}, \boldsymbol{b}_{i,r})}_{\text{normální, Bernoulli, Poisson, . . .}} \right\} \times \underbrace{\varphi\left(\boldsymbol{b}_i;\ \boldsymbol{\mu}_k^{(m)}, \mathbb{D}_k^{(m)}\right)}_{\text{normální}} d\boldsymbol{b}_i$$

- $\max_{\boldsymbol{\theta}\in\Theta} \log\{L(\boldsymbol{\theta})\}$ by šlo hledat EM algoritmem

- $\boldsymbol{\theta}^{(m)}$ . . . hodnota $\boldsymbol{\theta}$ v m-té iteraci EM algoritmu

  M-krok: maximalizuj vzhledem k $\boldsymbol{\theta}$

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \pi_{i,k}(\boldsymbol{\theta}^{(m)}) \log w_k +$$

$$\sum_{i=1}^{n}\sum_{k=1}^{K} \pi_{i,k}(\boldsymbol{\theta}^{(m)}) \log \left[\int_{\mathbb{R}^q} \left\{ \prod_{r=1}^{R}\prod_{j=1}^{n_{i,r}} \underbrace{p(y_{i,r,j} \mid \boldsymbol{\alpha}, \boldsymbol{b}_{i,r})}_{\text{normální, Bernoulli, Poisson, \ldots}} \right\} \times \right.$$

$$\left. \underbrace{\varphi(\boldsymbol{b}_i;\ \boldsymbol{\mu}_k, \mathbb{D}_k)}_{\text{normální}}\, d\boldsymbol{b}_i \right]$$

- Parametry modelu:

$$\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{w}', \boldsymbol{\mu}_1', \ldots, \boldsymbol{\mu}_K', \\ \text{vec}(\mathbb{D}_1)', \ldots, \text{vec}(\mathbb{D}_K)', \text{rozptyl GLMM})' \in \Theta$$

- Rušivé parametry:

$$\boldsymbol{\eta} = (\boldsymbol{b}_1', \ldots, \boldsymbol{b}_n', U_1, \ldots, U_n)' \in \mathbb{R}^q \times \cdots \times \mathbb{R}^q \times \{1, \ldots, K\}^n$$

- Věrohodnost:

$$L(\boldsymbol{\theta}, \boldsymbol{\eta}) = \prod_{i=1}^{n} p(\boldsymbol{y}_i \mid \boldsymbol{\theta}, \boldsymbol{\eta}) = \prod_{i=1}^{n} \prod_{r=1}^{R} \prod_{j=1}^{n_{i,r}} \underbrace{p(y_{i,r,j} \mid \boldsymbol{\alpha}, \boldsymbol{b}_{i,r})}_{\text{normální, Bernoulli, Poisson, } \ldots}$$

  - ✏ Věrohodnost ,,obyčejného“ GLM (pouze jednou M)

- Apriorní rozdělení:

$$p(\boldsymbol{\theta}, \boldsymbol{\eta}) = p(\boldsymbol{\eta} \mid \boldsymbol{\theta}) \times p(\boldsymbol{\theta})$$

$$= \prod_{i=1}^{n} \left\{ p(\boldsymbol{b}_i \mid U_i, \boldsymbol{\theta}) \times p(U_i \mid \boldsymbol{\theta}) \right\} \times p(\boldsymbol{\theta})$$

- $p(\boldsymbol{b}_i \mid U_i = k, \boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\mu}_k, \mathbb{D}_k)$
- $p(U_i \mid \boldsymbol{\theta}) \sim \mathrm{P}(U_i = k \mid \boldsymbol{w}) = w_k$
- $p(\boldsymbol{\theta}) \equiv$,,slušné“ apriorní rozdělení pro ,,klasické“ parametry

- Aposteriorní rozdělení:

$$p(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \boldsymbol{y}) = \frac{L(\boldsymbol{\theta}, \boldsymbol{\eta}) \times p(\boldsymbol{\theta}, \boldsymbol{\eta})}{\int L(\boldsymbol{\theta}, \boldsymbol{\eta}) \times p(\boldsymbol{\theta}, \boldsymbol{\eta}) \, d\boldsymbol{\theta} \, d\boldsymbol{\eta}}$$

- MCMC (detaily pro zájemce dnes po 22. hodině):
  - Náhodný výběr $(\boldsymbol{\theta}^{(1)}, \boldsymbol{\eta}^{(1)}), \ldots, (\boldsymbol{\theta}^{(M)}, \boldsymbol{\eta}^{(M)})$ z $p(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \boldsymbol{y})$

- Připomenutí: shlukujeme na základě

$$\mathrm{P}(U_i = k \,|\, \boldsymbol{y}) = \mathrm{E}_{\boldsymbol{\theta}}\big\{\mathrm{P}(U_i = k \,|\, \boldsymbol{y}, \boldsymbol{\theta}) \,\big|\, \boldsymbol{y}\big\} = \mathrm{E}_{\boldsymbol{\theta}}\big\{\pi_{i,k}(\boldsymbol{\theta}) \,\big|\, \boldsymbol{y}\big\}$$

- MCMC odhad:

$$\widehat{\mathrm{P}}(U_i = k \,|\, \boldsymbol{y}) = M^{-1} \sum_{m=1}^{M} \pi_{i,k}(\boldsymbol{\theta}^{(m)})$$

- Bylo:

$$\pi_{i,k}(\boldsymbol{\theta}^{(m)}) \;\propto\; w_k^{(m)} \int_{\mathbb{R}^q} \Big\{ \prod_{r=1}^{R} \prod_{j=1}^{n_{i,r}} \underbrace{p(y_{i,r,j} \,|\, \boldsymbol{\alpha}^{(m)}, \boldsymbol{b}_{i,r})}_{\text{normální, Bernoulli, Poisson, \ldots}} \Big\} \times \underbrace{\varphi\big(\boldsymbol{b}_i;\, \boldsymbol{\mu}_k^{(m)}, \mathbb{D}_k^{(m)}\big)}_{\text{normální}} \, d\boldsymbol{b}_i$$

- Připomenutí: shlukujeme na základě

$$\mathrm{P}(U_i = k \,|\, \boldsymbol{y}) = \mathrm{E}_{\boldsymbol{\theta}, \boldsymbol{b}_i} \left\{ \mathrm{P}(U_i = k \,|\, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{b}_i) \,\middle|\, \boldsymbol{y} \right\}$$

- MCMC odhad:

$$\widehat{\mathrm{P}}(U_i = k \,|\, \boldsymbol{y}) = M^{-1} \sum_{m=1}^{M} \mathrm{P}(U_i = k \,|\, \boldsymbol{y}, \boldsymbol{\theta}^{(m)}, \boldsymbol{b}_i^{(m)})$$

- Trocha počítání:

$$\mathrm{P}(U_i = k \,|\, \boldsymbol{y}, \boldsymbol{\theta}^{(m)}, \boldsymbol{b}_i^{(m)}) \;\propto w_k^{(m)} \prod_{r=1}^{R} \prod_{j=1}^{n_{i,r}} \underbrace{p(y_{i,r,j} \,|\, \boldsymbol{\alpha}^{(m)}, \boldsymbol{b}_{i,r})}_{\text{normální, Bernoulli, Poisson, \ldots}}$$

- Připomenutí: shlukujeme na základě

$$\mathrm{P}(U_i = k \,|\, \boldsymbol{y}) = \mathrm{P}(V_{i,k} = 1 \,|\, \boldsymbol{y})$$

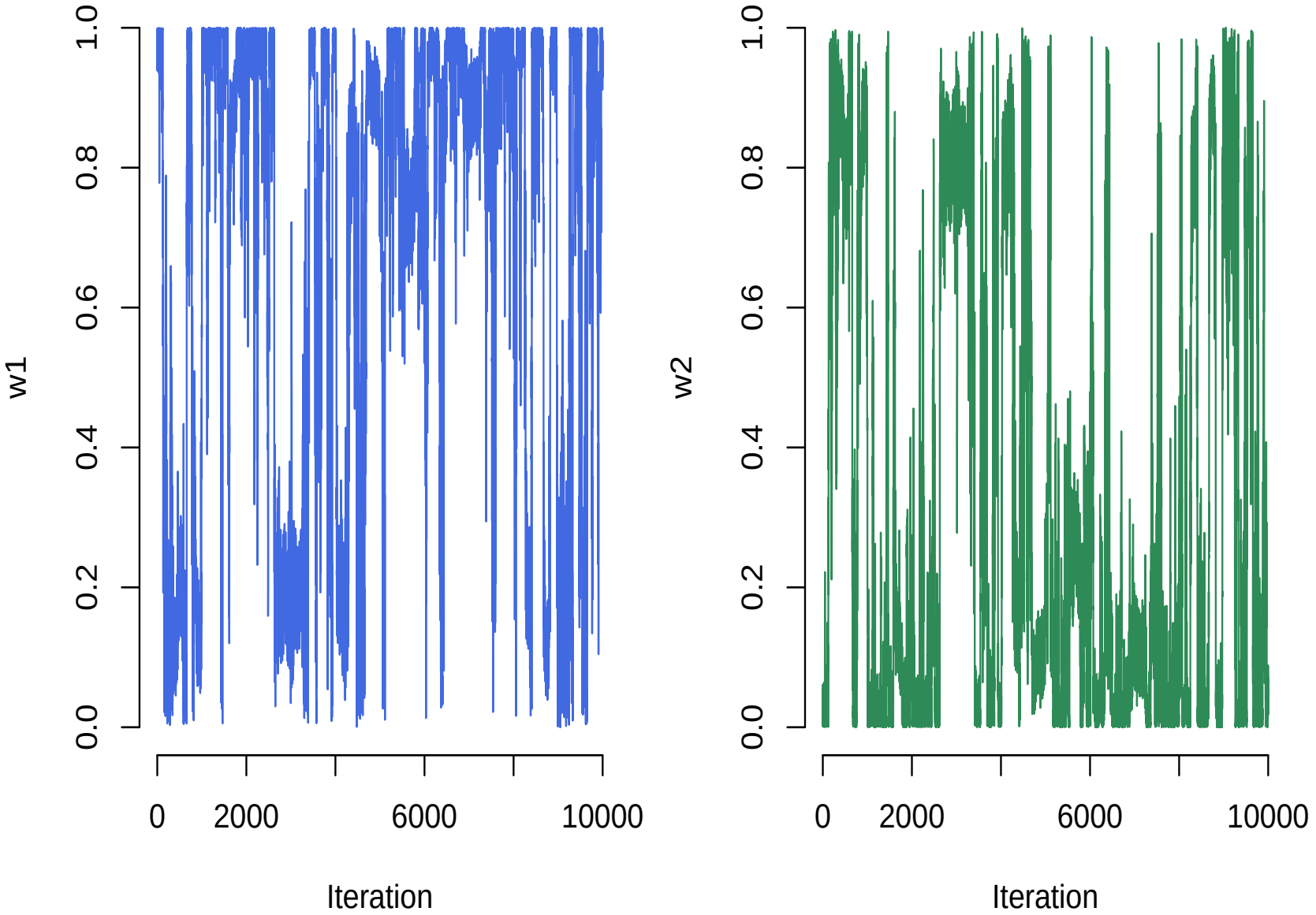
  - $\boldsymbol{V}_i = (V_{i,1}, \ldots, V_{i,K})'$, $V_{i,k} \in \{0, 1\}$, $\sum_{k=1}^{K} V_{i,k} = 1$
  - Jinými slovy: $V_{i,k} = \mathbb{I}(U_i = k)$

- MCMC odhad:

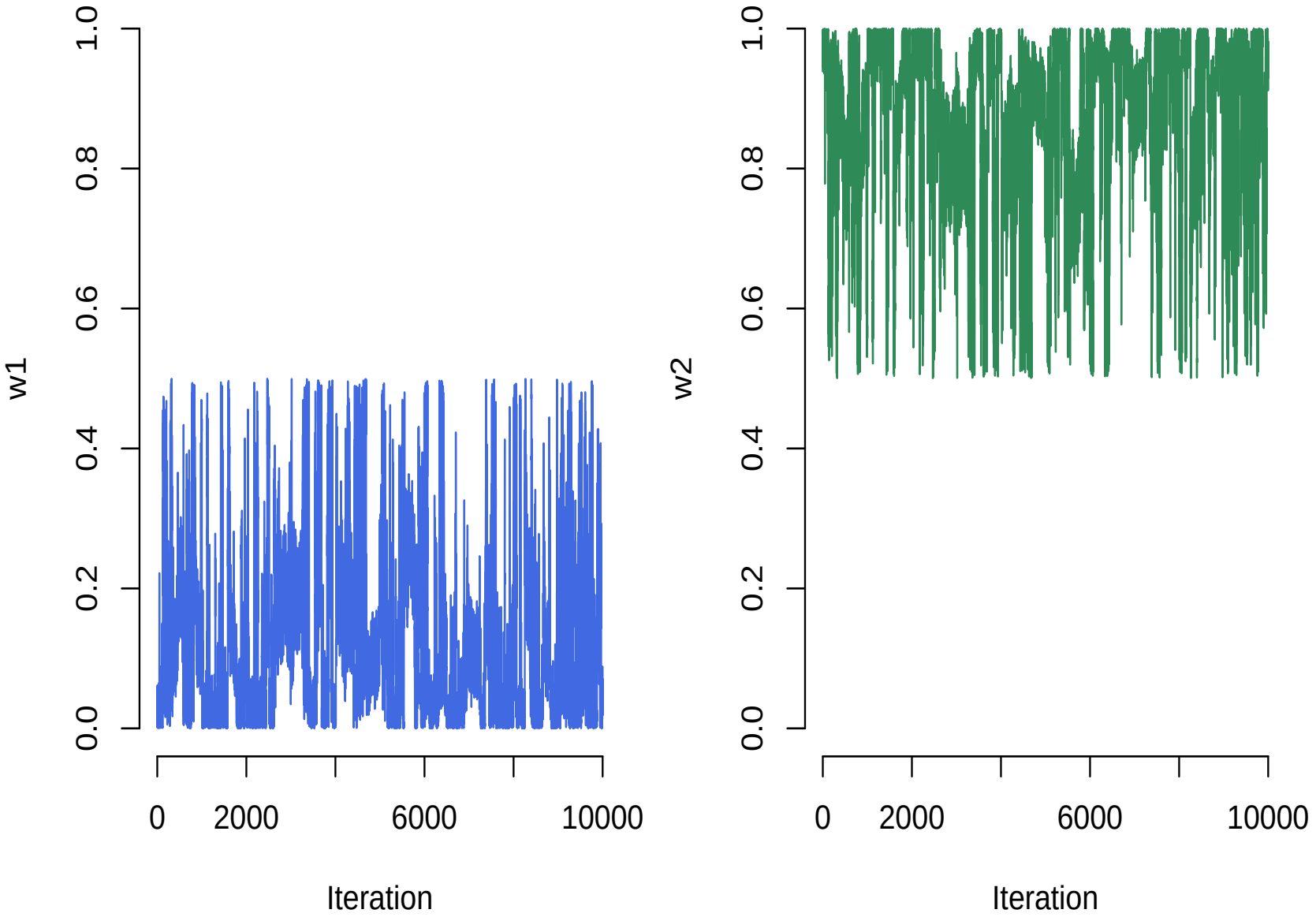
$$\widehat{\mathrm{P}}(U_i = k \,|\, \boldsymbol{y}) = M^{-1} \sum_{m=1}^{M} \mathbb{I}(U_i^{(m)} = k)$$

- ✧ Aposteriorní rozdělení je invariantní vůči změně očíslování komponent
- ✧ Existuje $K!$ různých očíslování
- ➠ Aposteriorní rozdělení má $K!$ symetrických lépe či hůře separovaných nosičů
- ✎ Je-li $(\boldsymbol{\theta}^{(m)}, \boldsymbol{\eta}^{(m)})$ stav MCMC v m-té iteraci, je potřeba vědět, ke kterému z $K!$ nosičů patří
- ✎ Jednoduchá zjednoznačnění aposteriorního rozdělení pomocí omezení typu $w_1 < \cdots < w_K$ nelze v rámci MCMC použít, jsou-li jednotlivé nosiče blízko sebe

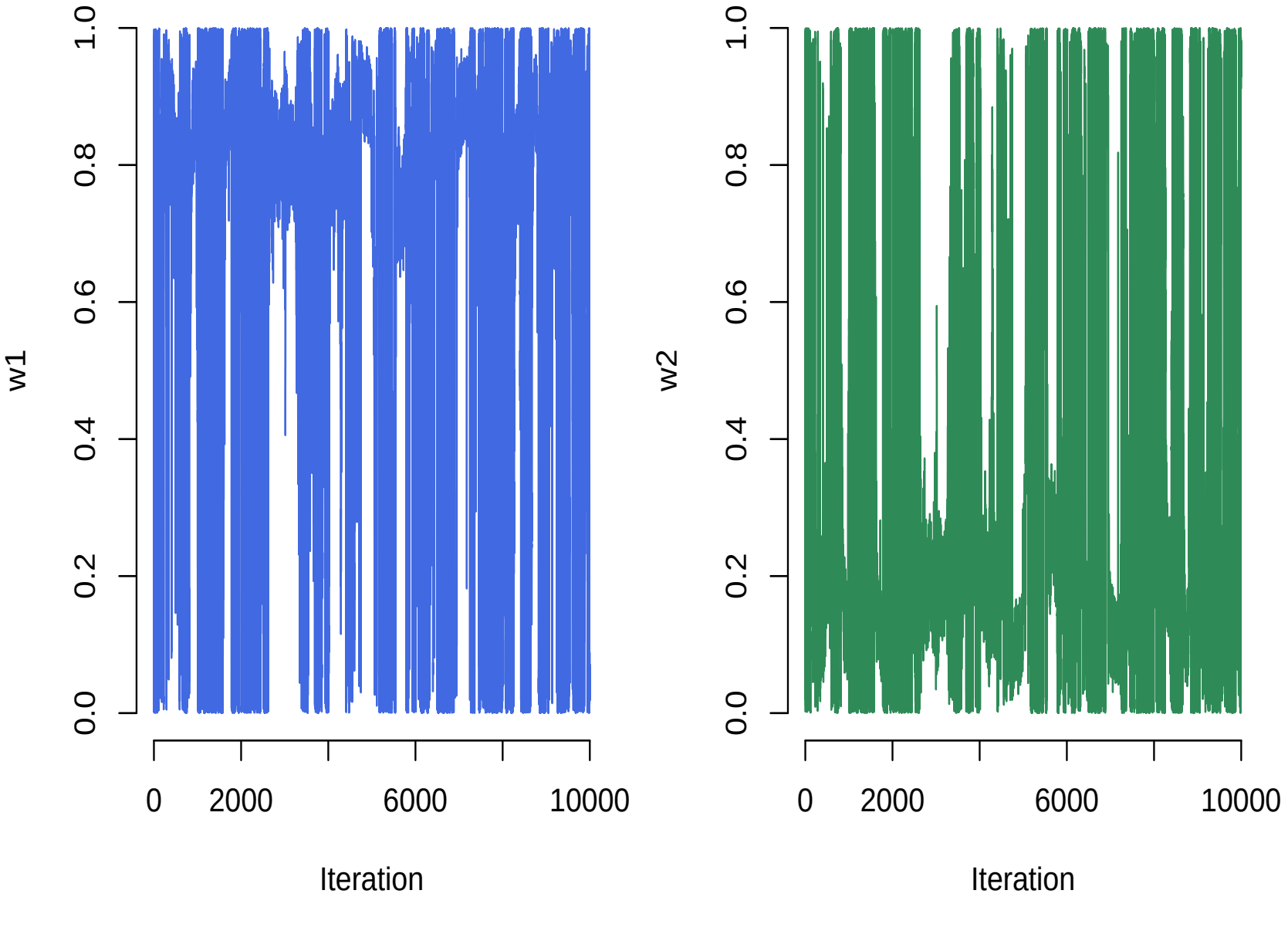
K = 2, MCMC bez omezení



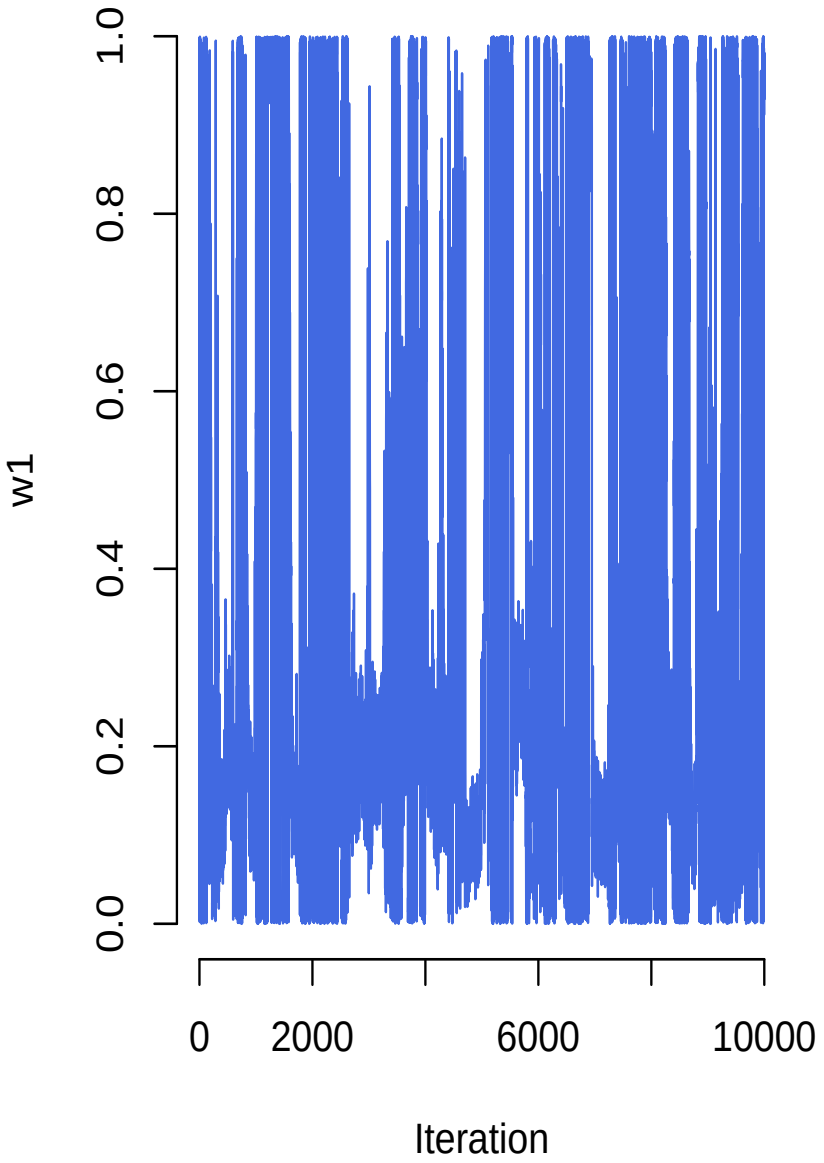
K = 2, identifikační omezení $w_1 < w_2$

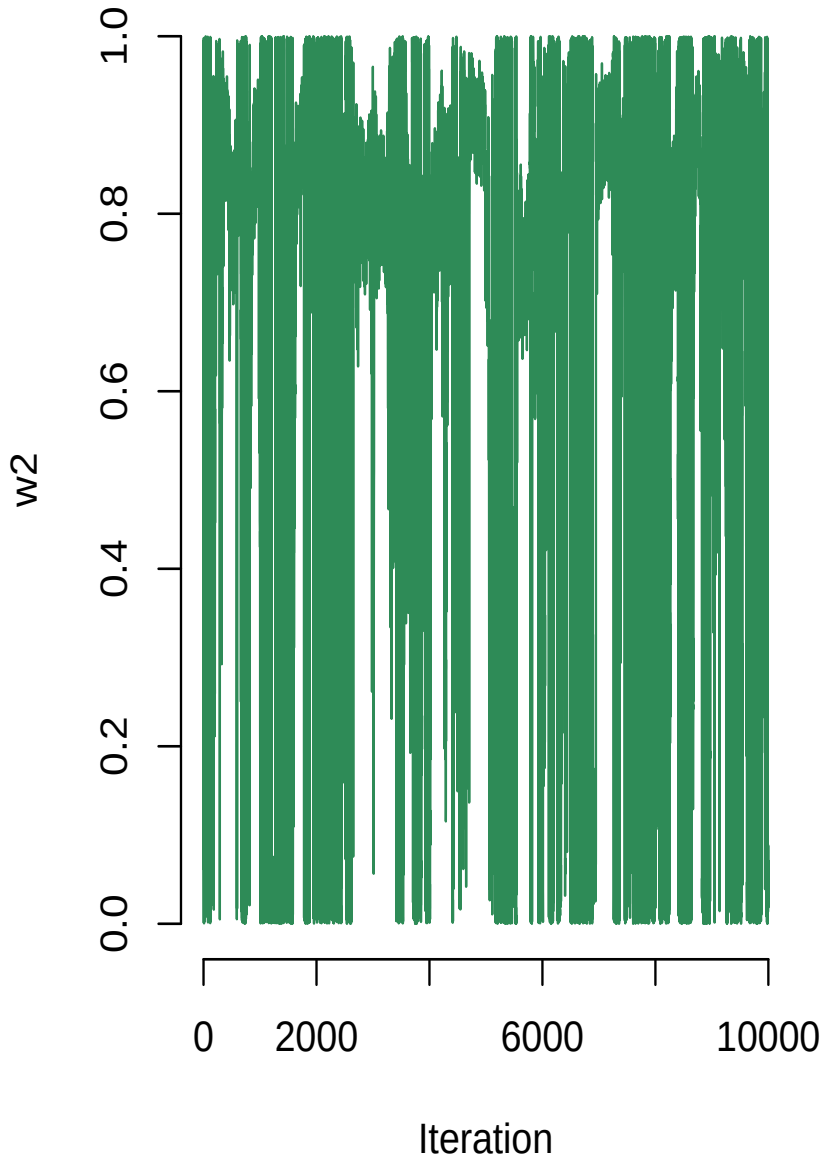


K = 2, identifikační omezení $\mu_{1,1} < \mu_{2,1}$

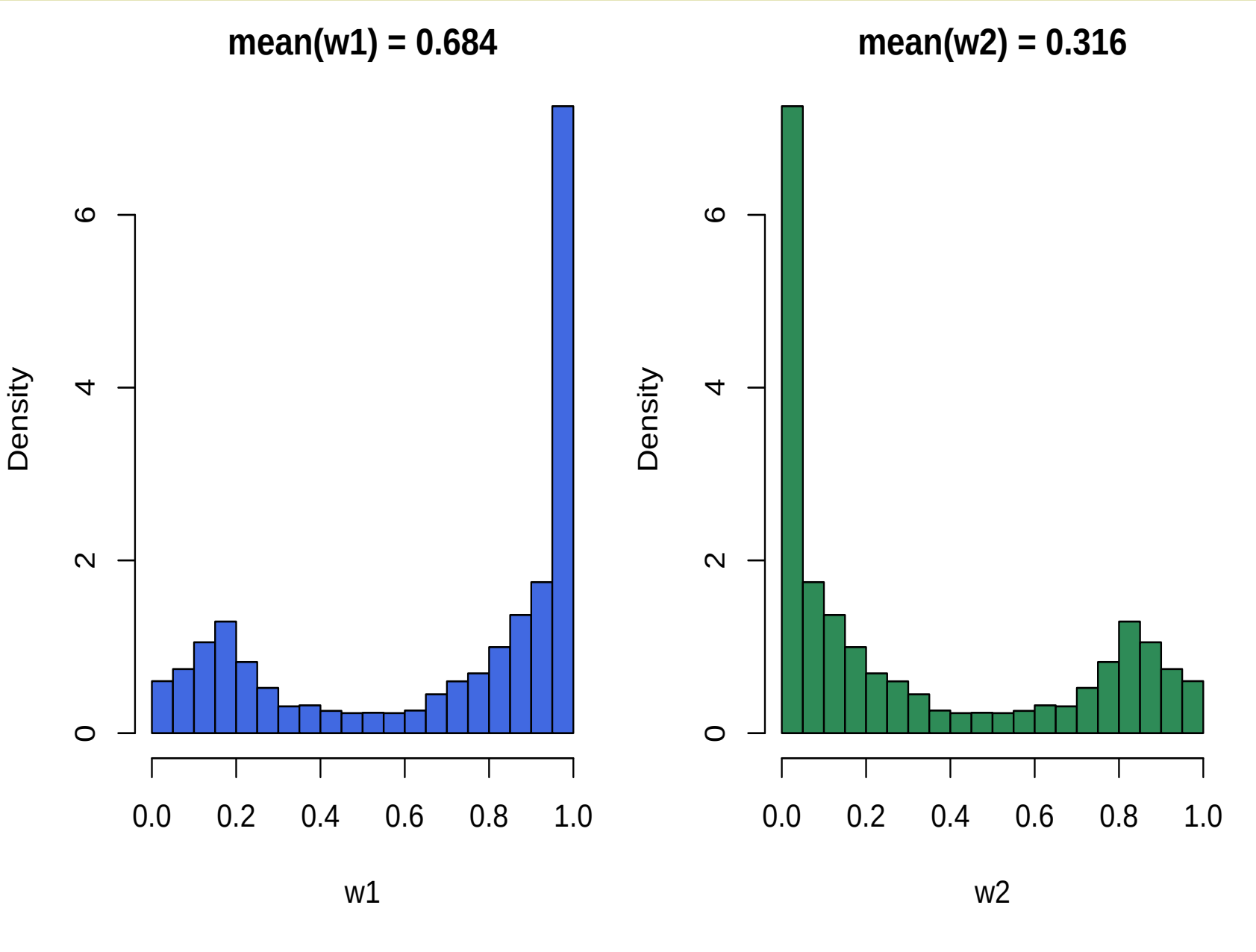


K = 2, identifikační omezení $\mu_{1,2} < \mu_{2,2}$

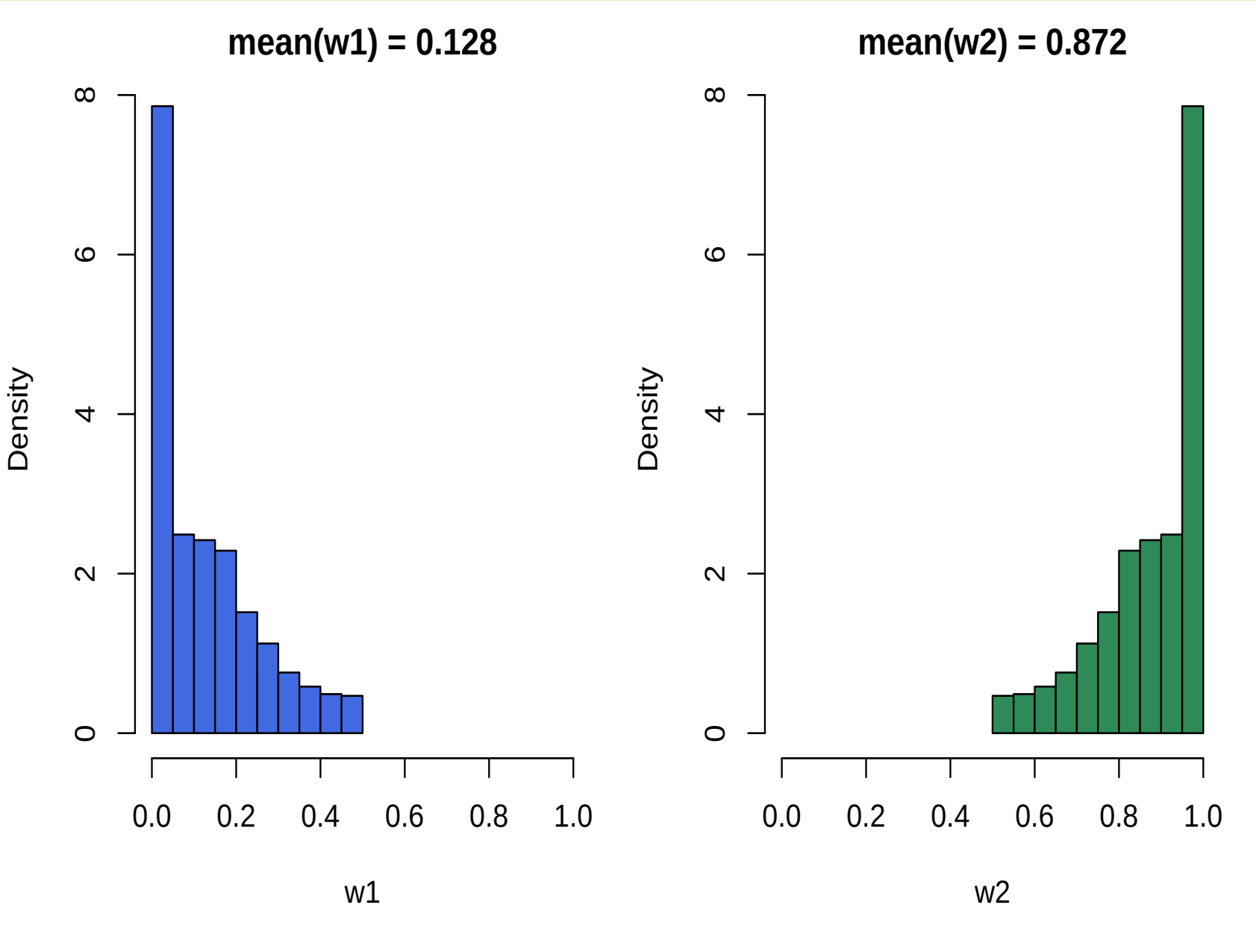




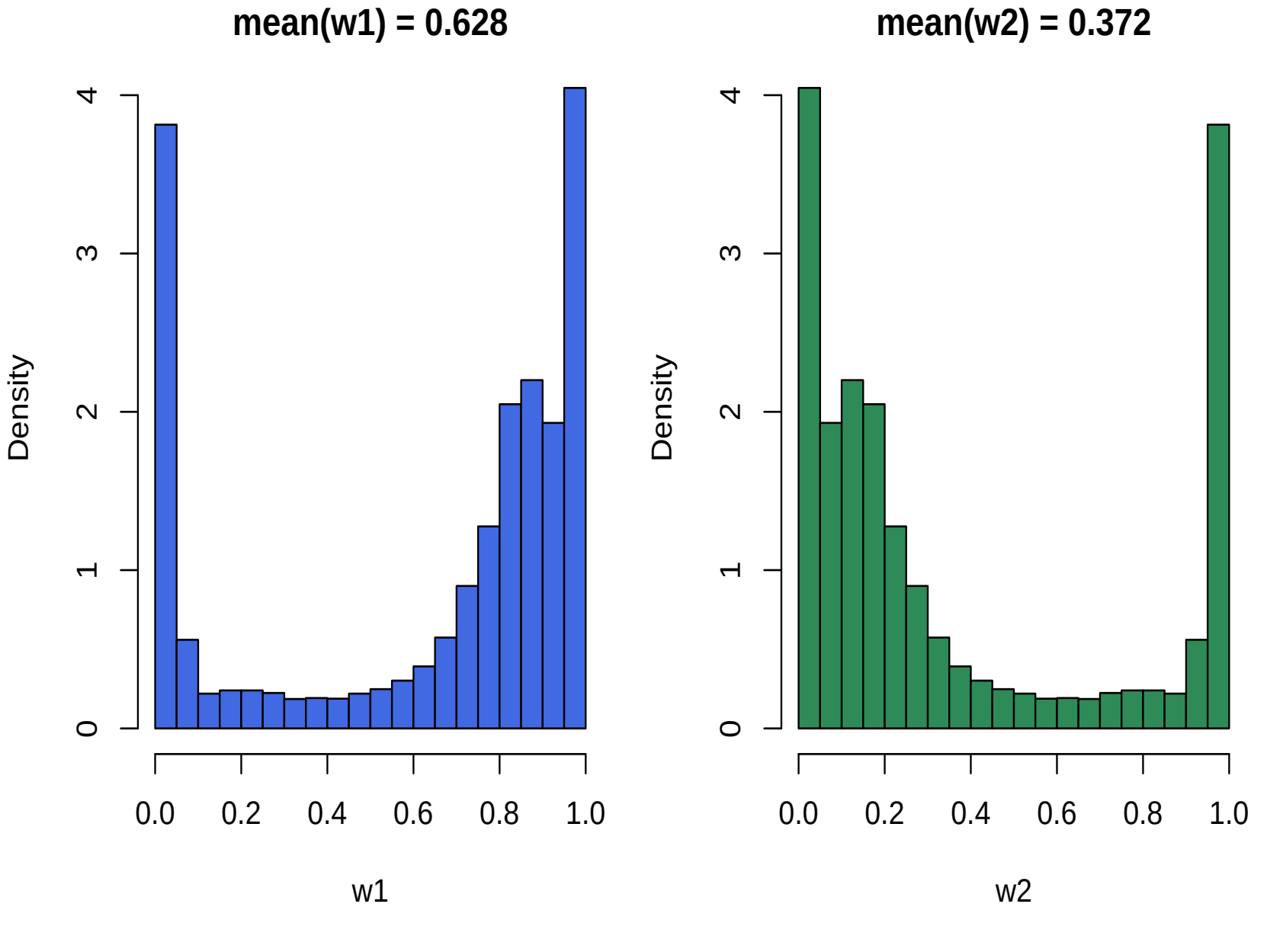
K = 2, MCMC bez omezení



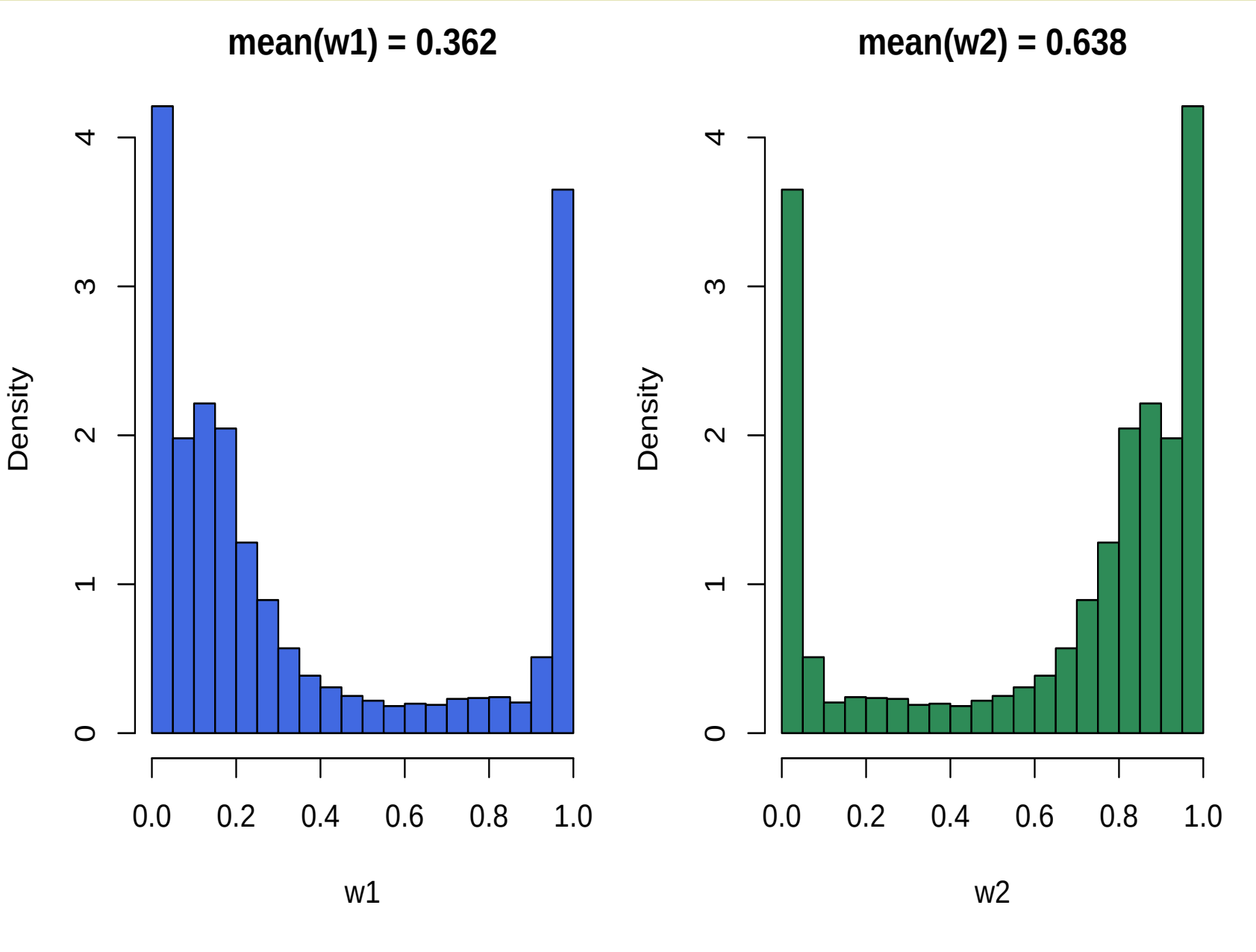
K = 2, identifikační omezení $w_1 < w_2$



K = 2, identifikační omezení $\mu_{1,1} < \mu_{2,1}$



K = 2, identifikační omezení $\mu_{1,2} < \mu_{2,2}$



- Aby šel výstup z MCMC použít ke shlukování, je potřeba použít sofistikovanější přístup k identifikaci jednoho z $K!$ nosičů
- Souhrnně na toto téma:
  Jasra, Holmes, Stephens (2005).
  Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling.
  Statistical Science, **20**, 50–67
- Se mnou na toto téma + aplikace na české děti či myši ošetřené etylen-glykolem:
  Zase někdy jindy. . .
  - ✏ Třeba na PASti ve středu odpoledne v Karlíně
  - ✏ Nebo v Řecku v srpnu na EMS

Část VI

Něco na závěr

Software

- R balíček mixAK
  - Byla o něm řeč na XV. ROBUSTu v Roháčích
  - Tenkrát tohle všechno ještě neuměl
- http://cran.R-project.org/package=mixAK

Poděkování

Za data děkuji:

Právnící: doc. K. Zvára (MFF UK), prof. D. Hendrych (PF UK)

Hepatitida B: H. Janssen, B. Hansen (Erasmus Medisch Centrum Rotterdam)

České děti: ing. J. Vignerová, dr. M. Brabec (SZÚ)

Myši s etylen-glykolem: National Toxicology Program (`http://ntp.niehs.nih.gov`), C. Faes (Univ. Hasselt)

Za finance děkuji:

GAČRu 201/09/P077

MŠMT ČR MSM 0021620839

Konec

Děkuji za pozornost!

Reklama

25th International Workshop on Statistical Modelling

4. – 9. července/júla 2010

Glasgow

http://mathstore.ac.uk/iwsm/

- Jednodenní krátký kurz:
  Giles Hooker – Analýza funkcionálních dat
- Studentská soutěž
- Abstrakty do 5. února/februára 2010 (zítra . . .)

☞ Alternativní program místo dnešního večírku???