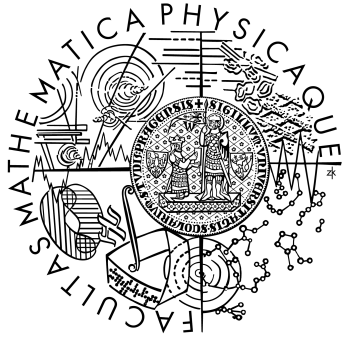
Katedra pravděpodobnosti a matematické statistiky



MATEMATICKO-FYZIKÁLNÍ
FAKULTA
Univerzita Karlova

doc. RNDr. Arnošt Komárek, Ph.D.

NMST431 Bayesovské metody

Zimní semestr 2021–22

1

Lineární model

Lineární model

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathbb{X} : \text{pevná matice } n \times k,$$
$$\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

$\Rightarrow$ $\boldsymbol{Y} \sim \mathcal{N}_n(\mathbb{X}\boldsymbol{\beta}, \sigma^2 I_n)$

- Parametry: $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top, \tau\right)^\top$, kde $\tau = \sigma^{-2} > 0$

Příklad: Vážení lehkých objektů

- Potřeba zjistit hmotnost dvou velice lehkých objektů.
  - β_1, β_2: hmotnost objektu A, resp. B
- Pokus (naměřené hodnoty v μg)
  - Y_i, $i = 1, \ldots, n$ $(n = 18)$:
    - 2$\times$ zvážen objekt A: 109, 85
    - 9$\times$ zvážen objekt B: 114, 121, 140, 122, 125, 129, 98, 134, 133
    - 7$\times$ zváženy oba objekty A a B: 217, 203, 243, 229, 233, 221, 221
- Každé měření má (náhodnou) chybu, o které budeme předpokládat, že má normální rozdělení s nulovou střední hodnotou a rozptylem σ^2.
- Lineární model pro $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$:

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\beta} = (\beta_1, \beta_2)^\top$$

  - Jak vypadá matice $\mathbb{X}$?

Věrohodnost

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \cdots \\
&= (2\pi)^{-\frac{n}{2}} \left|\mathbb{X}^\top \mathbb{X}\right|^{\frac{1}{2}} \tau^{\frac{n}{2}} \exp\left[-\frac{\tau}{2}\left\{\mathrm{SS}_e + (\boldsymbol{\beta} - \boldsymbol{b})^\top \mathbb{X}^\top \mathbb{X} (\boldsymbol{\beta} - \boldsymbol{b})\right\}\right] \\
&= (2\pi)^{-\frac{n}{2}} \left|\mathbb{X}^\top \mathbb{X}\right|^{\frac{1}{2}} \tau^{\frac{k}{2}} \exp\left\{-\frac{\tau}{2} (\boldsymbol{\beta} - \boldsymbol{b})^\top \mathbb{X}^\top \mathbb{X} (\boldsymbol{\beta} - \boldsymbol{b})\right\} \tau^{\frac{n-k+2}{2}-1} \exp\left(-\tau \frac{\mathrm{SS}_e}{2}\right)
\end{aligned}
$$

kde
- $\boldsymbol{b} = \boldsymbol{b}(\boldsymbol{y}) = \left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \mathbb{X}^\top \boldsymbol{y}$
- $\mathrm{SS}_e = \mathrm{SS}_e(\boldsymbol{y}) = (\boldsymbol{y} - \mathbb{X}\boldsymbol{b})^\top (\boldsymbol{y} - \mathbb{X}\boldsymbol{b})$
- $s = s(\boldsymbol{y}) = \sqrt{\dfrac{\mathrm{SS}_e}{n-k}}$

Příklad: Lehké objekty
- $\boldsymbol{b} = (98.89,\ 124.42)$
- $\mathrm{SS}_e = 2\,525.7$
- $s = 12.56$

Neinformativní apriorní rozdělení

- $p(\beta) \propto 1$, $\beta \in \mathbb{R}^k$
- $p(\sigma^2) \propto \frac{1}{\sigma^2}$, $\sigma^2 > 0$
  - odpovídá $p(\log \sigma^2) \propto c$, resp. $p(\log \sigma) \propto c$
- $p(\tau) \propto \frac{1}{\tau}$, $\tau > 0$
- β a τ apriori nezávislé

Sdružené aposteriorní rozdělení

$$p(\boldsymbol{\beta}, \tau \mid \boldsymbol{y}) \propto \cdots \propto$$

$$\tau^{\frac{n}{2}-1} \exp\left[-\frac{\tau}{2}\left\{\mathrm{SS}_e + (\boldsymbol{\beta} - \boldsymbol{b})^\top \mathbb{X}^\top \mathbb{X} (\boldsymbol{\beta} - \boldsymbol{b})\right\}\right], \qquad \tau > 0$$

- Za jakých podmínek se jedná skutečně o (nedegenerované) rozdělení?
- To jest, za jakých podmínek existuje integrál z výše uvedené funkce?

$$p(\boldsymbol{\beta}, \tau \mid \boldsymbol{y}) = p(\boldsymbol{\beta} \mid \tau, \boldsymbol{y}) \times p(\tau \mid \boldsymbol{y})$$

Podmíněné aposteriorní rozdělení β

$$p(\beta \mid \tau, \boldsymbol{y}) \propto \cdots \propto \exp\left[-\frac{\tau}{2}\left\{(\beta - \boldsymbol{b})^\top \mathbb{X}^\top \mathbb{X}(\beta - \boldsymbol{b})\right\}\right]$$

- Normující konstanta je proporcionální determinantu matice $\left|\mathbb{X}^\top \mathbb{X}\right|^{-\frac{1}{2}}$
- To jest, aby byla $p(\beta \mid \tau, \boldsymbol{y})$ hustotou nedegenerovaného rozdělení, musí mít matice $\mathbb{X}$ sloupcovou hodnost k.

  Potom

$$\beta \mid \tau, \boldsymbol{y} \sim \mathcal{N}_k\left(\boldsymbol{b}(\boldsymbol{y}), \tau^{-1}\left(\mathbb{X}^\top \mathbb{X}\right)^{-1}\right)$$

Marginální aposteriorní rozdělení τ

$$p(\tau \mid \boldsymbol{y}) \propto \cdots \propto \tau^{\frac{n-k}{2}-1} \exp\left(-\tau \frac{\mathrm{SS}_e}{2}\right), \qquad \tau > 0$$

- Aby existovala konečná normující konstanta, musí být $\frac{n-k}{2} > 0$, tj. $n > k$. Potom

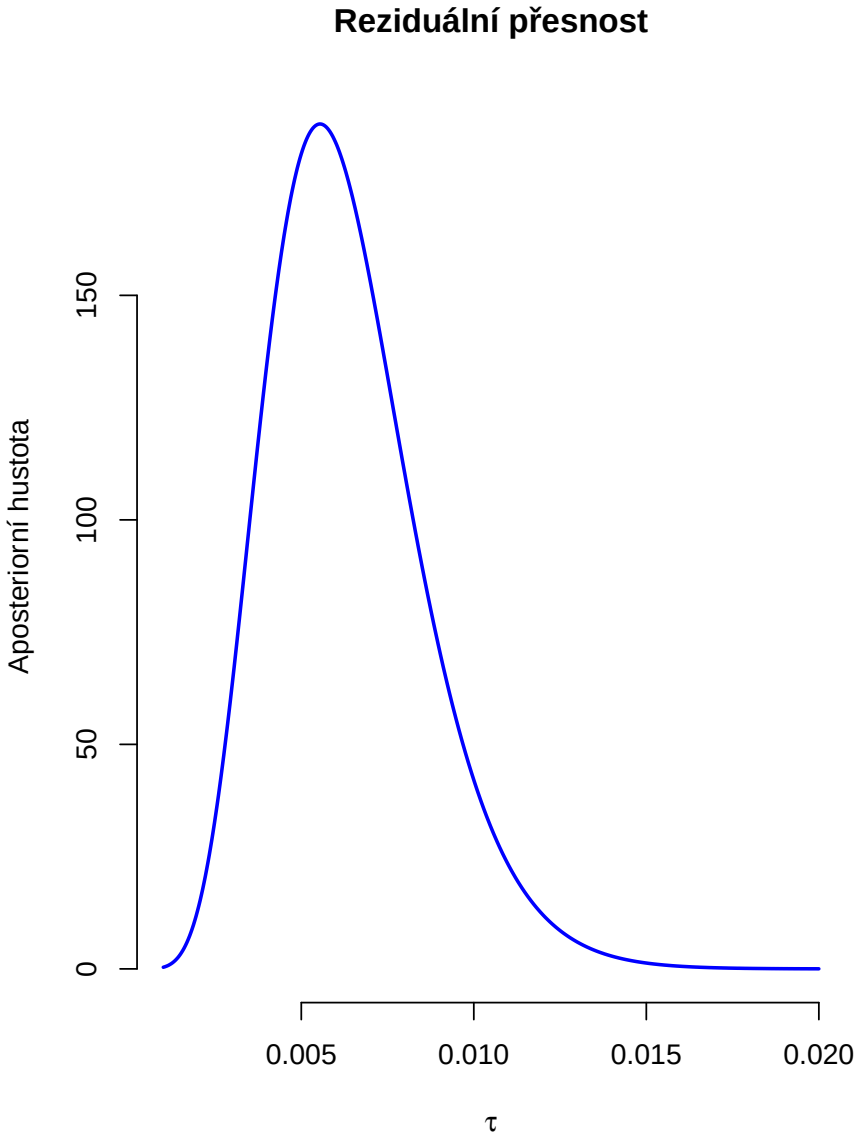
$$\tau \mid \boldsymbol{y} \sim \mathcal{G}\left(\frac{n-k}{2}, \frac{\mathrm{SS}_e(\boldsymbol{y})}{2}\right)$$

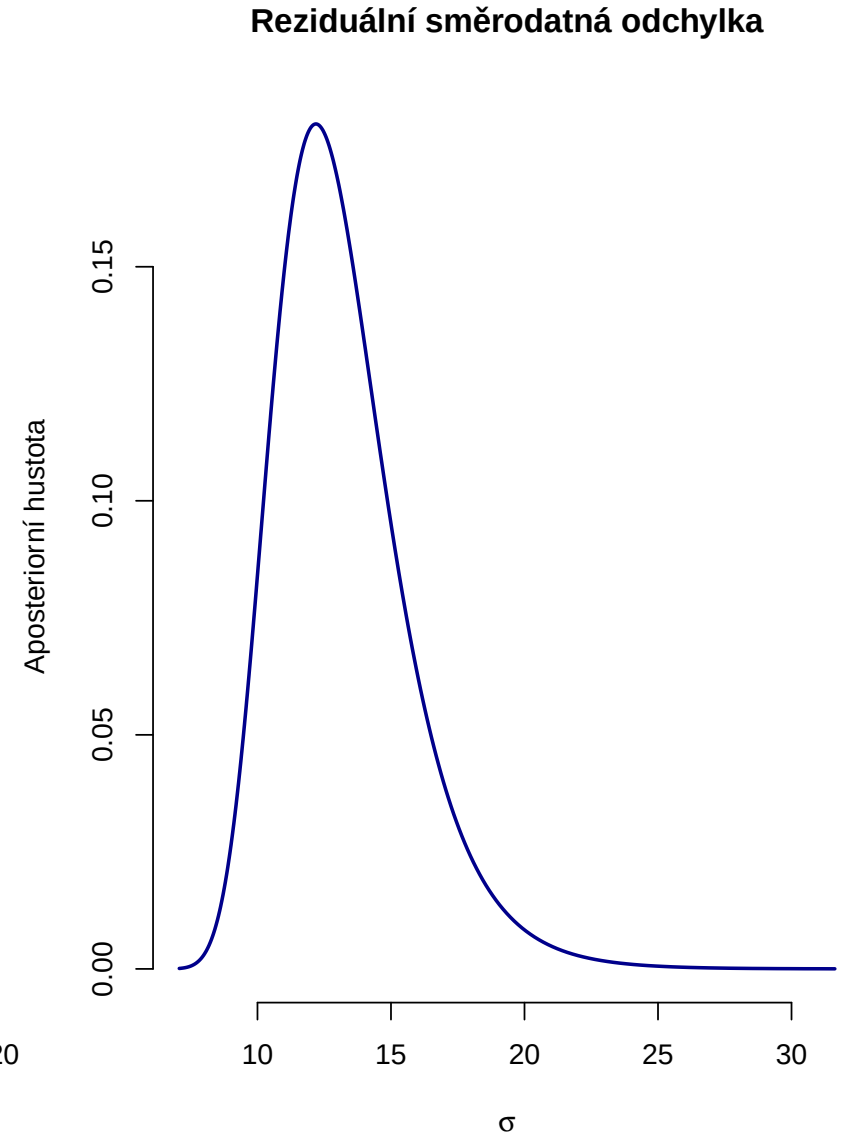
- Rutinní použití věty o transformaci

  ➠ aposteriorní hustota reziduálního rozptylu, resp. směrodatné odchylky

Marginální aposteriorní rozdělení τ a σ

Vážení lehkých objektů, $\tau \mid \boldsymbol{y} \sim \mathcal{G}(8,\, 1\,262.8)$





Momenty aposteriorních rozdělení τ, resp. σ^2, resp. σ

- $\mathbb{E}(\tau \mid \boldsymbol{Y} = \boldsymbol{y}) = \dfrac{n-k}{\mathrm{SS}_e}$
- $\mathrm{var}(\tau \mid \boldsymbol{Y} = \boldsymbol{y}) = \dfrac{2(n-k)}{\mathrm{SS}_e^2}$
- $\mathbb{E}(\sigma^2 \mid \boldsymbol{Y} = \boldsymbol{y}) = \dfrac{\mathrm{SS}_e}{n-k-2}$, pro $n-k>2$
- $\mathrm{var}(\sigma^2 \mid \boldsymbol{Y} = \boldsymbol{y}) = \dfrac{2\,\mathrm{SS}_e^2}{(n-k-2)^2\,(n-k-4)}$, pro $n-k>4$
- $\mathbb{E}(\sigma \mid \boldsymbol{Y} = \boldsymbol{y}) =$???
- $\mathrm{var}(\sigma \mid \boldsymbol{Y} = \boldsymbol{y}) =$???

Bayesovské odhady τ a σ

Vážení lehkých objektů, $\tau \mid \boldsymbol{y} \sim \mathcal{G}(8,\, 1\,262.8)$

(Možné) **bodové odhady**

$\widehat{\tau}_1 = \mathbb{E}(\tau \mid \boldsymbol{Y} = \boldsymbol{y}) = 0.00633$ $\quad$ $\widehat{\sigma}_1 = \mathbb{E}(\sigma \mid \boldsymbol{Y} = \boldsymbol{y}) = ???$

$\widehat{\tau}_2 = \text{med}(\tau \mid \boldsymbol{Y} = \boldsymbol{y}) = 0.00607$ $\quad$ $\widehat{\sigma}_2 = \text{med}(\sigma \mid \boldsymbol{Y} = \boldsymbol{y}) = 12.83$

95% věrohodnostní intervaly

ET interval $\quad \tau : (0.00273,\, 0.01142)$ $\quad \sigma : (9.36,\, 19.12)$

HPD interval $\quad \tau : (0.00235,\, 0.01079)$ $\quad \sigma : (8.87,\, 18.22)$

VSUVKA: Vícerozměrné Studentovo t-rozdělení

$\boldsymbol{T} \sim \mathrm{MVT}_{k,\nu}(\boldsymbol{\Sigma})$, jestliže

$$\boldsymbol{T} = \boldsymbol{U}\sqrt{\frac{\nu}{V}},$$

kde
- $\boldsymbol{\Sigma}$ je pozitivně definitní matice,
- $\boldsymbol{U} \sim \mathcal{N}_k(\boldsymbol{0}, \boldsymbol{\Sigma})$,
- $V \sim \chi^2_\nu$,
- $\boldsymbol{U}$ a V jsou nezávislé.

VSUVKA: Vícerozměrné Studentovo t-rozdělení

$\boldsymbol{T} \sim \mathrm{MVT}_{k,\nu}(\boldsymbol{\Sigma})$ má hustotu

$$p(\boldsymbol{t}) = \frac{\Gamma\left(\frac{\nu+k}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \nu^{\frac{k}{2}} \pi^{\frac{k}{2}}} \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \left\{1 + \frac{\boldsymbol{t}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{t}}{\nu}\right\}^{-\frac{\nu+k}{2}}, \qquad \boldsymbol{t} \in \mathbb{R}^k$$

- $\mathbb{E}\boldsymbol{T} = \boldsymbol{0}$, je-li $\nu > 1$,
- $\mathrm{var}\boldsymbol{T} = \frac{\nu}{\nu-2}\boldsymbol{\Sigma}$, je-li $\nu > 2$,
- $\mathrm{modus}\boldsymbol{T} = \boldsymbol{0}$.

VSUVKA: Posunuté vícerozměrné Studentovo t-rozdělení

Pro $\boldsymbol{\mu} \in \mathbb{R}^k$:

$$\boldsymbol{Z} = \boldsymbol{\mu} + \boldsymbol{T}, \text{ kde } \boldsymbol{T} \sim \mathrm{MVT}_{k,\nu}(\boldsymbol{\Sigma}) \text{ má hustotu}$$

$$p(\boldsymbol{z}) = \frac{\Gamma\left(\frac{\nu+k}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \nu^{\frac{k}{2}} \pi^{\frac{k}{2}}} \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \left\{1 + \frac{\left(\boldsymbol{z}-\boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{z}-\boldsymbol{\mu}\right)}{\nu}\right\}^{-\frac{\nu+k}{2}}, \qquad \boldsymbol{z} \in \mathbb{R}^k$$

- $\mathbb{E}\boldsymbol{Z} = \boldsymbol{\mu}$, je-li $\nu > 1$,
- $\mathrm{var}\boldsymbol{Z} = \frac{\nu}{\nu-2}\boldsymbol{\Sigma}$, je-li $\nu > 2$,
- $\mathrm{modus}\boldsymbol{Z} = \boldsymbol{\mu}$.

Marginální aposteriorní rozdělení β

$$p(\beta \mid \boldsymbol{y}) \propto \cdots \propto \left\{1 + \frac{s^{-2}\left(\beta - \boldsymbol{b}\right)^\top \mathbb{X}^\top \mathbb{X}\left(\beta - \boldsymbol{b}\right)}{n-k}\right\}^{-\frac{n-k+k}{2}}$$

- To jest,

$$\beta \mid \boldsymbol{y} \sim \boldsymbol{b}(\boldsymbol{y}) + \mathrm{MVT}_{k,\,n-k}\left(s^2(\boldsymbol{y})\left(\mathbb{X}^\top \mathbb{X}\right)^{-1}\right)$$

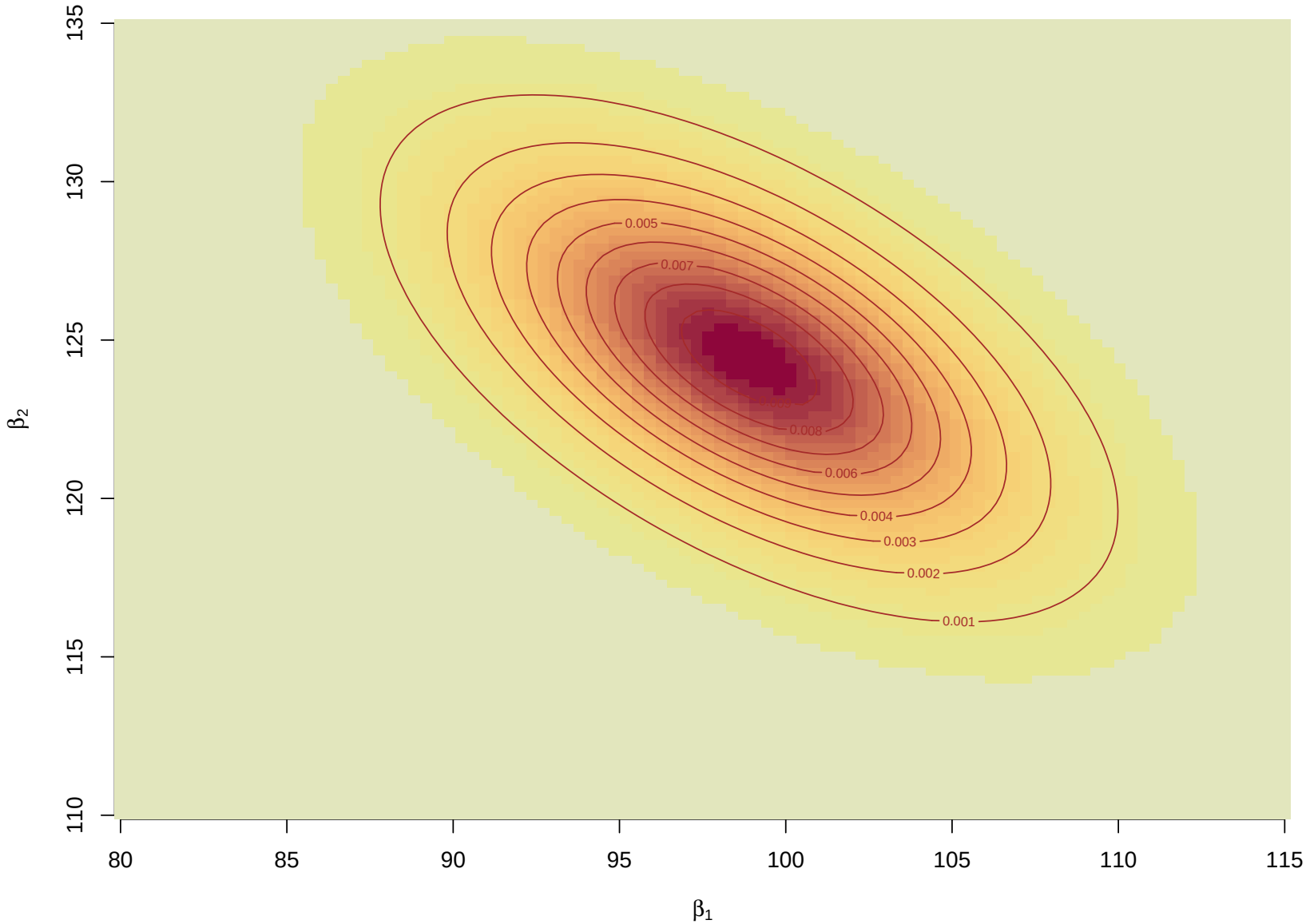
- Jaké je aposteriorní rozdělení pro β_j, $j = 1, \ldots, k$?

Momenty

- $\mathbb{E}\left(\beta \mid \boldsymbol{Y} = \boldsymbol{y}\right) = \boldsymbol{b}$, pro $n - k > 1$
- $\mathrm{var}\left(\beta \mid \boldsymbol{Y} = \boldsymbol{y}\right) = \frac{n-k}{n-k-2}\, s^2 \left(\mathbb{X}^\top \mathbb{X}\right)^{-1}$, pro $n - k > 2$

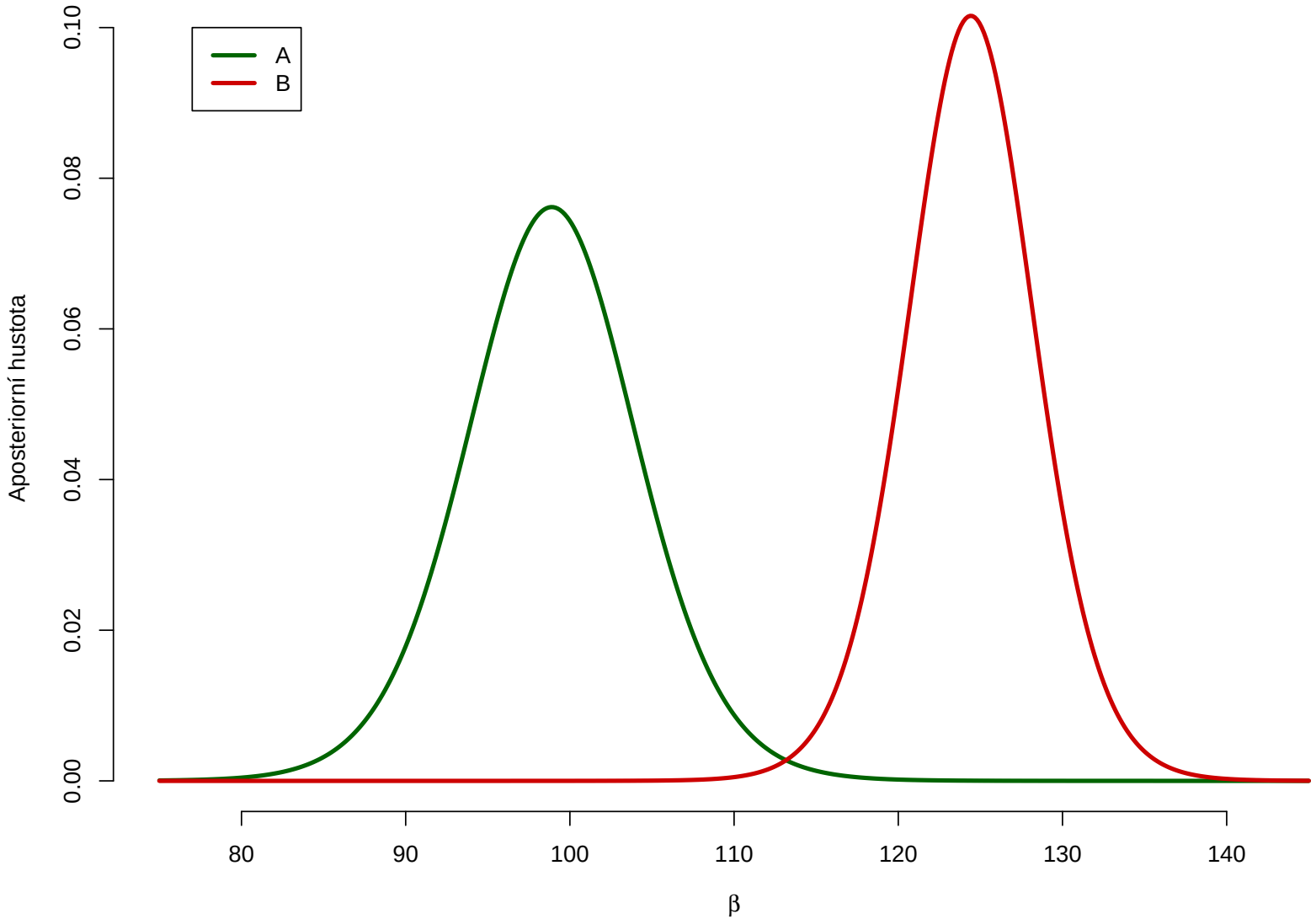
Marginální aposteriorní rozdělení β

Vážení lehkých objektů



Marginální aposteriorní rozdělení β_1 a β_2

Vážení lehkých objektů



Bayesovské odhady β_1 a β_2

Vážení lehkých objektů

(Možné) **bodové odhady**

$$\widehat{\beta}_1 = \mathbb{E}(\beta_1 \mid \boldsymbol{Y} = \boldsymbol{y}) = \text{med}(\beta_1 \mid \boldsymbol{Y} = \boldsymbol{y}) = 98.89$$

$$\widehat{\beta}_2 = \mathbb{E}(\beta_2 \mid \boldsymbol{Y} = \boldsymbol{y}) = \text{med}(\beta_2 \mid \boldsymbol{Y} = \boldsymbol{y}) = 124.42$$

95% věrohodnostní intervaly

ET i HPD interval $\beta_1 : (87.96,\ 109.83)$

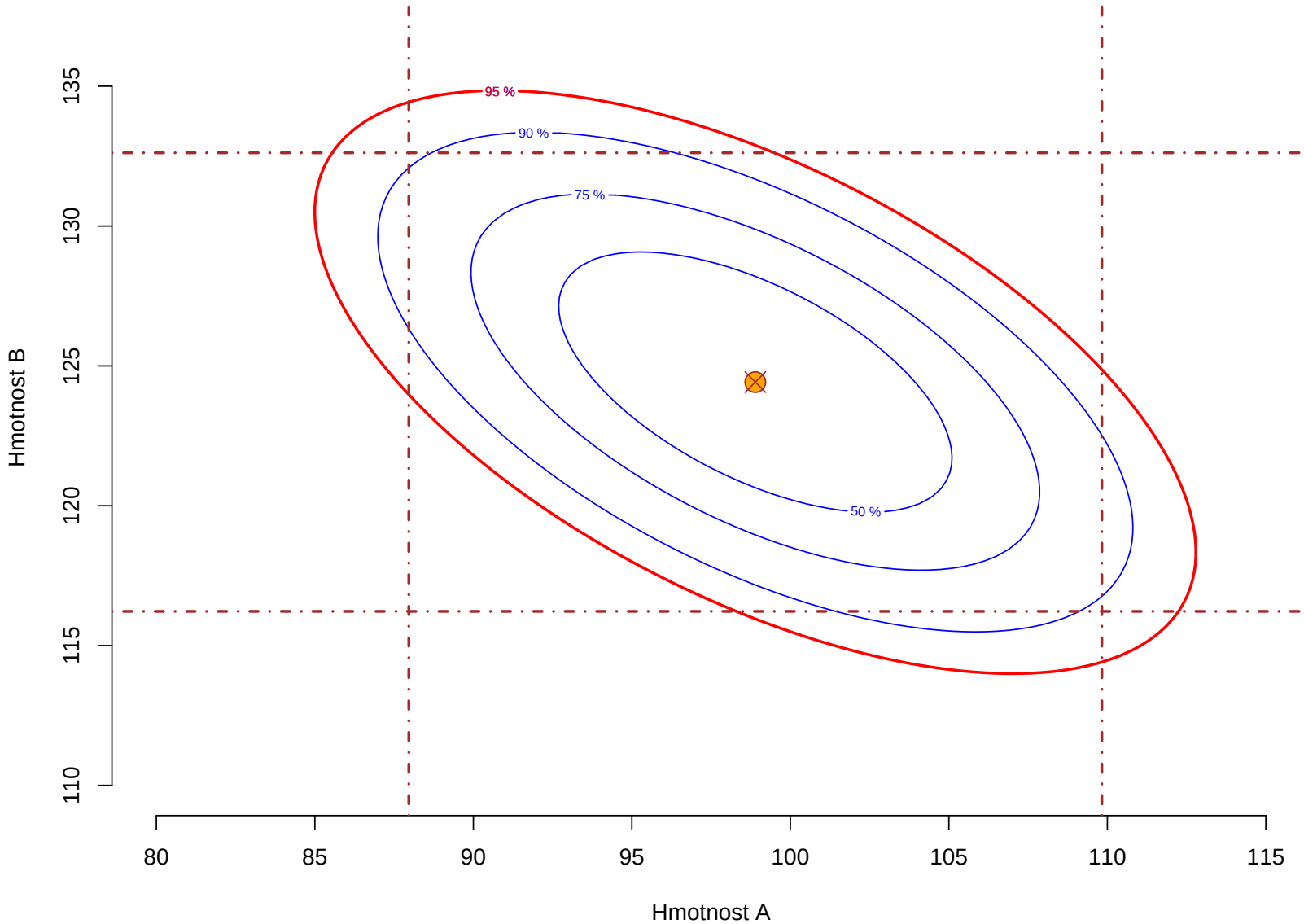
$\beta_2 : (116.22,\ 132.62)$

HPD věrohodnostní množina pro β

$$\frac{s^{-2}\left(\beta - \boldsymbol{b}\right)^{\top}\mathbb{X}^{\top}\mathbb{X}\left(\beta - \boldsymbol{b}\right)}{k} \leq F_{k,n-k}(1-\alpha)$$

HPD věrohodnostní množina pro β

Vážení lehkých objektů



Konjugovaný systém apriorních rozdělení

- $p(\beta, \tau) = p(\beta \mid \tau) \times p(\tau)$
- $p(\beta \mid \tau) \equiv \mathcal{N}_k(\beta_0, \tau^{-1}\mathbf{\Sigma}_0)$
- $p(\tau) \equiv \mathcal{G}(c_0, d_0)$
- β_0, $\mathbf{\Sigma}_0$, c_0, d_0 : pevné parametry apriorního rozdělení
  - Jaká je jejich interpretace?
  - Jakou apriorní informaci vyjadřují jejich konkrétní volby?

Snadné domácí cvičení ☺:

- Odvoďte $p(\beta, \tau \mid \boldsymbol{y})$, $p(\beta \mid \boldsymbol{y})$, $p(\tau \mid \boldsymbol{y})$ v případě, že je uvažováno výše uvedené apriorní rozdělení.

Užitečná maticová rovnost

(pro snadné domácí cvičení)

Nechť $\mathbb{A}$ a $\mathbb{B}$ jsou symetrické pozitivně semidefinitní matice dimenze $p \times p$ a alespoň jedna z nich nechť je pozitivně definitní. Nechť $\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b}$ jsou reálné vektory z $\mathbb{R}^p$. Potom

$$
\begin{aligned}
&(\boldsymbol{x}-\boldsymbol{a})^\top \mathbb{A}(\boldsymbol{x}-\boldsymbol{a}) + (\boldsymbol{x}-\boldsymbol{b})^\top \mathbb{B}(\boldsymbol{x}-\boldsymbol{b}) \\
&\quad = (\boldsymbol{x}-\boldsymbol{\mu})^\top \mathbb{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \\
&\qquad \underbrace{-(\mathbb{A}\boldsymbol{a}+\mathbb{B}\boldsymbol{b})^\top \mathbb{V}(\mathbb{A}\boldsymbol{a}+\mathbb{B}\boldsymbol{b}) + \boldsymbol{a}^\top \mathbb{A}\boldsymbol{a} + \boldsymbol{b}^\top \mathbb{B}\boldsymbol{b}}_{\text{nezávisí na } \boldsymbol{x}},
\end{aligned}
$$

kde

$$
\begin{aligned}
\mathbb{V} &= \left(\mathbb{A}+\mathbb{B}\right)^{-1} \\
\boldsymbol{\mu} &= \mathbb{V}\left(\mathbb{A}\boldsymbol{a} + \mathbb{B}\boldsymbol{b}\right)
\end{aligned}
$$

Konjugovaný systém apriorních rozdělení

Aposteriorní rozdělení

$$p(\boldsymbol{\beta}, \tau \,|\, \boldsymbol{y}) \propto \tau^{\frac{n}{2}+\frac{k}{2}+c_0-1} \exp\left\{-\tau\left(\frac{\mathrm{SS}_e}{2} + d_0\right)\right\}$$

$$\exp\left[-\frac{\tau}{2}\left\{(\boldsymbol{\beta}-\boldsymbol{b})^\top \mathbb{X}^\top \mathbb{X}(\boldsymbol{\beta}-\boldsymbol{b}) + (\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\}\right],$$

$$\tau \,|\, \boldsymbol{y} \sim \mathcal{G}\left(\frac{n}{2} + c_0,\ \frac{\mathrm{SS}_e(\boldsymbol{y})}{2} + d_0\right),$$

$$\boldsymbol{\beta} \,|\, \tau, \boldsymbol{y} \sim \mathcal{N}_k\left(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}), \tau^{-1}\,\widehat{\boldsymbol{\Sigma}}(\boldsymbol{y})\right), \qquad \widehat{\boldsymbol{\Sigma}}(\boldsymbol{y}) = \left(\mathbb{X}^\top\mathbb{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1},$$

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{y}) = \widehat{\boldsymbol{\Sigma}}(\boldsymbol{y})\left(\mathbb{X}^\top\mathbb{X}\,\boldsymbol{b}(\boldsymbol{y}) + \boldsymbol{\Sigma}_0^{-1}\,\boldsymbol{\beta}_0\right)$$

Interpretace parametrů apriorního rozdělení?

2

Aposteriorní inference založená na simulaci

Oddíl **2.1**

Úvod

Obecná formulace problému

- $\boldsymbol{\theta} = \left(\theta_1, \ldots, \theta_k\right)^\top$: vektor (bayesovských) parametrů
- Aposteriorní inference je založená na

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto L(\boldsymbol{\theta})\, p(\boldsymbol{\theta})$$

  - hustota vzhledem k σ-konečné míře λ na $(\mathbb{R}^k, \mathcal{B}^k)$,
  - v dalším budeme zkráceně namísto $p(\boldsymbol{\theta} \mid \boldsymbol{y}) d\lambda(\boldsymbol{\theta})$ psát $p(d\boldsymbol{\theta} \mid \boldsymbol{y})$.

Obecná formulace problému

- V rámci aposteriorní inference nás též pro nějaké měřitelné funkce t na parametrickém prostoru Θ zajímají zejména následující veličiny:
  - $\mathbb{E}_{p(d\theta \mid \boldsymbol{y})} t(\theta) = \int t(\theta)\, p(d\theta \mid \boldsymbol{y})$
  - kvantily $t(\theta)$ vzhledem k rozdělení $p(d\theta \mid \boldsymbol{y})$

    ⇛ medián, věrohodnostní intervaly,
  - aposteriorní rozdělení náhodné veličiny. $t(\theta)$
- Kromě triviálních případů vyžaduje integrování přes $p(d\theta \mid \boldsymbol{y})$.

Oddíl **2.2**

Monte Carlo integrace

Monte Carlo integrace

Snaha pro měřitelnou funkci $t : \mathbb{R}^k \longrightarrow \mathbb{R}$ numericky spočítat

$$\mathbb{E}_{p(d\boldsymbol{\theta} \mid \boldsymbol{y})} t(\boldsymbol{\theta}) = \int t(\boldsymbol{\theta})\, p(d\boldsymbol{\theta} \mid \boldsymbol{y}).$$

- Předpokládejme, že $\mathbb{E}_{p(d\boldsymbol{\theta} \mid \boldsymbol{y})} t(\boldsymbol{\theta})$ existuje konečné.
- Předpokládejme dále, že jsme schopni získat náhodný výběr

$$\mathcal{S}_M = \left\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}\right\} \qquad \text{z rozdělení } p(d\boldsymbol{\theta} \mid \boldsymbol{y}).$$

Potom $$\int t(\boldsymbol{\theta})\, p(d\boldsymbol{\theta} \mid \boldsymbol{y}) \approx \frac{1}{M} \sum_{m=1}^{M} t\left(\boldsymbol{\theta}^{(m)}\right) = \widehat{\mathbb{E}}_{p(d\boldsymbol{\theta} \mid \boldsymbol{y})} t(\boldsymbol{\theta}) := \widehat{t}_M.$$

- Proč a kdy to funguje?

Monte Carlo integrace

- Jestliže $\int \left|t(\boldsymbol{\theta})\right| p(d\boldsymbol{\theta}\,|\boldsymbol{y}) < \infty$, potom

$$\widehat{t}_M = \frac{1}{M}\sum_{m=1}^{M} t\left(\boldsymbol{\theta}^{(m)}\right) \xrightarrow{\text{a.s.}} \int t(\boldsymbol{\theta})\, p(d\boldsymbol{\theta}\,|\boldsymbol{y}) \quad \text{pro } M \to \infty$$

  (silný zákon velkých čísel).

- Dále, jestliže $\int \left\{t(\boldsymbol{\theta})\right\}^2 p(d\boldsymbol{\theta}\,|\boldsymbol{y}) < \infty$, potom

$$v_M := \text{var}\left\{\frac{1}{M}\sum_{m=1}^{M} t\left(\boldsymbol{\theta}^{(m)}\right)\right\} = \frac{1}{M}\text{var}_{p(d\boldsymbol{\theta}\,|\,\boldsymbol{y})}\, t(\boldsymbol{\theta})$$

$$= \frac{1}{M}\int \left\{t(\boldsymbol{\theta}) - \mathbb{E}_{p(\boldsymbol{\theta}\,|\,\boldsymbol{y})}\right\}^2 p(d\boldsymbol{\theta}\,|\,\boldsymbol{y})$$

  ohodnocuje rychlost konvergence a tudíž též přesnost aproximace při použití konečného M.

- $\sqrt{v_M}$= Monte Carlo chyba (*Monte Carlo Error*).

Monte Carlo integrace

Odhad Monte Carlo chyby

- Opět s využitím silného zákona velkých čísel se snadno ukáže (obdobně jako při důkazu konzistence výběrového rozptylu), že

$$\frac{1}{M-1}\sum_{m=1}^{M}\left\{t\left(\boldsymbol{\theta}^{(m)}\right)-\widehat{t}_M\right\}^2 \xrightarrow{\text{a.s.}} \text{var}_{p(d\boldsymbol{\theta}\mid \boldsymbol{y})}\, t(\boldsymbol{\theta}), \quad \text{pro } M\to\infty.$$

- Rozptyl v_M lze tudíž odhadnout pomocí

$$\widehat{v}_M=\frac{1}{M(M-1)}\sum_{m=1}^{M}\left\{t\left(\boldsymbol{\theta}^{(m)}\right)-\widehat{t}_M\right\}^2.$$

- $\sqrt{\widehat{v}_M}$ =odhad Monte Carlo chyby.

Rozdělení aproximace

- Za položených předpokladů platí též centrální limitní věta, která spolu s Cramérovou-Sluckého větou vede k

$$\frac{\widehat{t}_M - \mathbb{E}_{p(d\theta \mid \boldsymbol{y})} t(\boldsymbol{\theta})}{\sqrt{\widehat{v}_M}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

  - ohodnocení konvergence Monte Carlo integrace,
  - konstrukce konfidenčních mezí pro aproximaci.

Poznámka

- Potřebný náhodný výběr $\mathcal{S}_M$ obvykle na náš popud generuje počítač.
  - Velikost výběru M lze poměrně snadno ovlivnit (víceméně závisí pouze na výkonu počítače a době, po kterou jsme ochotni čekat na výsledek).
  - S patřičně velkým M lze i reálně dosáhnout požadované přesnosti při odhadu $\mathbb{E}_{p(d\theta \mid \boldsymbol{y})} t(\boldsymbol{\theta})$.

Oddíl **2.3**

Důležité speciální případy

Důležité speciální případy

- Předpokládejme, že $\boldsymbol{\theta} = \left(\theta_1, \ldots, \theta_k\right)^\top$ a $\mathcal{S}_M = \left\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}\right\}$ je náhodný výběr z $p(d\boldsymbol{\theta} \mid \boldsymbol{y})$

Důležitý speciální případ 1

$$t(\boldsymbol{\theta}) = \theta_j \quad (j = 1, \ldots, k)$$

- $\dfrac{1}{M} \sum_{m=1}^{M} \theta_j^{(m)}$ je MC odhadem pro $\mathbb{E}_{p(d\boldsymbol{\theta} \mid \boldsymbol{y})} \theta_j$.

Důležitý speciální případ 2

$$t(\boldsymbol{\theta}) = \mathbb{I}_{[\theta_j \le x]}(\theta_j) \text{ pro } x \in \mathbb{R} \quad (j = 1, \ldots, k)$$

- $\frac{1}{M} \sum_{m=1}^{M} \mathbb{I}_{[\theta_j \le x]}(\theta_j^{(m)}) = \frac{\#[\theta_j^{(m)} \le x]}{M}$ je MC odhadem pro

$$\mathbb{E}_{p(d\boldsymbol{\theta} \mid \boldsymbol{y})} \mathbb{I}_{[\theta_j \le x]}(\theta_j) = \mathrm{P}(\theta_j \le x \mid \boldsymbol{Y} = \boldsymbol{y}).$$

- MC odhad marginální aposteriorní distribuční funkce θ_j.
- Libovolný odhad hustoty (histogram, jádrový odhad, ...) založený na $\theta_j^{(1)}, \ldots, \theta_j^{(M)}$ je MC odhadem marginální aposteriorní hustoty $p(\theta_j \mid \boldsymbol{y})$,
  ➠ odsud lze mj. odhadnout HPD věrohodnostní intervaly.
- Výběrové kvantily založené na $\theta_j^{(1)}, \ldots, \theta_j^{(M)}$ odhadují kvantily marginálního aposteriorního rozdělení $p(d\theta_j \mid \boldsymbol{y})$
  ➠ odhady ET věrohodnostních intervalů.

Důležitý speciální případ 3

Veličina hlavního zájmu je $r(\boldsymbol{\theta})$, kde $r : \mathbb{R}^k \to \mathbb{R}$ je měřitelná funkce a $t(\boldsymbol{\theta}) = \mathbb{I}_{[r(\boldsymbol{\theta}) \leq x]}(\boldsymbol{\theta})$ pro $x \in \mathbb{R}$

- $\dfrac{1}{M} \sum_{m=1}^{M} \mathbb{I}_{[r(\boldsymbol{\theta}) \leq x]}(\boldsymbol{\theta}^{(m)}) = \dfrac{\#\left[r(\boldsymbol{\theta}^{(m)}) \leq x\right]}{M}$ je MC odhadem pro

$$\mathbb{E}_{p(d\boldsymbol{\theta} \mid \boldsymbol{y})} \mathbb{I}_{[r(\boldsymbol{\theta}) \leq x]}(\boldsymbol{\theta}) = \mathrm{P}\left\{ r(\boldsymbol{\theta}) \leq x \,\middle|\, \boldsymbol{Y} = \boldsymbol{y} \right\}.$$

- MC odhad marginální aposteriorní distribuční funkce $r(\boldsymbol{\theta})$.
- Libovolný odhad hustoty (histogram, jádrový odhad, ...) založený na $r(\boldsymbol{\theta}^{(1)}), \ldots, r(\boldsymbol{\theta}^{(M)})$ je MC odhadem marginální aposteriorní hustoty $p\left(r(\boldsymbol{\theta}) \,\middle|\, \boldsymbol{y}\right)$
  - ➠ odsud lze mj. odhadnout HPD věrohodnostní intervaly.
- Výběrové kvantily založené na $r(\boldsymbol{\theta}^{(1)}), \ldots, r(\boldsymbol{\theta}^{(M)})$ odhadují kvantily marginálního aposteriorního rozdělení $p\left(dr(\boldsymbol{\theta}) \,\middle|\, \boldsymbol{y}\right)$
  - ➠ odhady ET věrohodnostních intervalů.

Důležitý speciální případ 3, pokračování

Veličina hlavního zájmu je $r(\boldsymbol{\theta})$, kde $r : \mathbb{R}^k \to \mathbb{R}$ je měřitelná funkce

- Zvolíme-li $t(\boldsymbol{\theta}) = r(\boldsymbol{\theta})$, dostaneme MC odhad pro $\mathbb{E}_{p(d\boldsymbol{\theta} \mid \boldsymbol{y})} r(\boldsymbol{\theta})$.
- Při počítání MC odhadů veličin založených na aposteriorním rozdělení $r(\boldsymbol{\theta})$ již nemusíme ani integrovat, ani používat větu o transformaci.
- $r(\boldsymbol{\theta})$ může být prakticky libovolně "složitou" funkcí $\boldsymbol{\theta}$ a stále je vše upočítatelné i prakticky.

Drobná komplikace

Kde vzít onen náhodný výběr

$$\mathcal{S}_M = \left\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}\right\} \quad \text{z rozdělení } p(d\boldsymbol{\theta} \mid \boldsymbol{y})?$$

Oddíl **2.4**

Simulace ze zadaného rozdělení

Simulace ze zadaného rozdělení

- Vše dosud řečené o MC integraci lze samozřejmě použít k numerickému počítání integrálů vzhledem k libovolnému rozdělení.
  - Nejen aposteriornímu v rámci nějakého bayesovského modelu.
- V dalším se budeme zabývat problémem simulace ze zadaného rozdělení, které budeme reprezentovat hustotou $f(\boldsymbol{\theta})$ vzhledem k nějaké σ-konečné míře λ na $(\mathbb{R}^k, \mathcal{B}^k)$ a distribuční funkcí $F(\boldsymbol{\theta})$.
  - Nadále budeme zkráceně psát $f(\boldsymbol{\theta})d\lambda(\boldsymbol{\theta}) = f(d\boldsymbol{\theta})$.
  - V rámci bayesovských metod budeme obvykle používat s $f(d\boldsymbol{\theta}) = p(d\boldsymbol{\theta} \mid \boldsymbol{y})$.

- Nejprve se budeme zabývat případem jednorozměrného rozdělení $f(d\theta)$.

Věta 2.1 .

Nechť U je náhodná veličina s rovnoměrným rozdělením na intervalu $(0, 1)$. Nechť θ je náhodná veličina s distribuční funkcí F. Nechť pro $0 < u < 1$ je $F^{-1}(u) = \inf\{\theta : F(\theta) \geq u\}$ kvantilová funkce. Potom má náhodná veličina

$$Z = F^{-1}(U)$$

stejné rozdělení jako náhodná veličina θ, tj. Z má rozdělení s distribuční funkcí F.

Důkaz.

- Triviální, jestliže navíc předpokládáme, že F je spojitá a rostoucí na nosiči, tj. na množině Θ, pro kterou platí

$$\mathsf{P}(\theta \in \Theta) = 1, \quad \forall \widetilde{\Theta} \subset \Theta, \mathsf{P}(\theta \in \Theta \setminus \widetilde{\Theta}) > 0 \Rightarrow \mathsf{P}(\theta \in \widetilde{\Theta}) < 1.$$

- O něco složitější, jestliže spojitost a monotonii F na nosiči nepředpokládáme.

❑

Univerzální algoritmus pro simulování z jednorozměrného rozdělení

Generování z rozdělení θ s distribuční funkcí $F(\theta)$:

1. Vygeneruj $U \sim \mathcal{U}(0, 1)$.
2. Polož $\theta = F^{-1}(U)$.

- Efektivní, jestliže jsme schopni efektivně počítat hodnoty F^{-1}.

Specifické algoritmy pro simulování z jednorozměrných rozdělení

- Pro mnohá běžná rozdělení existují efektivnější metody založené obvykle na transformacích, z nichž mnohé byly propočítány v rámci Cvičení k předmětu Matematická statistika 1 (NMSA331).
- Příklad 1:
  $U_1, U_2 \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0, 1)$, potom pro

$$\theta_1 = \mu + \sigma \cos(2\pi U_1)\sqrt{-2\log U_2},$$

$$\theta_2 = \mu + \sigma \sin(2\pi U_1)\sqrt{-2\log U_2}$$

  platí $\theta_1, \theta_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$.
- Příklad 2:
  $U \sim \mathcal{U}(0, 1)$, potom pro

$$\theta = -\frac{1}{\lambda}\log(U)$$

  platí $\theta \sim \mathcal{E}xp(\lambda)$.

Více o simulování z rozdělení s plně specifikovanou hustotou (nejenom jednorozměrnou)

Podrobnější obrázek

- Přednáška Simulační metody (NMST535) vyučovaná v zimním semestru.
- Luc Devroye. *Non-Uniform Random Variate Generation.* New York: Springer-Verlag, 1986.

Praktické aplikace

- Statistické balíky využívají těchto metod ke generování z většiny běžných rozdělení.
  -  funkce `runif`, `rnorm`, `rexp`, ...
  - "Běžný" uživatel se nemusí příliš trápit s generováním z běžných (≡ "pojmenovaných") rozdělení a prostě použije zabudované možnosti svého software.

Přímá simulace založená na podmiňování

- Předpokládejme, že $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_k^\top\right)^\top$ a v rozkladu

$$\begin{aligned} f(d\boldsymbol{\theta}) &= f(d\boldsymbol{\theta}_1, \ldots, d\boldsymbol{\theta}_k) \\ &= f(d\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k)\, f(d\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_3, \ldots, \boldsymbol{\theta}_k) \cdots f(d\boldsymbol{\theta}_{k-1} \mid d\boldsymbol{\theta}_k)\, f(d\boldsymbol{\theta}_k) \end{aligned}$$

  jsme schopni snadno generovat ze všech (podmíněných) rozdělení $f(d\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k)$, $f(d\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_3, \ldots, \boldsymbol{\theta}_k)$, $\ldots$, $f(d\boldsymbol{\theta}_{k-1} \mid \boldsymbol{\theta}_k)$, $f(d\boldsymbol{\theta}_k)$.

- Algoritmus pro generování ze sdruženého rozdělení $f(d\boldsymbol{\theta})$ je potom následující:
  1. Vygeneruj $\boldsymbol{\theta}_k$ z $f(d\boldsymbol{\theta}_k)$.
  2. Vygeneruj $\boldsymbol{\theta}_{k-1}$ z $f(d\boldsymbol{\theta}_{k-1} \mid \boldsymbol{\theta}_k)$.
  3. $\vdots$
  4. Vygeneruj $\boldsymbol{\theta}_2$ z $f(d\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_3, \ldots, \boldsymbol{\theta}_k)$.
  5. Vygeneruj $\boldsymbol{\theta}_1$ z $f(d\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \ldots, \boldsymbol{\theta}_k)$.

Příklad: Lineární model s neinformativní apriorním rozdělením

Lineární model

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathbb{X} : \text{pevná matice } n \times k,$$
$$\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

- Parametry: $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top, \tau^\top\right)^\top$, kde $\tau = \sigma^{-2} > 0$.
- Věrohodnost: $\boldsymbol{Y} \sim \mathcal{N}_n(\mathbb{X}\boldsymbol{\beta}, \tau^{-1}\mathbf{I}_n)$.
- Neinformativní apriorní rozdělení:

$$p(\boldsymbol{\beta}) \propto 1, \qquad \boldsymbol{\beta} \in \mathbb{R}^k,$$
$$p(\tau) \propto \frac{1}{\tau}, \qquad \tau > 0.$$

Příklad: Lineární model s neinformativní apriorním rozdělením

Aposteriorní rozdělení

- Označme: $\boldsymbol{b} = \left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \mathbb{X}^\top \boldsymbol{y}$,

  $\mathrm{SS}_e = \left\| \boldsymbol{y} - \mathbb{X}\boldsymbol{b} \right\|^2$.

- Bylo odvozeno:

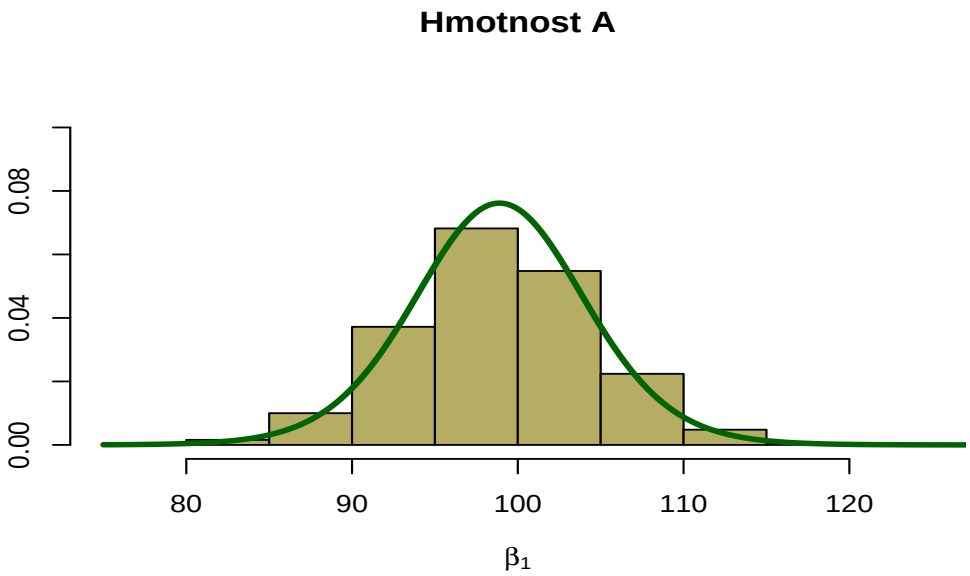
$$p(\boldsymbol{\beta}, \tau \mid \boldsymbol{y}) = p(\boldsymbol{\beta} \mid \tau, \boldsymbol{y}) \times p(\tau \mid \boldsymbol{y}),$$

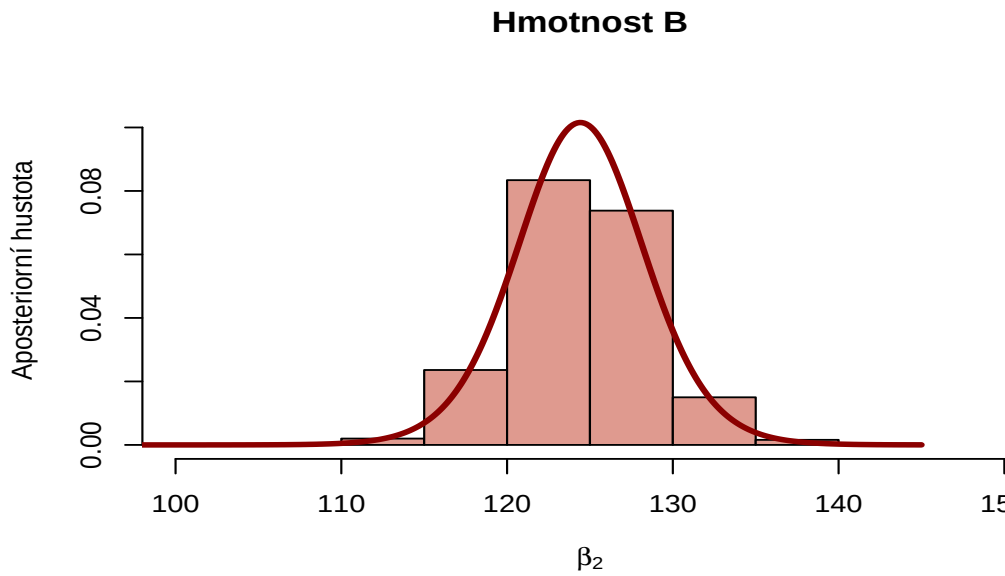
kde $p(\tau \mid \boldsymbol{y}) \sim \mathcal{G}\left(\frac{n-k}{2}, \frac{\mathrm{SS}_e}{2}\right)$,

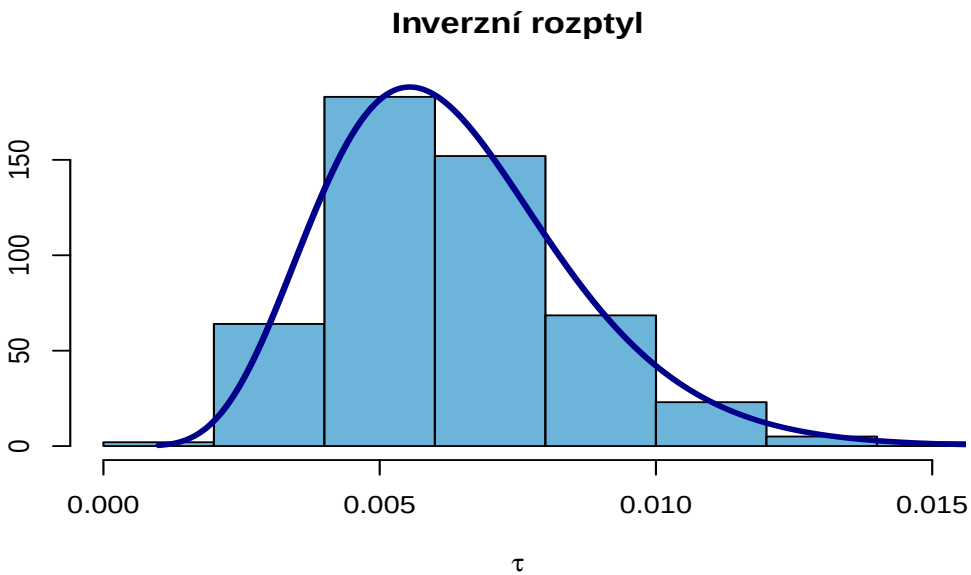
$$p(\boldsymbol{\beta} \mid \tau, \boldsymbol{y}) \sim \mathcal{N}_k\left(\boldsymbol{b}, \tau^{-1}\left(\mathbb{X}^\top \mathbb{X}\right)^{-1}\right).$$

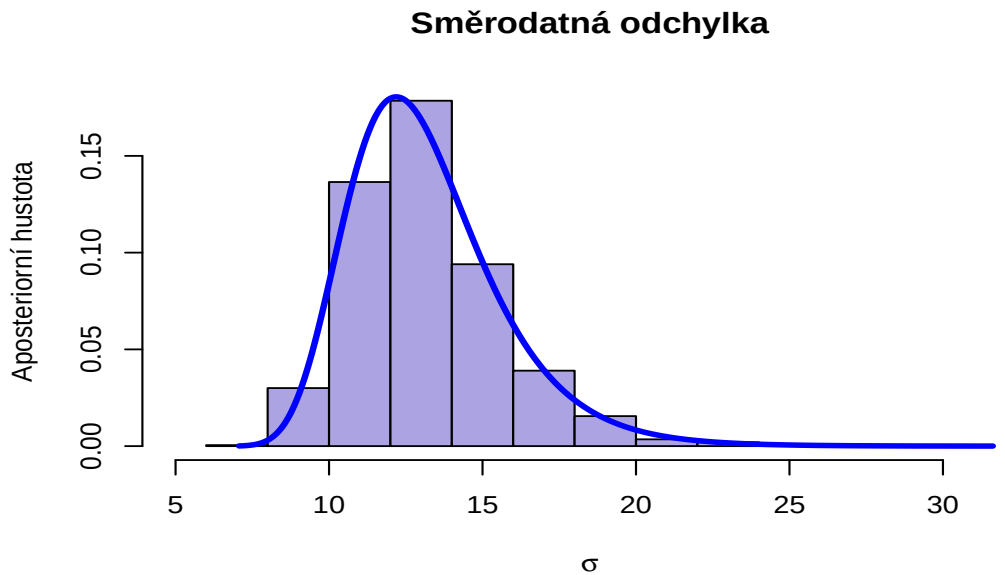
Příklad: Vážení lehkých objektů

Marginální aposteriorní hustoty (M=1 000)



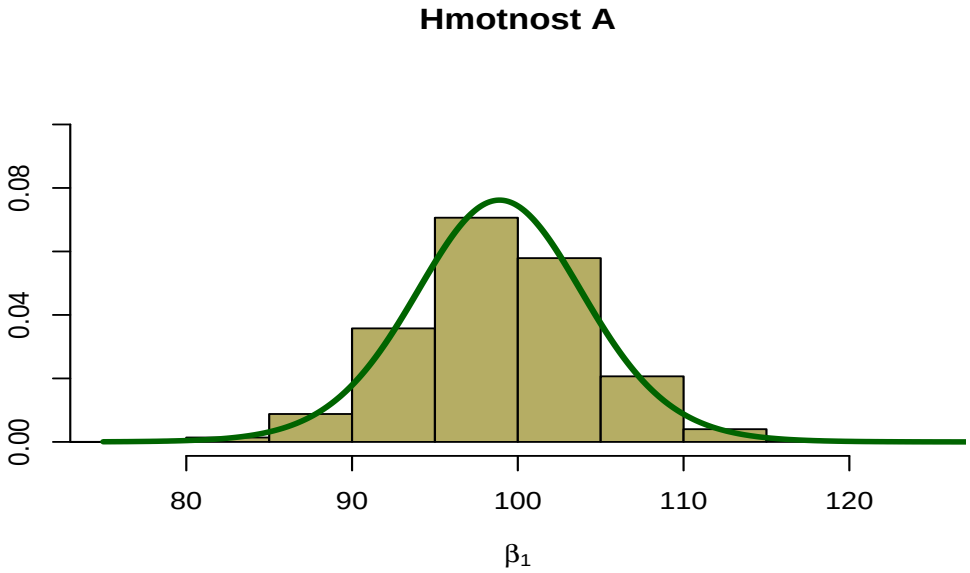


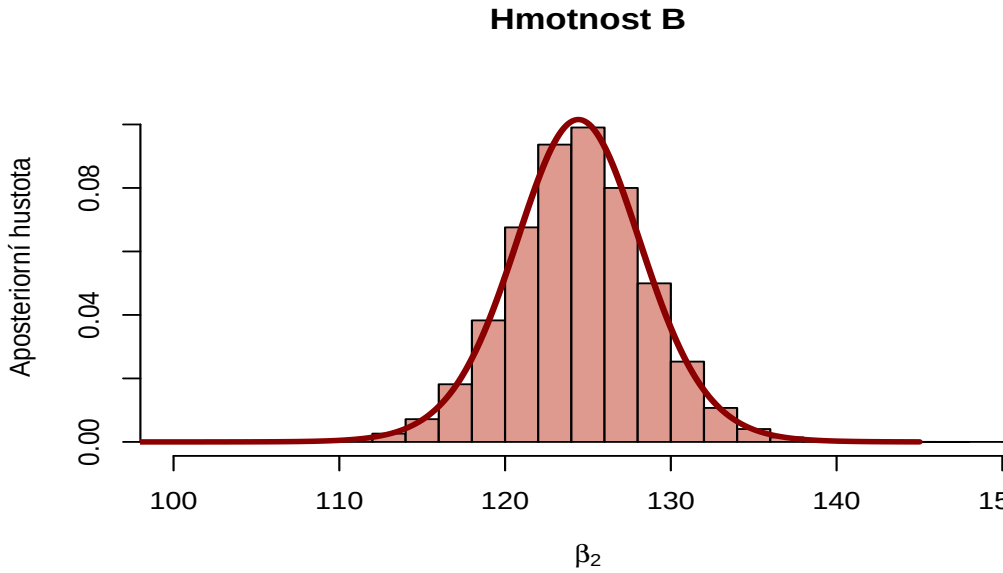


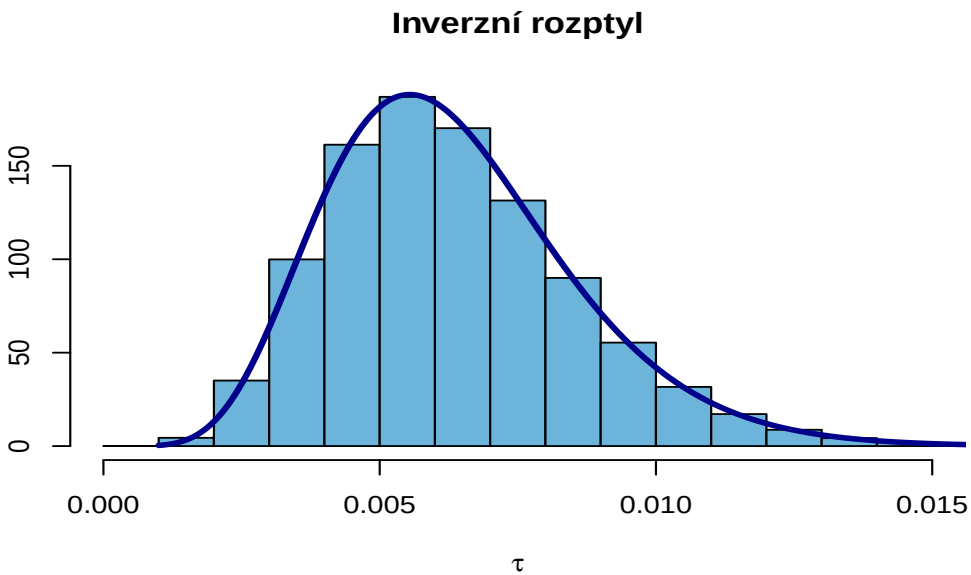


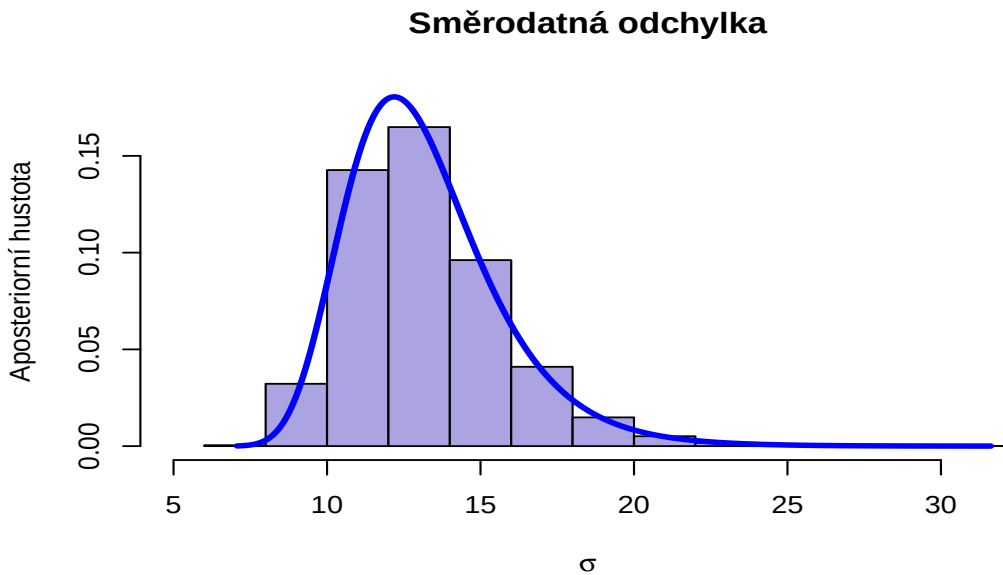
Příklad: Vážení lehkých objektů

Marginální aposteriorní hustoty (M=1 000 000)



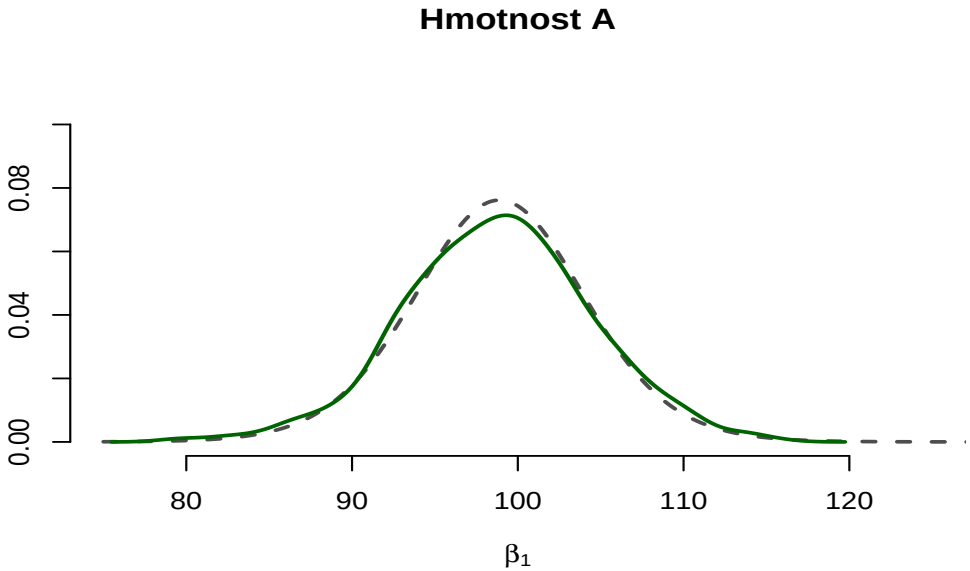


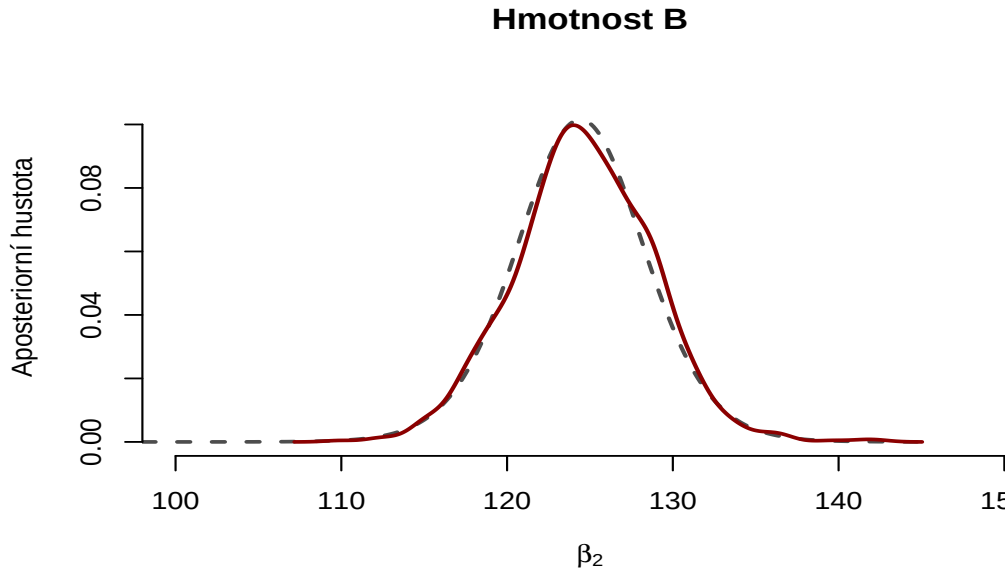


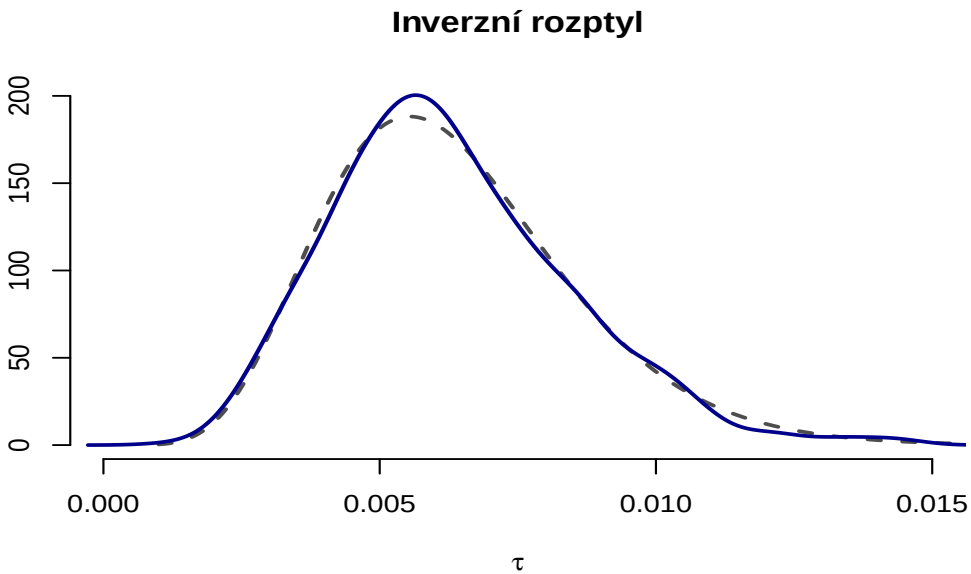


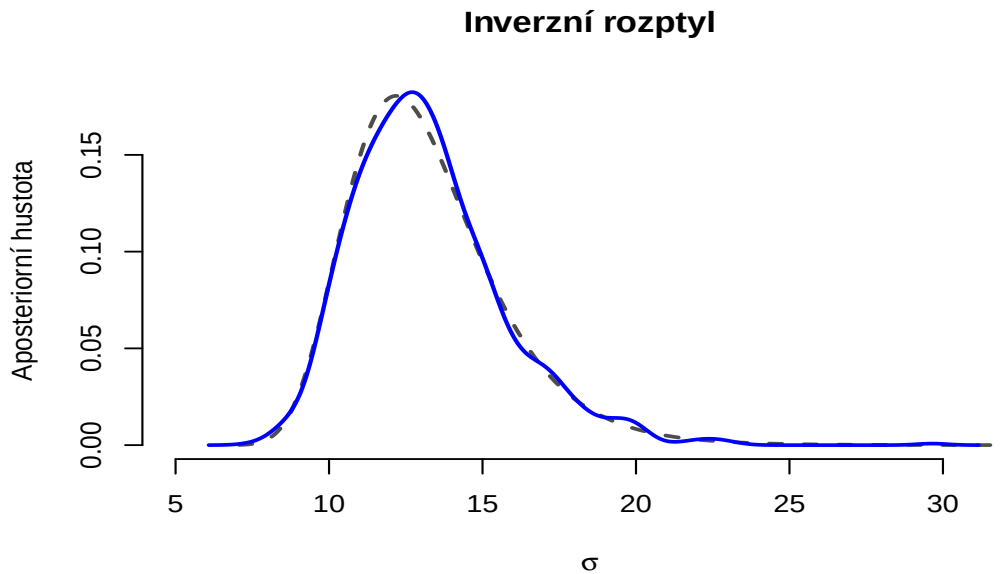
Příklad: Vážení lehkých objektů

Marginální aposteriorní hustoty (M=1 000)



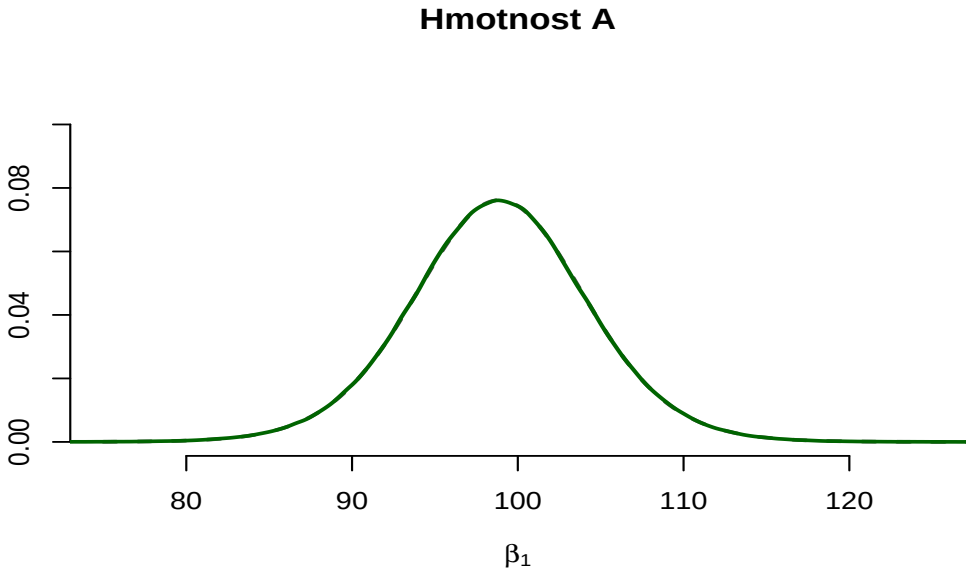


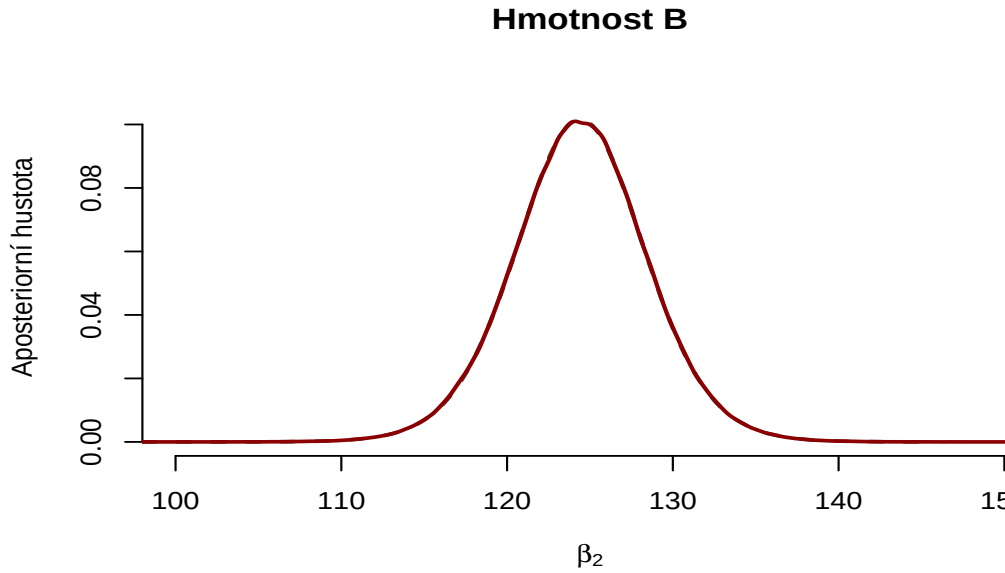


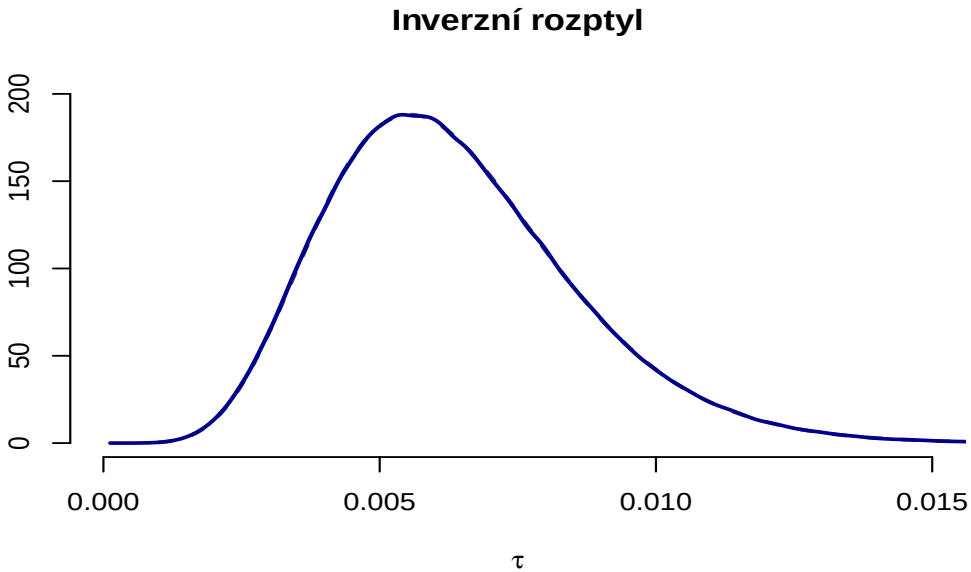


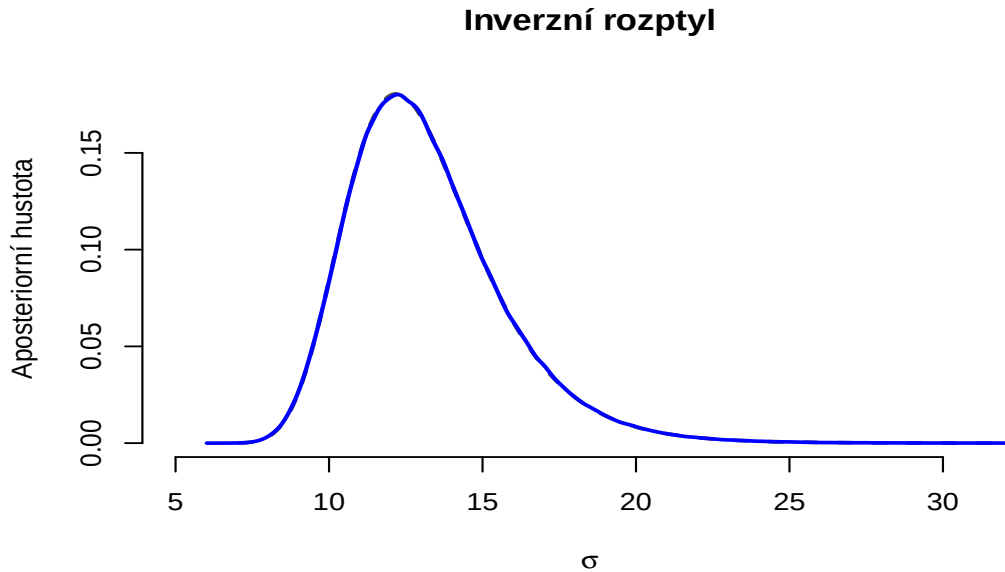
Příklad: Vážení lehkých objektů

Marginální aposteriorní hustoty (M=1 000 000)



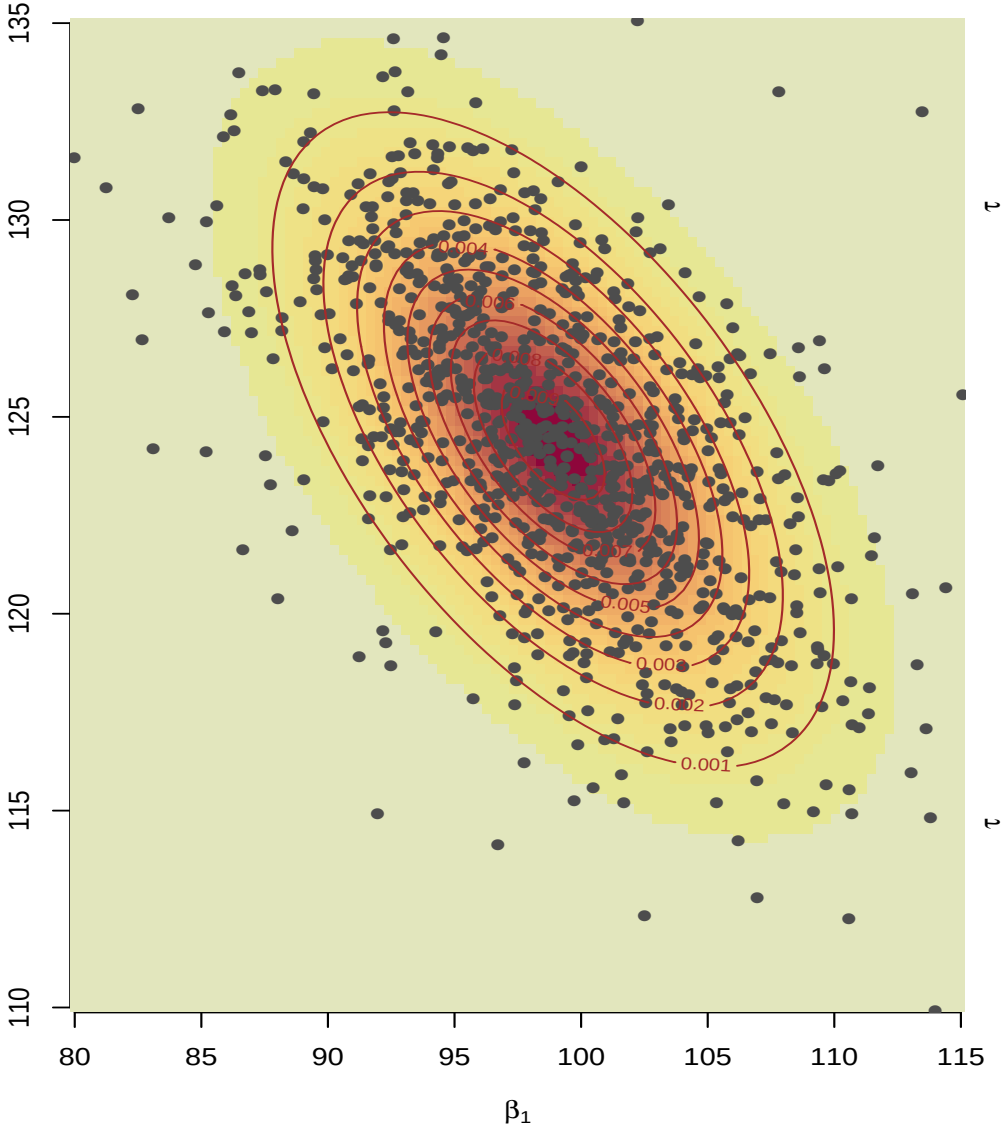


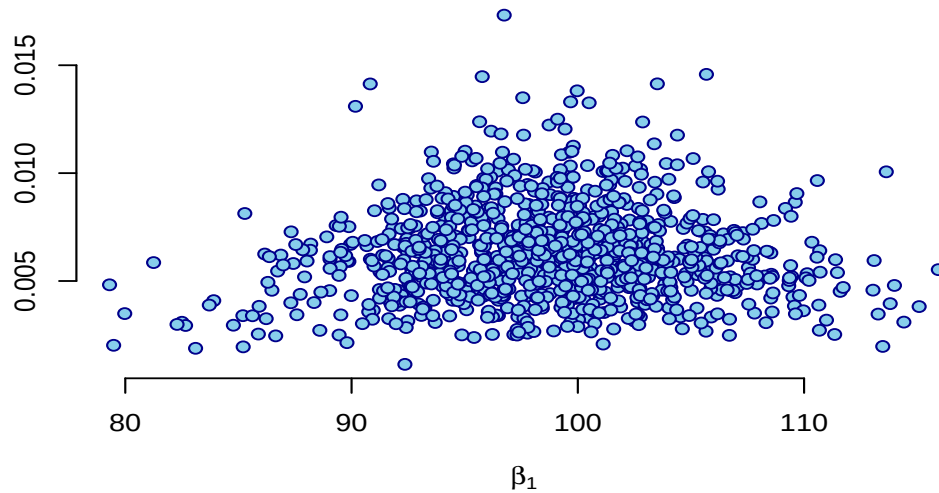


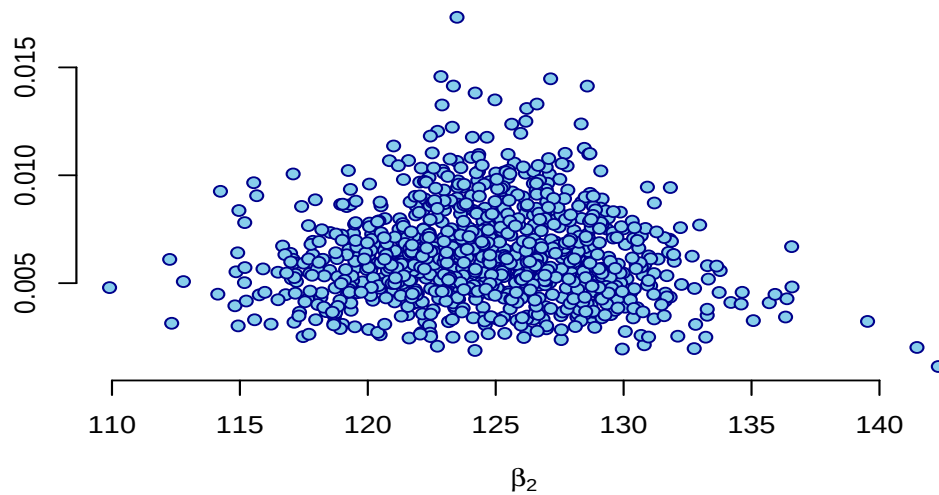


Příklad: Vážení lehkých objektů

Sdružené výběry z aposteriorního rozdělení (M=1 000)

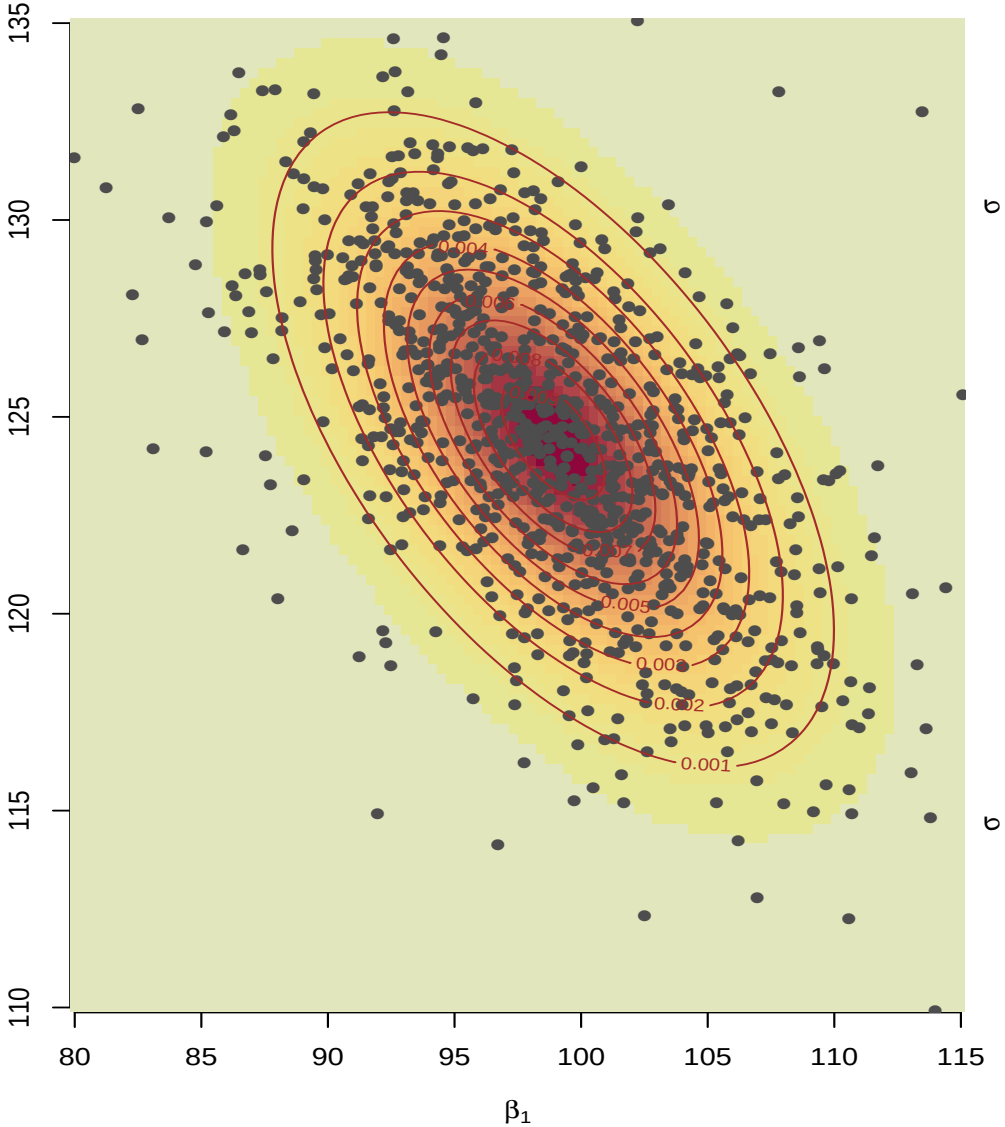


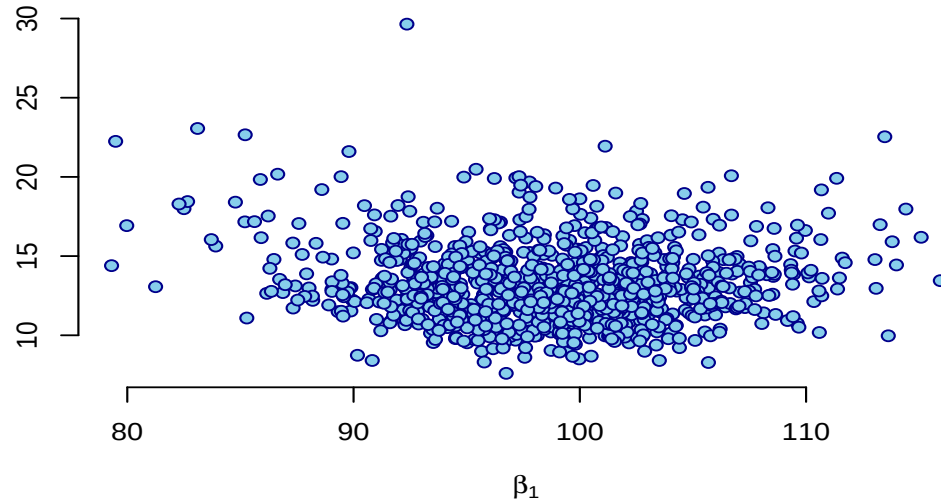


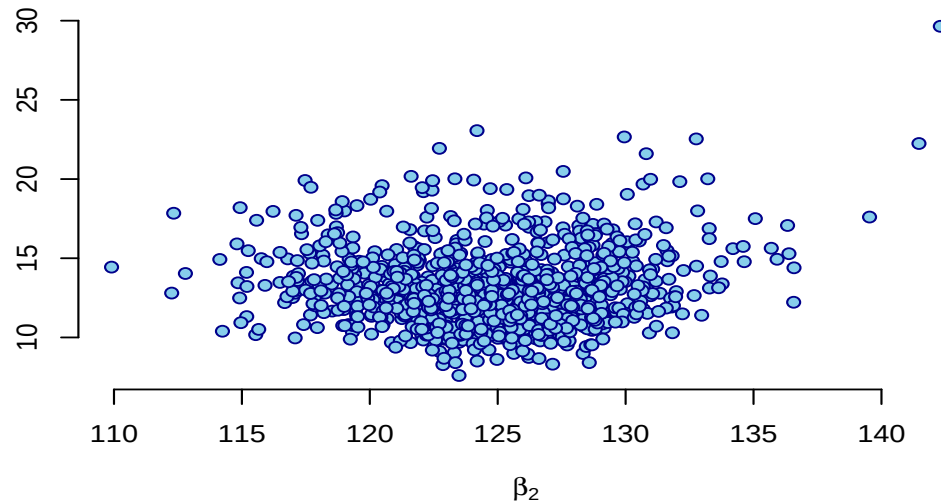


Příklad: Vážení lehkých objektů

Sdružené výběry z aposteriorního rozdělení (M=1 000)







Příklad: Vážení lehkých objektů

Aposteriorní inference pro β (M=1 000)

	β_1	β_2
Aposter. střední hodnota	98,8947	124,4211
MC odhad	98,7944	124,6197
MC chyba	0,1804	0,1312
Aposter. medián	98,8947	124,4211
MC odhad	98,7813	124,5673
95% ET věr. interval	(87,9641; 109,8253)	(116,2231; 132,6190)
MC odhad	(86,9761; 110,0815)	(116,7594; 132,6802)
95% HPD věr. interval	(87,9641; 109,8253)	(116,2231; 132,6190)
MC odhad	(87,9245; 110,7076)	(116,4892; 132,2210)

Příklad: Vážení lehkých objektů

Aposteriorní inference pro β (M=1 000 000)

	β_1	β_2
Aposter. střední hodnota	98,8947	124,4211
MC odhad	98,8874	124,4192
MC chyba	0,0055	0,0041
Aposter. medián	98,8947	124,4211
MC odhad	98,8849	124,4196
95% ET věr. interval	(87,9641; 109,8253)	(116,2231; 132,6190)
MC odhad	(87,9680; 109,8184)	(116,2239; 132,6113)
95% HPD věr. interval	(87,9641; 109,8253)	(116,2231; 132,6190)
MC odhad	(88,0765; 109,9202)	(116,1496; 132,5329)

Příklad: Vážení lehkých objektů

Aposteriorní inference pro τ a σ (M=1 000)

	τ	σ
Aposter. střední hodnota	0,00633	?
MC odhad	0,00629	13,207
MC chyba	0,0000689	0,0776
Aposter. medián	0,00607	?
MC odhad	0,00601	12,904
95% ET věr. interval	(0,00273; 0,01142)	?
MC odhad	(0,00272; 0,01097)	(9,547; 19,183)
95% HPD věr. interval	?	?
MC odhad	(0,00248; 0,01039)	(8,987; 18,186)

Příklad: Vážení lehkých objektů

Aposteriorní inference pro τ a σ (M=1 000 000)

	τ	σ
Aposter. střední hodnota	0,00633	?
MC odhad	0,00633	13,198
MC chyba	0,0000022	0,0025
Aposter. medián	0,00607	?
MC odhad	0,00607	12,838
95% ET věr. interval	(0,00273; 0,01142)	?
MC odhad	(0,00274; 0,01142)	(9,356; 19,116)
95% HPD věr. interval	?	?
MC odhad	(0,00237; 0,01081)	(8,872; 18,224)

3

Hierarchické modely

Oddíl **3.1**

Hierarchické apriorní rozdělení

Apriorní rozdělení

- Volba apriorního rozdělení může značným způsobem ovlivnit formu rozdělení aposteriorního.
- Nebezpečí zneužití bayesovské statistiky.
- Většina "bayesovských" aplikací z posledních cca 25 let
  - není motivována snahou využívat jakoukoliv apriorní informaci,
  - hlavní motivace: navržený model nelze (ani numericky) odhadnout frekventisticky (typicky pomocí maximální věrohodnosti), nicméně lze ho odhadnout pomocí simulací bayesovsky,
  - neexistuje žádná skutečná apriorní informace.

Apriorní rozdělení

- Pouze málokdy je apriorní informace dostatečně bohatá na to, abychom mohli zvolené apriorní rozdělení považovat za přesně a bez jakékoliv chyby definované.
- Potřeba vhodným způsobem vyjádřit nejistotu při volbě apriorního rozdělení.

➠ Bayesovský model s hierarchicky specifikovaným apriorním rozdělením

- rozklad apriorního rozdělení do několika úrovní podmíněných rozdělení,
- nejistota na libovolné úrovni je vyjádřena apriorním rozdělením v další úrovni.

Bayesovský model s hierarchicky specifikovaným apriorním rozdělením

Definice 3.1 Bayesovský model s hierarchicky specifikovaným apriorním rozdělením.

Bayesovský model s hierarchicky specifikovaným apriorním rozdělením je statistický model s věrohodností $L(\psi) = p(\boldsymbol{y} \mid \psi)$ a apriorním rozdělením $p(\psi)$, kde $p(\psi)$ je rozloženo na podmíněná rozdělení

$$p_0(\psi \mid \zeta_1),\ p_1(\zeta_1 \mid \zeta_2),\ \ldots,\ p_{m-1}(\zeta_{m-1} \mid \zeta_m)$$

a marginální rozdělení $p_m(\zeta_m)$ tak, že

$$p(\psi) = \int_{Z_1 \times \cdots \times Z_m} p_0(\psi \mid \zeta_1)\, p_1(\zeta_1 \mid \zeta_2) \cdots p_{m-1}(\zeta_{m-1} \mid \zeta_m)\, p_m(\zeta_m)\, d\zeta_1 \cdots d\zeta_m.$$

Parametry obsažené v ζ_i se nazývají hyperparametry i-té úrovně $(1 \leq i \leq m)$.

Příklad: Lineární model

Model

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathbb{X} : \text{pevná matice } n \times k,$$
$$\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

- Parametry: $\boldsymbol{\psi} = \left(\boldsymbol{\beta}^\top, \tau\right)^\top$, kde $\tau = \sigma^{-2} > 0$
- Věrohodnost: $L(\boldsymbol{\psi}) = p(\boldsymbol{y} \mid \boldsymbol{\psi}) \equiv \mathcal{N}_n(\mathbb{X}\boldsymbol{\beta}, \tau^{-1}\mathbf{I}_n)$
- Konjugované apriorní rozdělení:

$$p(\boldsymbol{\beta}, \tau) = p(\boldsymbol{\beta} \mid \tau) \times p(\tau)$$
$$p(\boldsymbol{\beta} \mid \tau) \equiv \mathcal{N}_k(\boldsymbol{\beta}_0, \tau^{-1}\boldsymbol{\Sigma}_0)$$
$$p(\tau) \equiv \mathcal{G}(c_0, d_0)$$

  - $\boldsymbol{\beta}_0$, $\boldsymbol{\Sigma}_0$, c_0, d_0 : pevné (hyper)parametry.

Příklad: Lineární model

- Volba parametrů gama apriorního rozdělení pro τ (resp. samotná volba gama rozdělení) mívá dosti značný vliv na výsledné aposteriorní rozdělení.
- Nepovažujme c_0 a/nebo d_0 za pevné konstanty, ale umožněme náhodnost při jejich výběru.

  ➠ hierarchický model
  - například:

$$p(\tau \mid d_0) \equiv \mathcal{G}(c_0, d_0)$$
$$p(d_0) \equiv \mathcal{G}(g_0, h_0)$$

  - c_0 : pevný (hyper)parametr
  - d_0 : náhodný hyperparametr 1. úrovně
  - g_0, h_0 : pevné (hyper)parametry (2. úrovně)

Oddíl **3.2**

Hierarchicky specifikovaná věrohodnost

Hierarchicky specifikovaná věrohodnost

- Hierarchická specifikace (též v nebayesovském kontextu) je často přirozeným způsobem jak zkonstruovat realistický pravděpodobnostní model pro popis reálné situace.
- Oblasti využití, které jste už potkali
  - Data získaná stratifikovaným výběrem či jinak shlukovaná *(grouped data)*.
  - Longitudinální (biostatistika), resp. panelová (ekonometrie) data.

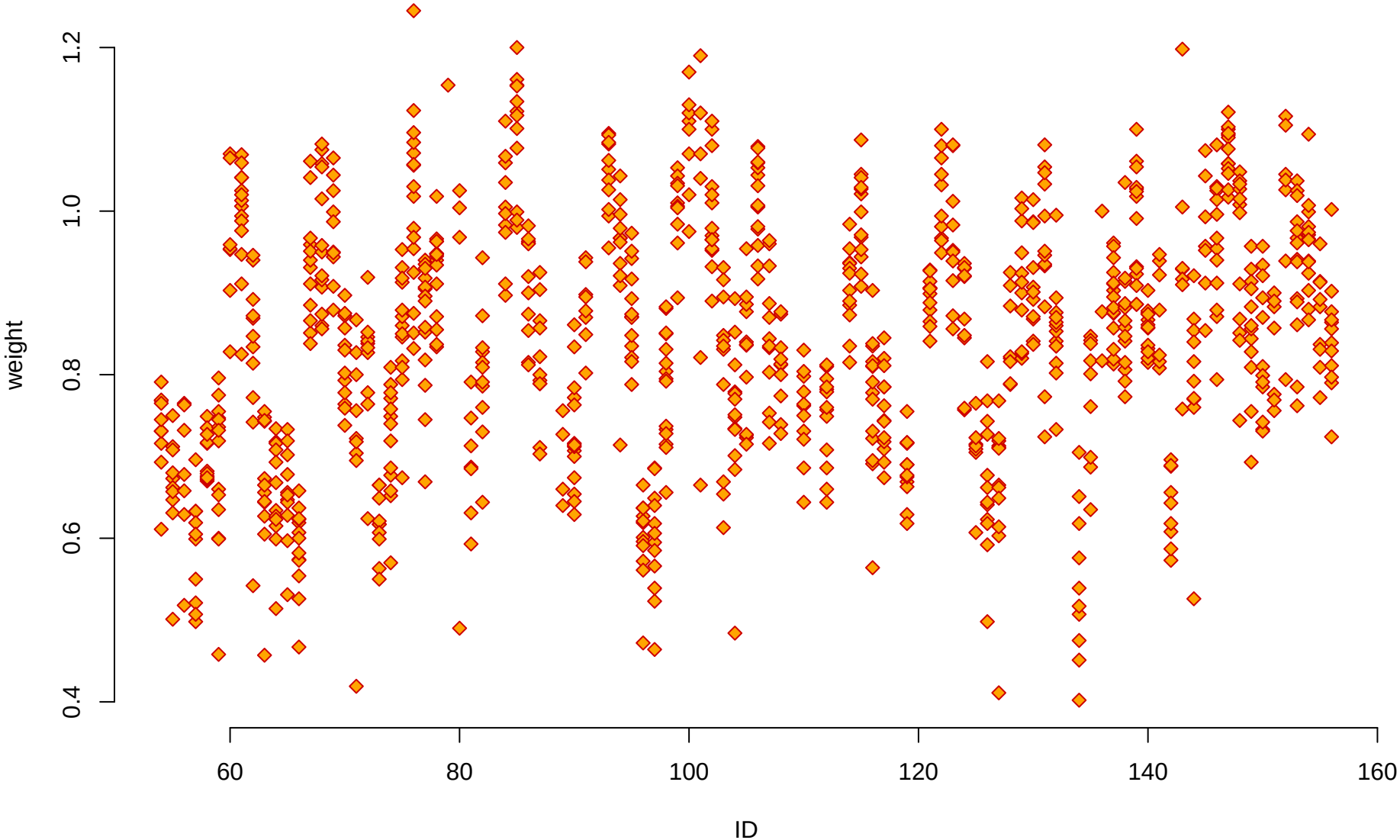
Příklad: NTP TER84073 pokus na myších

- Data z *National Toxicology Program*
- 94 těhotným myším byla podána v předem určených momentech určená množství etylenglykolu (EG)
- V 17. dnu těhotenství byly myši usmrceny a následně byla zaznamenána hmotnost zárodků
- Pravděpodobnostní reprezentace dat:

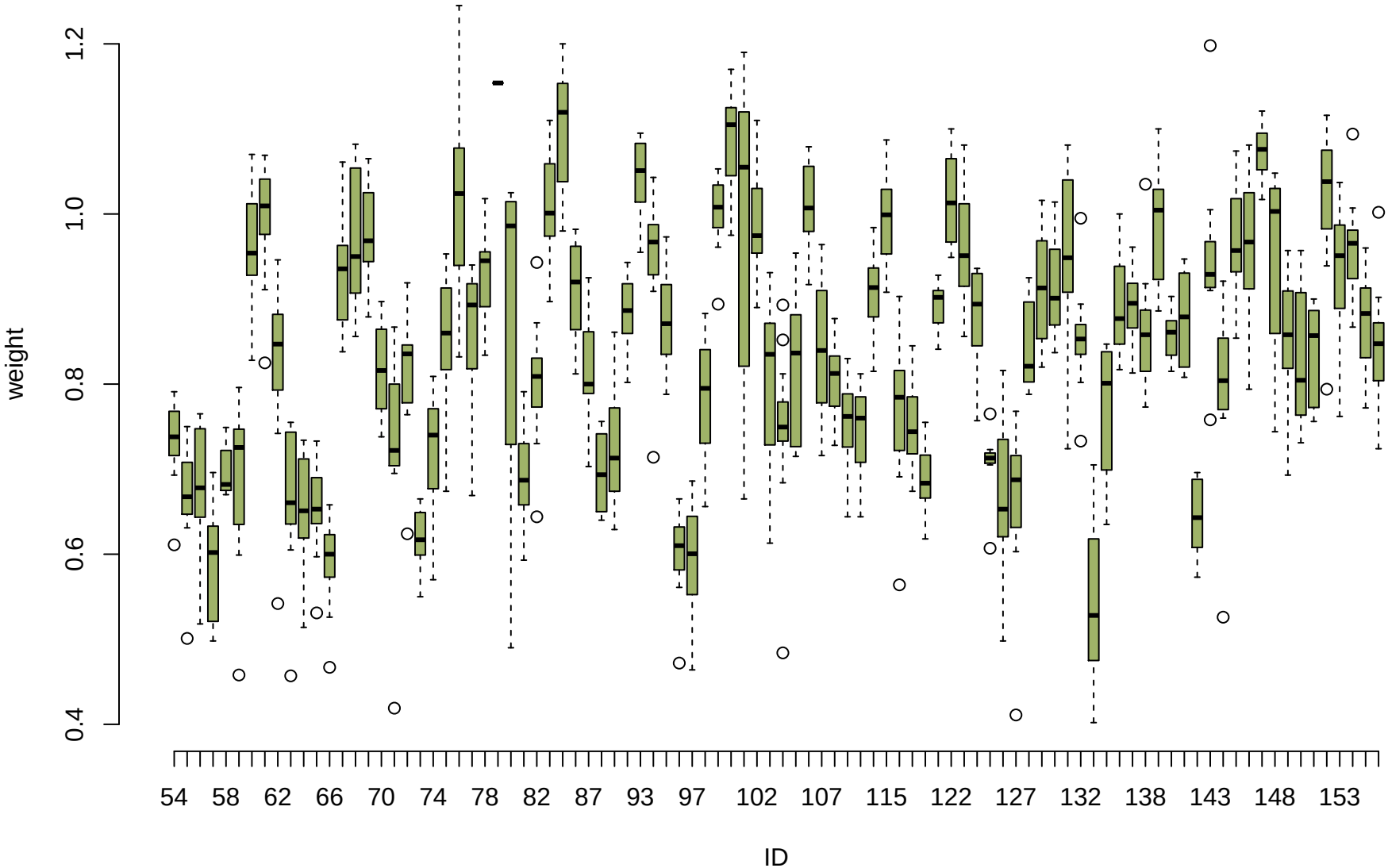
  $Y_{i,j}$ = hmotnost j-tého zárodku i-té myši, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$
- Primární cíl:

  Odhad a inference pro $\mu = \mathbb{E} Y_{i,j}$

Příklad: NTP TER84073 pokus na myších



Příklad: NTP TER84073 pokus na myších



Příklad: NTP TER84073 pokus na myších

- Možný pravděpodobnostní model:

$$Y_{i,j} \sim \mathcal{N}(\mu, \sigma^2).$$

- Je ospravedlnitelné předpokládat, že $Y_{1,1}, \ldots, Y_{N,n_N}$ jsou nezávislé?

Příklad: NTP TER84073 pokus na myších

- Realističtější pravděpodobnostní model:

$$Y_{i,j} \mid b_i \sim \mathcal{N}(b_i, \sigma^2),$$
$$b_i \sim \mathcal{N}(\mu, d^2).$$

- Pro každé i již lze předpokládat (podmíněnou) nezávislost $Y_{i,1} \mid b_i, \ldots, Y_{i,n_i} \mid b_i$.
- Lze též předpokládat nezávislost $b_1, \ldots, b_N$.
- b_i : střední hmotnost zárodku i-té myši.
- σ^2 : rozptyl hmotnosti zárodků u jednotlivé myši
  ➠ vnitroskupinová variabilita.
- μ : střední hmotnost zárodku v celé populaci

$$\mathbb{E}\, Y_{i,j} = \mathbb{E}\big\{\mathbb{E}(Y_{i,j} \mid b_i)\big\} = \mathbb{E} b_i = \mu.$$

Příklad: NTP TER84073 pokus na myších

- Realističtější pravděpodobnostní model:

$$Y_{i,j} \mid b_i \sim \mathcal{N}(b_i,\, \sigma^2),$$
$$b_i \sim \mathcal{N}(\mu,\, d^2).$$

- Modelujeme též jistým způsobem korelaci mezi zárodky jedné myši:

$$\mathrm{cov}(Y_{i,j},\, Y_{i,k}) = \cdots = d^2,$$
$$\mathrm{var}(Y_{i,j}) = \cdots = \sigma^2 + d^2.$$

➠ $\mathrm{cor}(Y_{i,j},\, Y_{i,k}) = \dfrac{d^2}{\sigma^2 + d^2}$

➠ vnitroskupinová korelace (*intraclass correlation*)

Delší další povídání

Hierarchické modely

Obecná poznámka

- Též v jiných modelech lze často rozlišit tři typy parametrů (v bayesovském smyslu):
  - skrytá data
    - ➠ v dalším budeme obvykle značit $\boldsymbol{\xi}$
  - "čisté" parametry $\equiv$ parametry též ve frekventistickém pojetí
    - ➠ v dalším budeme obvykle značit $\boldsymbol{\psi}$
  - náhodné hyperparametry
    - ➠ v dalším budeme obvykle značit $\boldsymbol{\zeta}$
- Parametry pro bayesovský model jsou potom $\boldsymbol{\theta} = \left(\boldsymbol{\xi}^\top, \boldsymbol{\psi}^\top, \boldsymbol{\zeta}^\top\right)^\top$.

Obecná poznámka, pokračování

- Sdružené apriorní rozdělení je zadáno rozkladem

$$p(\theta) = p(\xi, \psi, \zeta) = p(\xi \mid \psi)\, p(\psi \mid \zeta)\, p(\zeta)$$

  - $p(\xi \mid \psi)$: strukturální část apriorního rozdělení

    ➠ plyne z uvažovaného pravděpdobnostního modelu použitého pro popis situace
  - $p(\psi \mid \zeta)\, p(\zeta)$: “skutečné” apriorní rozdělení

“ANOVA” hierarchický model

Konkrétní apriorní rozdělení (jedna z možností)

- Např. konjugované apriorní rozdělení s dodatečnými hyperparametry, abychom se ochránili od nadměrného ovlivnění aposteriorního rozdělení rozdělením apriorním:

$$
\begin{aligned}
& p(\tau,\, \mu,\, q,\, b_0,\, d_0) \\
&= \underbrace{p(\mu \mid q)}_{\mathcal{N}(\mu_0,\, k_0^{-1}\, q^{-1})}\ \underbrace{p(q \mid b_0)}_{\mathcal{G}(a_0,\, b_0)}\ \underbrace{p(b_0)}_{\mathcal{G}(p_0,\, r_0)}\ \underbrace{p(\tau \mid d_0)}_{\mathcal{G}(c_0,\, d_0)}\ \underbrace{p(d_0)}_{\mathcal{G}(g_0,\, h_0)}
\end{aligned}
$$

- $b_0,\, d_0$: náhodné hyperparametry, tj. $\zeta = (b_0,\, d_0)^\top$
- $\mu_0,\, k_0,\, b_0,\, p_0,\, r_0,\, d_0,\, g_0,\, h_0$: pevné hyperparametry
- Graficky lze zpřehlednit pomocí DAGu (*directed acyclic graph*)
  - Kolečka: náhodné uzly
  - Čtverečky: nenáhodné uzly
  - Vyjádření (podmíněných) (ne)závislostí

“ANOVA” hierarchický model

Apriorní rozdělení (jedna z možností)

$$
\begin{aligned}
&p(\tau, \mu, q, b_0, d_0) \\
&= \underbrace{p(\mu \mid q)}_{\mathcal{N}(\mu_0,\, k_0^{-1}\, q^{-1})}\ \underbrace{p(q \mid b_0)}_{\mathcal{G}(a_0,\, b_0)}\ \underbrace{p(b_0)}_{\mathcal{G}(p_0,\, r_0)}\ \underbrace{p(\tau \mid d_0)}_{\mathcal{G}(c_0,\, d_0)}\ \underbrace{p(d_0)}_{\mathcal{G}(g_0,\, h_0)}
\end{aligned}
$$

Nenáhodné (pevné) (hyper)parametry:

- $\mu_0,\, k_0$
- $a_0,\, p_0,\, r_0$
- $c_0,\, g_0,\, h_0$

➠ Jak je volit?

“ANOVA” hierarchický model

Volba nenáhodných (hyper)parametrů

- Není-li k dispozici žádná rozumná apriorní informace, je snaha volit nenáhodné (hyper)parametry tak, aby výsledné apriorní rozdělení bylo co nejméně informativní *(weakly informative)*.
  - $p(\psi \mid \boldsymbol{y}) \propto L_F(\psi)\, p(\psi)$,
  - $p(\psi)$ představuje “nová” pozorování ve věrohodnosti.
- Snaha volit apriorní rozdělení tak, aby vliv těchto “nových” umělých pozorování na věrohodnost byl co možná nejmenší.
  - Snaha, aby co možná nejvíce platilo $p(\psi) \propto 1$ (viz též první část semestru).
  - Stačí, aby platilo relativně k $L_F(\psi)$.

➠ Konkrétní volba pevných hyperparametrů je často (částečně) motivována pozorovanými daty.

“ANOVA” hierarchický model

(Částečně) datově motivovaná volba pevných hyperparametrů

- μ_0 : apriorní střední hodnota pro $\mathbb{E}Y_{i,j} = \mathbb{E}b_i = \mu$
  - $\mu_0 \approx \overline{y}$
- k_0 : ovlivňuje apriorní inverzní rozptyl (přesnost) pro μ
  - k_0 blízké 0
- a_0, c_0, p_0, g_0 : “stupně volnosti” gama rozdělení
  - obvykle mezi 0 a 1
- r_0, h_0 : “rate” parametr gama rozdělení v poslední hierarchické úrovni
  - obvykle se volí blízké 0

“ANOVA” hierarchický model

Aposteriorní rozdělení

- $p(\theta \mid \boldsymbol{y})$ odvodíme standardním způsobem

  ➠ Jak?

- V tomto konkrétním případě lze při volbě konjugovaného systému ještě vše odvodit analyticky.

- Pro mnohé jiné volby apriorních rozdělení již analyticky odvodit nelze (problém spočítat integrál ve jmenovateli Bayesovy věty).

- Inference pomocí aposteriorního rozdělení obvykle založená na počítačové simulaci (Monte Carlo metody).

4

(Zobecněné) lineární smíšené modely

Oddíl **4.1**

Lineární smíšený model

Normální lineární smíšený model

- Normální lineární smíšený model

$$\boldsymbol{Y}_i = \mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \qquad i = 1, \ldots, N$$

$$\boldsymbol{b}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_q(\boldsymbol{\mu}, \mathbb{D})$$

$$\boldsymbol{\varepsilon}_i \sim \mathcal{N}_{n_i}(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$$

  - $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N, \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N$ vzájemně nezávislé

Normální lineární smíšený model

Hierarchický zápis

Normální lineární smíšený model zapsaný hierarchicky

$$\boldsymbol{Y}_i \mid \boldsymbol{b}_i \sim \mathcal{N}_{n_i}(\mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\boldsymbol{b}_i,\, \boldsymbol{\Sigma}_i), \qquad i = 1, \ldots, N$$

$$\boldsymbol{b}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_q(\boldsymbol{\mu},\, \mathbb{D})$$

- $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$ vzájemně nezávislé
- $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N$ vzájemně nezávislé

Oddíl **4.2**

Příklady

Příklad: NTP TER84073 pokus na myších

Normální lineární smíšený model

- Pravděpodobnostní reprezentace dat:

  $Y_{i,j} =$ hmotnost j-tého zárodku i-té myši, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$

- Normální LMM:

$$
\begin{aligned}
Y_{i,j} \mid b_i &\sim \mathcal{N}(b_i, \sigma^2) \\
\boldsymbol{Y}_i \mid b_i &\sim \mathcal{N}_{n_i}(b_i \mathbf{1}, \boldsymbol{\Sigma}_i) \\
b_i &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, d^2)
\end{aligned}
$$

- V obecném značení

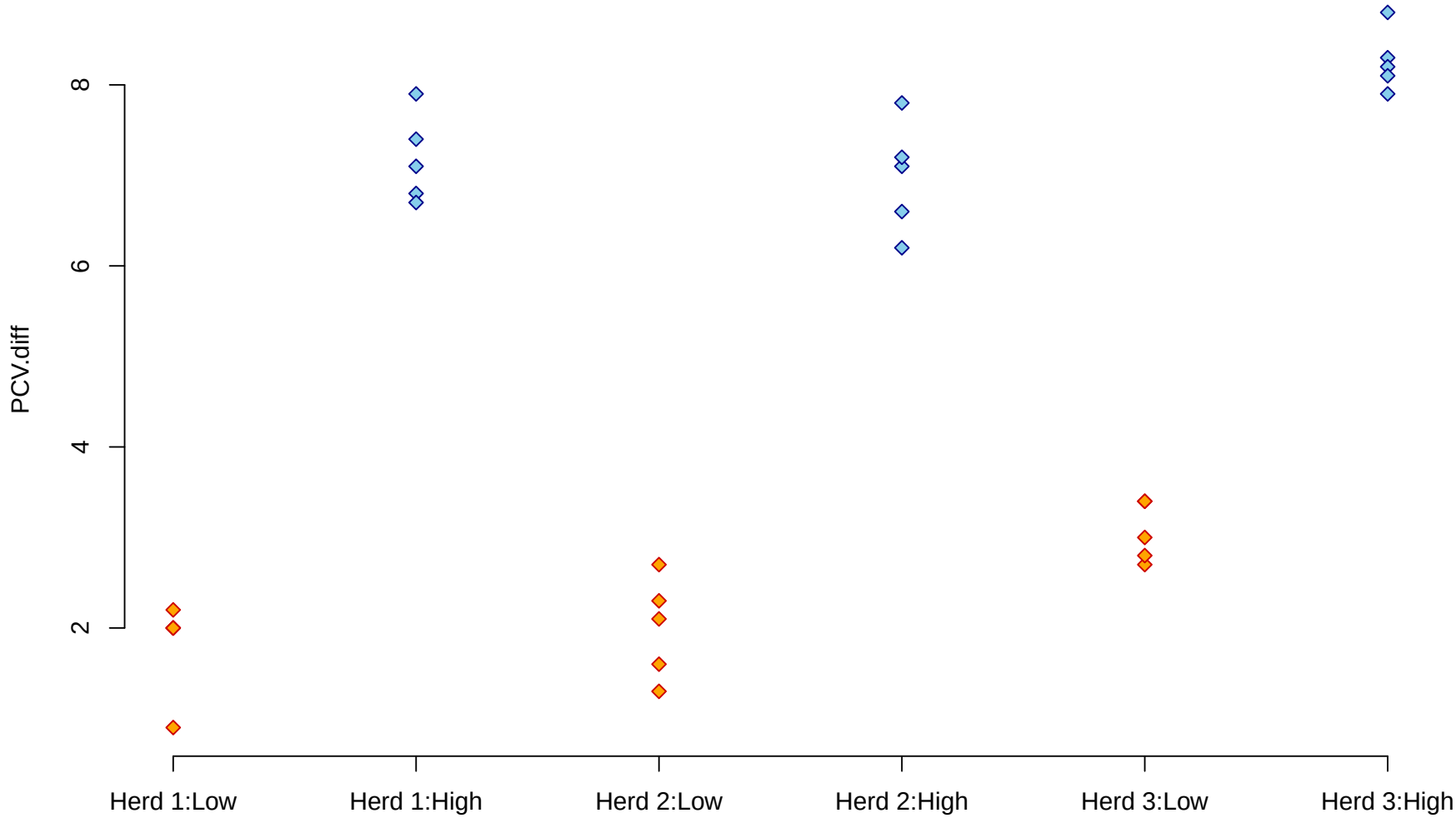
$$
\mathbb{X}_i \text{ není}, \quad \mathbb{Z}_i = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}_{n_i}, \quad \mathbb{D} = d^2
$$

Příklad: Vliv dávky Berenilu na léčbu trypanosomosis

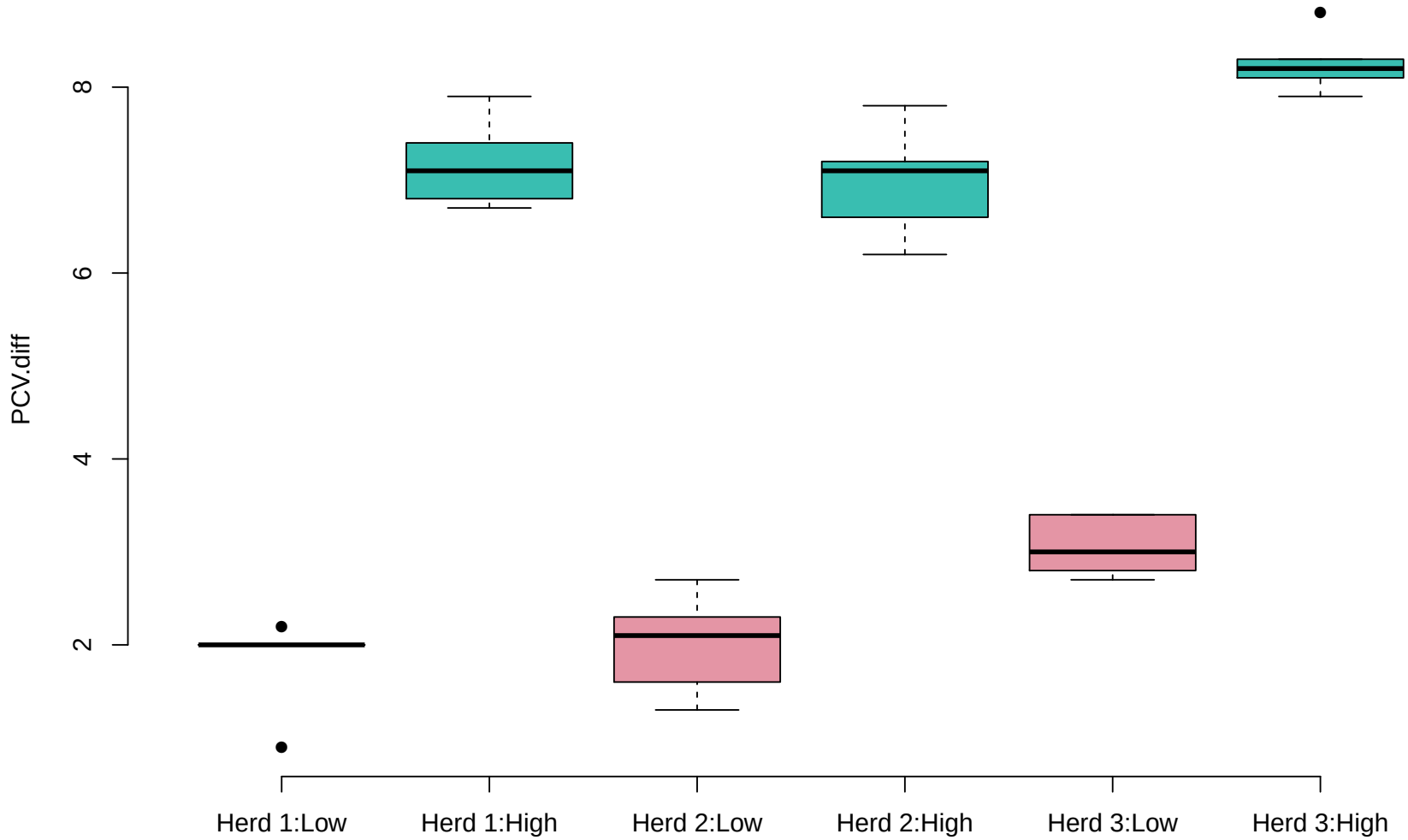
- Experiment mající za úkol vyhodnotit vliv dávky léku Berenil na úspěšnost léčby trypanosomosis u krav
- Úspěšnost léčby je měřena pomocí přírůstku PCV (packed cell volume) po určité době podávání léku
- Zajímá nás, zda existuje rozdíl v úspěšnosti léčby nízkou či vysokou dávkou Berenilu
- Experiment proběhl na kravách ze 3 stád (z různých farem)
- Každé stádo bylo rozděleno náhodně na dvě (přibližně stejné) části, přičemž krávy z jedné části stáda byly ošetřovány nízkou dávkou Berenilu, krávy z druhé části stáda byly ošetřovány vysokou dávkou Berenilu
- Pravděpodobnostní reprezentace dat:

$$Y_{i,j} = \text{PCV přírůstek u } j\text{-té krávy } i\text{-tého stáda}$$

Příklad: Vliv dávky Berenilu na léčbu trypanosomosis



Příklad: Vliv dávky Berenilu na léčbu trypanosomosis



Příklad: Vliv dávky Berenilu na léčbu trypanosomosis

Lineární smíšený model

- Jestliže lze předpokládat, že dávka má stejný účinek ve všech stádech, lze výsledek experimentu reprezentovat následujícím LMM

$$Y_{i,j} = b_i + x_{i,j}\,\beta + \varepsilon_{i,j}$$

- $x_{i,j} \begin{cases} 0, & \text{jestliže } (i,j)\text{-tá kráva léčena nízkou dávkou} \\ 1, & \text{jestliže } (i,j)\text{-tá kráva léčena vysokou dávkou} \end{cases}$
- b_i : střední účinek nízké dávky v i-tém stádu
  - náhodný efekt stáda
  - nepředpokládáme, že je stejné u všech stád
  - předpokládáme, že stáda jsou náhodně vybrána z populace stád
  - $\mathbb{E}b_i = \mu =$ střední účinek nízké dávky v populaci
- β : rozdíl mezi účinky vysoké a nízké dávky
  - konstantní (pevný efekt)
  - předpokládáme, že je stejné u všech stád
  - populační rozdíl mezi účinky vysoké a nízké dávky
- $\varepsilon_{i,j}$: náhodná odchylka (i,j)-té krávy od střední hodnoty její části (vysoká/nízká dávka) jejího stáda

Příklad: Vliv dávky Berenilu na léčbu trypanosomosis

Lineární smíšený model

$$Y_{i,j} = b_i + x_{i,j}\,\beta + \varepsilon_{i,j}$$

$$\mathbb{E} b_i = \mu$$

- Parametr hlavního zájmu: β

  $\equiv$ populační rozdíl v efektu léčby vysokou a nízkou dávkou

Příklad: Vliv dávky Berenilu na léčbu trypanosomosis

Normální lineární smíšený model

- Normální LMM:

$$
\begin{aligned}
Y_{i,j} \mid b_i &\sim \mathcal{N}(b_i + x_{i,j}\beta,\, \sigma^2) \\
\boldsymbol{Y}_i \mid b_i &\sim \mathcal{N}_{n_i}(\mathbb{Z}_i b_i + \mathbb{X}_i \beta,\, \sigma^2 \mathbf{I}_{n_i}) \\
b_i &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu,\, d^2)
\end{aligned}
$$

- V obecném značení

$$
\mathbb{X}_i = \begin{pmatrix} x_{i,1} \\ \vdots \\ x_{i,n_i} \end{pmatrix}, = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \mathbb{Z}_i = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}_{n_i}, \quad \mathbb{D} = d^2
$$

Příklad: Vliv dávky Berenilu na léčbu trypanosomosis

Lineární smíšený model 2

- Jestliže nelze předpokládat, že dávka má stejný účinek ve všech stádech, lze výsledek experimentu reprezentovat následujícím LMM

$$Y_{i,j} = b_{i,1} + z_{i,j}\, b_{i,2} + \varepsilon_{i,j}$$

- $z_{i,j} \begin{cases} 0, & \text{jestliže } (i,j)\text{-tá kráva léčena nízkou dávkou} \\ 1, & \text{jestliže } (i,j)\text{-tá kráva léčena vysokou dávkou} \end{cases}$
- $b_{i,1}$: střední účinek nízké dávky v i-tém stádu
  - náhodný efekt stáda (nepředpokládáme, že je stejné u všech stád)
  - předpokládáme, že stáda jsou náhodně vybrána z populace stád
  - $\mathbb{E}b_{i,1} = \mu_1 =$ střední účinek nízké dávky v populaci
- $b_{i,2}$: rozdíl mezi účinky vysoké a nízké dávky v i-tém stádu
  - náhodný efekt stáda (nepředpokládáme, že je stejné u všech stád)
  - $\mathbb{E}b_{i,2} = \mu_2 =$ střední rozdíl mezi účinky vysoké a nízké dávky v populaci
- $\varepsilon_{i,j}$: náhodná odchylka (i,j)-té krávy od střední hodnoty její části (vysoká/nízká dávka) jejího stáda

Příklad: Vliv dávky Berenilu na léčbu trypanosomosis

Lineární smíšený model 2

$$Y_{i,j} = b_{i,1} + b_{i,2}\, z_{i,j} + \varepsilon_{i,j}$$

$$\mathbb{E} b_{i,1} = \mu_1$$

$$\mathbb{E} b_{i,2} = \mu_2$$

- Parametr hlavního zájmu: μ_2

  $\equiv$ populační rozdíl v efektu léčby vysokou a nízkou dávkou

Příklad: Vliv dávky Berenilu na léčbu trypanosomosis

Normální lineární smíšený model 2

- Označme $\boldsymbol{b}_i = (b_{i,1},\, b_{i,2})$
- Normální LMM:

$$
\begin{aligned}
Y_{i,j} \mid \boldsymbol{b}_i &\sim \mathcal{N}(b_{i,1} + z_{i,j}\, b_{i,2},\, \sigma^2) \\
\boldsymbol{Y}_i \mid \boldsymbol{b}_i &\sim \mathcal{N}_{n_i}(\mathbb{Z}_i \boldsymbol{b}_i,\, \boldsymbol{\Sigma}_i) \\
\boldsymbol{b}_i &\overset{\text{i.i.d.}}{\sim} \mathcal{N}_2(\boldsymbol{\mu},\, \mathbb{D})
\end{aligned}
$$

- V obecném značení

$$
\mathbb{X}_i \text{ není}, \quad \mathbb{Z}_i = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}_{n_i}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \mathbb{D} = \begin{pmatrix} d_1^2 & d_{1,2} \\ d_{1,2} & d_2^2 \end{pmatrix}
$$

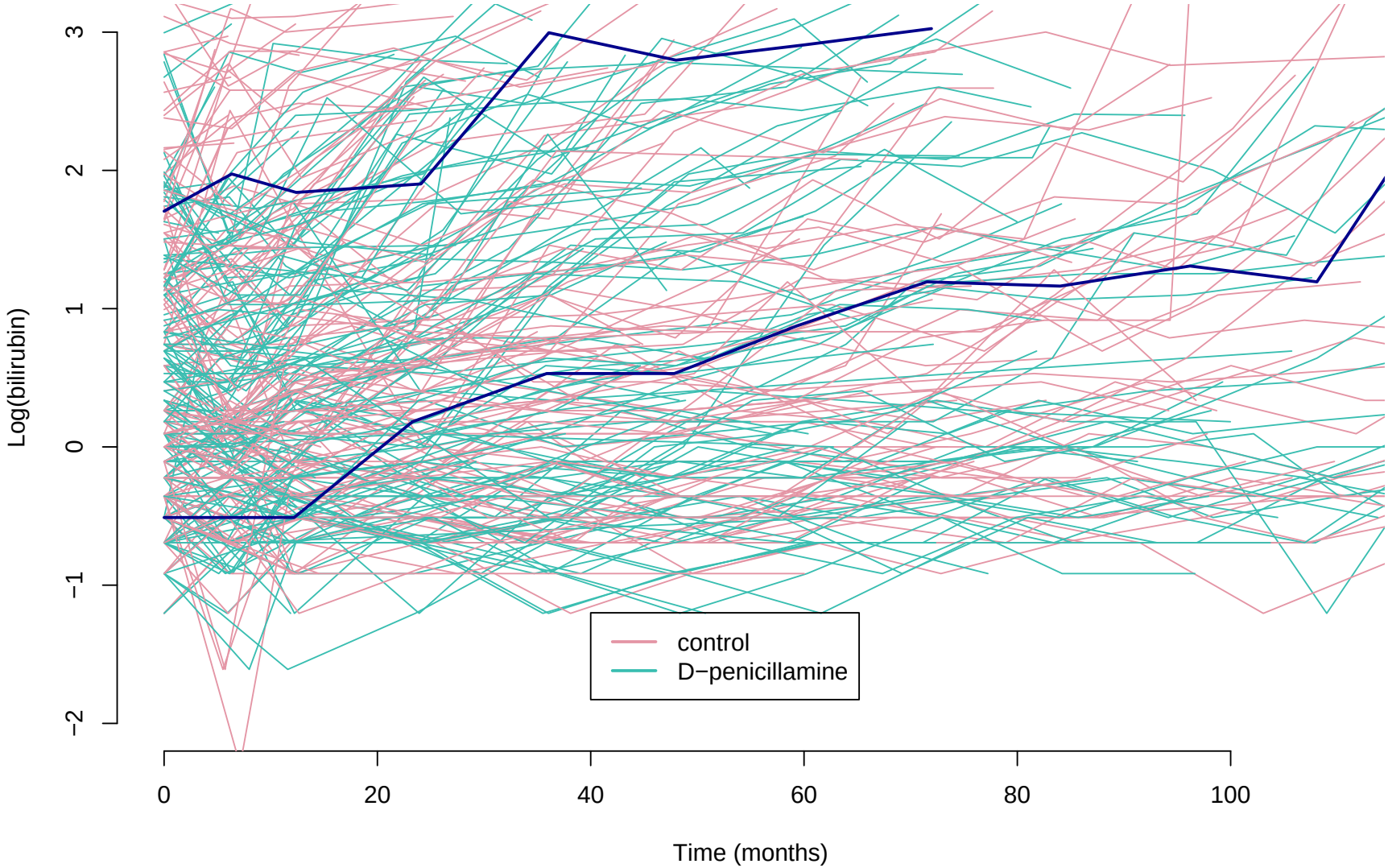
Příklad: Vývoj hladiny bilirubinu u pacientů s PBC

Longitudinální studie

- Studie, která proběhla v letech 1974–1984 na Mayo Clinic
- 312 pacientů s PBC (primary biliary cirrhosis)
- 158 pacientů léčeno D-penicillaminem
- 154 pacientů léčeno pouze standardní léčbou
- Jedním z cílů studie je srovnat skupiny (D-penicillamin vs. kontrolní) vzhledem k vývoji hladiny bilirubinu (jeden z ukazatelů vážnosti PBC)
- Pacienti chodili v zadaných (ne zcela pravidelných) intervalech na vyšetření, kde byla zjišťována (mimo jiného) hladina bilirubinu
- Medián doby sledování byl 6,3 roku (IQR 3,7 – 8,9 roku)
- Pravděpodobnostní reprezentace dat:

$$Y_{i,j} = \text{logaritmus hladiny bilirubinu } i\text{-tého pacienta v čase } t_{i,j}$$

Příklad: Vývoj hladiny bilirubinu u pacientů s PBC



Příklad: Vývoj hladiny bilirubinu u pacientů s PBC

Možný lineární smíšený model

- Vývoj hladiny logaritmického bilirubinu v čase se zdá být u každého pacienta lineární
- Možný model pro jednoho pacienta: přímka v čase
- Každý pacient však začíná v čase 0 na jiné úrovni
  - různé absolutní členy (intercepty) pro různé pacienty
- Růst/pokles log-bilirubinu může být jinak rychlý u jednotlivých pacientů
  - různé směrnice pro různé pacienty

⇛ $Y_{i,j} = b_{i,1} + b_{i,2}\, t_{i,j} + \varepsilon_{i,j}$

- $b_{i,1}$: absolutní člen přímky ($\equiv$ trend) i-tého pacienta
- $b_{i,2}$: směrnice přímky i-tého pacienta
- $\varepsilon_{i,j}$: náhodná odchylka od celkového trendu v čase $t_{i,j}$

Příklad: Vývoj hladiny bilirubinu u pacientů s PBC

Možný lineární smíšený model

$$Y_{i,j} = b_{i,1} + b_{i,2}\, t_{i,j} + \varepsilon_{i,j}$$

- $\boldsymbol{b}_i = \left(b_{i,1},\, b_{i,2}\right)^\top$: náhodný efekt

  ➠ reprezentuje výběr z populace pacientů
- $\mathbb{E} b_{i,1} = \mu_1$: střední (populační) absolutní člen
- $\mathbb{E} b_{i,2} = \mu_2$: střední (populační) směrnice

➠ $\mathbb{E}\, Y_{i,j} = \mu_1 + \mu_2\, t_{i,j}$

$\equiv$ střední vývoj log-bilirubinu v čase v populaci

Příklad: Vývoj hladiny bilirubinu u pacientů s PBC

Zahrnutí informace o skupině (kontrolní/D-penicillamine)

- Pacienti byli na začátku studie přiřazováni do skupin náhodně
  - ➠ Lze předpokládat stejnou střední hladinu log-bilirubinu v obou skupinách
  - ➠ Stejný střední absolutní člen v obou skupinách
- D-penicillamine může vést k jinak rychlé změně hladiny log-bilirubinu
  - ➠ Potřeba umožnit různé střední směrnice v jednotlivých skupinách

- Nechť $x_i \begin{cases} 0, & \text{jestliže } i\text{-tý pacient patří do kontrolní skupiny} \\ 1, & \text{jestliže } i\text{-tý pacient patří do D-penicillamine skupiny} \end{cases}$

➠ Model umožňující různé střední směrnice v jednotlivých skupinách

$$Y_{i,j} = b_{i,1} + b_{i,2}\, t_{i,j} + \beta\, t_{i,j}\, x_{i,j} + \varepsilon_{i,j}$$

Příklad: Vývoj hladiny bilirubinu u pacientů s PBC

Model zahrnující informaci o skupině (kontrolní/D-penicillamine)

$$Y_{i,j} = b_{i,1} + b_{i,2}\, t_{i,j} + \beta\, t_{i,j}\, x_{i,j} + \varepsilon_{i,j}$$

- $\boldsymbol{b}_i = \left(b_{i,1},\, b_{i,2}\right)^\top$: náhodný efekt

  ➠ reprezentuje výběr z populace pacientů

- $\mathbb{E} b_{i,1} = \mu_1$: střední (populační) absolutní člen

➠ Při $x_i = 0$ je $\mathbb{E} Y_{i,j} = \mu_1 + \mu_2\, t_{i,j}$
$\equiv$ střední vývoj log-bilirubinu v čase v kontrolní skupině

➠ Při $x_i = 1$ je $\mathbb{E} Y_{i,j} = \mu_1 + (\mu_2 + \beta)\, t_{i,j}$
$\equiv$ střední vývoj log-bilirubinu v čase ve skupině D-penicillamine

- $\mathbb{E} b_{i,2} = \mu_2$: střední (populační) směrnice v kontrolní skupině
- $\mathbb{E}(b_{i,2} + \beta) = \mu_2 + \beta$: střední (populační) směrnice ve skupině D-penicillamine

  ➠ β střední (populační) rozdíl směrnic mezi skupinou D-penicillamine a kontrolní skupinou

Příklad: Vývoj hladiny bilirubinu u pacientů s PBC

Normální lineární smíšený model

- Normální LMM:

$$Y_{i,j} \mid \boldsymbol{b}_i \sim \mathcal{N}(b_{i,1} + b_{i,2}\, t_{i,j} + \beta\, t_{i,j}\, x_{i,j},\, \sigma^2)$$

$$\boldsymbol{Y}_i \mid \boldsymbol{b}_i \sim \mathcal{N}_{n_i}(\mathbb{X}_i\beta + \mathbb{Z}_i\boldsymbol{b}_i,\, \boldsymbol{\Sigma}_i)$$

$$\boldsymbol{b}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_2(\boldsymbol{\mu},\, \mathbb{D})$$

- Jak vypadají $\mathbb{X}_i$, $\mathbb{Z}_i$, $\boldsymbol{\Sigma}_i$?

Oddíl **4.3**

Lineární smíšený model bayesovsky

Normální lineární smíšený model

Zápis 1

$$\boldsymbol{Y}_i = \mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \qquad i = 1, \ldots, N$$

$$\boldsymbol{b}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_q(\boldsymbol{\mu}, \mathbb{D})$$

$$\boldsymbol{\varepsilon}_i \sim \mathcal{N}_{n_i}(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$$

- $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N, \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N$ vzájemně nezávislé

Zápis 2

$$\boldsymbol{Y}_i \mid \boldsymbol{b}_i \sim \mathcal{N}_{n_i}(\mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\boldsymbol{b}_i, \boldsymbol{\Sigma}_i), \qquad i = 1, \ldots, N$$

$$\boldsymbol{b}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_q(\boldsymbol{\mu}, \mathbb{D})$$

- $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$ vzájemně nezávislé
- $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N$ vzájemně nezávislé

Normální lineární smíšený model

Parametry ve frekventistickém pojetí

“Čisté” parametry (parametry ve frekventistickém smyslu)

- β : pevné efekty
  - (populační) vliv regresorů zahrnutých v matici $\mathbb{X}$ na odezvu
- $\mu = \mathbb{E}\boldsymbol{b}_i$ $(i = 1, \ldots, N)$: střední hodnoty náhodných efektů
  - (populační) vliv regresorů zahrnutých v matici $\mathbb{Z}$ na odezvu
- $\boldsymbol{\Sigma}_i = \mathrm{var}(\boldsymbol{Y}_i \mid \boldsymbol{b}_i)$ $(i = 1, \ldots, N)$: “vnitroskupinová” varianční matice
  - často se předpokládá $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}_{n_i}$
    ⇛ podmíněná nezávislost
  - pomocí jiné struktury pro matici $\boldsymbol{\Sigma}_i$ lze modelovat i jiné struktury závislosti (AR(d), . . .)
- $\mathbb{D} = \mathrm{var}\boldsymbol{b}_i$ $(i = 1, \ldots, N)$: “meziskupinová” varianční matice
  - obvykle se nepředpokládá žádná speciální struktura $\mathbb{D}$ a pouze se požaduje, aby $\mathbb{D} > 0$ (pozitivně definitní matice)

Normální lineární smíšený model

Frekventistická věrohodnost

- **Frekventistické parametry:** $\psi = \left(\boldsymbol{\beta}^\top, \boldsymbol{\mu}^\top, \underbrace{\text{par}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N)}_{\text{často pouze } \sigma^2}, \text{par}(\mathbb{D})\right)^\top$
- **Frekventistická věrohodnost:**

$$L_F(\psi) = p(\boldsymbol{y} \mid \psi) = \cdots = \prod_{i=1}^{N} \varphi(\boldsymbol{y}_i \mid \mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\boldsymbol{\mu},\ \mathbb{V}_i),$$

kde $\mathbb{V}_i = \mathbb{Z}_i\mathbb{D}\mathbb{Z}_i^\top + \boldsymbol{\Sigma}_i$

- Při odhadu metodou maximální věrohodnosti potřeba maximalizovat $L_F(\psi)$ při omezeních $\boldsymbol{\Sigma}_i > 0$ (pro všechna $i = 1, \ldots, N$) a $\mathbb{D} > 0$
  -  balíčky `lme4`, `nlme`
  - SAS procedura `MIXED`

Normální lineární smíšený model

Skrytá data a věrohodnost bayesovského modelu

- Další náhodné složky modelu:

  $\equiv$ vektory náhodných efektů $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N$

  ➠ Další "parametry" při bayesovském přístupu

  $\equiv$ skrytá data

- **Skrytá data:** $\xi = \left(\boldsymbol{b}_1^\top, \ldots, \boldsymbol{b}_N^\top\right)^\top$

- **Věrohodnost bayesovského modelu:**

$$L(\xi, \psi) = p(\boldsymbol{y} \,|\, \xi, \psi) = \cdots = \prod_{i=1}^{N} \varphi(\boldsymbol{y}_i \,|\, \mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\boldsymbol{b}_i,\ \boldsymbol{\Sigma}_i)$$

Oddíl **4.4**

Zobecněný lineární smíšený model

Příklad: NTP TER84073 pokus na myších

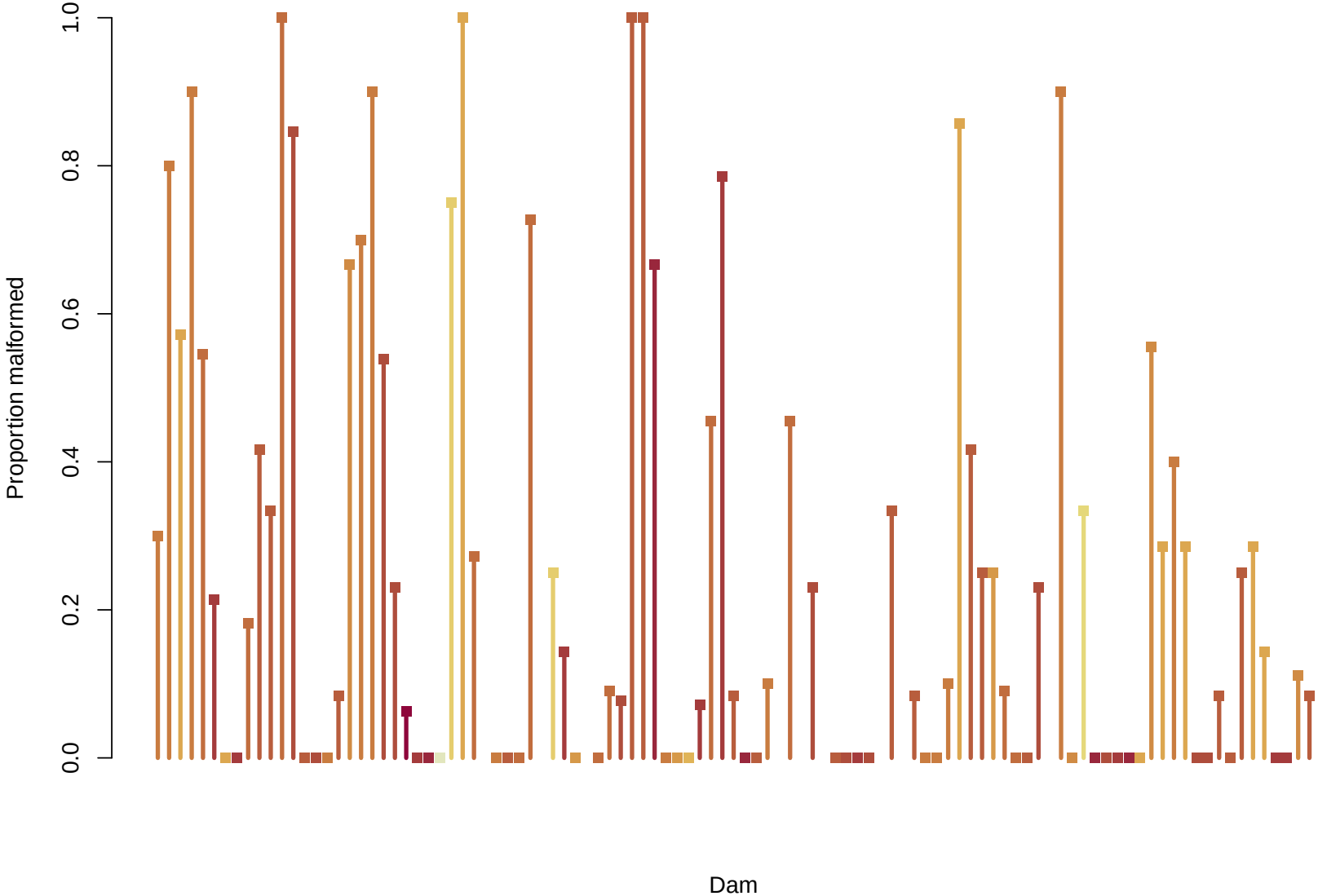
- Data z *National Toxicology Program*.
- 94 těhotným myším byla podána v předem určených momentech určená množství etylenglykolu (EG).
- V 17. dnu těhotenství byly myši usmrceny a následně byl zaznamenán počet zárodků ($n_i, i = 1, \ldots, 94$) a indikace, zda se u jednotlivých zárodků vyskytovala vývojová vada.
- Pravděpodobnostní reprezentace dat:

$$Y_{i,j} = \begin{cases} 0, & \text{jestliže } j\text{-tý zárodek } i\text{-té myši bez vývojové vady,} \\ 1, & \text{jestliže } j\text{-tý zárodek } i\text{-té myši s vývojovou vadou.} \end{cases}$$

- Primární cíl:
  Odhad a inference pro $\pi = \mathbb{E}\, Y_{i,j} = \mathrm{P}(Y_{i,j} = 1)$.

Příklad: NTP TER84073 pokus na myších

Pozorované proporce zárodků s vývojovými vadami



Příklad: NTP TER84073 pokus na myších

Možný model

Možný model ($j = 1, \ldots, n_i$ pro každé $i = 1, \ldots, N$)

- $Y_{i,j} \mid \pi_i \sim \mathcal{A}(\pi_i)$.
- $\pi_i \overset{\text{i.i.d.}}{\sim}$ z nějakého rozdělení.

➠ Reprezentuje fakt, že každá myš má jiné (genetické apod.) dispozice k tomu, aby se u jejích zárodků vyvinuly vývojové vady.

- Myši zahrnuté do studie jsou náhodným výběrem z populace myší a proto je rozumné předpokládat, že též π_i ($i = 1, \ldots, N$) jsou náhodné.
- Abychom se nemuseli starat o omezení $0 < \pi_i < 1$ ($i = 1, \ldots, N$), bude užitečné použít vhodnou reparametrizaci, např.

$$\pi_i = \frac{e^{b_i}}{1 + e^{b_i}}, \qquad b_i = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right)$$

a v modelu uvažovat $b_i \overset{\text{i.i.d.}}{\sim}$ z nějakého rozdělení.

➠ Logistická regrese s náhodnými efekty.

Příklad: NTP TER84073 pokus na myších

Logistická regrese s normálně rozdělenými náhodnými efekty

Možný model ($j = 1, \ldots, n_i$ pro každé $i = 1, \ldots, N$)

- $Y_{i,j} \mid b_i$ nezávislé s rozdělením $\mathcal{A}(\pi_i)$, kde $\pi_i = \frac{e^{b_i}}{1+e^{b_i}}$.
- b_i i.i.d. s rozdělením $\mathcal{N}(\mu, d^2)$.

Parametry (klasické)

- $\psi = (\mu, d^2)^\top$.

Věrohodnost (frekventistická)

$$L_F(\psi) = p(\boldsymbol{y} \mid \psi) = \prod_{i=1}^{N} p(\boldsymbol{y}_i \mid \psi) = \prod_{i=1}^{N} \int_{\mathbb{R}} \prod_{j=1}^{n_i} p(y_{i,j} \mid b_i, \psi)\, p(b_i \mid \psi) db_i$$

$$= \prod_{i=1}^{N} \int_{\mathbb{R}} \pi_i^{\sum_{j=1}^{n_i} y_{i,j}} (1-\pi_i)^{n_i - \sum_{j=1}^{n_i} y_{i,j}}\, \varphi(b_i \mid \mu, d^2) db_i$$

$$= \prod_{i=1}^{N} \int_{\mathbb{R}} \frac{e^{b_i \sum_{j=1}^{n_i} y_{i,j}}}{(1+e^{b_i})^{n_i}}\, \varphi(b_i \mid \mu, d^2) db_i.$$

Oddíl **4.5**

Apriorní rozdělení

Normální (zobecněný) lineární smíšený model

Apriorní rozdělení

- Využijeme rozklad, který sleduje hierarchickou strukturu modelu

$$p(\boldsymbol{\xi}, \boldsymbol{\psi}) = p(\boldsymbol{\xi} \mid \boldsymbol{\psi})\, p(\boldsymbol{\psi})$$

- První část

$$p(\boldsymbol{\xi} \mid \boldsymbol{\psi}) = p(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N \mid \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N, \mathbb{D}) =$$

$$p(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_N \mid \boldsymbol{\mu}, \mathbb{D}) = \cdots = \prod_{i=1}^{N} \varphi(\boldsymbol{b}_i \mid \boldsymbol{\mu}, \mathbb{D})$$

- Druhá část $p(\boldsymbol{\psi})$
  - "standardní" apriorní rozdělení pro "čisté" parametry
  - obvykle se při jeho specifikaci zavádějí na principu obecných hiearchických modelů další náhodné hyperparametry $\boldsymbol{\zeta}$ (snaha zabránit nadměrnému vlivu zvoleného apriorního rozdělení na rozdělení aposteriorní)

Normální (zobecněný) lineární smíšený model

Apriorní rozdělení

- **Parametry bayesovského modelu** $\boldsymbol{\theta} = (\boldsymbol{\xi}, \boldsymbol{\psi}^\top, \boldsymbol{\zeta}^\top)^\top$
  - $\boldsymbol{\xi} = (\boldsymbol{b}_1^\top, \ldots, \boldsymbol{b}_N^\top)^\top$
  - $\boldsymbol{\psi} = (\boldsymbol{\beta}^\top, \boldsymbol{\mu}^\top, \mathrm{par}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N), \mathrm{par}(\mathbb{D}))^\top$
  - $\boldsymbol{\zeta}$: případné další hyperparametry
- **Rozklad apriorního rozdělení:**

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\xi}, \boldsymbol{\psi}, \boldsymbol{\zeta}) = p(\boldsymbol{\xi} \mid \boldsymbol{\psi})\, p(\boldsymbol{\psi} \mid \boldsymbol{\zeta})\, p(\boldsymbol{\zeta})$$

$$= \left\{ \prod_{i=1}^{N} \varphi(\boldsymbol{b}_i \mid \boldsymbol{\mu}, \mathbb{D}) \right\} p(\boldsymbol{\psi} \mid \boldsymbol{\zeta})\, p(\boldsymbol{\zeta})$$

Normální (zobecněný) lineární smíšený model

Možné volby pro nestrukturální část apriorního rozdělení

- Potřeba zvolit $p(\psi)$, které často specifikujeme hierarchicky pomocí dalších hyperparametrů jako

$$p(\psi) = \int p(\psi \mid \zeta)\, p(\zeta)\, d\zeta,$$

  to jest $p(\psi, \zeta) = p(\psi \mid \zeta)\, p(\zeta)$

- $\psi = \left(\boldsymbol{\beta}^\top, \boldsymbol{\mu}^\top, \mathrm{par}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N), \mathrm{par}(\mathbb{D})\right)^\top$

- V dalším se budeme zabývat pouze situací, kdy $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}_{n_i}$, tj. $\mathrm{par}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N) = \sigma^2$

- Označíme dále

$$\tau = \sigma^{-2}$$

$$\mathbb{Q} = \mathbb{D}^{-1}$$

➠ $\psi = \left(\boldsymbol{\beta}^\top, \boldsymbol{\mu}^\top, \tau, \mathrm{par}(\mathbb{Q})\right)^\top$

Normální (zobecněný) lineární smíšený model

Možné volby pro nestrukturální část apriorního rozdělení

- Obvykle se předpokládá apriorní nezávislost pro jednotlivé sady "příbuzných" parametrů, např.

$$p(\boldsymbol{\beta},\ \boldsymbol{\mu},\ \tau,\ \mathbb{Q}) = p(\boldsymbol{\beta})\, p(\boldsymbol{\mu})\, p(\tau)\, p(\mathbb{Q}),$$

respektive

$$p(\boldsymbol{\beta},\ \boldsymbol{\mu},\ \tau,\ \mathbb{Q} \mid \boldsymbol{\zeta}) = p(\boldsymbol{\beta} \mid \boldsymbol{\zeta}^{(1)})\, p(\boldsymbol{\mu} \mid \boldsymbol{\zeta}^{(2)})\, p(\tau \mid \boldsymbol{\zeta}^{(3)})\, p(\mathbb{Q} \mid \boldsymbol{\zeta}^{(4)}),$$

kde $\boldsymbol{\zeta} = \left(\boldsymbol{\zeta}^{(1)\top}, \boldsymbol{\zeta}^{(2)\top}, \boldsymbol{\zeta}^{(3)\top}, \boldsymbol{\zeta}^{(4)\top}\right)^{\top}$

Normální (zobecněný) lineární smíšený model

Možné volby pro nestrukturální část apriorního rozdělení

β, μ mají interpretaci středních hodnot

- Smysluplné apriorní rozdělení:

$$p(\beta) \propto 1$$
$$p(\mu) \propto 1$$

- Jiné smysluplné apriorní rozdělení:

$$p(\beta) \sim \mathcal{N}(\beta_0, \mathbf{\Sigma}_0^\beta)$$
$$p(\mu) \sim \mathcal{N}(\mu_0, \mathbf{\Sigma}_0^\mu),$$

  - β_0, $\mathbf{\Sigma}_0^\beta$, μ_0, $\mathbf{\Sigma}_0^\mu$: pevné/náhodné hyperparametry
  - častá smysluplná volba:
    - $\beta_0 = \mathbf{0}$ (kromě abs. členu modelu)
    - $\mu_0 = \mathbf{0}$ (kromě abs. členu modelu)
    - $\mathbf{\Sigma}_0^\beta, \mathbf{\Sigma}_0^\mu$: diagonální matice s velkými (co je velké?) čísly na diagonále

Normální (zobecněný) lineární smíšený model

Možné volby pro nestrukturální část apriorního rozdělení

τ je inverzní rozptyl odchylek jednotlivých pozorování i-tého subjektu od střední úrovně (závislé na regresorech) tohoto subjektu

- Smysluplné apriorní rozdělení:

$$p(\tau) \sim \mathcal{G}(c_\tau, d_\tau)$$

  - c_τ, d_τ : pevné/náhodné hyperparametry
  - mimo jiné zajišťuje, že $\mathrm{P}(\tau > 0 \mid \boldsymbol{Y}) = 1$
  - častá smysluplná volba hyperparametrů:
    - c_τ : apriorní "stupně volnosti", tj. $c_\tau \in (0, 1]$ vede k slabě informativnímu rozdělení
    - pro připomenutí: $\mathbb{E}\tau = \dfrac{c_\tau}{d_\tau}, \qquad \mathrm{var}\tau = \dfrac{c_\tau}{d_\tau^2}$

      ➠ d_τ : "přesnost" apriorního gama rozdělení
    - d_τ blízké 0 může vést k slabě informativnímu rozdělení
    - volba d_τ však může poměrně značně ovlivnit tvar aposteriorního rozdělení

      ➠ d_τ se často volí jako náhodné s dalším (již pevně zvoleným) gama rozdělením jako apriorním

Normální (zobecněný) lineární smíšený model

Možné volby pro nestrukturální část apriorního rozdělení

$\mathbb{Q}$ je inverzní varianční matice "úrovní" jednotlivých subjektů

- Potřeba vícerozměrné apriorní rozdělení, které s pravděpodobností 1 vygeneruje pozitivně definitní matici

➠ Wishartovo rozdělení

Wishartovo rozdělení

$\mathbb{Q} \sim \mathcal{W}_p(\nu, \boldsymbol{\Xi})$, jestliže

$$\mathbb{Q} = \sum_{i=1}^{\nu} \boldsymbol{Z}_i \boldsymbol{Z}_i^\top,$$

kde
- $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_\nu \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Xi})$
- $\boldsymbol{\Xi}$ je pozitivně definitní měřítková matice
- $\nu > p - 1$ jsou stupně volnosti

- Zřejmě platí: $\mathrm{P}(\mathbb{Q} > 0) = 1$
- Jedná se o vícerozměrné rozšíření χ^2_ν rozdělení

Wishartovo rozdělení

$\mathbb{Q} \sim \mathcal{W}_p(\nu, \Xi)$ má hustotu

$$p(\mathbb{Q}) = \left\{ 2^{\frac{\nu p}{2}} \, \pi^{\frac{p(p-1)}{4}} \prod_{i=1}^{p} \Gamma\left(\frac{\nu + 1 - i}{2}\right) \right\}^{-1} |\Xi|^{-\frac{\nu}{2}}$$

$$|\mathbb{Q}|^{\frac{\nu-p-1}{2}} \exp\left\{ -\frac{1}{2} \operatorname{tr}\left(\Xi^{-1} \mathbb{Q}\right) \right\}, \qquad \mathbb{Q} \text{ pozitivně definitní}$$

- $\nu > p - 1$: "stupně volnosti" lze uvažovat i neceločíselné
  ➠ zobecnění klasického Wishartova rozdělení
- $\mathbb{E}\mathbb{Q} = \nu\Xi$
- $\mathcal{W}_1(\nu, 1) \equiv \mathcal{G}\left(\frac{\nu}{2}, \frac{1}{2}\right) \equiv \chi^2_\nu$
- $\mathcal{W}_1(\nu, \Xi) \equiv \mathcal{G}\left(\frac{\nu}{2}, \frac{\Xi^{-1}}{2}\right)$

Normální (zobecněný) lineární smíšený model

Možné volby pro nestrukturální část apriorního rozdělení

$\mathbb{Q}$ je inverzní varianční matice "úrovní" jednotlivých subjektů

- Smysluplné apriorní rozdělení:

$$p(\mathbb{Q}) \sim \mathcal{W}(\nu_Q, \Xi_Q)$$

  - Ξ_Q, ν_Q : pevné/náhodné hyperparametry
  - častá smysluplná volba hyperparametrů:
    - ν_Q : apriorní "stupně volnosti", tj. $\nu_\tau \in (p-1, p]$ vede k slabě informativnímu rozdělení
    - Ξ_Q se obvykle volí jako diagonální matice, např. $\Xi_Q = \text{diag}\left(\gamma_{Q,1}^{-1}, \ldots, \gamma_{Q,p}^{-1}\right)$
    - inverze měřítkové matice (Ξ_Q^{-1}) je "přesností" Wishartova rozdělení

      ➠ diagonální hodnoty Ξ^{-1} (tj. $\gamma_{Q,1}, \ldots, \gamma_{Q,p}$) blízké 0 mohou vést k slabě informativnímu apriornímu rozdělení
    - volby $\gamma_{Q,1}, \ldots, \gamma_{Q,p}$ však mohou poměrně značně ovlivnit tvar aposteriorního rozdělení

      ➠ $\gamma_{Q,1}, \ldots, \gamma_{Q,p}$ se proto často volí jako náhodné, apriorně vzájemně nezávislé, s dalšími (již pevně zvolenými) gama rozděleními jako apriorními

Normální (zobecněný) lineární smíšený model

Shrnutí

- **"čisté" parametry:** $\psi = \left(\beta^\top, \mu^\top, \tau, \mathrm{par}(\mathbb{Q})\right)^\top$
- **Skrytá data:** $\xi = \left(b_1^\top, \ldots, b_N^\top\right)^\top$
- **Věrohodnost bayesovského modelu:**

$$L(\xi, \psi) = p(y \mid \xi, \psi) = \cdots = \prod_{i=1}^{N} \varphi(y_i \mid \mathbb{X}_i\beta + \mathbb{Z}_i b_i,\ \tau^{-1}\mathbf{I}_{n_i})$$

- Dekompozice **apriorního rozdělení:**

$$p(\xi, \psi) = p(\xi \mid \psi)\, p(\psi) = \prod_{i=1}^{N} \varphi(b_i \mid \mu, \mathbb{Q}^{-1})\, p(\beta)\, p(\mu)\, p(\tau)\, p(\mathbb{Q})$$

Oddíl **4.6**

Aposterorní rozdělení

Normální (zobecněný) lineární smíšený model

Aposteriorní rozdělení

- Nejsou-li žádné náhodné hyperparametry:

$$p(\boldsymbol{\xi}, \boldsymbol{\psi} \mid \boldsymbol{y}) \propto L(\boldsymbol{\xi}, \boldsymbol{\psi})\, p(\boldsymbol{\xi} \mid \boldsymbol{\psi})\, p(\boldsymbol{\psi})$$

- Jsou-li některé hyperparametry (označené jako $\boldsymbol{\zeta}$) náhodné:

$$p(\boldsymbol{\xi}, \boldsymbol{\psi}, \boldsymbol{\zeta} \mid \boldsymbol{y}) \propto L(\boldsymbol{\xi}, \boldsymbol{\psi})\, p(\boldsymbol{\xi} \mid \boldsymbol{\psi})\, p(\boldsymbol{\psi} \mid \boldsymbol{\zeta})\, p(\boldsymbol{\zeta})$$

$$p(\boldsymbol{\xi}, \boldsymbol{\psi} \mid \boldsymbol{y}) \propto \int L(\boldsymbol{\xi}, \boldsymbol{\psi})\, p(\boldsymbol{\xi} \mid \boldsymbol{\psi})\, p(\boldsymbol{\psi} \mid \boldsymbol{\zeta})\, p(\boldsymbol{\zeta}) d\boldsymbol{\zeta}$$

Normální (zobecněný) lineární smíšený model

Aposteriorní rozdělení “čistých” parametrů

- Marginální aposteriorní rozdělení “čistých” parametrů:

$$p(\psi \mid \boldsymbol{y}) \propto \int L(\xi, \psi)\, p(\xi \mid \psi)\, p(\psi \mid \zeta)\, p(\zeta) d(\xi, \zeta)$$

- Díky hierarchické struktuře též platí

$$p(\psi \mid \boldsymbol{y}) \propto L_F(\psi)\, p(\psi),$$

neboť $\quad p(\psi) = \int p(\psi \mid \zeta) p(\zeta) d\zeta$

$$L_F(\psi) = \int L(\xi, \psi)\, p(\xi \mid \psi) d\xi$$

a platí Fubiniova věta

Oddíl **4.7**

Inference založená na aposteriorním rozdělení

Inference založená na aposteriorním rozdělení

- Sdružené aposteriorní rozdělení pro $\boldsymbol{\theta} = \left(\boldsymbol{\psi}^\top, \boldsymbol{\xi}^\top, \boldsymbol{\zeta}^\top\right)^\top$ vyjádříme snadno až na multiplikativní konstantu
- Primárně nás však zajímají hlavně marginální aposteriorní rozdělení pro sady parametrů nebo dokonce jednotlivé parametry
  - $p(\boldsymbol{\beta} \mid \boldsymbol{y})$, $p(\beta_j \mid \boldsymbol{y})$, $p(\boldsymbol{\mu} \mid \boldsymbol{y})$, $p(\tau \mid \boldsymbol{y})$, ...
  - z nich odvozujeme aposteriorní střední hodnotu, věrohodnostní množiny atp.
- Při výpočtu marginálních aposteriorních rozdělení se již nelze vyhnout integrování
  - analyticky proveditelné pro jednodušší modely s apriorními rozděleními vykazujícími alespoň nějaký stupeň konjugovanosti
  - analyticky pracné pro složitější modely i s apriorními rozděleními vykazujícími alespoň nějaký stupeň konjugovanosti
  - analyticky často neproveditelné

Inference založená na aposteriorním rozdělení

- Často nás zajímá též aposteriorní rozdělení pro nějakou měřitelnou funkci t – transformaci původních parametrů

- Příklad:
  - $Y_{i,j} \mid b_i \sim \mathcal{N}(b_i, \sigma^2)$, $b_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, d^2)$, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$
  - Bylo: $\operatorname{cor}(Y_{i,j}, Y_{i,k}) = d^2/(\sigma^2 + d^2) =: \varrho$

  ➠ Z $p(\mu, \sigma^2, d^2 \mid \boldsymbol{y})$ potřeba pomocí věty o transformaci a integrováním odvodit $p(\varrho \mid \boldsymbol{y})$

  - Vzpomeňte si na svoje úspěchy na tomto poli (obvykle v poměrně jednoduchých situacích) ze 3. ročníku…

- Odvození $p\big(t(\boldsymbol{\theta}) \mid \boldsymbol{y}\big)$ z $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ (které již samo o sobě obvykle známe až na multiplikativní konstantu) je často analyticky neproveditelné

Inference založená na aposteriorním rozdělení

- Vše výše řečené je důvodem toho, že bayesovská statistika byla až cca do začátku 90. let 20. století relativně málo prakticky využívána, a pokud ano, tak jenom v souvislosti s poměrně jednoduchými modely
- Řešení problému s nemožností odvozovat potřebné výrazy analyticky:
  - ⇛ aposteriorní inference založená na simulaci
    - k jejímu efektivnímu použití potřeba přiměřeně výkonné počítadlo
    - do cca začátku 90. let 20. století jenom omezeně dostupné/nedostupné (nejenom v ČSSR)

5

MCMC (Markov chain Monte Carlo) metody

Oddíl **5.1**

Úvod

Markov chain Monte Carlo

- Ne vždy (s reálnými aplikacemi jenom velice zřídka) jsme schopni získat náhodný výběr z požadovaného (aposteriorního) rozdělení.
- Alternativa: Konstrukce markovského řetězce, jehož stacionární rozdělení = požadované rozdělení.
- Za jistých předpokladů lze vygenerovaný dostatečně dlouhý markovský řetězec považovat přibližně za náhodný výběr z požadovaného rozdělení a s jako takovým s ním dále pracovat při přibližném výpočtu integrálů a jiných charakteristik tohoto rozdělení.

  ➠ **Markov chain Monte Carlo (MCMC)**

Připomenutí, čím se zabýváme

- Potřebujeme generovat v ideálním případě náhodný výběr z rozdělení s hustotou $f(\boldsymbol{\theta})$ vzhledem k σ-konečné míře λ.
  - V rámci bayesovských metod budeme obvykle používat s $f(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \boldsymbol{y})$.

- V dalším se předpokládají znalosti z předmětu Náhodné procesy 1 (NMSA334).

- Zde jste se zabývali zejména markovskými řetězci se spočetnou množinou stavů.

Markov chain Monte Carlo

- My potřebujeme uvažovat množinu stavů, která je obecnou podmnožinou $\mathbb{R}^k$.
  - Mnoho výsledků pro markovské řetězce se spočetnou množinou stavů lze zobecnit na markovské řetězce s obecnou množinou stavů.
  - V rámci této přednášky nebudeme téměř nic dokazovat, neboť potřebujeme MCMC pouze jako nástroj pro aposteriorní inferenci.
  - Důkazy a mnohé další užitečné poznatky a souvislosti lze nalézt v přednášce Metody Markov Chain Monte Carlo (NMTP539).
  - Existuje též nespočet knih věnujících se MCMC.
    - Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods, Second Edition*. New York: Springer-Verlag.
    - Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition*. Boca Raton: Chapman & Hall/CRC.

Markov chain Monte Carlo

- Budeme předpokládat, že $f(\theta)$ má hustotu vzhledem k σ-konečné míře λ na měřitelném prostoru $(\Theta, \mathcal{T})$.
- Typicky $\Theta \subset \mathbb{R}^k$.
- Obdobně jako dříve budeme zkráceně psát

$$f(\theta)d\lambda(\theta) = f(d\theta).$$

Oddíl **5.2**

Markovské procesy s obecnou množinou stavů

Definice 5.1 Markovské jádro (*Markov kernel*).

Měřitelné zobrazení $P : \Theta \times \mathcal{T} \longrightarrow [0, 1]$ se nazývá markovské jádro (*Markov kernel*) na $(\Theta, \mathcal{T})$, jestliže

1. pro každé $T \in \mathcal{T}$ je $P(\cdot, T)$ nezáporná měřitelná funkce na Θ,
2. pro každé $\theta \in \Theta$ je $P(\theta, \cdot)$ pravděpodobnostní míra na $\mathcal{T}$.

- Jedná se o zobecnění pravděpodobností přechodu.
- Markovské jádro určuje pravděpodobnost přechodu ze stavu θ do stavu v množině T.

Přechodová hustota

- Pro každé $\theta \in \Theta$ existuje hustota rozdělení $P(\theta, \cdot)$ (vzhledem k σ-konečné míře λ) $\equiv$ přechodová hustota, kterou budeme značit jako $p(\theta, \psi)$ a zkráceně psát

$$p(\theta, \psi) d\lambda(\psi) = p(\theta, d\psi),$$

$$\text{resp.} \qquad P(\theta, T) = \int_T p(\theta, \psi) d\lambda(\psi) = \int_T p(\theta, d\psi).$$

- Přechodovou hustotu lze chápat též jako podmíněnou hustotu ψ za podmínky θ (vzhledem k σ-konečné míře λ).

Definice 5.2 Homogenní markovský řetězec (*Homogeneous Markov chain*).

Řekneme, že náhodný proces $\{\boldsymbol{\theta}^{(m)} : m = 0, \dots\}$ je homogenní markovský řetězec (*homogeneous Markov chain*) s přechodovým jádrem (*transition kernel*) P a počátečním rozdělením $f_0(d\boldsymbol{\theta})$, pokud jeho konečně rozměrná rozdělení splňují pro každé $m \in \mathbb{N}_0$ a pro všechna $T_0, \dots, T_m \in \mathcal{T}$ podmínku

$$\mathrm{P}\left(\boldsymbol{\theta}^{(0)} \in T_0, \dots, \boldsymbol{\theta}^{(m)} \in T_m\right) =$$

$$\int_{T_0} \cdots \int_{T_{m-1}} P(\boldsymbol{\theta}^{(m-1)}, T_m) p(\boldsymbol{\theta}^{(m-2)}, d\boldsymbol{\theta}^{(m-1)}) \cdots p(\boldsymbol{\theta}^{(0)}, d\boldsymbol{\theta}^{(1)}) f_0(d\boldsymbol{\theta}^{(0)}).$$

- Jak chápat tuto definici?

Věta 5.1 .

Nechť $\{\boldsymbol{\theta}^{(m)} : m = 0, \dots\}$ *je homogenní markovský řetězec generovaný přechodovým jádrem* P *a* h *je omezená měřitelná funkce na* Θ*. Potom pro každé* $m \in \mathbb{N}_0$ *platí*

$$\mathbb{E}\left[h\left(\boldsymbol{\theta}^{(m+1)}\right) \,\middle|\, \boldsymbol{\theta}^{(m)}, \dots, \boldsymbol{\theta}^{(0)}\right] = \mathbb{E}\left[h\left(\boldsymbol{\theta}^{(m+1)}\right) \,\middle|\, \boldsymbol{\theta}^{(m)}\right].$$

- Pro libovolnou $T \in \mathcal{T}$ tedy volbou $h = \mathbb{I}_T$ platí markovská vlastnost:

$$\mathrm{P}\left(\boldsymbol{\theta}^{(m+1)} \in T \,\middle|\, \boldsymbol{\theta}^{(m)}, \dots, \boldsymbol{\theta}^{(0)}\right) = \mathrm{P}\left(\boldsymbol{\theta}^{(m+1)} \in T \,\middle|\, \boldsymbol{\theta}^{(m)}\right).$$

Markovské procesy s obecnou množinou stavů

Rozdělení markovského řetězce v čase $m+1$

- Nechť π je hustota vzhledem k σ-konečné míře λ na $(\Theta, \mathcal{T})$, tj. $\pi(d\boldsymbol{\theta})$ je pravděpodobnostní rozdělení.
- Pro homogenní markovský řetězec s přechodovým jádrem P zaveďme následující značení:

$$\pi P(T) = \int_{\Theta} P(\boldsymbol{\theta}, T)\pi(d\boldsymbol{\theta})$$

pro libovolnou $T \in \mathcal{T}$.

- Při použití různého značení máme:

$$\pi P(T) = \int_{\Theta}\int_{T} p(\boldsymbol{\theta}, d\boldsymbol{\psi})\pi(d\boldsymbol{\theta}) = \int_{\Theta}\int_{T} p(\boldsymbol{\theta}, \boldsymbol{\psi})\pi(\boldsymbol{\theta})d\lambda(\boldsymbol{\psi})d\lambda(\boldsymbol{\theta}).$$

Markovské procesy s obecnou množinou stavů

Rozdělení markovského řetězce v čase $m+1$

- Použitím Fubiniovy věty dostaneme

$$\begin{aligned}\pi P(T) &= \int_\Theta \int_T p(\theta, \psi)\pi(\theta)d\lambda(\psi)d\lambda(\theta) \\ &= \int_T \int_\Theta p(\theta, \psi)\pi(\theta)d\lambda(\theta)d\lambda(\psi).\end{aligned}$$

- $\pi P(d\psi)$ je opět pravděpodobnostní rozdělení na $(\Theta, \mathcal{T})$ a má hustotu

$$\int_\Theta p(\theta, \psi)\pi(\theta)d\lambda(\theta) = \int_\Theta p(\theta, \psi)\pi(d\theta)$$

  vzhledem k míře λ.

- Je-li $\pi(d\theta)$ rozdělením markovského řetězce v čase $m-1$, potom $\pi P(d\psi)$ je rozdělením markovského řetězce v čase m.

Definice 5.3 Stacionární rozdělení (*Stationary distribution*).

Pravděpodobnostní rozdělení $\pi(d\theta)$ se nazývá stacionárním rozdělením (*stationary distribution*) homogenního markovského řetězce s přechodovým jádrem P, jestliže

$$\pi P(d\theta) = \pi(d\theta),$$

to jest, jestliže pro libovolnou $T \in \mathcal{T}$ platí

$$\begin{aligned} \pi P(T) &= \pi(T), \\ \int_{\Theta} P(\theta, T)\pi(d\theta) &= \int_{T} \pi(d\theta). \end{aligned}$$

Definice 5.4 Reversibilita (*reversibility*) markovského řetězce.

Homogenní markovský řetězec s přechodovým jádrem P se nazývá reversibilní (*reversible*) vzhledem k pravděpodobnostnímu rozdělení π, jestliže pro libovolné $T, S \in \mathcal{T}$ platí

$$\int_T P(\theta, S)\pi(d\theta) = \int_S P(\psi, T)\pi(d\psi).$$

Reversibilita

- Na $\int_T P(\theta, S)\pi(d\theta)$ lze pohlížet jako na sdružené pravděpodobnostní rozdělení na $\left(\Theta \times \Theta, \mathcal{T} \otimes \mathcal{T}\right)$, které množinám $T, S \in \mathcal{T}$ přiřadí pravděpodobnost

$$Q_1(T, S) = \int_T P(\theta, S)\pi(d\theta) = \int_T \int_S p(\theta, d\psi)\pi(d\theta).$$

- Rozdělení Q_1 má hustotu

$$q_1(\theta, \psi) = p(\theta, \psi)\pi(\theta)$$

vzhledem k součinové míře $\lambda \otimes \lambda$, zkráceně

$$q_1(d\theta, d\psi) = p(\theta, d\psi)\pi(d\theta).$$

Reversibilita

- Na $\int_S P(\psi, T)\pi(d\psi)$ lze pohlížet jako na sdružené pravděpodobnostní rozdělení na $(\Theta \times \Theta, \mathcal{T} \otimes \mathcal{T})$, které množinám $T, S \in \mathcal{T}$ přiřadí pravděpodobnost

$$Q_2(S, T) = \int_S P(\psi, T)\pi(d\psi) = \int_S \int_T p(\psi, d\theta)\pi(d\psi).$$

- Rozdělení Q_2 má hustotu

$$q_2(\psi, \theta) = p(\psi, \theta)\pi(\psi)$$

vzhledem k součinové míře $\lambda \otimes \lambda$, zkráceně

$$q_2(d\psi, d\theta) = p(\psi, d\theta)\pi(d\psi).$$

Reversibilita

- Reversibilita vzhledem k π tedy znamená, že pro libovolné $T, S \in \mathcal{T}$ platí

$$Q_1(T, S) = Q_2(S, T).$$

- Zkráceně zapsáno

$$p(\theta, d\psi)\pi(d\theta) = p(\psi, d\theta)\pi(d\psi).$$

  - Nutno chápat jako rovnost rozdělení (pravděpodobnostních měr)!
  - Sdružené rozdělení stavů v časech m a $m+1$ je stejné jako sdružené rozdělení stavů v časech $m+1$ a m pro libovolné $m = 0, 1, \ldots$.

Markovské procesy s obecnou množinou stavů

Reversibilita a detailní podmínka rovnováhy

- $p(\boldsymbol{\theta}, d\boldsymbol{\psi})$ lze chápat též jako podmíněné rozdělení $\boldsymbol{\psi}$ za podmínky $\boldsymbol{\theta}$ mající hustotu $p(\boldsymbol{\theta}, \boldsymbol{\psi})$ (přechodová hustota) vzhledem k σ-konečné míře λ (v argumentu $\boldsymbol{\psi}$).
- Podmínku reversibility lze potom pomocí hustot přepsat jako

$$p(\boldsymbol{\theta}, \boldsymbol{\psi})\,\pi(\boldsymbol{\theta}) = p(\boldsymbol{\psi}, \boldsymbol{\theta})\,\pi(\boldsymbol{\psi}) \qquad \text{pro } \lambda\text{-s.v. } \boldsymbol{\theta}, \boldsymbol{\psi} \in \Theta.$$

➠ Detailní podmínka rovnováhy (*detailed balance condition*).

Věta 5.2 .

Je-li homogenní markovský řetězec reversibilní vzhledem k π, potom π je jeho stacionární rozdělení.

Důkaz. Stačí v definici reversibility položit $S = \Theta$.

❑

- To jest, reversibilita (splnění detailní podmínky rovnováhy) je postačující podmínkou pro stacionaritu.

Přechodové jádro m-tého řádu

Definice 5.5 Přechodové jádro m-tého řádu (*m-step transition probability kernel*).

Uvažujme homogenní markovský řetězec s přechodovým jádrem P a položme $P^0(\theta, T) = \delta_\theta(T)$. Přechodové jádro m-tého řádu (*m-step transition probability kernel*) je dáno induktivně vztahem

$$P^m(\theta, T) = \int_\Theta P(\psi, T)\, P^{m-1}(\theta, d\psi), \qquad m \in \mathbb{N}.$$

Věta 5.3 Chapmanova-Kolmogorovova rovnost.

V homogenním markovském řetězci s přechodovým jádrem P platí pro $n, m \in \mathbb{N}_0$ a $n \leq m$, $\theta \in \Theta$, $T \in \mathcal{T}$ vztah

$$P^m(\theta, T) = \int_\Theta P^{m-n}(\psi, T)\, P^n(\theta, d\psi).$$

Důkaz. V definici homogenního markovského řetězce stačí položit $f_0 = \delta_\theta$, $T_i = \Theta$ pro $i = 0, \ldots, m-1$, $T_m = T$. Definice P^n a P^{m-n} se použije pro prvních n a posledních $m - n$ integrandů.

❑

- Jak chápat Chapmanovu-Kolmogorovovu rovnost?

Definice 5.6 Limitní rozdělení (*limitting distribution*).

Pravděpodobnostní rozdělení π na $(\Theta, \mathcal{T})$ nazveme limitní rozdělení (limitting distribution) markovského řetězce $\{\boldsymbol{\theta}^{(m)} : m = 0, 1, \dots\}$ generovaného přechodovým jádrem P, jestliže

$$\lim_{m\to\infty} P^m(\boldsymbol{\theta}, T) = \pi(T) \quad \text{pro } \pi\text{-s.v. } \boldsymbol{\theta} \in \Theta \text{ a pro všechna } T \in \mathcal{T}.$$

- Poznámka. Je-li π limitním rozdělením, potom pro libovolné počáteční rozdělení f_0 a pro každou $T \in \mathcal{T}$ platí

$$\mathrm{P}(\boldsymbol{\theta}^{(m)} \in T) = \int_\Theta P^m(\boldsymbol{\theta}, T) f_0(d\boldsymbol{\theta}) \xrightarrow[m\to\infty]{} \int_\Theta \pi(T) f_0(d\boldsymbol{\theta}) = \pi(T).$$

Věta 5.4 Limitní a stacionární rozdělení.

Je-li π limitním rozdělením homogenního markovského řetězce, potom je π též stacionárním rozdělením tohoto řetězce.

Důkaz. Pro $T \in \mathcal{T}$ a π-s.v. $\boldsymbol{\theta} \in \Theta$ je

$$\pi(T) = \lim_{m\to\infty} P^m(\boldsymbol{\theta}, T) = \lim_{m\to\infty} \int_\Theta P(\boldsymbol{\psi}, T) P^{m-1}(\boldsymbol{\theta}, d\boldsymbol{\psi})$$

$$= \int_\Theta P(\boldsymbol{\psi}, T)\pi(d\boldsymbol{\psi}) = \pi P(T).$$

❑

Oddíl **5.3**

Principy MCMC

Připomenutí, čím se zabýváme

- $f(d\theta)$ je nějaké pravděpodobnostní rozdělení.
- Pro měřitelné funkce $t(\theta)$ potřebujeme aproximovat integrály typu

$$\int_\Theta t(\theta)\, f(d\theta) = \mathbb{E}_{f(d\theta)} t(\theta).$$

- Je-li $\mathcal{S}_M = \{\theta^{(1)}, \ldots, \theta^{(M)}\}$ náhodný výběr z rozdělení $f(d\theta)$, potom (za jistých předpokladů)

$$\int_\Theta t(\theta)\, f(d\theta) \approx \frac{1}{M} \sum_{m=1}^{M} t\left(\theta^{(m)}\right) = \widehat{\mathbb{E}}_{f(d\theta)} t(\theta) := \widehat{t}_M.$$

➠ Monte Carlo integrace

Principy MCMC

- Nechť $\{\boldsymbol{\theta}^{(m)} : m = 0, \ldots\}$ je homogenní markovský řetězec se stacionárním rozdělením $f(d\boldsymbol{\theta})$.
  - Víme: reversibilita vzhledem k $f(d\boldsymbol{\theta})$ implikuje stacionaritu vzhledem k $f(d\boldsymbol{\theta})$.
- Stačí tedy zvolit přechodové jádro markovského řetězce tak, aby přechodová hustota $p(\boldsymbol{\theta}, \boldsymbol{\psi})$ splňovala detailní podmínku rovnováhy vzhledem k $f(d\boldsymbol{\theta})$.
- Stačí tedy volit přechodovou hustotu tak, aby splňovala

$$p(\boldsymbol{\theta}, \boldsymbol{\psi})\, f(\boldsymbol{\theta}) = p(\boldsymbol{\psi}, \boldsymbol{\theta})\, f(\boldsymbol{\psi}) \qquad \forall\, \boldsymbol{\theta}, \boldsymbol{\psi} \in \Theta$$

---

a máme potřebný markovský řetězec.

➠ Toto není nikterak obtížné, jak bude záhy ukázáno.

Principy MCMC

- Předpokládejme, že se nám navíc podaří zajistit, že existuje limitní rozdělení uvažovaného markovského řetězce.
  - Víme: limitní rozdělení (existuje-li) = stacionární rozdělení $f(d\boldsymbol{\theta})$.
- Od jistého okamžiku (řekněme $B+1$) lze tedy
$$\mathcal{S}_M = \left\{\boldsymbol{\theta}^{(B+1)}, \ldots, \boldsymbol{\theta}^{(B+M)}\right\}$$
považovat za náhodné veličiny s rozdělením $f(d\boldsymbol{\theta})$.
- Začátku řetězce $\left\{\boldsymbol{\theta}^{(0)}, \ldots, \boldsymbol{\theta}^{(B)}\right\}$ se říká *burn-in period*.
- Nejde o náhodný výběr, neboť $\boldsymbol{\theta}^{(B+1)}, \ldots, \boldsymbol{\theta}^{(B+M)}$ nejsou nezávislé!

- Nicméně, jestliže $\int_\Theta |t(\theta)| f(d\theta) < \infty$ a jestliže dále platí jisté předpoklady, potom (**ergodická věta**):

$$\widehat{t}_M = \frac{1}{M} \sum_{m=1}^{M} t\left(\theta^{(B+m)}\right) \xrightarrow{\text{a.s.}} \int_\Theta t(\theta)\, f(d\theta) \quad \text{pro } M \to \infty.$$

- $\widehat{t}_M$ je tedy konzistentním odhadem pro $\int_\Theta t(\theta)\, f(d\theta) = \mathbb{E}_{f(d\theta)} t(\theta)$.
- Při splnění oněch jistých předpokladů lze též odhadnout

$$v_M = \text{var}_{f(d\theta)}\left(\widehat{t}_M\right)$$

a ohodnotit tak přesnost odhadu $\mathbb{E}_{f(d\theta)} t(\theta)$

(přesnost aproximace integrálu $\int_\Theta t(\theta)\, f(d\theta)$).

Ergodická věta

- Předpoklady pro platnost ergodické věty pro markovské řetězce s obecnou množinou stavů jsou zobecněními předpokladů ergodické věty pro markovské řetězce s diskrétní množinou stavů.
- Potřeba zobecnit (a rozšířit) následující pojmy:
  - nerozložitelnost (*irreducibility*),
  - neperiodicita (*aperiodicity*),
  - trvalý (*recurrent*) a pozitivně trvalý (*positive recurrent*) markovský řetězec.

  ➠ NMTP539: Metody Markov Chain Monte Carlo.
- Zajistit splnění těchto předpokladů v praktických aplikacích též není těžké.
- Co je tedy obtížné?

Největší obtíž při praktické aplikaci MCMC

- Zjistit, od kterého okamžiku již lze (s přiměřeně malou chybou) považovat rozdělení stavů vygenerovaného markovského řetězce za limitní $=$ stacionární $f(d\theta)$.
  - Jak velké má být B (délka *burn-in period*)?
  - Jedná se o konvergenci pravděpodobnostních měr a nelze ji tedy posoudit jednoduchým číslem jako třeba v případě numerického optimalizačního algoritmu!

Největší obtíže

Druhá největší obtíž při praktické aplikaci MCMC

- Připomeňme, že $\theta^{(B+1)}, \ldots, \theta^{(B+M)}$ nejsou obecně nezávislé a tudíž (i při předpokladu konvergence k limitnímu rozdělení) není nutně pravda

$$v_M = \mathrm{var}_{f(d\theta)}\left(\widehat{t}_M\right) = \frac{\mathrm{var}_{f(d\theta)}\left(t(\theta)\right)}{M}.$$

- Stavy markovského řetězce jsou typicky kladně (auto)korelovány a tudíž

$$v_M = \mathrm{var}_{f(d\theta)}\left(\widehat{t}_M\right) \geq \frac{\mathrm{var}_{f(d\theta)}\left(t(\theta)\right)}{M}.$$

- Je-li markovský řetězec zkonstruován tak, že vykazuje vysokou autokorelaci mezi jednotlivými stavy, může být v_M nepoužitelně vysoké i při poměrně vysoké hodnotě M.

Největší obtíže

- Snaha konstruovat markovský řetězec tak, aby autokorelace byla co nejnižší.
  - Nulová autokorelace $\equiv$ co do přesnosti se markovský řetězec chová stejně jako náhodný výběr, kde $v_M = \frac{1}{M} \operatorname{var}_{f(d\boldsymbol{\theta})}\big(t(\boldsymbol{\theta})\big)$.
- Konstrukce markovského řetězce s nízkou autokorelovaností snadná není a obtížnost této snahy závisí na konkrétní aplikaci.

Oddíl **5.4**

Gibbsův algoritmus

Gibbsův algoritmus

- Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayes restoration of image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.
  - Aplikace v oblasti restaurování digitálních obrázků.
- Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**(410), 398–409.
  - Aplikace v bayesovské statistice.

Předpoklady

Předpoklady:

- $\Theta = \prod_{i=1}^{k} \Theta_i$, $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_k^\top\right)^\top$
- Cílové (stacionární) rozdělení je $f(d\boldsymbol{\theta})$ a má hustotu $f(\boldsymbol{\theta})$ vzhledem k součinové míře $\lambda_1 \otimes \cdots \otimes \lambda_k$, přičemž λ_i je σ-konečná míra s $\lambda_i(\Theta_i) > 0$ $(i = 1, \ldots, k)$.
- $\Theta = \left\{\boldsymbol{\theta} : f(\boldsymbol{\theta}) > 0\right\}$.
- Jsme schopni (snadno) generovat z plně podmíněných (*full conditional*) rozdělení

$$f\left(d\boldsymbol{\theta}_i \,\middle|\, \boldsymbol{\theta}_{-i}\right) = f\left(d\boldsymbol{\theta}_i \,\middle|\, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_k\right).$$

Algoritmus:

1. Zvol počáteční stav $\boldsymbol{\theta}^{(0)} = \left(\boldsymbol{\theta}_1^{(0)\top}, \ldots, \boldsymbol{\theta}_k^{(0)\top}\right)^\top$, polož $m = 0$.

2. (i) generuj $\boldsymbol{\theta}_1^{(m+1)}$ z podmíněného rozdělení
$$f\left(d\boldsymbol{\theta}_1 \,\middle|\, \boldsymbol{\theta}_2^{(m)}, \ldots, \boldsymbol{\theta}_k^{(m)}\right).$$

   (ii) generuj $\boldsymbol{\theta}_2^{(m+1)}$ z podmíněného rozdělení
$$f\left(d\boldsymbol{\theta}_2 \,\middle|\, \boldsymbol{\theta}_1^{(m+1)}, \boldsymbol{\theta}_3^{(m)}, \ldots, \boldsymbol{\theta}_k^{(m)}\right).$$

   (iii) generuj $\boldsymbol{\theta}_3^{(m+1)}$ z podmíněného rozdělení
$$f\left(d\boldsymbol{\theta}_3 \,\middle|\, \boldsymbol{\theta}_1^{(m+1)}, \boldsymbol{\theta}_2^{(m+1)}, \boldsymbol{\theta}_4^{(m)}, \ldots, \boldsymbol{\theta}_k^{(m)}\right).$$

   $\vdots$

   (k) generuj $\boldsymbol{\theta}_k^{(m+1)}$ z podmíněného rozdělení
$$f\left(d\boldsymbol{\theta}_k \,\middle|\, \boldsymbol{\theta}_1^{(m+1)}, \ldots, \boldsymbol{\theta}_{k-1}^{(m+1)}\right).$$

3. Zvětši m o jedničku a jdi na 2. krok algoritmu.

Přechodová hustota

$$p(\boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^{k} f(\boldsymbol{\psi}_i \mid \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{i-1}, \boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_k).$$

- Odpovídá přechodovému jádru

$$P(\boldsymbol{\theta}, T) = \int_T p(\boldsymbol{\theta}, d\boldsymbol{\psi}) = \int_T p(\boldsymbol{\theta}, \boldsymbol{\psi}) d\lambda(\boldsymbol{\psi}).$$

Věta 5.5 .

Rozdělení $f(d\theta)$ je stacionárním rozdělením markovského řetězce generovaného Gibbsovým algoritmem.

Důkaz. Viz tabule. ❑

- Pokud bude existovat limitní rozdělení, musí se jednat o rozdělení stacionární a tedy cílové $f(d\theta)$.

Gibbsův algoritmus

Existence limitního rozdělení, ergodicita

- Ergodicitu (existenci limitního rozdělení) lze dokázat například při splnění předpokladů, které byly uvedeny na začátku povídání o Gibbsově algoritmu, to jest
  - $\Theta = \prod_{i=1}^{k} \Theta_i$, $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_k^\top\right)^\top$.
  - Cílové (stacionární) rozdělení je $f(d\boldsymbol{\theta})$ a má hustotu $f(\boldsymbol{\theta})$ vzhledem k součinové míře $\lambda_1 \otimes \cdots \otimes \lambda_k$, přičemž λ_i je σ-konečná míra s $\lambda_i(\Theta_i) > 0$ $(i = 1, \ldots, k)$.
  - $\Theta = \left\{\boldsymbol{\theta} : f(\boldsymbol{\theta}) > 0\right\}$.
- Pro standardní statistické aplikace je toto obvykle splněno.
- Při rutinním použití Gibbsova algoritmu nicméně zůstává nemalým problémem zjistit, zda použitá konečná realizace markovského řetězce již dostatečně dobře odpovídá limitnímu $=$ stacionárnímu $=$ cílovému rozdělení.
- Při nevhodném použití Gibbsova algoritmu (viz dále) nemusí ani velmi dlouhá realizace markovského řetězce dostatečně dobře aproximovat limitní rozdělení!

Gibbsův algoritmus

Reversibilita

- Lze ukázat, že markovský řetězec generovaný Gibbsovým algoritmem nesplňuje detailní podmínku rovnováhy, tj. řetězec není reversibilní vzhledem k rozdělení f.

- Reversibility lze dosáhnout několika způsoby:
  - Generujeme střídavě odpředu a odzadu.
  - Pořadí vybíráme náhodně.
    - V každém podkroku Gibbsova algoritmu generujeme i-tý podvektor s pravděpodobností p_i ($0 < p_i < 1, \sum_{i=1}^{k} p_i = 1$).
    - Častá volba je $p_i = 1/k$ (rovnoměrné rozdělení).
    - *Random scan* Gibbsův algoritmus.

Gibbsův algoritmus

Autokorelace

- V principu lze generovat ze všech jednorozměrných podmíněných rozdělení.
  - V případě, že složky θ jsou v cílovém rozdělení $f(d\theta)$ významně korelovány, vede generování z jednorozměrných podmíněných rozdělení k markovskému řetězci s velkou autokorelací.
  - Ideální situace je stav, kdy podvektory $\theta_1, \ldots, \theta_k$ jsou v cílovém rozdělení $f(d\theta)$ co možná nejméně korelovány.

Plně podmíněná rozdělení

- Při odvozování plně podmíněných rozdělení je vhodné si uvědomit a využívat základní fakt a to

$$f\left(\theta_i \,\middle|\, \theta_{-i}\right) \propto f(\theta),$$

přičemž $\propto$ nyní znamená, že vše, co neobsahuje θ_i je konstantou.

- V případě hierarchického modelu, kde je $f(\theta)$ zadáno jako součin postupně podmíněných rozdělení, pak $f\left(\theta_i \,\middle|\, \theta_{-i}\right)$ závisí pouze na těch podmíněných rozděleních ze specifikace $f(\theta)$, kde jde v uvažované hierarchické struktuře:
  - o "potomky" θ_i,
  - o sourozence θ_i (nejsou-li v $f(d\theta)$ nezávislí),
  - o "rodiče" θ_i.

Příklad: Lineární model s neinformativním apriorním rozdělením

Lineární model

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathbb{X} : \text{pevná matice } n \times k,$$
$$\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

- Parametry: $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top, \tau\right)^\top$, kde $\tau = \sigma^{-2} > 0$.
- Věrohodnost: $\boldsymbol{Y} \sim \mathcal{N}_n(\mathbb{X}\boldsymbol{\beta}, \tau^{-1}\mathbf{I}_n)$.
- Neinformativní apriorní rozdělení:

$$p(\boldsymbol{\beta}) \propto 1, \qquad \boldsymbol{\beta} \in \mathbb{R}^k,$$
$$p(\tau) \propto \frac{1}{\tau}, \qquad \tau > 0.$$

Příklad: Lineární model s neinformativním apriorním rozdělením

Věrohodnost

- Označme: $\boldsymbol{b} = \left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \mathbb{X}^\top \boldsymbol{y}$,

  $\mathrm{SS}_e = \left\| \boldsymbol{y} - \mathbb{X}\boldsymbol{b} \right\|^2$.

- Věrohodnost:

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= p(\boldsymbol{y} \mid \boldsymbol{\beta}, \tau) \\
&= (2\pi)^{-\frac{n}{2}} \tau^{\frac{n}{2}} \exp\left[-\frac{\tau}{2} \left\{ \mathrm{SS}_e + (\boldsymbol{\beta} - \boldsymbol{b})^\top \mathbb{X}^\top \mathbb{X} (\boldsymbol{\beta} - \boldsymbol{b}) \right\} \right] \\
&= (2\pi)^{-\frac{n}{2}} \tau^{\frac{n}{2}} \exp\left\{ -\frac{\tau}{2} (\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta}) \right\}.
\end{aligned}
$$

Příklad: Lineární model s neinformativním apriorním rozdělením

Aposteriorní rozdělení

- Bylo odvozeno:

$$p(\boldsymbol{\beta}, \tau \mid \boldsymbol{y}) = p(\boldsymbol{\beta} \mid \tau, \boldsymbol{y}) \times p(\tau \mid \boldsymbol{y}),$$

---

kde $p(\tau \mid \boldsymbol{y}) \sim \mathcal{G}\left(\frac{n-k}{2}, \frac{\mathrm{SS}_e}{2}\right)$,

$$p(\boldsymbol{\beta} \mid \tau, \boldsymbol{y}) \sim \mathcal{N}_k\left(\boldsymbol{b}, \tau^{-1}\left(\mathbb{X}^\top \mathbb{X}\right)^{-1}\right).$$

- Dále bylo odvozeno: $p(\boldsymbol{\beta} \mid \boldsymbol{y}) \sim \mathrm{MVT}_{k,n-k}\left(\boldsymbol{b}, \frac{\mathrm{SS}_e}{n-k}\left(\mathbb{X}^\top \mathbb{X}\right)^{-1}\right)$.
- Pomocí Gibbsova algoritmu sestrojíme markovský řetězec, který bude mít rozdělení $p(\boldsymbol{\beta}, \tau \mid \boldsymbol{y})$ jako stacionární i limitní.

Příklad: Lineární model s neinformativním apriorním rozdělením

Plně podmíněná rozdělení

- Označme $\mathbb{W} = \mathbb{X}^\top \mathbb{X}$ s prvky $w_{i,j}$ $(i, j = 1, \ldots, k)$.

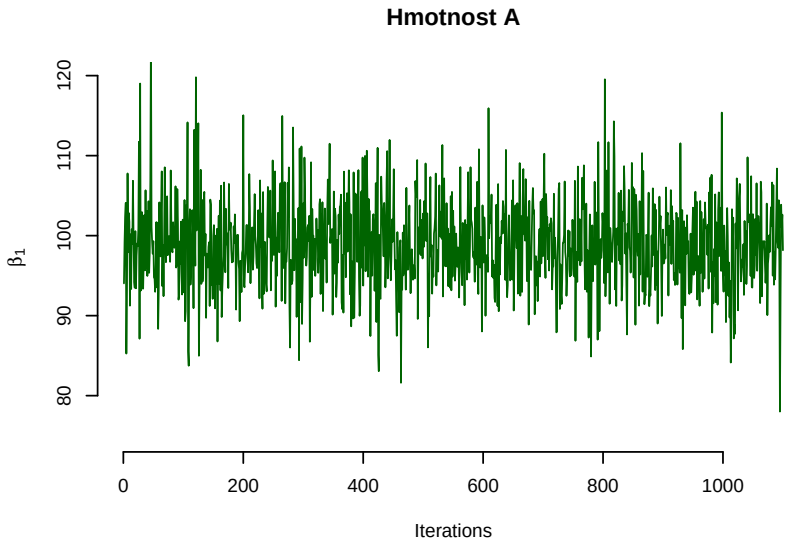
$$p(\boldsymbol{\beta} \mid \cdots) = p(\boldsymbol{\beta} \mid \tau, \boldsymbol{y}) \sim \mathcal{N}_k\left(\boldsymbol{b}, \tau^{-1}(\mathbb{X}^\top \mathbb{X})^{-1}\right),$$

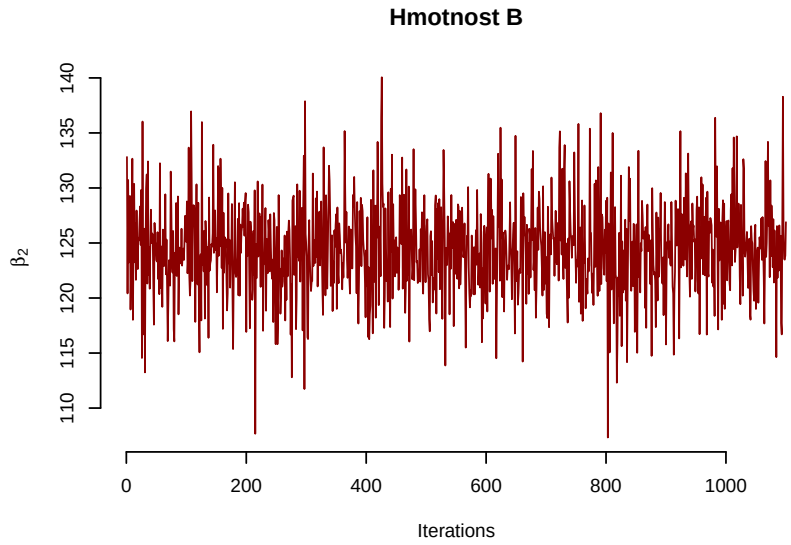
$$p(\beta_i \mid \cdots) = p(\beta_i \mid \beta_{-i}, \tau, \boldsymbol{y}) \sim \mathcal{N}\left(b_i - \sum_{j \neq i} \frac{w_{i,j}}{w_{i,i}}(\beta_j - b_j),\ (\tau\, w_{i,i})^{-1}\right),$$

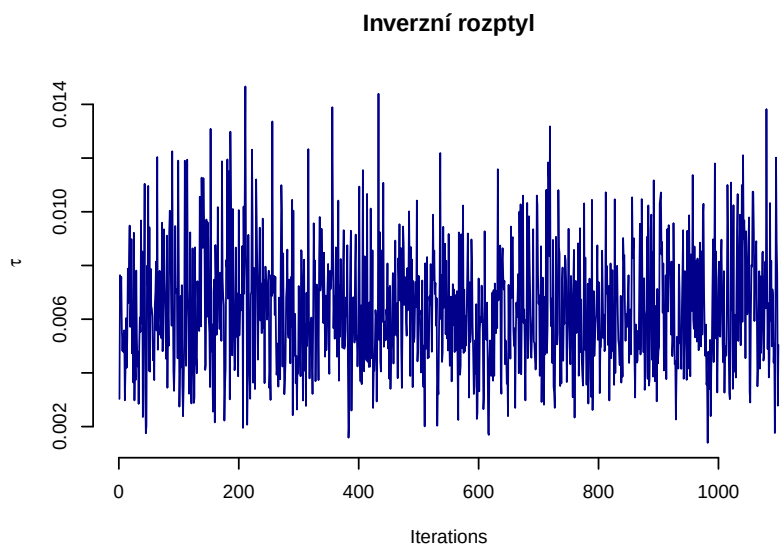
$$p(\tau \mid \cdots) = p(\tau \mid \boldsymbol{\beta}, \boldsymbol{y}) \sim \mathcal{G}\left(\frac{n}{2},\ \frac{(\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta})}{2}\right).$$

Příklad: Vážení lehkých objektů

Blokový Gibbsův algoritmus: Generované hodnoty (B=100, M=1 000)

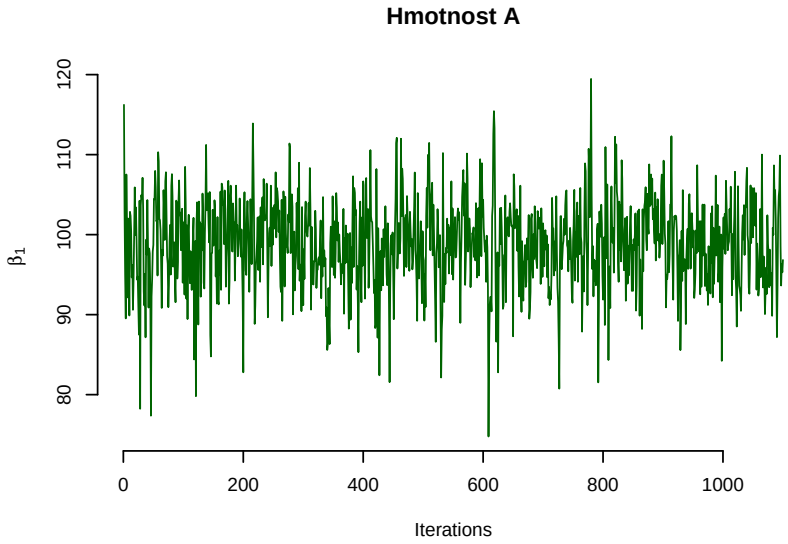


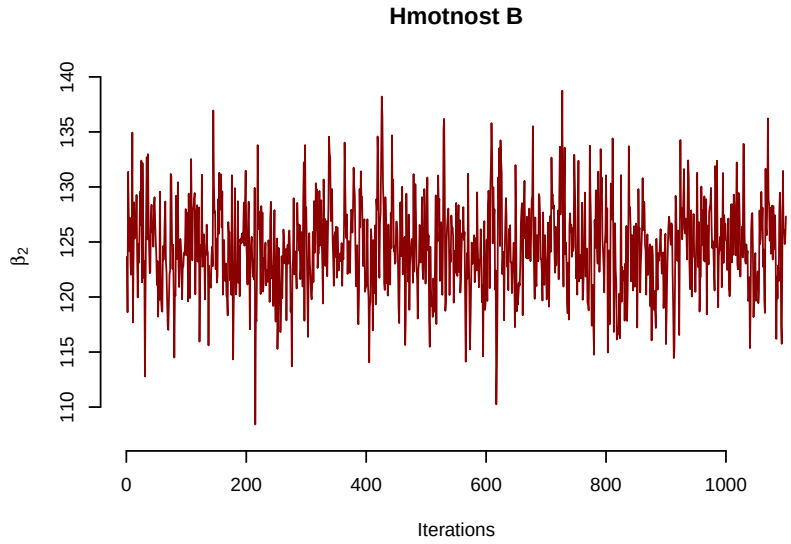


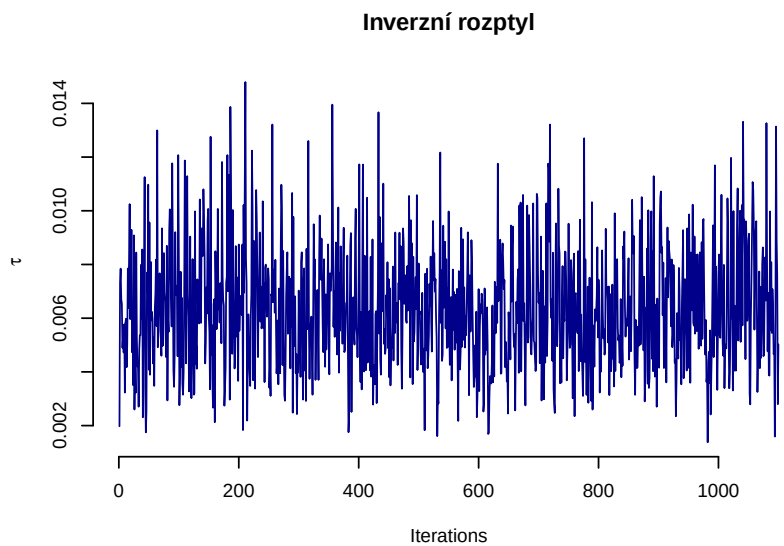


Příklad: Vážení lehkých objektů

Single site Gibbsův algoritmus: Generované hodnoty (B=100, M=1 000)

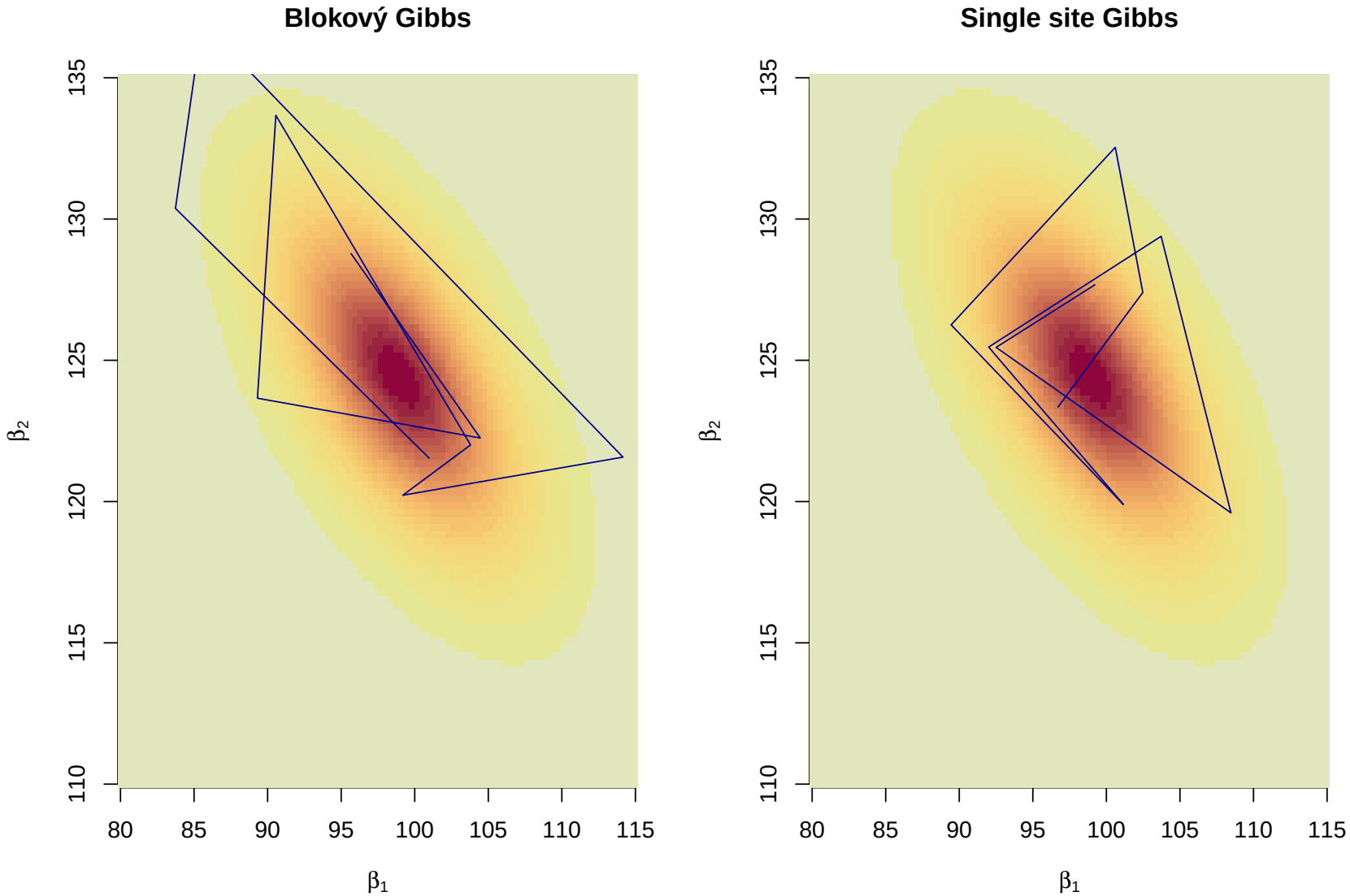






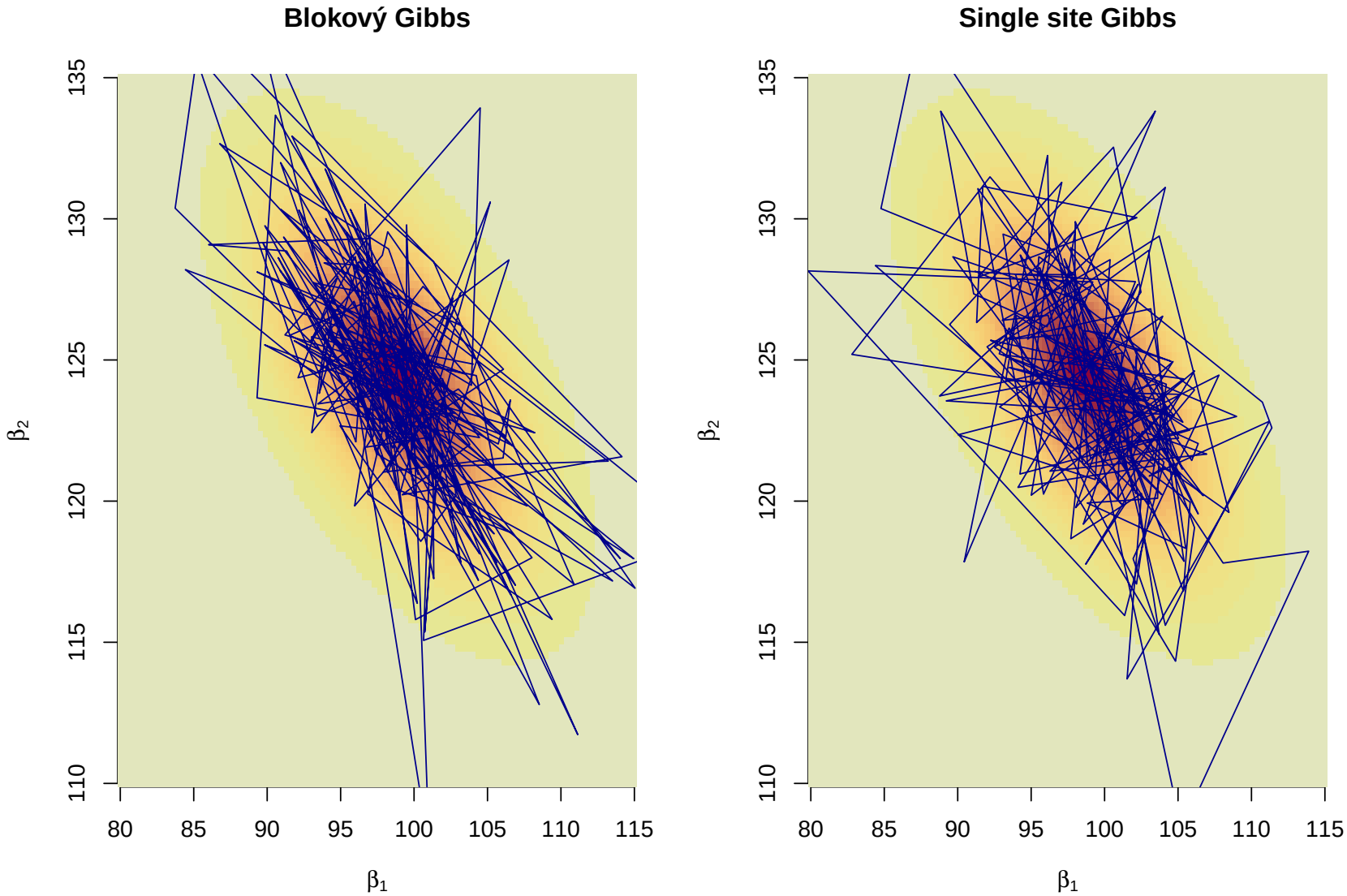
Příklad: Vážení lehkých objektů

Gibbsův algoritmus: Generované hodnoty β (iterace 101 – 110)



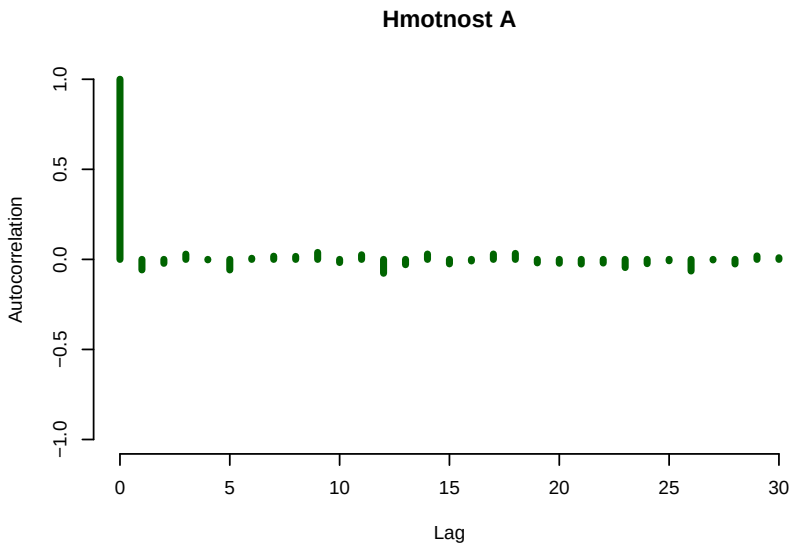
Příklad: Vážení lehkých objektů

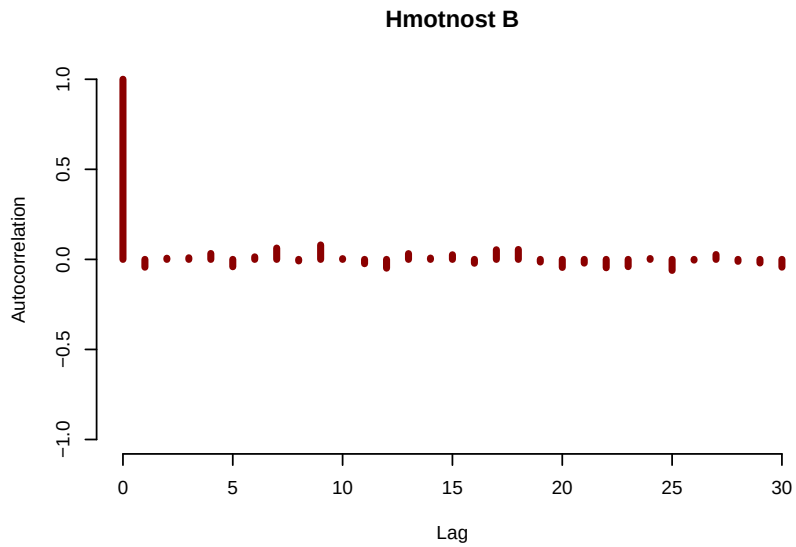
Gibbsův algoritmus: Generované hodnoty β (iterace 101 – 300)

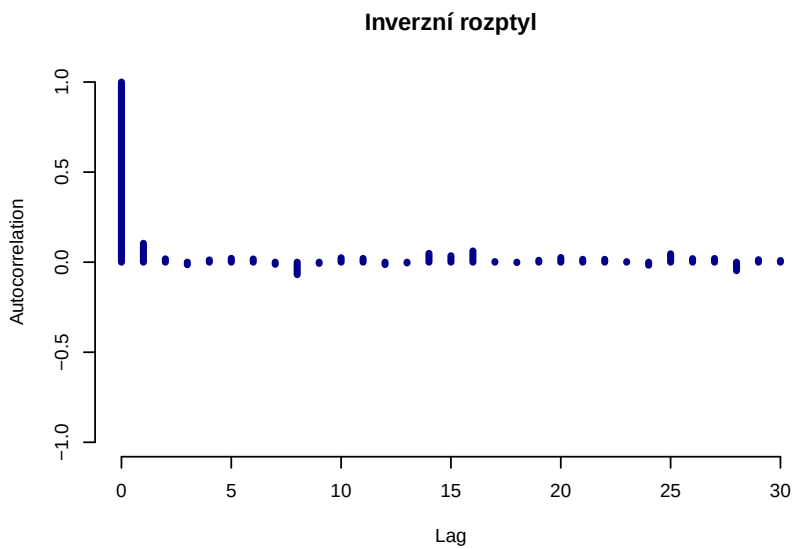


Příklad: Vážení lehkých objektů

Blokový Gibbsův algoritmus: Odhady autokorelačních funkcí (B=100, M=1 000)

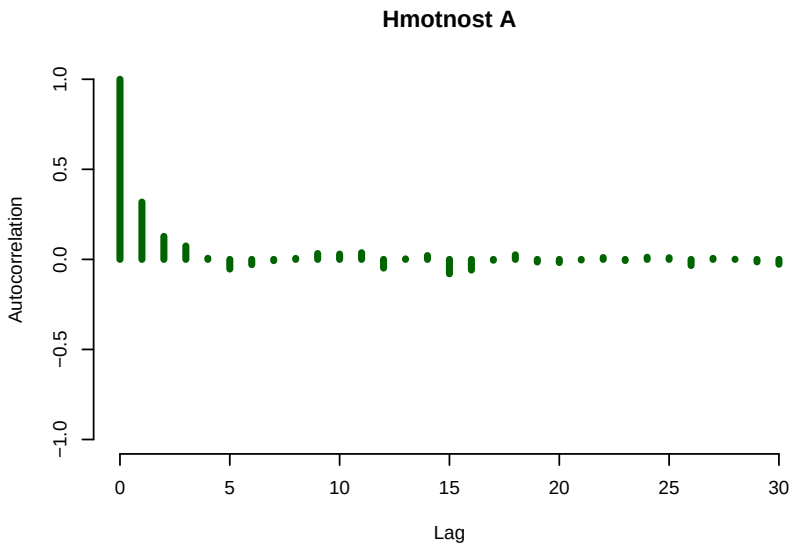


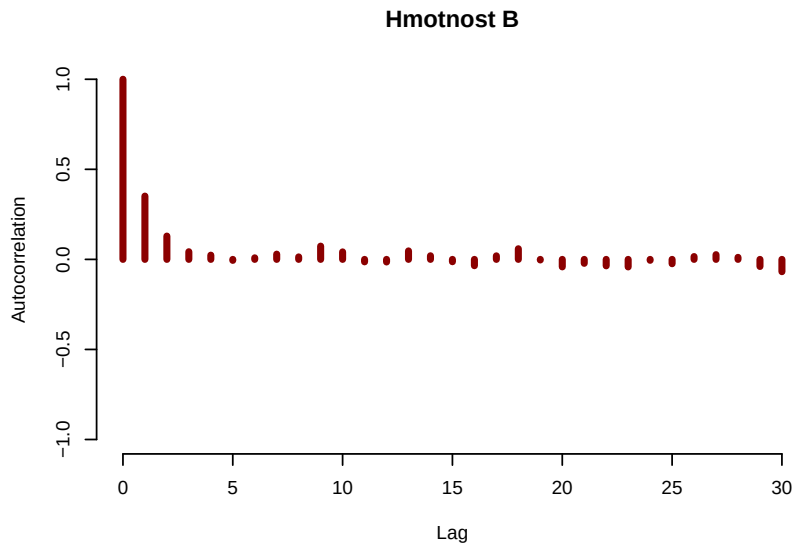


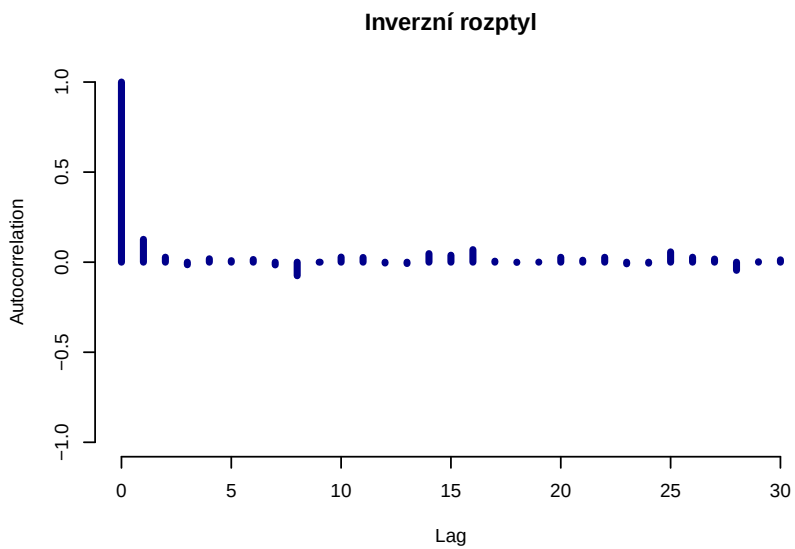


Příklad: Vážení lehkých objektů

Single site Gibbsův algoritmus: Odhady autokorelačních funkcí (B=100, M=1 000)

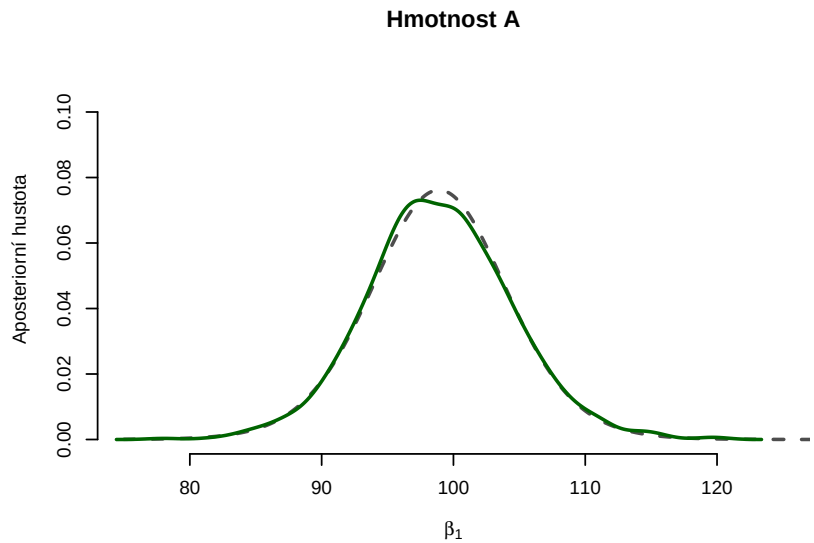


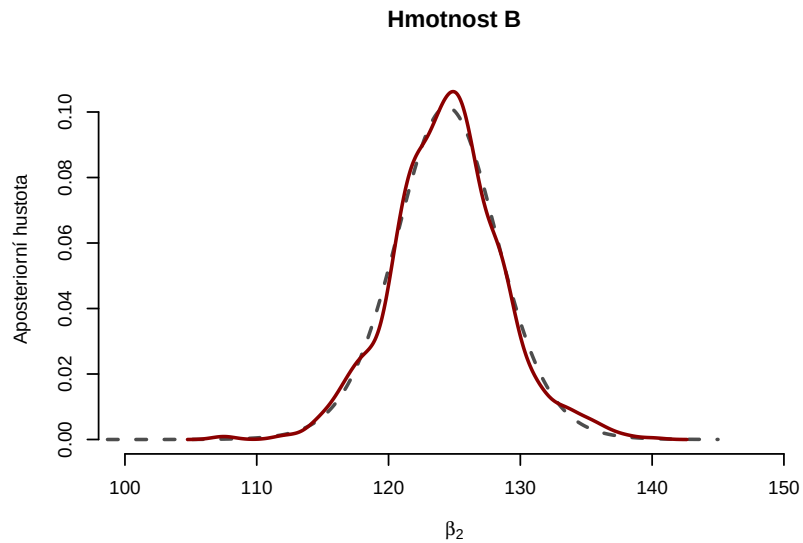


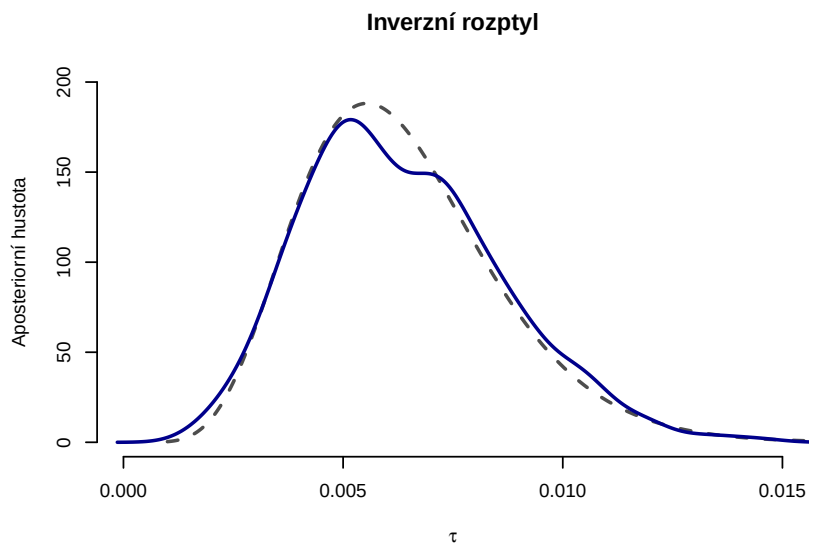


Příklad: Vážení lehkých objektů

Blokový Gibbsův algoritmus: Odhady aposteriorních hustot (B=100, M=1 000)

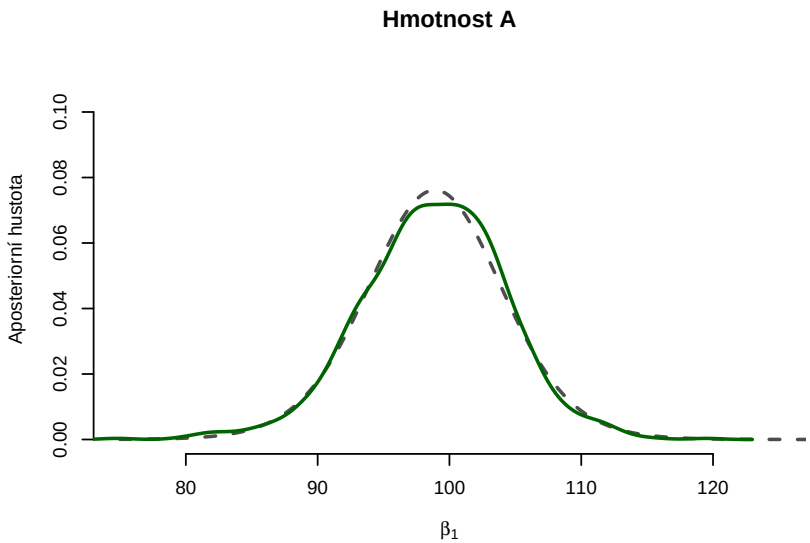


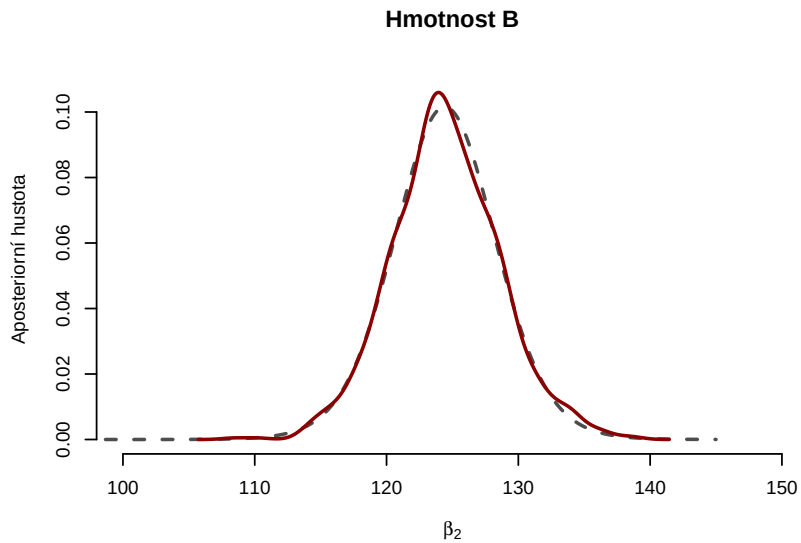


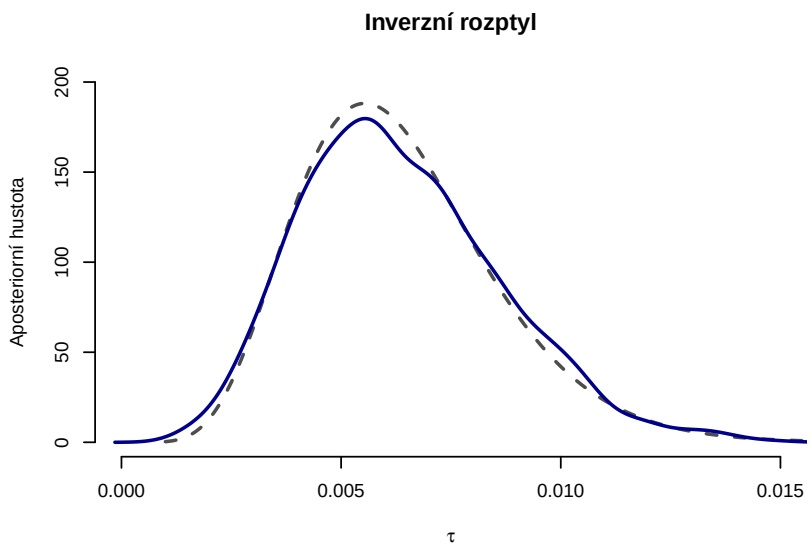


Příklad: Vážení lehkých objektů

Single site Gibbsův algoritmus: Odhady aposteriorních hustot (B=100, M=1 000)







Příklad: Vážení lehkých objektů

Aposteriorní inference pro β (B=100, M=1 000)

Blokový Gibbsův algoritmus

	β_1	β_2
Aposter. střední hodnota	98,8947	124,4211
MCMC odhad	98,9084	124,4561
MC chyba (naivní)	0,1744	0,1323
MC chyba	0,1662	0,1398
Aposter. medián	98,8947	124,4211
MCMC odhad	98,7209	124,4226
95% ET věr. interval	(87,9641; 109,8253)	(116,2231; 132,6190)
MCMC odhad	(86,9793; 110,2375)	(116,2702; 133,5293)
95% HPD věr. interval	(87,9641; 109,8253)	(116,2231; 132,6190)
MCMC odhad	(88,8421; 110,7950)	(114,8493; 132,0199)

Příklad: Vážení lehkých objektů

Aposteriorní inference pro β (B=100, M=1 000)

Single site Gibbsův algoritmus

	β_1	β_2
Aposter. střední hodnota	98,8947	124,4211
MCMC odhad	98,8051	124,4914
MC chyba (naivní)	0,1729	0,1302
MC chyba	0,2535	0,2015
Aposter. medián	98,8947	124,4211
MCMC odhad	98,9217	124,3193
95% ET věr. interval	(87,9641; 109,8253)	(116,2231; 132,6190)
MCMC odhad	(87,4650; 109,2495)	(116,3649; 133,4410)
95% HPD věr. interval	(87,9641; 109,8253)	(116,2231; 132,6190)
MCMC odhad	(87,8757; 109,4290)	(117,1579; 133,9186)

Oddíl **5.5**

Metropolisův-Hastingsův algoritmus

Metropolisův-Hastingsův algoritmus

- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., and Teller, A. H. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1091.
  - Aplikace ve statistické fyzice.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**(1), 97–109.
  - Zobecnění algoritmu.
  - Uvážení též čistě statistických problémů.

Předpoklady:

- Parametrický prostor Θ
- Cílové (stacionární) rozdělení je $f(d\boldsymbol{\theta})$ a má hustotu $f(\boldsymbol{\theta})$ vzhledem k σ-konečné míře λ s $\lambda(\Theta) > 0$.
- $\Theta = \{\boldsymbol{\theta} : f(\boldsymbol{\theta}) > 0\}$

Algoritmus

Algoritmus:

1. Zvol počáteční stav $\boldsymbol{\theta}^{(0)}$, polož $m = 0$.
2. Generuj návrh $\boldsymbol{\psi}$ z rozdělení $q(\boldsymbol{\theta}^{(m)}, d\boldsymbol{\psi})$ s hustotou $q(\boldsymbol{\theta}^{(m)}, \boldsymbol{\psi})$ (vzhledem k σ-konečné míře λ).
3. Spočti pravděpodobnost přijetí návrhu (*proposal acceptance probability*)

$$\alpha(\boldsymbol{\theta}^{(m)}, \boldsymbol{\psi}) = \begin{cases} \min\left\{ \dfrac{f(\boldsymbol{\psi})\, q(\boldsymbol{\psi}, \boldsymbol{\theta}^{(m)})}{f(\boldsymbol{\theta}^{(m)})\, q(\boldsymbol{\theta}^{(m)}, \boldsymbol{\psi})}, 1 \right\} & \text{pro } f(\boldsymbol{\theta}^{(m)})\, q(\boldsymbol{\theta}^{(m)}, \boldsymbol{\psi}) > 0, \\ 1 & \text{jinak.} \end{cases}$$

4. Generuj $U \sim \mathcal{U}(0, 1)$

$$\boldsymbol{\theta}^{(m+1)} = \begin{cases} \boldsymbol{\psi}, & \text{jestliže } U < \alpha(\boldsymbol{\theta}^{(m)}, \boldsymbol{\psi}), \\ \boldsymbol{\theta}^{(m)}, & \text{jestliže } U \geq \alpha(\boldsymbol{\theta}^{(m)}, \boldsymbol{\psi}). \end{cases}$$

5. Zvětši m o jedničku a jdi na 2. krok algoritmu.

Poznámky

Poznámky:

- Pro aplikaci MH algoritmu není potřeba znát normující konstantu cílové hustoty $f(\boldsymbol{\theta})$.
  - Ideální pro použití v bayesovské statistice.
- Návrhová hustota $q(\boldsymbol{\theta}, \boldsymbol{\psi})$ může být libovolná.
  - Nevhodná volba $q(\boldsymbol{\theta}, \boldsymbol{\psi})$ však vede k vysoké autokorelaci a s tím spojené neefektivitě.
  - Příliš "ambiciózní" $q(\boldsymbol{\theta}, \boldsymbol{\psi})$ vede k malým pravděpodobnostem přijetí návrhu a řetězec pak dlouho setrvává v jednom stavu

    ⇛ vysoká autokorelace.
  - Příliš "opatrné" $q(\boldsymbol{\theta}, \boldsymbol{\psi})$ vede sice k vysokým pravděpodobnostem přijetí návrhu, ale řetězec se přesouvá jenom velice pomalu

    ⇛ vysoká autokorelace.
- Optimální proporce přijatých návrhů *(acceptance rate)* závisí na konkrétní situaci.

Metropolisův-Hastingsův algoritmus

Poznámky

- Symetrická návrhová hustota, tj. $q(\boldsymbol{\theta}, \boldsymbol{\psi}) = q(\boldsymbol{\psi}, \boldsymbol{\theta}) \quad \forall \boldsymbol{\theta}, \boldsymbol{\psi} \in \Theta$
  ➠ Metropolisův algoritmus.

- Hlavní část pravděpodobnosti přijetí

$$\alpha^*(\boldsymbol{\theta}^{(m)}, \boldsymbol{\psi}) = \frac{f(\boldsymbol{\psi})\, q(\boldsymbol{\psi}, \boldsymbol{\theta}^{(m)})}{f(\boldsymbol{\theta}^{(m)})\, q(\boldsymbol{\theta}^{(m)}, \boldsymbol{\psi})}$$

obvykle počítáme v logaritmickém měřítku, tj.

$$\log\{\alpha^*(\boldsymbol{\theta}^{(m)}, \boldsymbol{\psi})\} = \log\{f(\boldsymbol{\psi})\} + \log\{q(\boldsymbol{\psi}, \boldsymbol{\theta}^{(m)})\} - \log\{f(\boldsymbol{\theta}^{(m)})\} - \log\{q(\boldsymbol{\theta}^{(m)}, \boldsymbol{\psi})\}$$

➠ vyhneme se mnoha numerickým obtížím při počítání s čísly, jež mohou být blízká nule.

Návrhové rozdělení

Možné volby **návrhových rozdělení (*proposal distribution*)**

- **Nezávislý výběr (*independent sampler*)**

$$q(\boldsymbol{\theta}, \boldsymbol{\psi}) = q_0(\boldsymbol{\psi}) \qquad \forall \boldsymbol{\theta} \in \Theta.$$

- q_0 : nějaká hustota vzhledem k σ-konečné míře λ s nosičem na Θ.
- Návrhová hustota nezávisí na současném stavu.
- Za q_0 je vhodné volit rozdělení s těžšími chvosty (vícerozměrné t-rozdělení, ...).
- Ideální stav: $q_0(\boldsymbol{\psi}) = f(\boldsymbol{\psi})$

  ➠ generujeme přímo náhodný výběr z cílového rozdělení $f(d\boldsymbol{\theta})$.

Návrhové rozdělení

Možné volby **návrhových rozdělení (*proposal distribution*)**

- **Náhodná procházka (*random walk*)**

$$q(\boldsymbol{\theta}, \boldsymbol{\psi}) = q_0(\boldsymbol{\psi} - \boldsymbol{\theta}) \qquad \forall \boldsymbol{\theta} \in \Theta.$$

- q_0 : nějaká hustota vzhledem k σ-konečné míře λ s nosičem na Θ.
- Návrh: $\boldsymbol{\psi} = \boldsymbol{\theta} + \boldsymbol{Z}$, kde $\boldsymbol{Z}$ má hustotu q_0.
- Častá volba: $q_0 \equiv$ (vícerozměrné) normální, respektive t-rozdělení s nulovou střední hodnotou a obvykle diagonální varianční/měřítkovou maticí.

➠ Potřeba vhodně zvolit rozptyly.

- Je-li q_0 symetrická (tj. $q_0(\boldsymbol{z}) = q_0(-\boldsymbol{z})$), potom $q(\boldsymbol{\theta}, \boldsymbol{\psi}) = q(\boldsymbol{\psi}, \boldsymbol{\theta})$ a při počítání pravděpodobnosti přijetí nemusíme vůbec počítat hodnoty hustoty q_0 (resp. návrhové hustoty q).

Věta 5.6 .

Rozdělení $f(d\theta)$ je stacionárním rozdělením markovského řetězce generovaného Metropolisovým-Hastingsovým algoritmem.

Důkaz. Viz přednáška NMTP539 Metody Markov Chain Monte Carlo. ❑

- Pokud bude existovat limitní rozdělení, musí se jednat o rozdělení stacionární a tedy cílové $f(d\theta)$.

Metropolisův-Hastingsův algoritmus

Existence limitního rozdělení, ergodicita

- Pro důkaz ergodicity (existence limitního rozdělení) je potřeba učinit několik předpokladů o návrhové hustotě q.
- Ergodicita je např, zajištěna v případě, kdy

$$q(\boldsymbol{\theta}, \boldsymbol{\psi}) = q_0(\boldsymbol{\psi} - \boldsymbol{\theta}), \qquad q_0(\boldsymbol{z}) = q_0(-\boldsymbol{z})$$

(symetrická náhodná procházka)

a $q_0(\boldsymbol{z}) > 0$ pro všechna $\mathbb{R}^d$ (např. vícerozměrné normální nebo t-rozdělení).

- Podrobnosti viz přednáška NMTP539.

Oddíl **5.6**

Hybridní algoritmy

Hybridní algoritmy

V Bayesovských aplikacích jsme obvykle v následující situaci.

- $\Theta = \prod_{i=1}^{k} \Theta_i$, $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_k^\top\right)^\top$.
- Cílové (stacionární) rozdělení je $f(d\boldsymbol{\theta})$ a má hustotu $f(\boldsymbol{\theta})$ vzhledem k součinové míře $\lambda_1 \otimes \cdots \otimes \lambda_k$, přičemž λ_i je σ-konečná míra s $\lambda_i(\Theta_i) > 0$ $(i = 1, \ldots, k)$.
- $\Theta = \left\{\boldsymbol{\theta} : f(\boldsymbol{\theta}) > 0\right\}$.
- Jsme schopni snadno vyjádřit všechna plně podmíněná rozdělení $f\left(d\boldsymbol{\theta}_i \,\middle|\, \boldsymbol{\theta}_{-i}\right)$ až na multiplikativní konstantu, tj. víme, že pro hustoty platí

$$f\left(\boldsymbol{\theta}_i \,\middle|\, \boldsymbol{\theta}_{-i}\right) \propto f^*\left(\boldsymbol{\theta}_i \,\middle|\, \boldsymbol{\theta}_{-i}\right),$$

  přičemž funkci $f^*\left(\cdot \,\middle|\, \boldsymbol{\theta}_{-i}\right)$ jsme schopni vyjádřit analyticky.

Hybridní algoritmy

- Markovský řetězec se stacionárním rozdělením $f(d\theta)$ lze sestavit tak, že jako základ vezmeme Gibbsův algoritmus a v případě, že pro nějaké i nejsme schopni (snadno) generovat z plně podmíněného rozdělení $f\left(d\theta_i \,\middle|\, \theta_{-i}\right)$, generujeme θ_i pomocí Metropolisova-Hastingsova algoritmu

  ➠ *Metropolis within Gibbs algorithm*.

- Celková přechodová hustota je součinem přechodových hustot pro Gibbsův, resp. Metropolisův-Hastingsův algoritmus.

- K tomu, aby celková procedura vedla k ergodickému řetězci s požadovaným limitním rozdělením $f(d\theta)$ stačí, aby přechodová hustota v každém kroku vedla k ergodickému řetězci s požadovaným limitním rozdělením.

Hybridní algoritmy

Některé další metody pro generování z plně podmíněných rozdělení

- **Adaptive rejection sampling (ARS)**
  - Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337–348.
  - Potřeba, aby hustota, z které chceme generovat byla log-konkávní.
  - ⇛ Poměrně častý případ, viz rozdělení z exponenciální třídy rozdělení.
- **Adaptive rejection Metropolis sampling (ARMS)**
  - Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics*, **44**, 455–472.
  - Zobecnění ARS metody na situace, kdy hustota, z které generujeme není log-konkávní.
- **Slice sampling**
  - Neal, R. M. (2003). Slice sampling (with Discussion). *The Annals of Statistics*, **31**, 705–767.
  - Efektivní v případě, že hustota, z které generujeme je unimodální.

Příklad: Lineární model s neinformativním apriorním rozdělením

Lineární model

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathbb{X} : \text{pevná matice } n \times k,$$
$$\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

- Parametry: $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top, \tau\right)^\top$, kde $\tau = \sigma^{-2} > 0$.
- Věrohodnost: $\boldsymbol{Y} \sim \mathcal{N}_n(\mathbb{X}\boldsymbol{\beta}, \tau^{-1}\mathbf{I}_n)$.
- Neinformativní apriorní rozdělení:

$$p(\boldsymbol{\beta}) \propto 1, \qquad \boldsymbol{\beta} \in \mathbb{R}^k,$$
$$p(\tau) \propto \frac{1}{\tau}, \qquad \tau > 0.$$

Příklad: Lineární model s neinformativním apriorním rozdělením

Věrohodnost

- Označme: $\boldsymbol{b} = \left(\mathbb{X}^\top \mathbb{X}\right)^{-1} \mathbb{X}^\top \boldsymbol{y}$,

  $\mathrm{SS}_e = \left\| \boldsymbol{y} - \mathbb{X}\boldsymbol{b} \right\|^2$.

- Věrohodnost:

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= p(\boldsymbol{y} \mid \boldsymbol{\beta}, \tau) \\
&= (2\pi)^{-\frac{n}{2}} \tau^{\frac{n}{2}} \exp\left[-\frac{\tau}{2} \left\{ \mathrm{SS}_e + (\boldsymbol{\beta} - \boldsymbol{b})^\top \mathbb{X}^\top \mathbb{X} (\boldsymbol{\beta} - \boldsymbol{b}) \right\} \right] \\
&= (2\pi)^{-\frac{n}{2}} \tau^{\frac{n}{2}} \exp\left\{ -\frac{\tau}{2} \left\| \boldsymbol{y} - \mathbb{X}\boldsymbol{\beta} \right\|^2 \right\}.
\end{aligned}
$$

Příklad: Lineární model s neinformativním apriorním rozdělením

Aposteriorní rozdělení

- Bylo odvozeno:

$$p(\boldsymbol{\beta}, \tau \mid \boldsymbol{y}) = p(\boldsymbol{\beta} \mid \tau, \boldsymbol{y}) \times p(\tau \mid \boldsymbol{y}),$$

kde $p(\tau \mid \boldsymbol{y}) \sim \mathcal{G}\left(\frac{n-k}{2}, \frac{\mathrm{SS}_e}{2}\right)$,

$p(\boldsymbol{\beta} \mid \tau, \boldsymbol{y}) \sim \mathcal{N}_k\left(\boldsymbol{b}, \tau^{-1}\left(\mathbb{X}^\top \mathbb{X}\right)^{-1}\right)$.

- Dále bylo odvozeno: $p(\boldsymbol{\beta} \mid \boldsymbol{y}) \sim \mathrm{MVT}_{k,\, n-k}\left(\boldsymbol{b}, \frac{\mathrm{SS}_e}{n-k}\left(\mathbb{X}^\top \mathbb{X}\right)^{-1}\right)$
- Pomocí algoritmu Metropolis within Gibbs algoritmu sestrojíme markovský řetězec, který bude mít rozdělení $p(\boldsymbol{\beta}, \tau \mid \boldsymbol{y})$ jako stacionární i limitní.

Příklad: Lineární model s neinformativním apriorním rozdělením

Plně podmíněná rozdělení

$$p(\boldsymbol{\beta} \mid \cdots) = p(\boldsymbol{\beta} \mid \tau, \boldsymbol{y}) \sim \mathcal{N}_k\left(\boldsymbol{b}, \tau^{-1}\left(\mathbb{X}^\top \mathbb{X}\right)^{-1}\right),$$

$$p(\tau \mid \cdots) = p(\tau \mid \boldsymbol{\beta}, \boldsymbol{y}) \sim \mathcal{G}\left(\frac{n}{2}, \frac{\left\|\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta}\right\|^2}{2}\right).$$

Příklad: Lineární model s neinformativním apriorním rozdělením

Metropolis within Gibbs algoritmus

- Regresní parametry β budeme generovat pomocí symetrické náhodné procházky s návrhovou hustotou

$$q(\beta_1, \beta_2) = q_0(\beta_2 - \beta_1),$$

$$\text{kde } q_0 \sim \mathcal{N}_k(\mathbf{0}, \mathbb{D}_{prop}), \qquad \mathbb{D}_{prop} = \text{diag}(d^2_{1,prop}, \ldots, d^2_{p,prop}).$$

  - V kroku $m + 1$ algoritmu navrhneme β_{prop} vygenerované z rozdělení $\mathcal{N}_k(\beta^{(m)}, \mathbb{D}_{prop})$.

- Hlavní část pravděpodobnosti přijetetí návrhu je

$$\alpha^*(\beta^{(m)}, \beta_{prop}) = \frac{p(\beta_{prop} \mid \cdots)\, q(\beta_{prop}, \beta^{(m)})}{p(\beta^{(m)} \mid \cdots)\, q(\beta^{(m)}, \beta_{prop})} = \frac{p(\beta_{prop} \mid \cdots)}{p(\beta^{(m)} \mid \cdots)}$$

$$= \exp\left[-\frac{\tau^{(m)}}{2}\left\{(\beta_{prop} - \boldsymbol{b})^\top \mathbb{X}^\top \mathbb{X}(\beta_{prop} - \boldsymbol{b}) - (\beta^{(m)} - \boldsymbol{b})^\top \mathbb{X}^\top \mathbb{X}(\beta^{(m)} - \boldsymbol{b})\right\}\right].$$

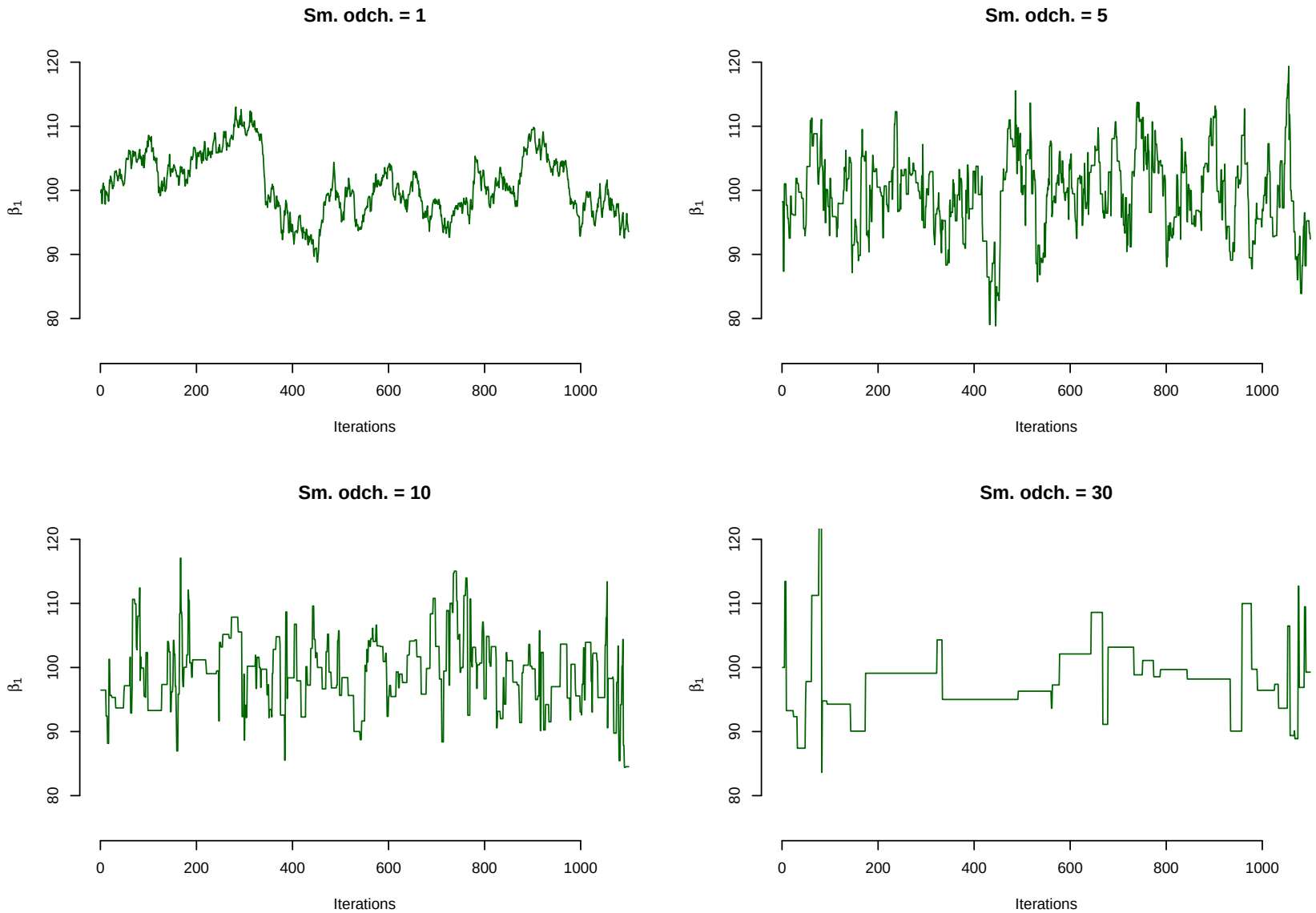
Příklad: Vážení lehkých objektů

Metropolis within Gibbs algoritmus

- Budou čtyři ukázky vygenerované při použití různých variančních matic v návrhové normální hustotě.
  1. $\mathbb{D}_{prop} = \text{diag}(1, 1)$
     ➠ proporce přijatých návrhů 0,87.
  2. $\mathbb{D}_{prop} = \text{diag}(5^2, 5^2)$
     ➠ proporce přijatých návrhů 0,47.
  3. $\mathbb{D}_{prop} = \text{diag}(10^2, 10^2)$
     ➠ proporce přijatých návrhů 0,22.
  4. $\mathbb{D}_{prop} = \text{diag}(30^2, 30^2)$
     ➠ proporce přijatých návrhů 0,04.

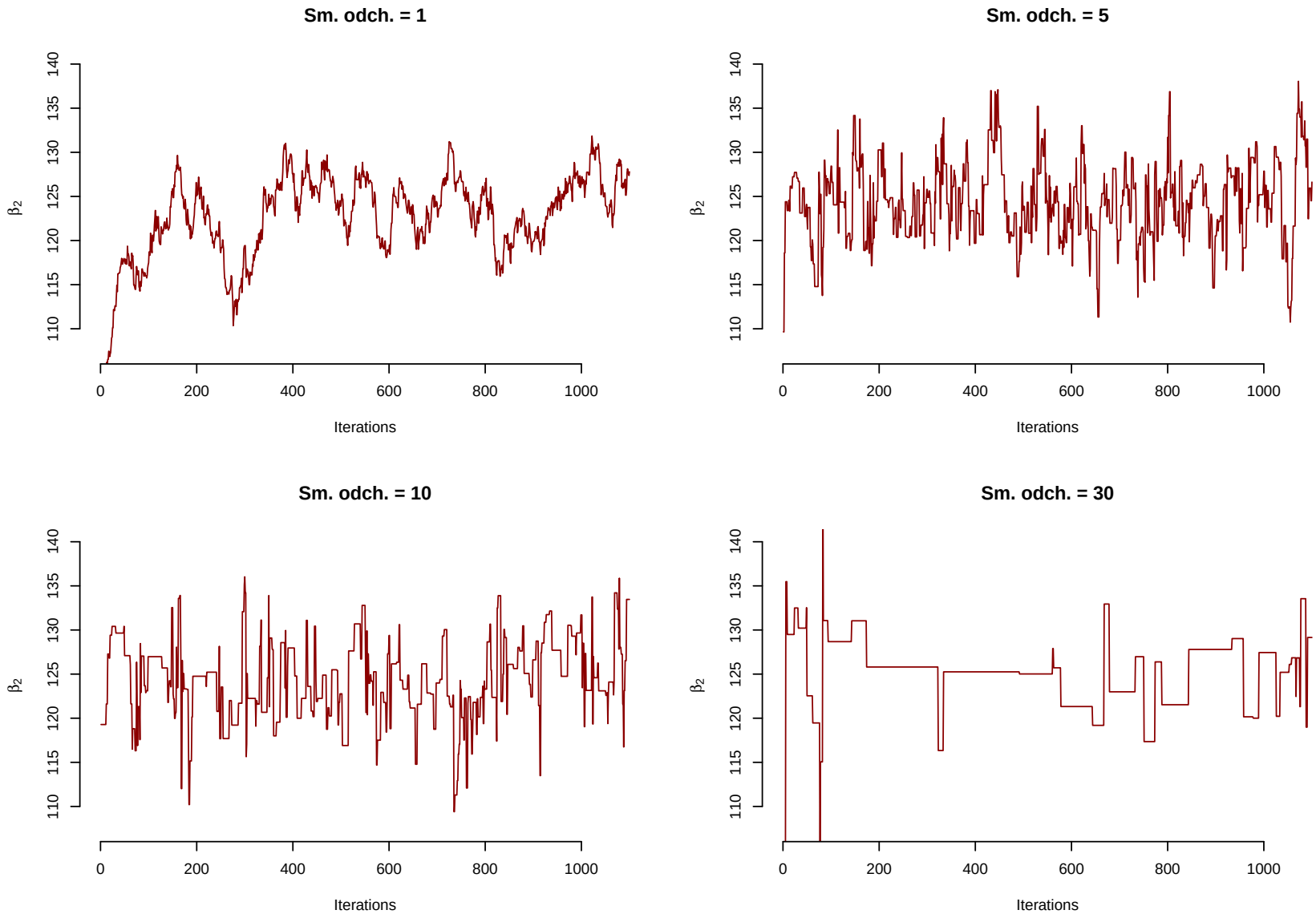
Příklad: Vážení lehkých objektů

Metropolis within Gibbs algoritmus: Generované hodnoty β_1 (B=100, M=1 000)



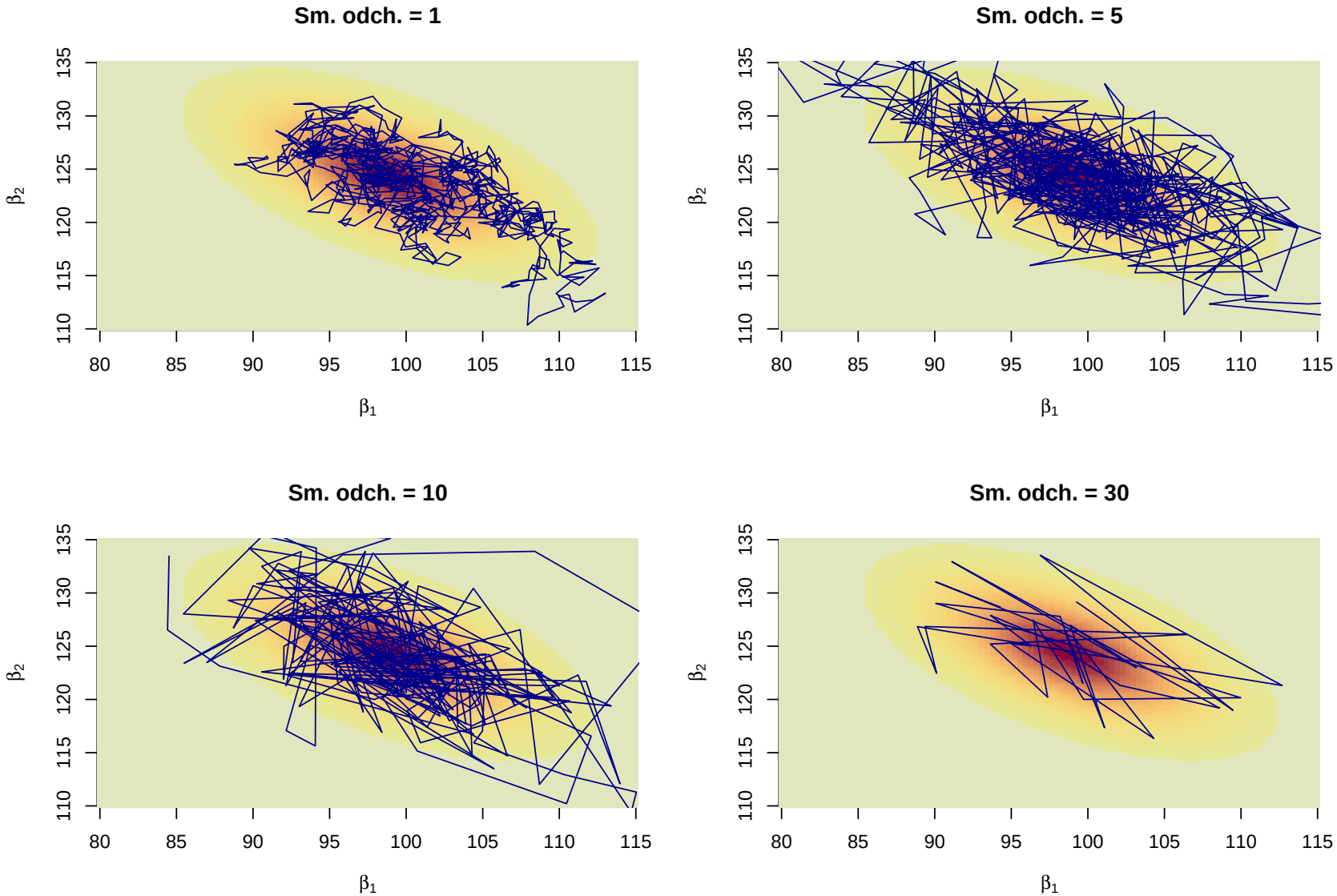
Příklad: Vážení lehkých objektů

Metropolis within Gibbs algoritmus: Generované hodnoty β_2 (B=100, M=1 000)



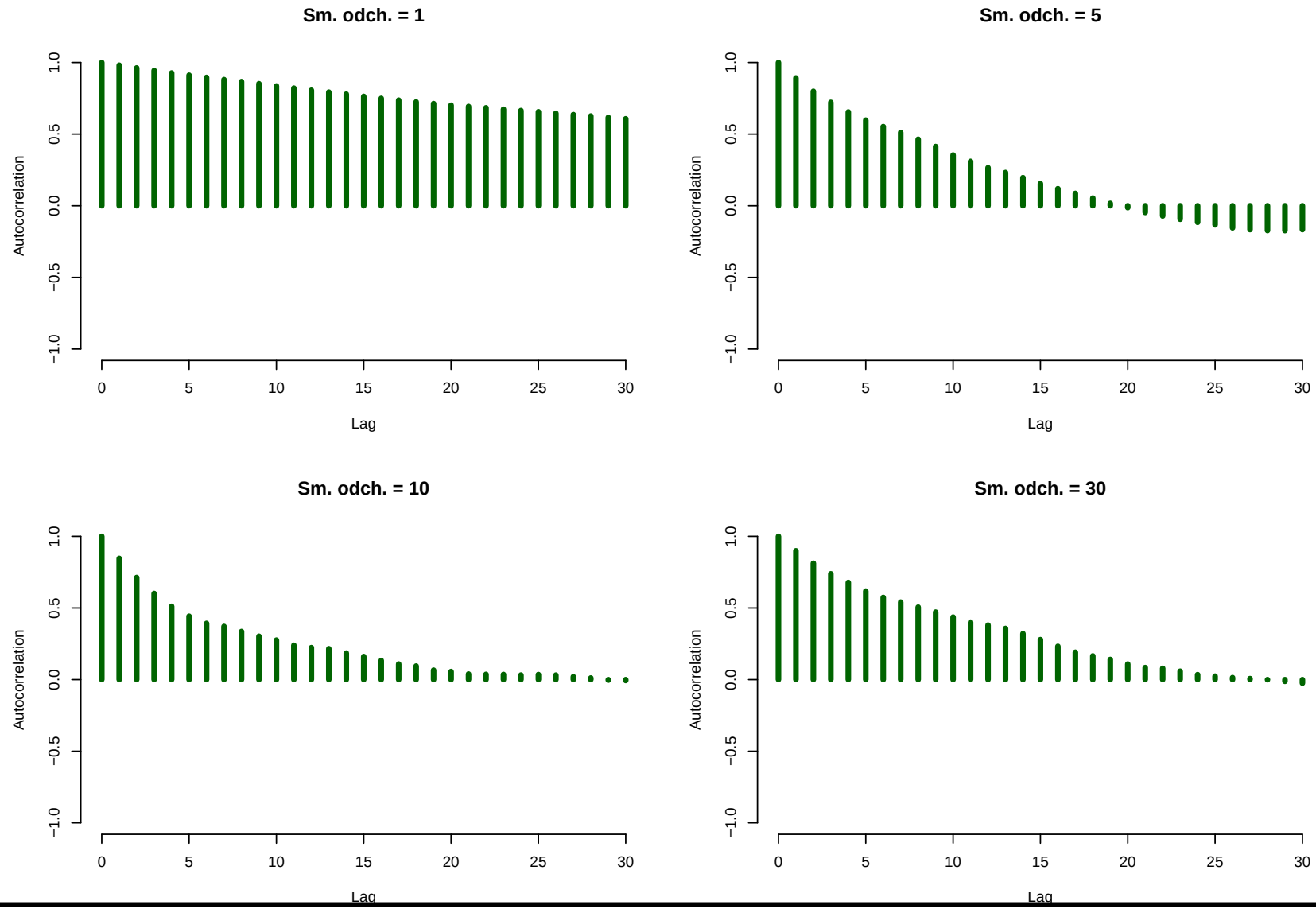
Příklad: Vážení lehkých objektů

Metropolis within Gibbs algoritmus: Generované hodnoty β (iterace 101 – 1 100)



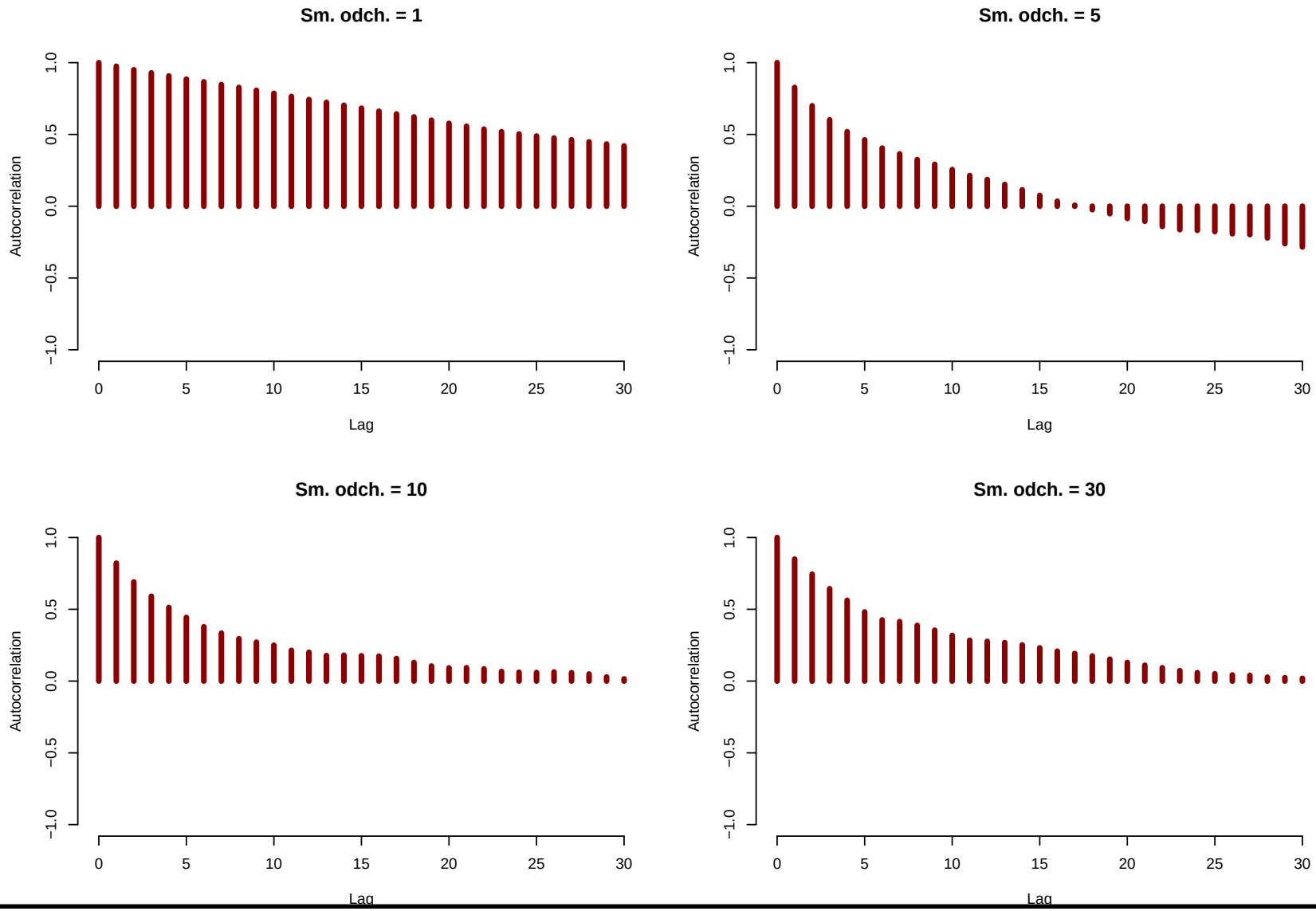
Příklad: Vážení lehkých objektů

Metropolis within Gibbs algoritmus: Odhady autokorelačních funkcí pro β_1 (B=100, M=1 000)



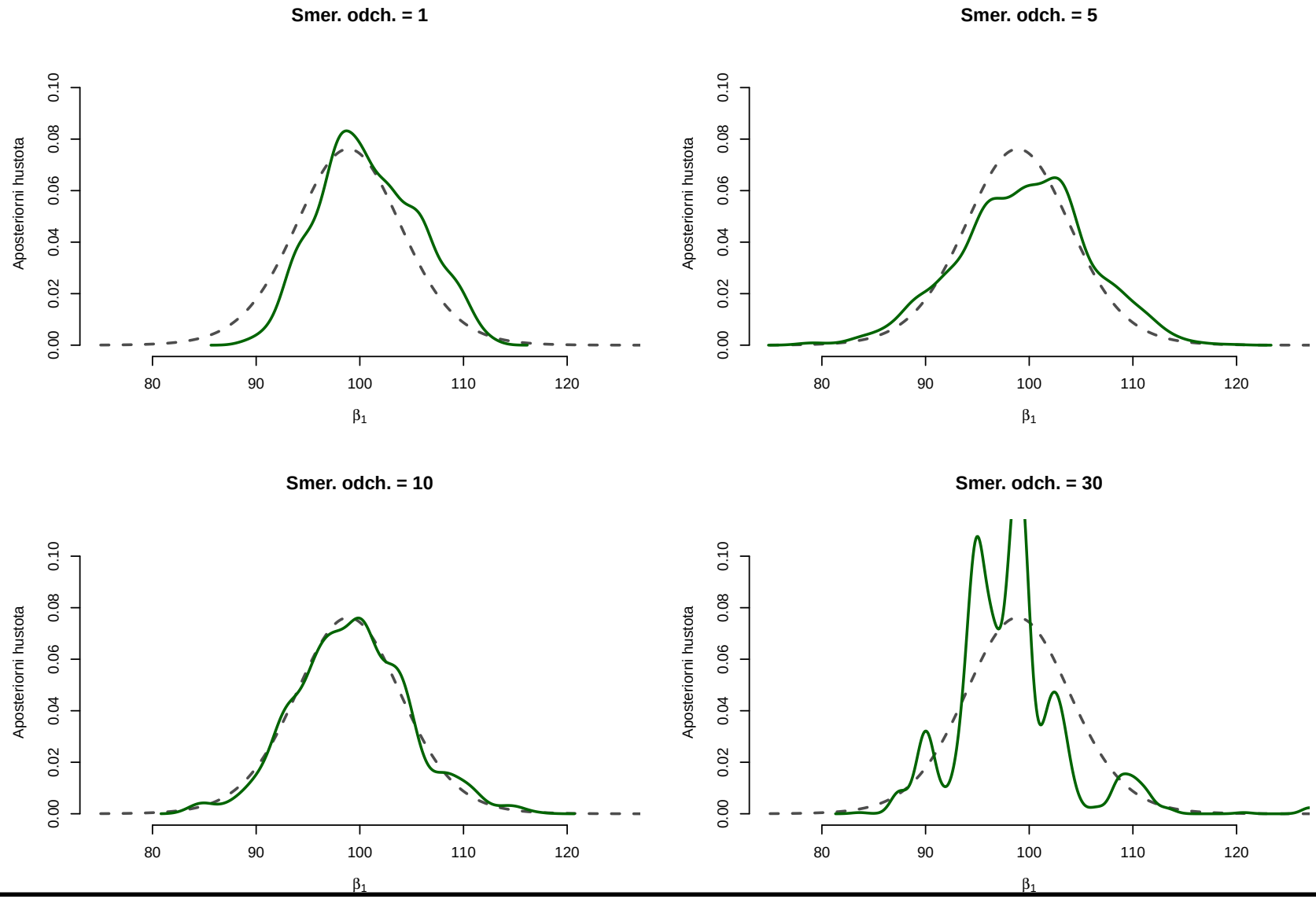
Příklad: Vážení lehkých objektů

Metropolis within Gibbs algoritmus: Odhady autokorelačních funkcí pro β_2 (B=100, M=1 000)



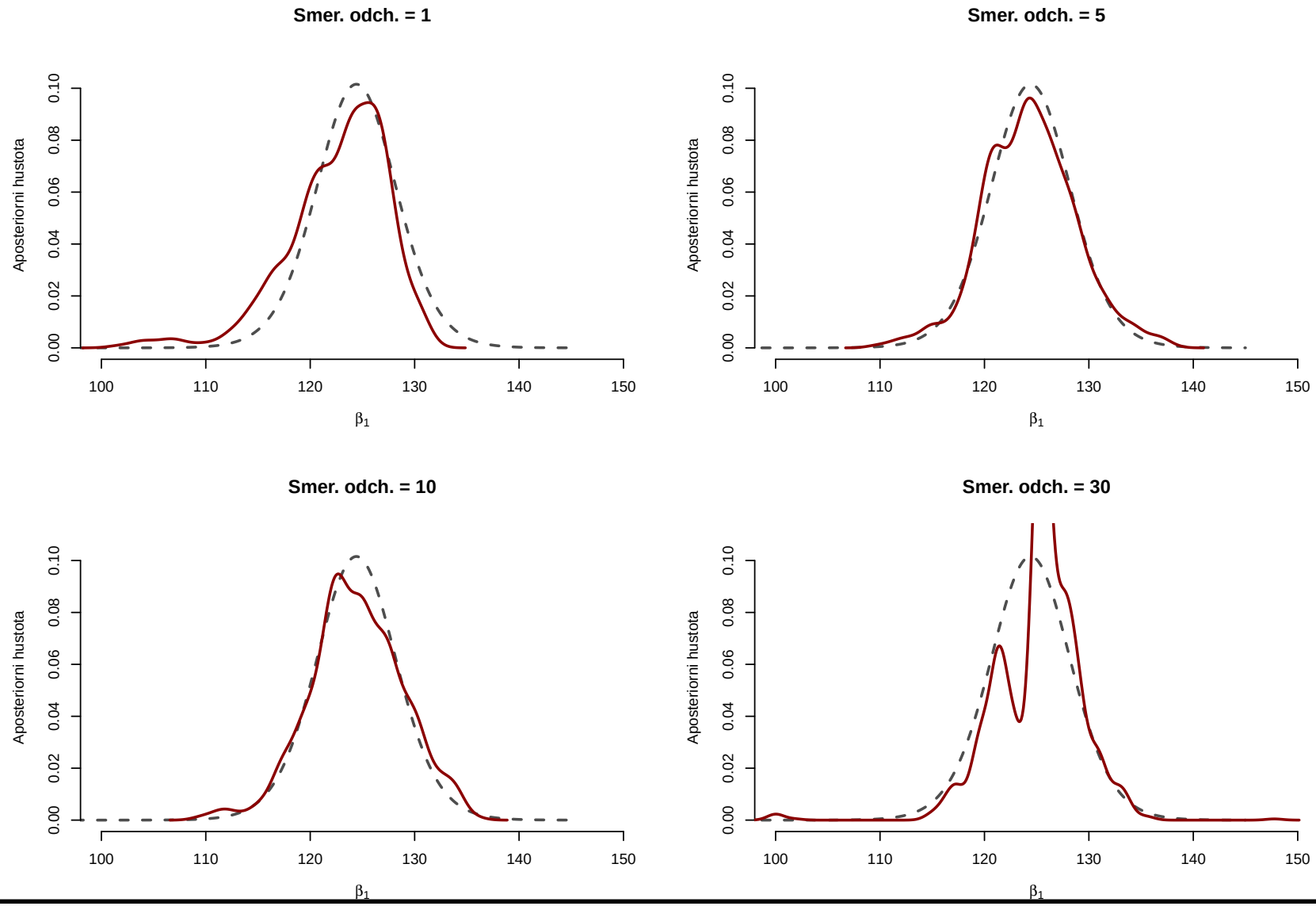
Příklad: Vážení lehkých objektů

Metropolis within Gibbs algoritmus: Odhady aposteriorních hustot pro β_1 (B=100, M=1 000)



Příklad: Vážení lehkých objektů

Metropolis within Gibbs algoritmus: Odhady aposteriorních hustot pro β_2 (B=100, M=1 000)



6

Bayesovské rozšiřování dat

Oddíl **6.1**

Úvod

Rozšiřování dat (data augmentation)

- V bayesovské statistice potřebujeme typicky generovat z aposteriorního rozdělení s hustotou

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{L_{obs}(\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int_{\Theta} L_{obs}(\boldsymbol{\theta})\, p(\boldsymbol{\theta}) d\lambda(\boldsymbol{\theta})} \propto L_{obs}(\boldsymbol{\theta})\, p(\boldsymbol{\theta})$$

  vzhledem k σ-konečné míře λ kde
  - $L_{obs}(\boldsymbol{\theta}) = p(\boldsymbol{y} \mid \boldsymbol{\theta})$: věrohodnost (pozorovaných dat)
  - $p(\boldsymbol{\theta})$: apriorní rozdělení
- Pro použití MCMC metod obvykle:
  - není potřeba znát normující konstantu $\int_{\Theta} L_{obs}(\boldsymbol{\theta})\, p(\boldsymbol{\theta}) d\lambda(\boldsymbol{\theta})$;
  - je však vhodné a výhodné, aby výraz $L_{obs}(\boldsymbol{\theta})\, p(\boldsymbol{\theta})$ byl snadno spočítatelný pro libovolné $\boldsymbol{\theta} \in \Theta$.
- Často se však stává, že zejména $L_{obs}(\boldsymbol{\theta})$ není jednoduchého tvaru.
  - nejčastější komplikace: při vyjadřování $L_{obs}(\boldsymbol{\theta})$ je nutné integrovat.

Příklad 1: Lineární smíšený model

Model (pro $i = 1, \ldots, N$)

- $\boldsymbol{Y}_i \mid \boldsymbol{b}_i$ nezávislé s rozdělením $\mathcal{N}_{n_i}(\mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\boldsymbol{b}_i,\ \sigma^2\mathbf{I}_{n_i})$
- $\boldsymbol{b}_i$ i.i.d. s rozdělením $\mathcal{N}_q(\boldsymbol{\mu},\ \mathbb{D})$

Parametry

- $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top,\ \boldsymbol{\mu}^\top,\ \sigma^2,\ \mathrm{vec}(\mathbb{D})\right)^\top$

Věrohodnost (pozorovaných dat)

$$L_{obs}(\boldsymbol{\theta}) = p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{y}_i \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} \int_{\mathbb{R}^q} p(\boldsymbol{y}_i \mid \boldsymbol{b}_i,\ \boldsymbol{\theta})\, p(\boldsymbol{b}_i \mid \boldsymbol{\theta}) d\boldsymbol{b}_i$$

$$= \prod_{i=1}^{N} \int_{\mathbb{R}^q} \varphi(\boldsymbol{y}_i \mid \mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\boldsymbol{b}_i,\ \sigma^2\mathbf{I}_{n_i})\, \varphi(\boldsymbol{b}_i \mid \boldsymbol{\mu},\ \mathbb{D}) d\boldsymbol{b}_i$$

Příklad 2: Logistická regrese s normálně rozdělenými náhodnými efekty

Model ($j = 1, \ldots, n_i$ pro každé $i = 1, \ldots, N$)

- $Y_{i,j} \mid b_i$ nezávislé s rozdělením $\mathcal{A}(\pi_i)$, kde $\pi_i = \frac{e^{b_i}}{1+e^{b_i}}$.
- b_i i.i.d. s rozdělením $\mathcal{N}(\mu, d^2)$.

Parametry

- $\boldsymbol{\theta} = (\mu, d^2)^\top$.

Věrohodnost (pozorovaných dat)

$$L_{obs}(\boldsymbol{\theta}) = p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{y}_i \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} \int_{\mathbb{R}} \prod_{j=1}^{n_i} p(y_{i,j} \mid b_i, \boldsymbol{\theta})\, p(b_i \mid \boldsymbol{\theta}) db_i$$

$$= \prod_{i=1}^{N} \int_{\mathbb{R}} \pi_i^{\sum_{j=1}^{n_i} y_{i,j}} (1-\pi_i)^{n_i - \sum_{j=1}^{n_i} y_{i,j}}\, \varphi(b_i \mid \mu, d^2) db_i$$

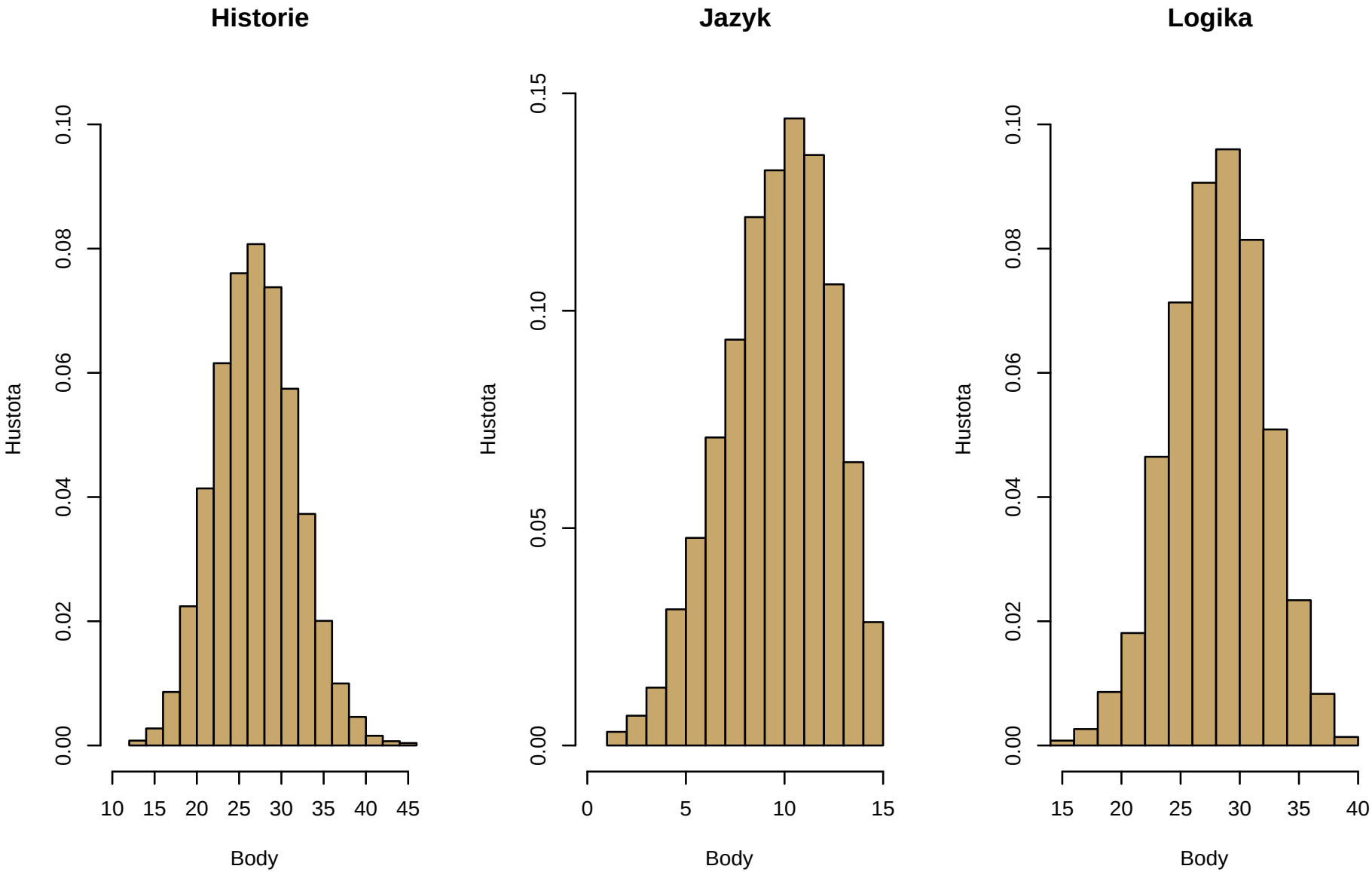
$$= \prod_{i=1}^{N} \int_{\mathbb{R}} \frac{e^{b_i \sum_{j=1}^{n_i} y_{i,j}}}{(1+e^{b_i})^{n_i}}\, \varphi(b_i \mid \mu, d^2) db_i.$$

Příklad 3: Výsledky přijímacích zkoušek

- V roce 1999 pořádala jistá (právnická) fakulta jedné české VŠ (nebylo to v Plzni) 11 řádných a jeden náhradní termín (termín č. 12) přijímaček.
- Někdo si povšimnul, že u 12. náhradního termínu byla proporce přijatých studentů mnohem vyšší než u všech předchozích termínů.
  - Chytří studenti se hromadně omlouvali z řádného termínu přijímaček a přišli až na ten náhradní?
  - Výrazně více stimulující ovzduší v učebnách během 12. termínu?
  - Zázrak?
- Ukázalo, že zadání otázek pro 12. termín přijímaček záhadně uniklo (z uzamčeného trezoru) a bylo (v určitých kruzích) ke koupi již před konáním tohoto termínu.
- Celá událost byla dle tehdejšího rektora dané VŠ i děkana dotčené fakulty dílem *gangsterské mafie* stojící mimo fakultu.
  Viz http://www.cibulka.com/nnoviny/nn2000/nn1900/obsah/05.htm.

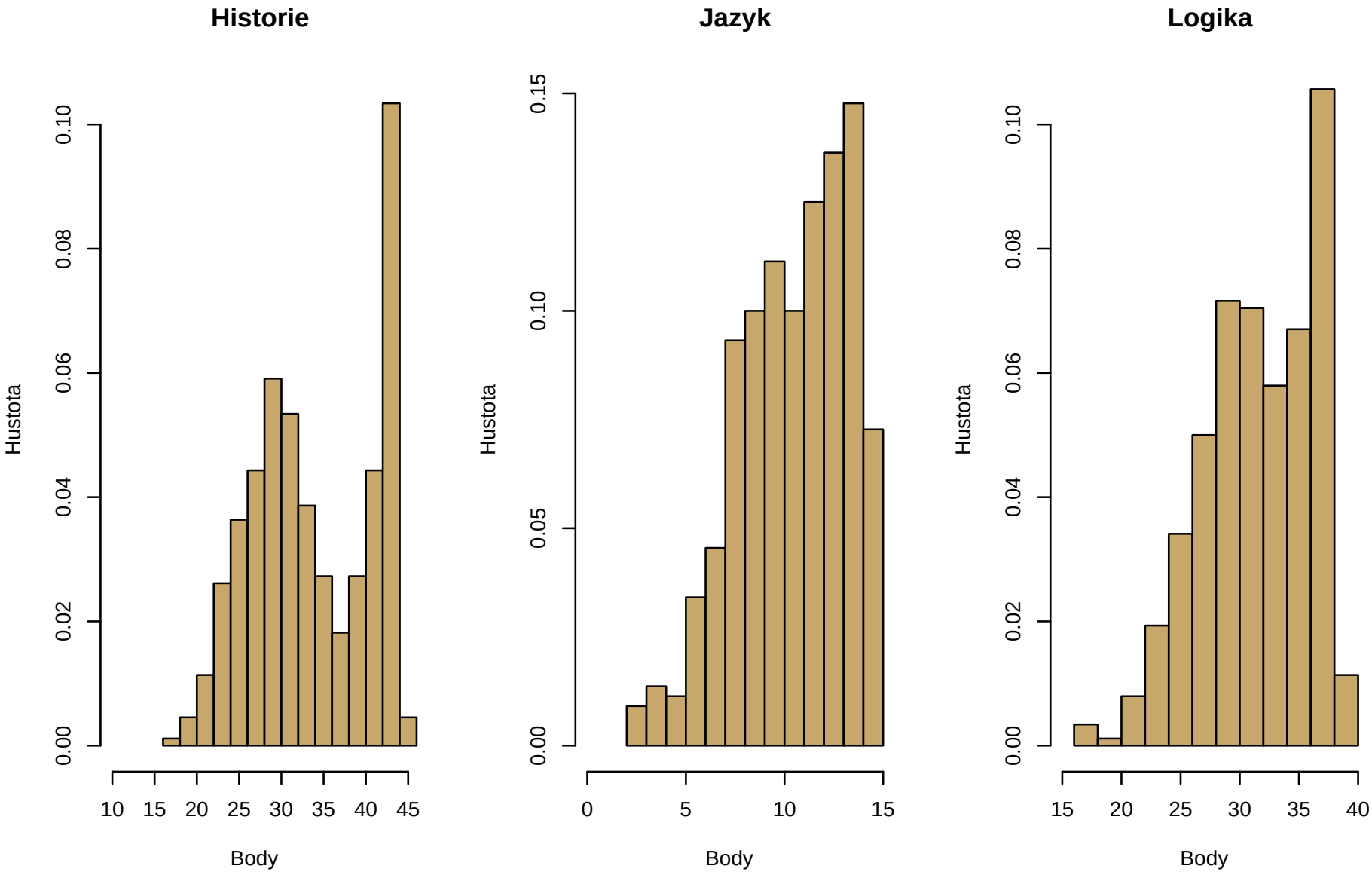
Příklad 3: Výsledky přijímacích zkoušek

Termíny č. 1–11



Příklad 3: Výsledky přijímacích zkoušek

Termín č. 12



Příklad 3: Výsledky přijímacích zkoušek

Možný model pro výsledky termínu č. 12

Možný model (pro $i = 1, \ldots, N$)

- $\boldsymbol{Y}_i = \left(Y_{i,1},\ Y_{i,2},\ Y_{i,3}\right)^\top$: bodové zisky $i$tého studenta u jednotlivých částí zkoušky.

- Studenti pocházejí ze dvou populací:
  1. běžní studenti (proporce w_1);
  2. studenti napojení na gangsterskou mafii (proporce w_2, $w_1 + w_2 = 1$).

- Budeme předpokládat, že (sdružené) rozdělení bodových zisků je v každé populaci normální se střední hodnotou $\boldsymbol{\mu}_1$, resp. $\boldsymbol{\mu}_2$ a varianční maticí $\boldsymbol{\Sigma}_1$, resp. $\boldsymbol{\Sigma}_2$.

$\Rightarrow$ Rozdělení $\boldsymbol{Y}_i$: směs normálních rozdělení s hustotou

$$p(\boldsymbol{y}_i \,|\, \boldsymbol{\theta}) \;=\; w_1\, \varphi(\boldsymbol{y}_i \,|\, \boldsymbol{\mu}_1,\, \boldsymbol{\Sigma}_1) \;+\; w_2\, \varphi(\boldsymbol{y}_i \,|\, \boldsymbol{\mu}_2,\, \boldsymbol{\Sigma}_2)$$

Příklad 3: Výsledky přijímacích zkoušek

Možný model pro výsledky termínu č. 12

- Potřeba odhadnout:
  - váhy (proporce) w_1, w_2,
  - střední hodnoty $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$,
  - varianční matice $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$.

➠ $\boldsymbol{\theta} = \left(w_1,\, w_2,\, \boldsymbol{\mu}_1^\top,\, \boldsymbol{\mu}_2^\top,\, \mathrm{vec}(\boldsymbol{\Sigma}_1),\, \mathrm{vec}(\boldsymbol{\Sigma}_2)\right)^\top$

Věrohodnost (pozorovaných dat)

$$L_{obs}(\boldsymbol{\theta}) = p(\boldsymbol{y} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{y}_i \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{N} \left\{ \sum_{k=1}^{2} w_k \, \varphi(\boldsymbol{y}_i \,|\, \boldsymbol{\mu}_k,\, \boldsymbol{\Sigma}_k) \right\}.$$

Oddíl **6.2**

Principy

Rozšiřování dat (data augmentation)

- Viděli jsme, že pozorovaná věrohodnost $L_{obs}(\boldsymbol{\theta}) = p(\boldsymbol{y} \mid \boldsymbol{\theta})$ (která tvoří základ aposteriorní hustoty $p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto L_{obs}(\boldsymbol{\theta})\, p(\boldsymbol{\theta})$) není vždy snadno vyjádřitelná jako součin "pěkných" funkcí.
- Někdy není $L_{obs}(\boldsymbol{\theta})$ dokonce ani analyticky vyjádřitelná:
  - logistická regrese s náhodnými efekty,
  - zobecněné lineární smíšené modely (GLMM).
- Nicméně často se "věrohodnost" výrazně zjednoduší, jestliže budeme uvažovat více parametrů:
  - označme je $\boldsymbol{\psi}$;
  - "věrohodnost" je pak $L_{augm}(\boldsymbol{\psi}, \boldsymbol{\theta}) = p(\boldsymbol{y} \mid \boldsymbol{\psi}, \boldsymbol{\theta})$;
  - $L_{augm}(\boldsymbol{\psi}, \boldsymbol{\theta})$ budeme nazývat rozšířená věrohodnost *(augmented likelihood)*.

Rozšiřování dat (data augmentation)

Principy

- Paramery ψ mají často význam nepozorovatelných (resp. pouze nepřímo pozorovatelných) dat.
  - odsud termín rozšiřování dat (*data augmentation*);
  - termín pochází z článku Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**(398), 528–550.
- $\boldsymbol{y}$: pozorovaná/pozorovatelná data (*observed data*).
- $(\boldsymbol{y}, \psi)$: úplná/rozšířená data (*complete data*).

Rozšiřování dat (data augmentation)

Principy

- Primárně nás zajímá $p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \underbrace{p(\boldsymbol{y} \mid \boldsymbol{\theta})}_{L_{obs}(\boldsymbol{\theta})}\, p(\boldsymbol{\theta})$,

  kde $p(\boldsymbol{y} \mid \boldsymbol{\theta}) = L_{obs}(\boldsymbol{\theta})$ plyne z předpokládaného (ne nutně hierarchického) modelu.

- Řekněme, že pro vhodné ψ se se sdruženou hustotou

$$p(\psi, \boldsymbol{\theta} \mid \boldsymbol{y}) \;\propto\; p(\boldsymbol{y} \mid \psi, \boldsymbol{\theta})\, p(\psi, \boldsymbol{\theta}) = \underbrace{p(\boldsymbol{y} \mid \psi, \boldsymbol{\theta})}_{L_{augm}(\psi, \boldsymbol{\theta})}\; p(\psi \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$$

  mnohem lépe pracuje.

  - $L_{augm}(\psi, \boldsymbol{\theta})$: model (věrohodnost) pro pozorovaná data, jestliže na doplněná data pohlížíme jako na další parametry modelu.
  - $p(\psi \mid \boldsymbol{\theta})$: model (věrohodnost) pro doplněná data.

- Řekněme, že zajistíme, aby platilo

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \int p(\psi, \boldsymbol{\theta} \mid \boldsymbol{y}) d\lambda(\psi)$$

tj., aby $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ bylo marginální hustotou rozdělení $\boldsymbol{\theta} \mid \boldsymbol{Y} = \boldsymbol{y}$ odpovídající sdruženému rozdělení $(\psi, \boldsymbol{\theta}) \mid \boldsymbol{Y} = \boldsymbol{y}$.

Principy

- Provádíme-li posteriorní inferenci na základě simulace, vygenerujeme náhodný výběr/markovský řetězec

$$\mathcal{S}_{(\psi,\,\theta),M} = \left\{ \left(\psi^{(1)},\, \theta^{(1)}\right), \ldots, \left(\psi^{(M)},\, \theta^{(M)}\right) \right\}$$

  s limitním rozdělením majícím hustotu $p(\psi,\, \theta \,|\, \boldsymbol{y})$.

- Jestliže $p(\theta \,|\, \boldsymbol{y})$ je marginální hustotou odpovídající sdružené hustotě $p(\psi,\, \theta \,|\, \boldsymbol{y})$, potom

$$\mathcal{S}_{\theta,M} = \left\{ \theta^{(1)}, \ldots, \theta^{(M)} \right\}$$

je náhodný výběr/markovský řetězec s limitním rozdělením majícím hustotu $p(\theta \,|\, \boldsymbol{y})$.

Rozšiřování dat (data augmentation)

Principy

- Máme: $p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto L_{obs}(\boldsymbol{\theta})\, p(\boldsymbol{\theta})$,

  $p(\boldsymbol{\psi}, \boldsymbol{\theta} \mid \boldsymbol{y}) \propto L_{augm}(\boldsymbol{\psi}, \boldsymbol{\theta})\, p(\boldsymbol{\psi} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$.

  - $L_{obs}(\boldsymbol{\theta})$ plyne z předpokládaného modelu pro pozorovaná data.
  - $L_{augm}(\boldsymbol{\psi}, \boldsymbol{\theta})\, p(\boldsymbol{\psi} \mid \boldsymbol{\theta})$ plyne z uvažovaného rozšiřování dat.
    - $L_{augm}(\boldsymbol{\psi}, \boldsymbol{\theta})$: model (věrohodnost) pro pozorovaná data, jestliže na doplněná data pohlížíme jako na další parametry modelu.
    - $p(\boldsymbol{\psi} \mid \boldsymbol{\theta})$: model (věrohodnost) pro doplněná data.

- Rozšiřování je potřeba udělat tak, aby $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ bylo marginální hustotou odpovídající sdružené hustotě $p(\boldsymbol{\psi}, \boldsymbol{\theta} \mid \boldsymbol{y})$.

- Rozšiřování je tedy potřeba udělat tak, aby

$$L_{obs}(\boldsymbol{\theta}) \propto \int L_{augm}(\boldsymbol{\psi}, \boldsymbol{\theta})\, p(\boldsymbol{\psi} \mid \boldsymbol{\theta}) d\lambda(\boldsymbol{\psi}).$$

Rozšiřování dat (data augmentation)

Principy

- K tomu, aby $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ bylo marginální hustotou odpovídající sdružené hustotě $p(\boldsymbol{\psi}, \boldsymbol{\theta} \mid \boldsymbol{y})$ stačí, aby platilo

$$L_{obs}(\boldsymbol{\theta}) = p(\boldsymbol{y} \mid \boldsymbol{\theta})$$
$$\propto \int p(\boldsymbol{y} \mid \boldsymbol{\psi}, \boldsymbol{\theta}) p(\boldsymbol{\psi} \mid \boldsymbol{\theta}) d\lambda(\boldsymbol{\psi}) = \int L_{augm}(\boldsymbol{\psi}, \boldsymbol{\theta})\, p(\boldsymbol{\psi} \mid \boldsymbol{\theta}) d\lambda(\boldsymbol{\psi}).$$

- Výraz

$$p(\boldsymbol{y} \mid \boldsymbol{\psi}, \boldsymbol{\theta}) p(\boldsymbol{\psi} \mid \boldsymbol{\theta}) = L_{augm}(\boldsymbol{\psi}, \boldsymbol{\theta})\, p(\boldsymbol{\psi} \mid \boldsymbol{\theta})$$

  je roven $p(\boldsymbol{y}, \boldsymbol{\psi} \mid \boldsymbol{\theta})$ a lze ho tedy interpretovat jako věrohodnost, jestliže bychom pozorovali úplná data.

- $L_{compl}(\boldsymbol{\theta}) := L_{augm}(\boldsymbol{\psi}, \boldsymbol{\theta})\, p(\boldsymbol{\psi} \mid \boldsymbol{\theta})$

  $=$ věrohodnost úplných dat.

Rozšiřování dat (data augmentation)

Principy

- K tomu, aby $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ bylo marginální hustotou odpovídající sdružené hustotě $p(\boldsymbol{\psi}, \boldsymbol{\theta} \mid \boldsymbol{y})$ tedy stačí specifikovat (rozšířený) model zahrnující nepozorovaná data $\boldsymbol{\psi}$ tak, aby si odpovídaly jednotlivé věrohodnosti.

➠ Přirozeně zajištěno v případě hierarchických modelů, kde rozšířená data $\boldsymbol{\psi}$ mají typicky přesně danou roli v popisu pravděpodobnostního mechanizmu, o kterém předpokládáme, že generuje pozorovaná data $\boldsymbol{y}$.

Specifikace hierarchického modelu:

1. $L_{augm}(\boldsymbol{\psi}, \boldsymbol{\theta}) = p(\boldsymbol{y} \mid \boldsymbol{\psi}, \boldsymbol{\theta})$: 1. hierarchická úroveň

   (model pro pozorovaná data za podmínky nepozorovaných dat).

2. $p(\boldsymbol{\psi} \mid \boldsymbol{\theta})$: 2. hierarchická úroveň (model pro nepozorovaná data).

3. $L_{obs}(\boldsymbol{\theta})$ (marginální model pro pozorovaná data)

   "dopočítává" se jako $L_{obs}(\boldsymbol{\theta}) = \int L_{augm}(\boldsymbol{\psi}, \boldsymbol{\theta})\, p(\boldsymbol{\psi} \mid \boldsymbol{\theta}) d\lambda(\boldsymbol{\psi})$.

Rozšiřování dat (data augmentation)

Shrnutí terminologie

Data, parametry

- $\boldsymbol{y}$: pozorovaná data;
- θ : (frekventistické) parametry
  - inference o nich je naším primárním cílem;
- ψ : nepozorovaná (pouze nepřímo pozorovaná) data, dodatečné parametry.

Věrohodnosti

- $L_{obs}(\theta) = p(\boldsymbol{y} \mid \theta)$: věrohodnost pozorovaných dat;
- $L_{augm}(\psi, \theta) = p(\boldsymbol{y} \mid \psi, \theta)$: rozšířená věrohodnost pozorov. dat;
- $p(\psi \mid \theta)$: věrohodnost doplněných dat;
- $L_{compl}(\theta) = p(\psi, \boldsymbol{y} \mid \theta) = L_{augm}(\psi, \theta)\, p(\psi \mid \theta)$:

  věrohodnost úplných dat.

- Jednotlivé věrohodnosti potřeba specifikovat tak, aby

  $$L_{obs}(\theta) = \int L_{augm}(\psi, \theta)\, p(\psi \mid \theta) d\lambda(\psi).$$

Oddíl **6.3**

Příklady

Příklad 1: Lineární smíšený model

Model (pro $i = 1, \ldots, N$)

- $\boldsymbol{Y}_i \mid \boldsymbol{b}_i$ nezávislé s rozdělením $\mathcal{N}_{n_i}(\mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\boldsymbol{b}_i,\ \sigma^2\mathbf{I}_{n_i})$,
- $\boldsymbol{b}_i$ i.i.d. s rozdělením $\mathcal{N}_q(\boldsymbol{\mu},\ \mathbb{D})$.

Parametry

- $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top,\ \boldsymbol{\mu}^\top,\ \sigma^2,\ \mathrm{vec}(\mathbb{D})\right)^\top$.

Nepřímo pozorovatelná (doplněná) data

- $\boldsymbol{\psi} = \left(\boldsymbol{b}_1^\top, \ldots,\ \boldsymbol{b}_N^\top\right)^\top$.

Příklad 1: Lineární smíšený model

Věrohodnost **pozorovaných dat**

$$L_{obs}(\boldsymbol{\theta}) = \prod_{i=1}^{N} \int_{\mathbb{R}^q} \varphi(\boldsymbol{y}_i \,|\, \mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\boldsymbol{b}_i,\, \sigma^2\mathbf{I}_{n_i})\, \varphi(\boldsymbol{b}_i \,|\, \boldsymbol{\mu},\, \mathbb{D}) d\boldsymbol{b}_i$$

$$= \int_{\mathbb{R}^q} \cdots \int_{\mathbb{R}^q} \left\{ \prod_{i=1}^{N} \varphi(\boldsymbol{y}_i \,|\, \mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\boldsymbol{b}_i,\, \sigma^2\mathbf{I}_{n_i}) \right\} \left\{ \prod_{i=1}^{N} \varphi(\boldsymbol{b}_i \,|\, \boldsymbol{\mu},\, \mathbb{D}) \right\} d\boldsymbol{b}_1 \cdots d\boldsymbol{b}_N.$$

Rozšířená věrohodnost

$$L_{augm}(\boldsymbol{\psi},\, \boldsymbol{\theta}) = \prod_{i=1}^{N} \varphi(\boldsymbol{y}_i \,|\, \mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\boldsymbol{b}_i,\, \sigma^2\mathbf{I}_{n_i}).$$

Věrohodnost **doplněných dat**

$$p(\boldsymbol{\psi} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{N} \varphi(\boldsymbol{b}_i \,|\, \boldsymbol{\mu},\, \mathbb{D}).$$

Příklad 2: Logistická regrese s normálně rozdělenými náhodnými efekty

Model ($j = 1, \ldots, n_i$ pro každé $i = 1, \ldots, N$)

- $Y_{i,j} \mid b_i$ nezávislé s rozdělením $\mathcal{A}(\pi_i)$, kde $\pi_i = \frac{e^{b_i}}{1+e^{b_i}}$,
- b_i i.i.d. s rozdělením $\mathcal{N}(\mu, d^2)$.

Parametry

- $\theta = (\mu, d^2)^\top$.

Nepřímo pozorovatelná (doplněná) data

- $\psi = (b_1, \ldots, b_N)^\top$.

Příklad 2: Logistická regrese s normálně rozdělenými náhodnými efekty

Věrohodnost **pozorovaných dat**

$$L_{obs}(\theta) = \prod_{i=1}^{N} \int_{\mathbb{R}} \frac{e^{b_i \sum_{j=1}^{n_i} y_{i,j}}}{(1+e^{b_i})^{n_i}} \varphi(b_i \mid \mu,\, d^2) db_i$$

$$= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \left\{ \prod_{i=1}^{N} \frac{e^{b_i \sum_{j=1}^{n_i} y_{i,j}}}{(1+e^{b_i})^{n_i}} \right\} \left\{ \prod_{i=1}^{N} \varphi(b_i \mid \mu,\, d^2) \right\} db_1 \cdots db_N.$$

Rozšířená věrohodnost

$$L_{augm}(\psi,\, \theta) = \prod_{i=1}^{N} \frac{e^{b_i \sum_{j=1}^{n_i} y_{i,j}}}{(1+e^{b_i})^{n_i}}.$$

Věrohodnost **doplněných dat**

$$p(\psi \mid \theta) = \prod_{i=1}^{N} \varphi(b_i \mid \mu,\, d^2).$$

Příklad 3: Normální směsový model ($K > 1$ skupin)

Model ($i = 1, \ldots, N$)

- $\boldsymbol{Y}_i$ i.i.d. se směsovým rozdělením s hustotou

$$p(\boldsymbol{y}_i \,|\, \boldsymbol{\theta}) \;=\; \sum_{k=1}^{K} w_k \, \varphi(\boldsymbol{y}_i \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Parametry

- $\boldsymbol{\theta} \;\equiv\; \boldsymbol{w}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K$

  $\boldsymbol{w} = \left(w_1, \ldots, w_K\right)^\top, 0 < w_k < 1, \sum_{k=1}^{K} w_k = 1$

Nepřímo pozorovatelná (doplněná) data

- ???

Příklad 3: Normální směsový model

Věrohodnost **pozorovaných dat**

$$L_{obs}(\boldsymbol{\theta}) = \prod_{i=1}^{N} \left\{ \sum_{k=1}^{K} \varphi(\boldsymbol{y}_i \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\, w_k \right\}.$$

Rozšířená věrohodnost

$$L_{augm}(\boldsymbol{\psi}, \boldsymbol{\theta}) = \prod_{i=1}^{N} \varphi(\boldsymbol{y}_i \,|\, \boldsymbol{\mu}_{Z_i}, \boldsymbol{\Sigma}_{Z_i}).$$

Věrohodnost **doplněných dat**

$$p(\boldsymbol{\psi} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{N} w_{Z_i} = \prod_{i=1}^{N} \prod_{k=1}^{K} w_k^{\mathbb{I}(Z_i = k)}.$$

Příklad 3: Normální směsový model

- Též nyní platí, že

$$L_{obs}(\boldsymbol{\theta}) = \int L_{augm}(\boldsymbol{\psi}, \boldsymbol{\theta}) p(\boldsymbol{\psi} \mid \boldsymbol{\theta}) d\lambda(\boldsymbol{\psi}).$$

- λ je nyní součinová čítací míra na $\{1, \ldots, K\}^N$ a tedy

$$\begin{aligned}
&\int L_{augm}(\boldsymbol{\psi}, \boldsymbol{\theta}) p(\boldsymbol{\psi} \mid \boldsymbol{\theta}) d\lambda(\boldsymbol{\psi}) \\
&\quad = \sum_{z_1=1}^{K} \cdots \sum_{z_N=1}^{K} \left\{ \prod_{i=1}^{N} \varphi(\boldsymbol{y}_i \mid \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) \right\} \left\{ \prod_{i=1}^{N} \underbrace{\mathrm{P}(Z_i = z_i \mid \boldsymbol{\theta})}_{w_{z_i}} \right\} \\
&\quad = \prod_{i=1}^{N} \left\{ \sum_{z_i=1}^{K} \varphi(\boldsymbol{y}_i \mid \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) w_{z_i} \right\}.
\end{aligned}$$

Oddíl **6.4**

Poznámky

Poznámky

- Ani rozšiřování dat není úplně bez komplikací
- Zvětšujeme (často poměrně výrazně) dimenzi parametrického prostoru.
  - Generujeme-li z aposteriorního rozdělení pomocí MCMC, může být poměrně obtížné sestrojit markovský řetězec s nízkou autokorelací a rychle konvergující k limitnímu rozdělení.
- Při použití rozšiřování dat pracujeme primárně s aposteriorním rozdělením

$$p(\psi, \boldsymbol{\theta} \mid \boldsymbol{y}) \propto L_{augm}(\psi, \boldsymbol{\theta})\, p(\psi \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}).$$

  - Též zde lze apriorní rozdělení $p(\boldsymbol{\theta})$ specifikovat hierarchicky za pomoci náhodných hyperparametrů ζ s apriorní hustotou $p(\zeta)$.
  - Fakticky potom pracujeme s aposteriorním rozdělením

$$p(\psi, \boldsymbol{\theta}, \zeta \mid \boldsymbol{y}) \propto L_{augm}(\psi, \boldsymbol{\theta})\, p(\psi \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \zeta)\, p(\zeta).$$

Rozšiřování dat (data augmentation)

Další oblasti využití

- Modely pro cenzorovaná data:
  - nejenom cenzorování zprava, ale též obecnější intervalové cenzorování,
  - nejenom neinformativní, ale též informativní cenzorování.
- A mnohé jiné. . .

7

Bayesian Model Selection

Oddíl **7.1**

Bayes factor

Bayesian model

Data: $\boldsymbol{Y}$,

Likelihood: $p(\boldsymbol{y} \mid \boldsymbol{\theta}) = L(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$,

Prior distribution: $p(\boldsymbol{\theta})$.

Definice 7.1 Integrated (marginal) likelihood.

Marginal density of $\boldsymbol{Y}$ following from the joint distribution of $(\boldsymbol{Y}, \boldsymbol{\theta})$ is called the integrated (marginal) likelihood, i.e.,

$$p(\boldsymbol{y}) = \int_{\Theta} p(\boldsymbol{y}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = \int_{\Theta} \underbrace{p(\boldsymbol{y} \mid \boldsymbol{\theta})}_{L(\boldsymbol{\theta})} p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}.$$

Remarks

- Marginal likelihood is a likelihood of the model where the values of the unknown parameters are averaged over their prior distribution.
- It is also the denominator from the Bayes theorem.
- Also reported as model evidence.

Model selection

- Interest in selecting a model from a set of candidate models $\mathrm{M}_1, \ldots, \mathrm{M}_r$.
- Model M_k, $k = 1, \ldots, r$:

  Likelihood: $p_k(\boldsymbol{y} \mid \boldsymbol{\theta}_k) = L(\boldsymbol{\theta}_k), \quad \boldsymbol{\theta}_k \in \Theta_k \subset \mathbb{R}^{p_k},$

  Prior distribution: $p_k(\boldsymbol{\theta}_k),$

  Integrated likelihood: $p_k(\boldsymbol{y}) = \int_\Theta p_k(\boldsymbol{y} \mid \boldsymbol{\theta}_k)\, p_k(\boldsymbol{\theta}_k) \mathrm{d}\boldsymbol{\theta}_k.$

- Integrated likelihood $p_k(\boldsymbol{y})$ can also be intepreted as distribution of data under validity of model M_k:

$$p_k(\boldsymbol{y}) = p(\boldsymbol{y} \mid \mathrm{M}_k), \quad k = 1, \ldots, r.$$

Model selection

- Let $\mathrm{P}(\mathrm{M}_1), \ldots, \mathrm{P}(\mathrm{M}_r)$ be the prior probabilities of models $\mathrm{M}_1, \ldots, \mathrm{M}_k$:

$$0 < \mathrm{P}(\mathrm{M}_k) < 1, \quad k = 1, \ldots, r \qquad \sum_{k=1}^{r} \mathrm{P}(\mathrm{M}_k) = 1.$$

  - For example (but not necessarily): $\mathrm{P}(\mathrm{M}_k) = \frac{1}{r}$, $k = 1, \ldots, r$.

- Model selection in Bayesian context can be based on posterior probabilities of models $\mathrm{M}_1, \ldots, \mathrm{M}_r$:

$$\mathrm{P}(\mathrm{M}_k \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \mathrm{M}_k)\, \mathrm{P}(\mathrm{M}_k)}{\sum_{l=1}^{r} p(\boldsymbol{y} \mid \mathcal{M}_l)\, \mathrm{P}(\mathrm{M}_l)}, \quad k = 1, \ldots, r.$$

  - Choose model with the maximal posterior probability.

- "Small" complication: Integrated likelihood $p_k(\boldsymbol{y}) = p(\boldsymbol{y} \mid \mathrm{M}_k)$ must be calculated for each model which requires calculation of (usually complicated/intractable) integral.

Definice 7.2 Bayes factor.

Bayes factor of the two models M_k and M_j is defined as the odds of the two integrated likelihoods, i.e.,

$$\mathrm{BF}(\mathrm{M}_k, \mathrm{M}_j) = \frac{p(\boldsymbol{y} \mid \mathrm{M}_k)}{p(\boldsymbol{y} \mid \mathrm{M}_j)} = \frac{p_k(\boldsymbol{y})}{p_j(\boldsymbol{y})}.$$

Remarks

- Bayes factor measures the evidence for model M_k versus model M_j.
- Posterior odds of model M_k versus model M_j:

$$= \frac{\mathrm{P}(\mathrm{M}_k \mid \boldsymbol{y})}{\mathrm{P}(\mathrm{M}_j \mid \boldsymbol{y})} = \frac{p(\boldsymbol{y} \mid \mathrm{M}_k)\,\mathrm{P}(\mathrm{M}_k)}{p(\boldsymbol{y} \mid \mathrm{M}_j)\,\mathrm{P}(\mathrm{M}_j)} = \mathrm{BF}(\mathrm{M}_k, \mathrm{M}_j) \underbrace{\frac{\mathrm{P}(\mathrm{M}_k)}{\mathrm{P}(\mathrm{M}_j)}}_{\text{prior odds}(\mathrm{M}_k, \mathrm{M}_j)}.$$

- With the uniform prior distribution for the competing models:

$$\text{posterior odds}(\mathrm{M}_k, \mathrm{M}_j) = \mathrm{BF}(\mathrm{M}_k, \mathrm{M}_j).$$

Jeffreys' scale of evidence for Bayes factor

Bayes factor(M_k, M_j)	Interpretation
$BF(M_k, M_j) < 1$	Negative support for M_k
$1 \leq BF(M_k, M_j) < 3$	Barely worth mentioning evidence for M_k
$3 \leq BF(M_k, M_j) < 10$	Substantial evidence for M_k
$10 \leq BF(M_k, M_j) < 30$	Strong evidence for M_k
$30 \leq BF(M_k, M_j) < 100$	Very strong evidence for M_k
$100 \leq BF(M_k, M_j)$	Decisive evidence for M_k

Problems with Bayes factor

- The integrated likelihoods $p_k(\boldsymbol{y})$ which enter the Bayes factor are, in fact, the means (expected values) of the likelihood (under model M_k) with respect to the prior distribution (under model M_k).
- $p_k(\boldsymbol{y})$ is not well defined when the prior distribution $p_k(\boldsymbol{\theta}_k)$ is improper.
- Bayes factor is numerically unstable when proper but diffuse (weakly informative) prior distributions used.
- There exist numerous approaches that were suggested in the literature to overcome above problems.

Further reading

- Robert E. Kass, Adrian E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association*. **90**(430), 773–795.
- Tomohiro Ando (2010). *Bayesian Model Selection and Statistical Modeling*. Boca Raton: Chapman & Hall/CRC. ISBN 978-1-4398-3614-9.

Oddíl **7.2**

Posterior predictive distribution

Bayesian model

Data: $\boldsymbol{Y}$,

Likelihood: $p(\boldsymbol{y} \mid \boldsymbol{\theta}) = L(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$,

Prior distribution: $p(\boldsymbol{\theta})$.

Posterior predictive distribution

- Let $\boldsymbol{Y}_{new}$ be the random vector generated according to the same probabilistic mechanism as the data random vector $\boldsymbol{Y}$.
- In a Bayesian setting, it will always be assumed that $\boldsymbol{Y}$ and $\boldsymbol{Y}_{new}$ are (conditionally) independent given $\boldsymbol{\theta}$.
- $\boldsymbol{Y}_{new} \equiv$ new (replicated) data.

Definice 7.3 Posterior predictive distribution.

Posterior distribution of the random vector $\boldsymbol{Y}_{new}$, i.e., $p(\boldsymbol{y}_{new} \mid \boldsymbol{y})$, is called the posterior predictive distribution.

We have

$$p(\boldsymbol{y}_{new} \mid \boldsymbol{y}) = \int_{\Theta} p(\boldsymbol{y}_{new}, \boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta} = \int_{\Theta} p(\boldsymbol{y}_{new} \mid \boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}$$

$$= \int_{\Theta} \underbrace{p(\boldsymbol{y}_{new} \mid \boldsymbol{\theta})}_{L_{new}(\boldsymbol{\theta})} p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}.$$

Posterior predictive distribution versus integrated likelihood

Integrated likelihood

$$p(\boldsymbol{y}) = \int_{\Theta} L(\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$$

$\equiv$ Distribution of data when the unknown parameters are averaged over their prior distribution.

➠ Evidence of the model before unknown parameters being estimated.

Posterior predictive distribution

$$p(\boldsymbol{y}_{new} \mid \boldsymbol{y}) = \int_{\Theta} L_{new}(\boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \boldsymbol{y})\, \mathrm{d}\boldsymbol{\theta}$$

$\equiv$ Distribution of (new) data when the unknown parameters are averaged over their posterior distribution.

➠ Evidence of the model after using the data $\boldsymbol{Y}$ for inference on unknown $\boldsymbol{\theta}$.

Oddíl **7.3**

Kullback-Leibler distance and deviance of the model

Kullback-Leibler distance and deviance of the model

Scetch of a theory will follow now which explains why the likelihood (or some of its derivatives) of the model can be considered as evidence of that model.

Definice 7.4 Kullback-Leibler distance.

Let Q_1 and Q_2 be two distributions with densities q_1 and q_2 (with respect to some σ-finite measure). The Kullback-Leibler distance (divergence) of Q_2 from Q_1 is defined as

$$\mathrm{KL}(Q_2,\, Q_1) = \mathbb{E}_{Q_1} \log\left\{\frac{q_1(\boldsymbol{Y})}{q_2(\boldsymbol{Y})}\right\} = \int q_1(\boldsymbol{y}) \log\left\{\frac{q_1(\boldsymbol{y})}{q_2(\boldsymbol{y})}\right\} \mathrm{d}\boldsymbol{y}.$$

- We have: $\mathrm{KL}(Q_2,\, Q_1) = \mathbb{E}_{Q_1} \log\{q_1(\boldsymbol{Y})\} - \mathbb{E}_{Q_1} \log\{q_2(\boldsymbol{Y})\}$.
- Can also be shown: $\mathrm{KL}(Q_2,\, Q_1) \geq 0$,
  $\mathrm{KL}(Q_2,\, Q_2) = 0$.

In context of statistical modelling

- Let Q (with a density q) be the true (unknown) distribution of data $\boldsymbol{Y}$.
- $L(\theta) = p(\cdot \mid \theta)$: likelihood (model) for data (which possibly depends on a parameter vector θ).

Then

$$\mathrm{KL}\big(L(\theta),\, Q\big) = \underbrace{\mathbb{E}_Q \log\big\{q(\boldsymbol{Y})\big\}}_{\text{const for all models}} - \mathbb{E}_Q \log\big\{p(\boldsymbol{Y} \mid \theta)\big\}.$$

- Up to an additive constant, the term $-\mathbb{E}_Q \log\big\{p(\boldsymbol{Y} \mid \theta)\big\}$ is the Kullback-Leibler distance of the used model from the truth.

Deviance of a model

Definice 7.5 Deviance of a model.

For given model with the likelihood $L(\theta) = p(\boldsymbol{y} \mid \theta)$, a quantity

$$D(\theta;\, \boldsymbol{y}) = -2\, \log\{p(\boldsymbol{y} \mid \theta)\} = -2\, \log\{L(\theta)\}$$

is called the deviance of the model.

Remarks

- If Q is the true (unknown) distribution of data $\boldsymbol{Y}$, we have

$$2\,\mathrm{KL}\big(L(\theta),\, Q\big) = \mathbb{E}_Q\{D(\theta;\, \boldsymbol{Y})\} + \text{const.}$$

- Factor 2 in the definition of the deviance is used to get direct link to the statistic of the likelihood-ratio test.
- Deviance: suitable measure of the model quality (small deviance $\equiv$ better model).

Deviance of a model

Typically

Data: $\boldsymbol{Y} = \left(\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_n^\top\right)^\top,$

Model (likelihood): $L(\boldsymbol{\theta}) = p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p_i(\boldsymbol{y}_i \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\theta}),$

$\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ (conditionally) independent given $\boldsymbol{\theta}$.

Deviance

$$D(\boldsymbol{\theta};\, \boldsymbol{y}) = -2 \log\left\{\prod_{i=1}^{n} p_i(\boldsymbol{y}_i \mid \boldsymbol{\theta})\right\} = -2 \sum_{i=1}^{n} \log\left\{p_i(\boldsymbol{y}_i \mid \boldsymbol{\theta})\right\}$$

$$= \sum_{i=1}^{n} \underbrace{\left[-2 \log\left\{p_i(\boldsymbol{y}_i \mid \boldsymbol{\theta})\right\}\right]}_{D_i(\boldsymbol{\theta};\, \boldsymbol{y}_i)}.$$

Oddíl **7.4**

Measures of predictive ability of the model

Measures of predictive ability of the model

Our aim will now be to specify some criteria to evaluate/measure the ability of the model to make accurate predictions of new (replicated) data.

Those criteria can then be used for model selection.

Statistical model

(Observed) data: $\boldsymbol{Y} = \left(\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_n^\top\right)^\top$.

(New, not yet observed) data: $\boldsymbol{Y}_{new} = \left(\boldsymbol{Y}_{new,1}^\top, \ldots, \boldsymbol{Y}_{new,n}^\top\right)^\top$.

$\boldsymbol{Y}$ and $\boldsymbol{Y}_{new}$ generated by the same probabilistic mechanism.

Model (likelihood): $L(\theta) = p(\cdot \mid \theta) = \prod_{i=1}^{n} p_i(\cdot \mid \theta)$.

$\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, $\boldsymbol{Y}_{new,1}, \ldots, \boldsymbol{Y}_{new,n}$ (conditionally) independent given θ,

$$p(\boldsymbol{y} \mid \theta) = \prod_{i=1}^{n} p_i(\boldsymbol{y}_i \mid \theta).$$

$$p(\boldsymbol{y}_{new} \mid \theta, \boldsymbol{y}) = p(\boldsymbol{y}_{new} \mid \theta) = \prod_{i=1}^{n} p_i(\boldsymbol{y}_{new,i} \mid \theta).$$

Bayesian inference

Prior distribution: $p(\boldsymbol{\theta})$.

➠ Integrated likelihood: $p(\boldsymbol{y}) = \int_{\Theta} p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$

Evidence of the model before unknown parameters being estimated.

Inference on unknown $\boldsymbol{\theta}$ based on the posterior distribution:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta}) \, p(\boldsymbol{\theta})}{p(\boldsymbol{y})}.$$

➠ Posterior predictive distribution: $p(\boldsymbol{y}_{new} \mid \boldsymbol{y}) = \int_{\Theta} p(\boldsymbol{y}_{new} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) \, \mathrm{d}\boldsymbol{\theta}$

Evidence of the model after the observed data $\boldsymbol{Y} = \boldsymbol{y}$ used to infer on unknown parameters $\boldsymbol{\theta}$.

Posterior predictive deviance

Posterior predictive distribution:

$$p(\boldsymbol{y}_{new} \mid \boldsymbol{y}) = \int_{\Theta} p(\boldsymbol{y}_{new} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) \, \mathrm{d}\boldsymbol{\theta} = \int_{\Theta} \prod_{i=1}^{n} p_i(\boldsymbol{y}_{new,i} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) \, \mathrm{d}\boldsymbol{\theta}.$$

Definice 7.6 Posterior predictive deviance.

Quantity

$$\overline{D}_{pred} = \mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y})} D(\boldsymbol{\theta}; \boldsymbol{y}_{new}) = \int_{\Theta} \underbrace{\left[-2 \log\left\{ p(\boldsymbol{y}_{new} \mid \boldsymbol{\theta}) \right\} \right]}_{D(\boldsymbol{\theta}; \boldsymbol{y}_{new})} p(\boldsymbol{\theta} \mid \boldsymbol{y}) \, \mathrm{d}\boldsymbol{\theta}$$

will be called the posterior predictive deviance.

➠ Suitable measure of prediction error (loss of prediction) when predicting $\boldsymbol{Y}_{new} = \boldsymbol{y}_{new}$ using a (Bayesian) model estimated using data $\boldsymbol{Y} = \boldsymbol{y}$.

Posterior predictive deviance

We have

$$\overline{D}_{pred} = \mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y})} D(\boldsymbol{\theta};\, \boldsymbol{y}_{new}) = \mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y})} \left\{ \sum_{i=1}^{n} D_i(\boldsymbol{\theta};\, \boldsymbol{y}_{new,i}) \right\}$$

$$= \sum_{i=1}^{n} \underbrace{\mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y})}\, D_i(\boldsymbol{\theta};\, \boldsymbol{y}_{new,i})}_{\overline{D}_{pred,i}} = \sum_{i=1}^{n} \int_{\Theta} \underbrace{\left[-2 \log\{p_i(\boldsymbol{y}_{new,i} \mid \boldsymbol{\theta})\} \right]}_{D_i(\boldsymbol{\theta},\, \boldsymbol{y}_{new,i})} p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}.$$

- To calculate $\overline{D}_{pred}$ in practice (to be able to use it for model selection), we need the value of "new" data $\boldsymbol{Y}_{new}$.

Posterior predictive deviance

$$\overline{D}_{pred} = \sum_{i=1}^{n} \overline{D}_{pred,i} = \sum_{i=1}^{n} \mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y})} D_i(\boldsymbol{\theta}, \boldsymbol{y}_{new,i}).$$

$\equiv$ Values of new data $\boldsymbol{Y}_{new} = \boldsymbol{y}_{new}$ needed.

➠ Measure of the loss of prediction.

- Use **cross-validation** to estimate a value of each $\overline{D}_{pred,i}$, $i = 1, \ldots, n$:

$$\overline{D}_{pred,i} \approx \overline{D}^{CV}_{pred,i} = \mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y}_{-i})} D_i(\boldsymbol{\theta}, \boldsymbol{y}_i) = \int_{\Theta} D_i(\boldsymbol{\theta}, \boldsymbol{y}_i)\, p(\boldsymbol{\theta} \mid \boldsymbol{y}_{-i})\, \mathrm{d}\boldsymbol{\theta}.$$

Cross-validated posterior predictive deviance

Definice 7.7 Cross-validated posterior predictive deviance.

Quantity

$$\overline{D}_{pred}^{CV} = \sum_{i=1}^{n} \underbrace{\mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y}_{-i})} D_i(\boldsymbol{\theta};\, \boldsymbol{y}_i)}_{\overline{D}_{pred,i}^{CV}}$$

$$= \sum_{i=1}^{n} \int_{\Theta} \underbrace{\left[-2 \log\{p_i(\boldsymbol{y}_i \mid \boldsymbol{\theta})\}\right]}_{D_i(\boldsymbol{\theta},\, \boldsymbol{y}_i)} p(\boldsymbol{\theta} \mid \boldsymbol{y}_{-i}) \mathrm{d}\boldsymbol{\theta}.$$

will be called the cross-validated posterior predictive deviance.

- With MCMC based Bayesian inference, (relatively) easily estimable if we have time to run the MCMC n-times (always with one observation left out).

Definice 7.8 Posterior expected deviance.

Quantity

$$\overline{D} = \mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y})}\, D(\boldsymbol{\theta};\, \boldsymbol{y}) = \int_{\Theta} \underbrace{\left[-2\, \log\{p(\boldsymbol{y} \mid \boldsymbol{\theta})\}\right]}_{D(\boldsymbol{\theta};\, \boldsymbol{y})} p(\boldsymbol{\theta} \mid \boldsymbol{y})\, \mathrm{d}\boldsymbol{\theta}$$

will be called the posterior expected deviance.

We have

$$\overline{D} = \mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y})} D(\boldsymbol{\theta};\, \boldsymbol{y}) = \mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y})} \sum_{i=1}^{n} D_i(\boldsymbol{\theta};\, \boldsymbol{y}_i)$$

$$= \sum_{i=1}^{n} \underbrace{\mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y})}\, D_i(\boldsymbol{\theta};\, \boldsymbol{y}_i)}_{\overline{D}_i} = \sum_{i=1}^{n} \int_{\Theta} \underbrace{\left[-2\, \log\{p_i(\boldsymbol{y}_i \mid \boldsymbol{\theta})\}\right]}_{D_i(\boldsymbol{\theta},\, \boldsymbol{y}_i)} p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}.$$

Posterior expected deviance

$$\overline{D} = \sum_{i=1}^{n} \overline{D}_i = \sum_{i=1}^{n} \mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y})} D_i(\boldsymbol{\theta}, \boldsymbol{y}_i).$$

$\equiv$ Only the observed data $\boldsymbol{Y} = \boldsymbol{y}$ needed.

➠ With MCMC based Bayesian inference, (relatively) easily estimable.

➠ Underestimates the **cross-validated posterior predictive deviance** which is

$$\overline{D}^{CV}_{pred} = \sum_{i=1}^{n} \overline{D}^{CV}_{pred,i} = \sum_{i=1}^{n} \mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y}_{-i})} D_i(\boldsymbol{\theta}, \boldsymbol{y}_i).$$

Věta 7.1 .

For all $i = 1, \ldots, n$

$$\overline{D}^{CV}_{pred,i} - \overline{D}_i \quad \geq \quad 0.$$

Reminder

$$\overline{D}^{CV}_{pred,i} = \mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y}_{-i})} D_i(\boldsymbol{\theta}, \boldsymbol{y}_i) = -2\, \mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y}_{-i})} \log p_i(\boldsymbol{y}_i \mid \boldsymbol{\theta}),$$

$$\overline{D}_i = \mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y})} D_i(\boldsymbol{\theta}, \boldsymbol{y}_i) = -2\, \mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y})} \log p_i(\boldsymbol{y}_i \mid \boldsymbol{\theta}).$$

Cross-validation and posterior expected deviance

$$
\begin{aligned}
\mathrm{KL}_1 &:= \mathrm{KL}\Big(p(\boldsymbol{\theta} \mid \boldsymbol{y}),\, p(\boldsymbol{\theta} \mid \boldsymbol{y}_{-i})\Big) \\
&= \int_{\Theta} p(\boldsymbol{\theta} \mid \boldsymbol{y}_{-i}) \log\left\{ \frac{p(\boldsymbol{\theta} \mid \boldsymbol{y}_{-i})}{p(\boldsymbol{\theta} \mid \boldsymbol{y})} \right\} \mathrm{d}\boldsymbol{\theta} \\
&= \int_{\Theta} p(\boldsymbol{\theta} \mid \boldsymbol{y}_{-i}) \log\left\{ \frac{p(\boldsymbol{y}_{-i} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\;\; p(\boldsymbol{y})}{p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\;\; p(\boldsymbol{y}_{-i})} \right\} \mathrm{d}\boldsymbol{\theta} \\
&= \int_{\Theta} p(\boldsymbol{\theta} \mid \boldsymbol{y}_{-i}) \log\left\{ \frac{p(\boldsymbol{y})}{p_i(\boldsymbol{y}_i \mid \boldsymbol{\theta})\;\; p(\boldsymbol{y}_{-i})} \right\} \mathrm{d}\boldsymbol{\theta} \\
&= -\int_{\Theta} p(\boldsymbol{\theta} \mid \boldsymbol{y}_{-i}) \log\big\{ p_i(\boldsymbol{y}_i \mid \boldsymbol{\theta}) \big\} \mathrm{d}\boldsymbol{\theta} \;+\; \log\left\{ \frac{p(\boldsymbol{y})}{p(\boldsymbol{y}_{-i})} \right\} \\
&= \frac{1}{2}\, \mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y}_{-i})}\, D_i(\boldsymbol{\theta},\, \boldsymbol{y}_i) \;+\; \log\left\{ \frac{p(\boldsymbol{y})}{p(\boldsymbol{y}_{-i})} \right\} \\
&= \frac{1}{2}\, \overline{D}^{CV}_{pred,i} \;+\; \log\left\{ \frac{p(\boldsymbol{y})}{p(\boldsymbol{y}_{-i})} \right\}.
\end{aligned}
$$

Cross-validation and posterior expected deviance

$$
\begin{aligned}
\mathrm{KL}_2 &:= \mathrm{KL}\Big(p(\boldsymbol{\theta} \mid \boldsymbol{y}_{-i}),\, p(\boldsymbol{\theta} \mid \boldsymbol{y})\Big) \\
&= \int_{\Theta} p(\boldsymbol{\theta} \mid \boldsymbol{y}) \log\left\{ \frac{p(\boldsymbol{\theta} \mid \boldsymbol{y})}{p(\boldsymbol{\theta} \mid \boldsymbol{y}_{-i})} \right\} \mathrm{d}\boldsymbol{\theta} \\
&= \int_{\Theta} p(\boldsymbol{\theta} \mid \boldsymbol{y}) \log\left\{ \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\;\; p(\boldsymbol{y}_{-i})}{p(\boldsymbol{y}_{-i} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\;\; p(\boldsymbol{y})} \right\} \mathrm{d}\boldsymbol{\theta} \\
&= \int_{\Theta} p(\boldsymbol{\theta} \mid \boldsymbol{y}) \log\left\{ \frac{p_i(\boldsymbol{y}_i \mid \boldsymbol{\theta})\;\; p(\boldsymbol{y}_{-i})}{p(\boldsymbol{y})} \right\} \mathrm{d}\boldsymbol{\theta} \\
&= \int_{\Theta} p(\boldsymbol{\theta} \mid \boldsymbol{y}) \log\big\{ p_i(\boldsymbol{y}_i \mid \boldsymbol{\theta}) \big\} \mathrm{d}\boldsymbol{\theta} \;-\; \log\left\{ \frac{p(\boldsymbol{y})}{p(\boldsymbol{y}_{-i})} \right\} \\
&= -\frac{1}{2}\, \mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y})}\, D_i(\boldsymbol{\theta},\, \boldsymbol{y}_i) \;-\; \log\left\{ \frac{p(\boldsymbol{y})}{p(\boldsymbol{y}_{-i})} \right\} \\
&= -\frac{1}{2}\, \overline{D}_i \;-\; \log\left\{ \frac{p(\boldsymbol{y})}{p(\boldsymbol{y}_{-i})} \right\}.
\end{aligned}
$$

Cross-validation and posterior expected deviance

That is,

$$\overline{D}^{CV}_{pred,i} - \overline{D}_i = 2\left(\mathrm{KL}_1 + \mathrm{KL}_2\right) \geq 0.$$

Definice 7.9 Expected optimism.

Quantity

$$p_{opt,i} = \mathbb{E}\left(\overline{D}^{CV}_{pred,i} - \overline{D}_i \,\middle|\, \boldsymbol{Y}_{-i}\right), \quad i = 1, \ldots, n$$

will be called the expected optimism when the loss of prediction of the ith observation is evaluated by $\overline{D}_i$ (ith contribution to the posterior expected deviance) rather than by $\overline{D}^{CV}_{pred,i}$ (ith cross-validated posterior predictive deviance).

Definice 7.10 Penalized expected deviance (PED).

Quantity

$$\text{PED} = \underbrace{\sum_{i=1}^{n} \overline{D}_i}_{\overline{D}} + \underbrace{\sum_{i=1}^{n} p_{opt,i}}_{p_{opt}} = \sum_{i=1}^{n} \underbrace{\left(\overline{D}_i + p_{opt,i}\right)}_{\text{PED}_i}$$

will be called the penalized expected deviance (PED).
Quantity

$$p_{opt} = \sum_{i=1}^{n} p_{opt,i}$$

will be called the overall optimism, quantity

$$\text{PED}_i = \overline{D}_i + p_{opt,i}, \quad i = 1, \ldots, n,$$

will be called contribution of the ith observation to the penalized expected deviance.

Penalized expected deviance and cross-validated posterior predictive deviance

Věta 7.2 Penalized expected deviance and cross-validated posterior predictive deviance.

The following holds for each $i = 1, \ldots, n$*:*

$$\mathbb{E}\left(PED_i \,\middle|\, \boldsymbol{Y}_{-i}\right) = \mathbb{E}\left(\overline{D}^{CV}_{pred,i} \,\middle|\, \boldsymbol{Y}_{-i}\right).$$

➠ With respect to cross-validation

$$\text{PED} = \sum_{i=1}^{n} \text{PED}_i$$

is equivalent to

$$\overline{D}^{CV}_{pred} = \sum_{i=1}^{n} \overline{D}^{CV}_{pred,i}.$$

Penalized expected deviance and cross-validated posterior predictive deviance

Důkaz.

$$\mathbb{E}\left(\mathrm{PED}_i \,\middle|\, \boldsymbol{Y}_{-i}\right) = \mathbb{E}\left\{ \overline{D}_i + \underbrace{\mathbb{E}\left(\overline{D}^{CV}_{pred,i} - \overline{D}_i \,\middle|\, \boldsymbol{Y}_{-i}\right)}_{p_{opt,i}} \,\middle|\, \boldsymbol{Y}_{-i} \right\}$$

$$= \mathbb{E}\left(\overline{D}_i \,\middle|\, \boldsymbol{Y}_{-i}\right) + \mathbb{E}\left(\overline{D}^{CV}_{pred,i} \,\middle|\, \boldsymbol{Y}_{-i}\right) - \mathbb{E}\left(\overline{D}_i \,\middle|\, \boldsymbol{Y}_{-i}\right)$$

$$= \mathbb{E}\left(\overline{D}^{CV}_{pred,i} \,\middle|\, \boldsymbol{Y}_{-i}\right).$$

❑

Practical use of PED

- The last complication when using the PED for model comparison: calculation of the expected optimisms:

$$p_{opt,i} = \mathbb{E}\left(\overline{D}^{CV}_{pred,i} - \overline{D}_i \,\middle|\, \boldsymbol{Y}_{-i}\right), \quad i = 1, \ldots, n.$$

- With MCMC based Bayesian inference, all expected optimisms $p_{opt,i}$, $i = 1, \ldots, n$, can be estimated using two parallel Markov chains (with $p(\theta \mid \boldsymbol{y})$ as their limit distribution).

Deviance information criterion

- For some classes of models, e.g., when $p_i(\boldsymbol{y}_i \mid \boldsymbol{\theta})$, $i = 1, \ldots, n$, belongs to exponential family, the overall optimism can be estimated as

$$p_{opt} = \sum_{i=1}^{n} p_{opt,i} \approx p_D = \overline{D} - D\Big(\widehat{\boldsymbol{\theta}}(\boldsymbol{y}); \boldsymbol{y}\Big),$$

  where $\widehat{\boldsymbol{\theta}}(\boldsymbol{y}) = \mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y})}\boldsymbol{\theta}$ (posterior mean of $\boldsymbol{\theta}$).

- Terminology: $D\Big(\widehat{\boldsymbol{\theta}}(\boldsymbol{y}); \boldsymbol{y}\Big)$: plug-in deviance;
  p_D: effective number of parameters (measure of model complexity).

- "Small" inconvenience: the value of both the plug-in deviance and the effective number of parameters depends on the parameterization of the model.

- PED with p_D used in place of p_{opt}

  ➠ **Deviance information criterion (DIC)**.

Deviance information criterion (DIC)

$$\begin{aligned}\text{DIC} &= \overline{D} + p_D \\ &= \overline{D} + \left\{ \overline{D} - D\left(\widehat{\theta}(\boldsymbol{y}); \boldsymbol{y}\right) \right\} \\ &= 2\,\overline{D} - D\left(\widehat{\theta}(\boldsymbol{y}); \boldsymbol{y}\right).\end{aligned}$$

- DIC $\equiv$ approximation to $\overline{D}^{CV}_{pred}$ which was defined to evaluate the loss of prediction.
- Model with lower DIC is better.
- DIC is nowadays somehow overused/misused (even in situations when it is not justifiable)!

Further reading

- David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, Angelika Van Der Linde (2002). Bayesian measures of model complexity and fit (with Discussion). *Journal of the Royal Statistical Society, Series B*, **64**(4), 583–639.

- Martyn Plummer (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics*, **9**(3), 523–539.

- David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, Angelika Van Der Linde (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society, Series B*, **76**(3), 485–493.