

Dept. of Probability and Mathematical Statistics



FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

doc. RNDr. Arnošt Komárek, Ph.D.

NMST431 Bayesian Methods

Spring term 2023–24

1

Introduction

Section **1.1**

History



Thomas Bayes
(\approx 1701 – 7 April 1761)

- English statistician, philosopher and Presbyterian minister
- Mathematical book: *An Introduction to the Doctrine of Fluxions, and a Defence of the Mathematicians Against the Objections of the Author of The Analyst* (published anonymously in 1736): defense of the logical foundation of Isaac Newton's calculus ("fluxion") against the criticism by George Berkeley, a bishop and noted philosopher
- His notes (including Bayes theorem) edited and published posthumously by Richard Price (1723–1791, Welsh moral philosopher, Non-conformist minister and mathematician)

$$P(A_k | B) = \frac{P(B | A_k) P(A_k)}{\sum_{j=1}^m P(B | A_j) P(A_j)}, \quad k = 1, \dots, m,$$

where A_1, \dots, A_m events such that

$$P(A_j \cap A_s) = 0, \quad j \neq s, \quad P(\cup_{j=1}^m A_j) = 1.$$

Conditional probability

$$P(A_k | B) = \frac{P(A_k \cap B)}{P(B)}, \quad P(B) > 0.$$

In this context:

$P(A_k)$: **prior** (*apriori*) probability of event A_k ,

$P(A_k | B)$: **posterior** (*aposteriori*) probability of event A_k , given the fact that the event B occurred.

Bayes theorem (for events)

- Derived by Thomas Bayes for $P(A_k) = \frac{1}{m}$ for all $k = 1, \dots, m$.
- Published in 1763 (by R. Price), further generalized by P. C. Laplace (1773).
- And then globally ignored in 20's of the 21st century.

The (testing) business must go on

- $A_1 = A$, $A_2 = A^c \equiv$ *infected* or not (by virus, ...).
- B (or B^c) \equiv some test *positive* (or *negative*).
- $P(\text{TEST} + \mid \text{INFECT}) =$ *sensitivity* (of the test).
- $P(\text{TEST} - \mid \text{NOT INFECT}) =$ *specificity* (of the test).
- $P(\text{INFECT}) =$ *prevalence/incidence* (of the infection)

$$P(\text{INFECT} \mid \text{TEST}+) = \frac{\text{sens} \cdot \text{prev}}{\text{sens} \cdot \text{prev} + (1 - \text{spec}) \cdot (1 - \text{prev})}$$

$$P(\text{INFECT} \mid \text{TEST}-) = \frac{(1 - \text{sens}) \cdot \text{prev}}{(1 - \text{sens}) \cdot \text{prev} + \text{spec} \cdot (1 - \text{prev})}$$

depends on **prevalence**

(**prior** information available before seeing data \equiv test result).

The (testing) business must go on

specificity $P(T - \text{NOT INF})$	sensitivity $P(T + \text{INF})$	preval. $P(\text{INF})$	false positives $P(\text{NOT INF} T+)$	false negat. $P(\text{INF} T-)$
0.99	0.99	0.10	0.083	0.001
		0.01	0.500	0.000
0.90	0.50	0.10	0.643	0.058
		0.01	0.952	0.006
0.95	0.50	0.10	0.474	0.055
		0.01	0.908	0.005
0.99	0.50	0.10	0.153	0.053
		0.01	0.664	0.005
0.90	0.75	0.10	0.545	0.030
		0.01	0.930	0.003
0.95	0.75	0.10	0.375	0.028
		0.01	0.868	0.003
0.99	0.75	0.10	0.107	0.027
		0.01	0.569	0.003

Screening by stupids or helping helpful diagnostic by MD

specificity $P(T - \text{NOT INF})$	sensitivity $P(T + \text{INF})$	preval. $P(\text{INF})$	false positives $P(\text{NOT INF} T+)$	false negat. $P(\text{INF} T-)$
0.99	0.99	0.70 0.01	0.004 0.500	0.023 0.000
0.95	0.95	0.70 0.01	0.022 0.839	0.109 0.001
0.90	0.90	0.70 0.01	0.045 0.917	0.206 0.001

- Bayes theorem published in 1763 (by R. Price), further generalized by P. C. Laplace (1773).
- Further development: only in 30's of the 20th century (de Finetti – he did not write in English. . .).
- Then after WW II in context of theory of *statistical decision*
 - still only very simple applications;
 - mainly used in the crypto community.

Bayesian methods as **disinformation** by mainstream of the 20th century statistics

Big (statistical) names of the first half of the 20th century:

- **Karl Pearson (1857–1936):**
 - did not like the Bayesian concept,
 - hated R. A. Fisher,
 - formally introduced the (frequentist) p-value concept;
- **Ronald A. Fisher (1890–1962):**
 - popularized and supported the p-value concept,
 - hated K. Pearson,
 - did not like the Bayesian way of thinking,
“Theory of inverse probability is founded upon error and must be whole rejected.”
- **Jerzy Neyman (1894–1981):**
 - transformed UC Berkley to anti-Bayes camp.

“The scientific concensus” of that time.

Bayesian methods in **disent** of the 20th century statistics

It always takes some time to change disinformation to information. . .

- **Harold Jeffreys (1891–1989)**: primarily geophysicist, criticized the (frequentist) p-value concept, kept Bayes alive in Cambridge;
- **Alan Turing (1912–1954)**: developed “Banburismus” – a Bayesian method for ENIGMA code breaking, later used the Bayes rule to search for German U-boats, use of the Bayes rule brought to the group of the U.S. code breakers;
- **Andrej N. Kolmogorov (1903–1987)**: used the Bayes rule to direct Soviet artillery during WW II;
- **After WW II**: Secret life and survival of the Bayes rule mainly in the crypto community;
- **Irwing John “Jack” Good (1916–2009)**: cryptologist at Bletchley Park (with AT), after WW II – U. of Manchester, Virginia Tech.
- **Leonard J. Savage (1917–1971)**: the book “*The Foundations of Statistics*” (1954) – theory of subjective and personal probability.
- **Dennis V. Lindley (1923–2013)**: taught the Bayesian methods at UC

- Considerable progress: (\approx) since 90's of the 20th century jointly with rapid development of computers (allowing for practical use of Monte Carlo methods).
 - also (very) complex applications/statistical models hardly tractable by frequentist methods
 - unfortunately, also exponential increase of incompetent use of statistics by nonstatisticians. . .
 - computers produce something. . .

Bayesian methods: Dobrý sluha, ale zlý pán. . .

Pouze do povolanych rukou. . .

The screenshot shows the Science journal website interface. At the top, there is a navigation bar with 'Science' logo, 'Current issue', 'First release papers', 'Archive', 'About', and a 'Submit manuscript' button. Below the navigation bar, the article title 'Inferring the effectiveness of government interventions against COVID-19' is prominently displayed. The authors listed are Jan M. Brauner, Sören Minderhann, Mirvank Sharma, David Johnston, John Salvatier, Tomas Gavendai, and Anna B. Stephenson. The article is dated 15 Dec 2020 and has a DOI of 10.1126/science.abd9338. The article's abstract is visible, starting with 'Early in 2020, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) transmission was curbed in many countries by imposing combinations of nonpharmaceutical interventions. Sufficient data on transmission have now accumulated to discern the effectiveness of individual interventions. Brauner *et al.* amassed and curated data from 41 countries as input to a model to identify the individual nonpharmaceutical interventions that were the most effective at curtailing transmission during the early pandemic. Limiting gatherings to fewer than 10 people, closing high-exposure businesses, and closing schools and universities were each more effective than stay-at-home orders, which were of modest effect in slowing transmission.' The article is categorized under 'RESEARCH ARTICLE' and has a 'Structured Abstract' section. On the left side, there is a sidebar with a list of related topics: 'How to hold down transmission', 'Structured Abstract', 'Abstract', 'Cross-country NPI effectiveness modeling', 'Effectiveness of individual NPIs', 'Effectiveness of NPI combinations', 'Seroitivity and validation', 'Discussion', 'Materials and methods', 'Acknowledgments', 'Supplementary Material', and 'References and Notes'. On the right side, there is a vertical toolbar with icons for search, share, and other actions.

Section **1.2**

Basics

Setting for (simpler) statistical problems

Data: $\mathbf{Y}_1, \dots, \mathbf{Y}_N$ (d -dimensional random vectors).

Model (obvious and frequent part): $\mathbf{Y}_1, \dots, \mathbf{Y}_N \stackrel{\text{i.i.d.}}{\sim} \mathbf{Y}$.

Model (less obvious and more important part): $\mathbf{Y} \sim F(\mathbf{y}; \boldsymbol{\theta})$,

$$\mathbf{y} \in \mathbb{R}^d, F \in \mathcal{F}, \boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^\top \in \Theta \subseteq \mathbb{R}^k.$$

Usually: F absolutely continuous w.r.t. Lebesgue/count measure having a density f , i.e.,

MODEL $\equiv \mathbf{Y}_1, \dots, \mathbf{Y}_N \stackrel{\text{i.i.d.}}{\sim} \mathbf{Y}, \mathbf{Y} \sim f(\mathbf{y}; \boldsymbol{\theta}), \mathbf{y} \in \mathbb{R}^d, \boldsymbol{\theta} \in \Theta$.

If density w.r.t. count measure then $f(\mathbf{y}; \boldsymbol{\theta}) = P(\mathbf{Y} = \mathbf{y}; \boldsymbol{\theta})$.

→ **Likelihood** (given observed data $\mathbf{y}_1, \dots, \mathbf{y}_N$)

$$L(\boldsymbol{\theta}) = \prod_{i=1}^N f(\mathbf{y}_i; \boldsymbol{\theta}) = \prod_{i=1}^N L_i(\boldsymbol{\theta})$$

Frequentist (classical) approach

$\theta \in \Theta$ is unknown constant.

Principal tasks: point/interval estimation, hypothesis testing,

$$\text{e.g., } \hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} L(\theta) \text{ (maximum likelihood – ML).}$$

Statistical properties of estimators, tests (evaluation of uncertainty):

What happens if

- (a) we **repeatedly** obtain (sample) data (i.e., a full set $\mathbf{Y}_1, \dots, \mathbf{Y}_N$) **under the same stochastic conditions (model)** as the original data (e.g., unbiasedness);
- (b) we **add** more data ($\mathbf{Y}_{N+1}, \mathbf{Y}_{N+2}, \dots$) **under the same stochastic conditions (model)** as the original data (e.g., consistency, asymptotics).

Asymptotics quite crucial for practical usage as only rarely (for very, very simple models) one can derive all needed properties when $N < \infty$.

Frequentist (classical) approach

How reliable are results based on real data (always $N < \infty$, often $N \ll \infty$) that rely on asymptotics ($N = \infty$)?

Well, good question, not always an easy answer. . .

Bayesian approach

We postulate that some (stochastic) information is available about $\theta \in \Theta$ before any data arrive

→ prior distribution (*apriorní rozdělení*) above Θ .

Principal tasks: point/interval estimation, hypothesis testing, based on a (stochastic) “combination” of prior distribution and information provided by data.

Statistical properties of estimators (evaluation of uncertainty):

What are possible values of θ

given prior information and given data at hand.

No repeated sampling of data behind.

Only minor theory for what happens if $N \rightarrow \infty$

→ not really needed for practical problems.

Bayesian approach

Only minor theory for what happens if $N \rightarrow \infty$

→ not really needed for practical problems.

But, there is no free lunch, see later.

→ One of reasons why Bayesian statistics not much used until 90's of 20th century.

Notation

$p(\cdot)$: a generic symbol for a density (w.r.t. some σ -finite measure);

$p(\cdot | \cdot)$: a generic symbol for a conditional density
(w.r.t. some σ -finite measure);

$\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^\top$: a random vector with a density $p(\boldsymbol{\theta})$ w.r.t. a σ -finite measure λ on $(\Theta, \mathcal{B}(\Theta))$, $\Theta \subseteq \mathbb{R}^k$: non-empty Borel set, $\mathcal{B}(\Theta)$: Borel subsets of Θ ;

$\mathbf{Y} = (Y_1, \dots, Y_{N^*})^\top$: a random vector with a **conditional density** $p(\mathbf{y} | \boldsymbol{\theta})$ w.r.t. some σ -finite measure ν on $(\mathbb{R}^{N^*}, \mathcal{B}^{N^*})$, i.e., for any measurable sets B and C

$$P(\boldsymbol{\theta} \in B, \mathbf{Y} \in C) = \int_B \left(\int_C p(\mathbf{y} | \boldsymbol{\theta}) d\nu(\mathbf{y}) \right) p(\boldsymbol{\theta}) d\lambda(\boldsymbol{\theta}).$$

Typically

$\Theta \subseteq \mathbb{R}^k$, almost always product of (open) intervals.

Measure λ (for the model parameter): almost always Lebesgue,
symbol λ omitted from all subsequent integrals.

$p(\theta)$: **prior** distribution for model parameters.

$\mathbf{Y} \equiv (\mathbf{Y}_1, \dots, \mathbf{Y}_N) \equiv$ **Data**, $N^* = N \cdot d$.

Measure ν (for the data): Lebesgue or count or combination (product measure).

$p(\mathbf{y} | \theta) = \prod_{i=1}^N f_i(\mathbf{y}_i; \theta) \equiv$ **Model** for data.

Theorem 1.1 Bayes.

The conditional density $p(\boldsymbol{\theta} | \mathbf{y})$ of the random vector $\boldsymbol{\theta}$ given $\mathbf{Y} = \mathbf{y}$ is given as

$$p(\boldsymbol{\theta} | \mathbf{y}) = \begin{cases} \frac{p(\mathbf{y} | \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int_{\Theta} p(\mathbf{y} | \boldsymbol{\theta}^*) p(\boldsymbol{\theta}^*) d\boldsymbol{\theta}^*}, & \int_{\Theta} p(\mathbf{y} | \boldsymbol{\theta}^*) p(\boldsymbol{\theta}^*) d\boldsymbol{\theta}^* > 0, \\ 0, & \text{otherwise.} \end{cases}$$

$p(\boldsymbol{\theta} | \mathbf{y})$: **posterior** distribution (*aposteriorní rozdělení*) on $(\Theta, \mathcal{B}(\Theta))$.

Bayes theorem, more important part: $p(\boldsymbol{\theta} | \mathbf{y}) \propto p(\mathbf{y} | \boldsymbol{\theta}) p(\boldsymbol{\theta})$,
 \propto constants w.r.t. $\boldsymbol{\theta}$.

$$\int_{\Theta} p(\mathbf{y} | \boldsymbol{\theta}^*) p(\boldsymbol{\theta}^*) d\boldsymbol{\theta}^* = \int_{\Theta} p(\mathbf{y}, \boldsymbol{\theta}^*) d\boldsymbol{\theta} = p(\mathbf{y}):$$

marginal density of $\mathbf{Y} \equiv$ marginal/integrated likelihood

just a normalizing constant for $p(\boldsymbol{\theta} | \mathbf{y})$

→ meaning?

This is also the lunch price. . .

Basic theoretical problems

- Choice of a prior distribution ($p(\theta)$),
see next part.
- Point and interval estimation, hypothesis testing, ...
based on the posterior distribution $p(\theta | \mathbf{y})$.

Point and interval estimation

For a measurable set $B \subseteq \Theta$: $P(\theta \in B | \mathbf{Y} = \mathbf{y}) = \int_B p(\theta | \mathbf{y}) d\theta$.

If it exists, perhaps

$$\hat{\theta} := \int_{\Theta} \theta p(\theta | \mathbf{y}) d\theta = \mathbb{E}(\theta | \mathbf{Y} = \mathbf{y})?$$

Extension of the lunch price?

If $\Theta = \mathbb{R}$, $\theta = \theta$, let for $0 < \alpha < 1$, θ_L and θ_U satisfy

$$\int_{-\infty}^{\theta_L} p(\theta | \mathbf{y}) d\theta = \int_{\theta_U}^{\infty} p(\theta | \mathbf{y}) d\theta = \frac{\alpha}{2},$$

then

$$P(\theta \in (\theta_L, \theta_U) | \mathbf{Y} = \mathbf{y}) = \int_{\theta_L}^{\theta_U} p(\theta | \mathbf{y}) d\theta = 1 - \alpha.$$

Perhaps, $(\theta_L, \theta_U) \equiv$ interval estimate for θ ?

→ **credible** interval (*věrohodnostní interval*).

Major practical problems

Nasty (mostly analytically not tractable) **integrals**:

Denominator from the Bayes theorem (marginal likelihood)

$$p(\mathbf{y}) = \int_{\Theta} p(\mathbf{y} | \theta^*) p(\theta^*) d\theta^*.$$

Moments, quantiles, ... from the posterior distribution, e.g.,

$$\mathbb{E}(\theta | \mathbf{Y} = \mathbf{y}) = \int_{\Theta} \theta p(\theta | \mathbf{y}) d\theta.$$

See the second part of the semester.

Exercise 1.1 (The Bayesian Choice, Exercise 1.7).

An examination has 15 questions, each with 3 possible answers. Assume that 70% of the students taking the examination are prepared and answer correctly each question with probability 0.8; the remaining 30% answer at random.

- (i) Characterize the distribution of S , score of a student if one point is attributed to each correct answer.*
- (ii) Eight correct answers are necessary to pass the examination. Given that a student has passed the examination, what is the probability that (s)he/it was prepared?*

Exercise 1.2 (The Bayesian Choice, Example 1.2.2 (Bayes, 1764)).

A billiard ball W is rolled on a line of length one, with a uniform probability of stopping anywhere. It stops at U . A second ball O is then rolled N times under the same assumptions and Y denotes the number of times the ball O stopped on the left of the ball W . *Given Y , what inference can we make on U ?*

2

Choice of the prior distribution

Section **2.1**

Introduction

Posterior distribution (typically)

$$\begin{aligned} p(\boldsymbol{\theta} | \mathbf{y}) &\propto p(\mathbf{y} | \boldsymbol{\theta}) p(\boldsymbol{\theta}) \\ &= \left\{ \prod_{i=1}^N p(\mathbf{y}_i | \boldsymbol{\theta}) \right\} p(\boldsymbol{\theta}) \\ &= \left\{ \prod_{i=1}^N L_i(\boldsymbol{\theta}) \right\} L_{\text{prior}}(\boldsymbol{\theta}). \end{aligned}$$

Prior distribution

≡ information on $\boldsymbol{\theta}$ from one more virtual observation/another dataset to the likelihood.

Prior distribution

Knowledge on possible values of unknown parameters θ **before** any data arrive, **before** experiment or study are conducted.

- **Objective**

- information supported by physical, . . . theory;
- information from older data, past experiments, e.g. probab. mass of $p(\theta)$ concentrated on a confidence/credible set on θ based on older data.

- **Subjective**

- expert opinion;
- belief, philosophy.

Prior distribution, client oriented approach



Quite some space to adjust the results “as needed/requested by client”

$$p(\theta | \mathbf{y}) \propto p(\mathbf{y} | \theta) p(\theta), \quad \theta \in \Theta.$$

- Support of $p(\theta | \mathbf{y}) \subseteq$ support of $p(\theta)$.
- Client does not like negative values of θ .
Náš zákazník, náš pán. $p(\theta) \sim \text{Ga}(1, 1)$
 $\Rightarrow P(\theta > 0 | \mathbf{Y} = \mathbf{y}) = 1$ (for any data \mathbf{y}).
- Client wishes $\hat{\theta} \in (1, 2)$. *Služebníček.* $p(\theta) \sim \text{Unif}(1, 2)$
 $\Rightarrow P(\theta \in (1, 2) | \mathbf{Y} = \mathbf{y}) = 1$ (for any data \mathbf{y}).
- Client desires $\hat{\theta} = 1$. *Nic není nemožné.* $p(\theta) \sim \text{Dirac}(1)$
 $\Rightarrow P(\theta = 1 | \mathbf{Y} = \mathbf{y}) = 1$ (for any data \mathbf{y}).

Prior distribution, client oriented approach



or



?

Prior distribution

For majority of problems, nothing/not much is known in advance on θ .

Approaches/strategies have been developed to specify $p(\theta)$ such that

- for given model $p(\mathbf{y} | \theta)$, $p(\theta)$ leads to “more tractable” calculation of the marginal likelihood $p(\mathbf{y}) = \int p(\mathbf{y} | \theta) p(\theta) d\theta$;
- allow for specification of $p(\theta)$ which expresses “no” or at least “a weak” prior information.

Section **2.2**

Conjugate systems

Definition 2.1 Conjugate system of prior distributions.

The system $\mathcal{Q} = \{q(\boldsymbol{\theta}; \boldsymbol{\eta}) : \boldsymbol{\eta} \in \mathcal{H}\}$ of distributions on $\Theta \subseteq \mathbb{R}^k$, where $\boldsymbol{\eta} \in \mathcal{H} \subseteq \mathbb{R}^r$ are hyperparameters (that index the system) is said to be **conjugated** with the model $p(\mathbf{y} | \boldsymbol{\theta})$ if and only if for any $\boldsymbol{\eta} \in \mathcal{H}$ and any $\mathbf{y} \in \mathbb{R}^d$ that satisfy

$$0 < \int_{\Theta} p(\mathbf{y} | \boldsymbol{\theta}) q(\boldsymbol{\theta}; \boldsymbol{\eta}) d\boldsymbol{\theta} < \infty, \quad (2.1)$$

the distribution $p(\boldsymbol{\theta} | \mathbf{y}; \boldsymbol{\eta}) \propto p(\mathbf{y} | \boldsymbol{\theta}) q(\boldsymbol{\theta}; \boldsymbol{\eta})$ belongs to \mathcal{Q} as well.

- Condition (2.1), its part > 0 , means that we only care about data $\mathbf{Y} = \mathbf{y}$ that belong to their support given by the model.
- Meaning: take prior $p(\boldsymbol{\theta}) = q(\boldsymbol{\theta}; \boldsymbol{\eta}_{\text{prior}})$ for some choice of $\boldsymbol{\eta}_{\text{prior}} \in \mathcal{H}$, if it's conjugated with the model $p(\mathbf{y} | \boldsymbol{\theta})$ then the respective posterior $p(\boldsymbol{\theta} | \mathbf{y}) = q(\boldsymbol{\theta}; \boldsymbol{\eta}_{\text{poster}}(\mathbf{y}))$ for some $\boldsymbol{\eta}_{\text{poster}}(\mathbf{y}) \in \mathcal{H}$.

Example 2.1 (Normal model with known variance).

Data: $Y_1, \dots, Y_N \stackrel{i.i.d.}{\sim} \mathcal{N}(\theta, \sigma_0^2)$, $\theta \in \mathbb{R}$, $0 < \sigma_0^2 < \infty$ known.

$$p(\mathbf{y} | \theta) = \frac{1}{(2\pi\sigma_0^2)^{N/2}} \exp\left\{-\frac{\sum_{i=1}^N (y_i - \theta)^2}{2\sigma_0^2}\right\}$$

$$\propto \exp\left\{-\frac{\sum_{i=1}^N (y_i - \theta)^2}{2\sigma_0^2}\right\}$$

$$q(\theta; \boldsymbol{\eta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\theta - \mu)^2}{2\sigma^2}\right\} \propto \exp\left\{-\frac{(\theta - \mu)^2}{2\sigma^2}\right\}$$

$$\sim \mathcal{N}(\mu, \sigma^2), \quad \boldsymbol{\eta} = (\mu, \sigma^2)^\top, \quad \mu \in \mathbb{R}, \quad 0 < \sigma^2 < \infty.$$

Conjugate system of prior distributions

Some calculation: $p(\theta | \mathbf{y}; \eta) \sim \mathcal{N}(\mu_{\text{poster}}, \sigma_{\text{poster}}^2)$, where

$$\mu_{\text{poster}} = \frac{\frac{N}{\sigma_0^2} \bar{y} + \frac{1}{\sigma^2} \mu}{\frac{N}{\sigma_0^2} + \frac{1}{\sigma^2}}, \quad \bar{y} = \frac{1}{N} \sum_{i=1}^N y_i,$$

$$\sigma_{\text{poster}}^2 = \left(\frac{N}{\sigma_0^2} + \frac{1}{\sigma^2} \right)^{-1}.$$

Conjugate system of prior distributions

For given model, the conjugate system is not unique.

Possible method of construction: based on **sufficient** statistics
(*postačujících statistikách*).

Construction using factorization theorem and sufficient statistic

Suppose

$$p(\mathbf{y} | \theta) = g_1(\mathbf{T}(\mathbf{y}); \theta) g_2(\mathbf{y}), \quad \theta \in \Theta,$$

where g_1 and g_2 are non-negative measurable functions and $\mathbf{T}(\mathbf{Y})$ is an r -dimensional sufficient statistic for given model. Let

$$\mathcal{H} = \left\{ \eta \in \mathbb{R}^r : 0 < \int_{\Theta} g_1(\eta; \theta) d\theta < \infty \right\}.$$

Further, let

$$q(\theta; \eta) = \frac{g_1(\eta; \theta)}{\int_{\Theta} g_1(\eta; \theta) d\theta}, \quad \theta \in \Theta, \eta \in \mathcal{H}.$$

The system

$$\mathcal{Q} = \{q(\theta; \eta) : \eta \in \mathcal{H}\}$$

is conjugated with the model $p(\mathbf{y} | \theta)$, $\theta \in \Theta$ (under additional mild conditions on g_1).

Some more examples for i.i.d. univariate data, $Y_1, \dots, Y_N \stackrel{\text{i.i.d.}}{\sim} f(y; \theta)$

Bernoulli distribution, $\theta = P(Y_i = 1) \in (0, 1)$

$$p(\mathbf{y} | \theta) = \theta^{\sum_{i=1}^N y_i} (1 - \theta)^{N - \sum_{i=1}^N y_i},$$

$$q(\theta; a, b) \propto \theta^{a-1} (1 - \theta)^{b-1} \\ \sim \text{Be}(a, b), \quad \eta \equiv a > 0, b > 0,$$

$$p(\theta | \mathbf{y}; a, b) \sim \text{Be}\left(a + \sum_{i=1}^N y_i, b + N - \sum_{i=1}^N y_i\right).$$

Some more examples for i.i.d. univariate data, $Y_1, \dots, Y_N \stackrel{\text{i.i.d.}}{\sim} f(y; \theta)$

Poisson distribution $\text{Po}(\theta)$, $\theta > 0$

$$p(\mathbf{y} | \theta) = \exp(-\theta N) \frac{\theta^{\sum_{i=1}^N y_i}}{\prod_{i=1}^N y_i!},$$

$$\begin{aligned} q(\theta; a, b) &\propto \theta^{a-1} \exp(-\theta b) \\ &\sim \text{Ga}(a, b), \quad \eta \equiv a > 0, b > 0, \end{aligned}$$

$$p(\theta | \mathbf{y}; a, b) \sim \text{Ga}\left(a + \sum_{i=1}^N y_i, b + N\right).$$

Some more examples for i.i.d. univariate data, $Y_1, \dots, Y_N \stackrel{\text{i.i.d.}}{\sim} f(y; \theta)$

Normal distribution $\mathcal{N}(\mu_0, \theta^{-1})$, $\theta > 0$ (inverse variance), $\mu_0 \in \mathbb{R}$ known

$$p(\mathbf{y} | \theta) \propto \theta^{N/2} \exp\left\{-\frac{\theta}{2} \sum_{i=1}^N (y_i - \mu_0)^2\right\},$$

$$\begin{aligned} q(\theta; a, b) &\propto \theta^{a-1} \exp(-\theta b) \\ &\sim \text{Ga}(a, b), \quad \eta \equiv a > 0, b > 0, \end{aligned}$$

$$p(\theta | \mathbf{y}; a, b) \sim \text{Ga}\left(a + \frac{N}{2}, b + \frac{1}{2} \sum_{i=1}^N (y_i - \mu_0)^2\right).$$

Conjugate system of prior distributions

Some more examples for i.i.d. univariate data, $Y_1, \dots, Y_N \stackrel{\text{i.i.d.}}{\sim} f(y; \theta)$

Normal distribution $\mathcal{N}(\mu, \tau^{-1})$, $\theta = (\mu, \tau)^\top$, $\mu \in \mathbb{R}$, $\tau > 0$ (inverse variance)

$$p(\mathbf{y} | \theta) \propto \tau^{1/2} \exp\left\{-\frac{\tau}{2} \sum_{i=1}^N (y_i - \mu)^2\right\},$$

$$q(\theta; \eta) \propto ???$$

$$p(\theta | \mathbf{y}; \eta) \sim ???$$



Conjugate system of prior distributions

In a particular situation (analysis of a given dataset): value $\eta \in \mathcal{H}$ of **hyperparameters** must be chosen.

How?

Well, good question, not an obvious (and easy) answer.

Section **2.3**

Hyperparameters in a prior

Empirical Bayes methods

A way on how to choose values of **hyperparameters** based on historical data.

Assume for historical data \mathbf{Y}_{old} the model $p(\mathbf{y}_{old} | \theta)$, $\theta \in \Theta$.

Consider the prior $q(\theta; \eta)$ with the hyperparameter $\eta \in \mathcal{H}$.

→ The integrated/marginal likelihood of historical data is

$$p(\mathbf{y}_{old}; \eta) = \int_{\Theta} p(\mathbf{y}_{old} | \theta) q(\theta; \eta) d\theta$$

→ Likelihood of historical data which depends on unknown (hyper)parameter η .

Use some classical method (ML, moments, ...) to estimate η using the historical data and the model $p(\mathbf{y}_{old}; \eta)$

→ $\hat{\eta}$.

Prior for “new” data: $p(\theta) = q(\theta; \hat{\eta})$.

Example 2.2 (Random sample from a normal distribution with known variance).

Historical data and the model: $\mathbf{Y}_{old} \equiv Y_1, \dots, Y_N \stackrel{i.i.d.}{\sim} \mathcal{N}(\theta, \sigma_0^2)$,
 $\theta \in \mathbb{R}$: parameter, $0 < \sigma_0^2 < \infty$: known variance.

Prior with a hyperparameter: $q(\theta; \eta) \equiv \mathcal{N}(\mu_q, \sigma_q^2)$,
 $\eta = (\mu_q, \sigma_q^2)^\top$, $\mu_q \in \mathbb{R}$, $0 < \sigma_q^2 < \infty$.

Integrated likelihood $p(\mathbf{y}_{old}; \eta) \equiv \mathbf{Y}_{old} \sim \mathcal{N}_N(\mu_q \mathbf{1}_N, \sigma_0^2 \mathbf{I}_N + \sigma_q^2 \mathbf{1}_N \mathbf{1}_N^\top)$.

E.g., estimates for μ_q and σ_q^2 motivated by the sample mean and the sample variance:

$$\hat{\mu}_q = \frac{1}{N} \sum_{i=1}^N Y_i, \quad \hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \hat{\mu}_q)^2, \quad \hat{\sigma}_q^2 = \max(\hat{\sigma}^2 - \sigma_0^2, 0).$$

Prior for the analysis with the “new” data is then $p(\theta) \sim \mathcal{N}(\hat{\mu}_q, \hat{\sigma}_q^2)$.

Aneb přenesení odpovědnosti na jinou úroveň řízení.

Idea: We express uncertainty in a choice of hyperparameters in a stochastic way

→ by considering them as additional parameters of the model and giving them also a prior.

In other words: $q(\theta; \eta) \equiv p(\theta | \eta)$ and some hyperprior $p(\eta)$, $\eta \in \mathcal{H}$ is given.

Prior for data and the model $p(\mathbf{y} | \theta)$ at hand is then

$$p(\theta) = \int_{\mathcal{H}} p(\theta | \eta) p(\eta) d\eta = \int_{\mathcal{H}} q(\theta; \eta) p(\eta) d\eta.$$

Example 2.3 (Random sample from a normal distribution with known variance).

Data and the model: $\mathbf{Y} \equiv Y_1, \dots, Y_N \stackrel{i.i.d.}{\sim} \mathcal{N}(\theta, \sigma_0^2)$,
 $\theta \in \mathbb{R}$: parameter, $0 < \sigma_0^2 < \infty$: known variance.

Prior with a hyperparameter: $q(\theta; \eta) \sim \mathcal{N}(\eta, \sigma_q^2)$,
 $\eta \in \mathbb{R}$, $0 < \sigma_q^2 < \infty$ given (known).

Hyperprior: $p(\eta) \sim \mathcal{N}(\mu_\eta, \sigma_\eta^2)$, $\mu_\eta \in \mathbb{R}$, $0 < \sigma_\eta^2 < \infty$ both given (known).

Prior for the analysis:

$$p(\theta) = \int_{\mathbb{R}} q(\theta; \eta) p(\eta) d\eta \sim \mathcal{N}(\mu_\eta, \sigma_q^2 + \sigma_\eta^2).$$

Both θ (primary parameters) as well as η (hyperparameters) can be viewed as just parameters of the Bayesian model having the (joint) prior

$$p(\theta, \eta) = p(\theta | \eta) p(\eta),$$

where the model for data does not depend on η , i.e., where

$$p(\mathbf{y} | \theta, \eta) = p(\mathbf{y} | \theta).$$

The joint posterior $p(\theta, \eta | \mathbf{y}) \propto p(\mathbf{y} | \theta) p(\theta, \eta)$ leads to the marginal distribution

$$p(\theta | \mathbf{y}) \propto \int_{\mathcal{H}} p(\theta, \eta | \mathbf{y}) d\eta$$

which is equal to the posterior distribution $p(\theta | \mathbf{y})$ obtained by taking the (same) model $p(\mathbf{y} | \theta)$ for data and the hierarchical prior

$$p(\theta) = \int_{\mathcal{H}} p(\theta | \eta) p(\eta) d\eta.$$

That is, the “hyperparameter” integral $\int_{\mathcal{H}}$ can be calculated either in the prior phase or in the posterior phase. It does not matter.

Later (MCMC methods), we will see that the “hyperparameter” integral does not have to be calculated at all to get many useful things and that it might be easier to work with “more” dimensional posterior $p(\theta, \eta | \mathbf{y})$ rather than to work directly with $p(\theta | \mathbf{y})$.

Aneb rozmněnění odpovědnosti, až už není odpovědný nikdo za nic.

M levels of hyperparameters: $\eta_1 \in \mathcal{H}_1, \dots, \eta_M \in \mathcal{H}_M$.

Hierarchically specified **joint** prior

$$\begin{aligned} p(\theta, \eta_1, \dots, \eta_M) &= p(\theta \mid \eta_1, \dots, \eta_M) \cdot p(\eta_1 \mid \eta_2, \dots, \eta_M) \cdots p(\eta_{M-1} \mid \eta_M) \cdot p(\eta_M) \\ &\stackrel{\text{assume}}{=} p(\theta \mid \eta_1) \cdot p(\eta_1 \mid \eta_2) \cdots p(\eta_{M-1} \mid \eta_M) \cdot p(\eta_M). \end{aligned}$$

Prior for the **primary** parameters θ :

$$\begin{aligned} p(\theta) &= \int_{\mathcal{H}_1} \cdots \int_{\mathcal{H}_{M-1}} \int_{\mathcal{H}_M} \\ & p(\theta \mid \eta_1) \cdot p(\eta_1 \mid \eta_2) \cdots p(\eta_{M-1} \mid \eta_M) \cdot p(\eta_M) d\eta_M d\eta_{M-1} \cdots d\eta_1. \end{aligned}$$

Section **2.4**

Noninformative prior

Noninformative prior (princip neurčitosti)

Uniform distribution over the whole parameter space Θ

$$q(\theta) \propto 1, \quad \theta \in \Theta.$$

It can be

$$\int_{\Theta} q(\theta) d\theta = \infty$$

→ **improper density** (*nevlastní hustota*).

Improper prior can be used with the given model $p(\mathbf{y} | \theta)$ as soon as (if and only if) the integrated likelihood exists (finite), i.e., as soon as

$$p(\mathbf{y}) = \int_{\Theta} p(\mathbf{y} | \theta) 1 d\theta < \infty$$

leading to *proper* posterior

$$p(\theta | \mathbf{y}) = \frac{p(\mathbf{y} | \theta)}{p(\mathbf{y})}.$$

Noninformative with one parameterization is not necessarily noninformative with other parameterization.

Example 2.4 (Noninformative prior for standard deviation).

Consider $\theta = \sigma$, standard deviation in a i.i.d. sample from a distribution with a finite variance, $0 < \sigma < \infty$.

Let $\psi = \log(\sigma)$, $\psi \in \mathbb{R}$.

$$p(\sigma) \propto 1, \quad 0 < \sigma < \infty$$

$$\rightarrow p(\psi) \propto \exp(\psi), \quad \psi \in \mathbb{R}.$$

$$p(\psi) \propto 1, \quad \psi \in \mathbb{R}$$

$$\rightarrow p(\sigma) \propto 1/\sigma, \quad 0 < \sigma < \infty.$$

Section **2.5**
Jeffreys prior

Jeffreys prior

Prior distribution which “does not depend on parameterization”.

For models satisfying **regularity conditions** (known from the MLE theory).

Definition 2.2 Regularity conditions.

The model $p(\mathbf{y} | \boldsymbol{\theta})$, $\mathbf{y} \in N^*$ (having a density with respect to the σ -finite measure ν), $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$ satisfies the **regularity conditions** if

- (i) Θ is a non-empty open set in \mathbb{R}^k .
- (ii) The set $M = \{\mathbf{y} : p(\mathbf{y} | \boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta}$.
- (iii) For almost all $\mathbf{y} \in \mathbb{R}^{N^*}$, for all $\boldsymbol{\theta} \in \Theta$ and for each $j = 1, \dots, k$ there exist a finite partial derivative $p'_j(\mathbf{y} | \boldsymbol{\theta}) := \frac{\partial}{\partial \theta_j} p(\mathbf{y} | \boldsymbol{\theta})$.
- (iv) For each $\boldsymbol{\theta} \in \Theta$ and each $j = 1, \dots, k$ $\int_M p'_j(\mathbf{y} | \boldsymbol{\theta}) d\nu(\mathbf{y}) = 0$.
- (v) For each $\boldsymbol{\theta} \in \Theta$ and each pair (j, l) there exist a finite integral

$$J_{j,l}(\boldsymbol{\theta}) = \int_M \frac{p'_j(\mathbf{y} | \boldsymbol{\theta}) p'_l(\mathbf{y} | \boldsymbol{\theta})}{p^2(\mathbf{y} | \boldsymbol{\theta})} p(\mathbf{y} | \boldsymbol{\theta}) d\nu(\mathbf{y}).$$

- (vi) For each $\boldsymbol{\theta} \in \Theta$, the matrix $\mathbb{J}(\boldsymbol{\theta}) = (J_{j,l}(\boldsymbol{\theta}))_{j,l=1,\dots,k}$ is positive definite.

Notes

- Matrix $\mathbb{J}(\boldsymbol{\theta})$ is called the **Fisher information matrix**.
- In case of i.i.d. data $\mathbf{Y} \equiv (\mathbf{Y}_1, \dots, \mathbf{Y}_N)$, $\mathbf{Y}_i \stackrel{\text{i.i.d.}}{\sim} p_0(\mathbf{y}_i | \boldsymbol{\theta})$, $\mathbf{y}_i \in \mathbb{R}^d$, $i = 1, \dots, N$, $N^* = Nd$, we have

$$\mathbb{J}(\boldsymbol{\theta}) = N \mathbb{J}_0(\boldsymbol{\theta}),$$

where the matrix $\mathbb{J}_0(\boldsymbol{\theta})$ is based on the density $p_0(\cdot | \boldsymbol{\theta})$.

- For each $\boldsymbol{\theta} \in \Theta$ and each $j = 1, \dots, k$

$$\frac{p'_j(\mathbf{y} | \boldsymbol{\theta})}{p(\mathbf{y} | \boldsymbol{\theta})} = \frac{\partial}{\partial \theta_j} \log \{ p(\mathbf{y} | \boldsymbol{\theta}) \} =: U_j(\boldsymbol{\theta}; \mathbf{y}).$$

- Vector

$$\mathbf{U}(\boldsymbol{\theta}; \mathbf{y}) = (U_1(\boldsymbol{\theta}; \mathbf{y}), \dots, U_k(\boldsymbol{\theta}; \mathbf{y}))^\top = \frac{\partial}{\partial \boldsymbol{\theta}} \log \{ p(\mathbf{y} | \boldsymbol{\theta}) \}$$

is called the **score vector**.

Notes, cont'd

- For each $j, l = 1, \dots, k$, $J_{j,l}(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{y}|\boldsymbol{\theta})} U_j(\boldsymbol{\theta}) U_l(\boldsymbol{\theta})$, i.e.,

$$\mathbb{J}(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{y}|\boldsymbol{\theta})} \mathbf{U}(\boldsymbol{\theta}; \mathbf{Y}) \mathbf{U}^\top(\boldsymbol{\theta}; \mathbf{Y}).$$

- Under regularity conditions also

$$\mathbb{J}(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{y}|\boldsymbol{\theta})} \left[- \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \log \{ p(\mathbf{Y} | \boldsymbol{\theta}) \} \right].$$

- Matrix

$$\mathbb{I}(\boldsymbol{\theta}; \mathbf{y}) := - \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \log \{ p(\mathbf{y} | \boldsymbol{\theta}) \}$$

is called the **observed information matrix**.

- Asymptotic normality of the MLE $\hat{\boldsymbol{\theta}}_{ML} = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} \log \{ p(\mathbf{y} | \boldsymbol{\theta}) \}$ related to above matrices.

Theorem 2.1 Jeffreys.

Let the model $p(\mathbf{y} | \theta)$, $\theta \in \Theta$ satisfy the regularity conditions with the Fisher information matrix $\mathbb{J}(\theta)$. Let for $\mathbf{y} \in \mathbb{R}^{N^*}$

$$c(\mathbf{y}) = \int_{\Theta} p(\mathbf{y} | \theta) \left[\det\{\mathbb{J}(\theta)\} \right]^{1/2} d\theta.$$

Let H be a regular injective function (prosté zobrazení) $\Theta \rightarrow \Psi \in \mathcal{B}^k$. Let $\psi = H(\theta)$ and $p^*(\mathbf{y} | \psi) = p(\mathbf{y} | H^{-1}(\psi))$. The following then holds.

- (i) The model $p^*(\mathbf{y} | \psi)$, $\psi \in \Psi$ satisfies the regularity conditions.
- (ii) Let $\mathbb{J}^*(\psi)$ denote the Fisher information matrix of this model. Then for any $B \subseteq \Theta$, $B \in \mathcal{B}^k$ and any $\mathbf{y} \in \mathbb{R}^{N^*}$ such that $c(\mathbf{y}) > 0$

$$\frac{\int_B p(\mathbf{y} | \theta) \left[\det\{\mathbb{J}(\theta)\} \right]^{1/2} d\theta}{c(\mathbf{y})} = \frac{\int_{H(B)} p^*(\mathbf{y} | \psi) \left[\det\{\mathbb{J}^*(\psi)\} \right]^{1/2} d\psi}{c(\mathbf{y})}.$$

Proof.

- ▶ Regularity of the system $p^*(\mathbf{y} | \psi)$, $\psi \in \Psi$ is obvious except $\mathbb{J}^*(\psi) > 0$ for any $\psi \in \Psi$ (will be shown below).
- ▶ While noting that $\psi = \psi(\theta) = H(\theta)$, we have for $j = 1, \dots, k$:

$$\begin{aligned} \frac{\partial \log\{p(\mathbf{y} | \theta)\}}{\partial \theta_j} &= \frac{\partial \log\{p^*(\mathbf{y} | \psi(\theta))\}}{\partial \theta_j} \\ &= \sum_{u=1}^k \frac{\partial \log\{p^*(\mathbf{y} | \psi)\}}{\partial \psi_u} \cdot \frac{\partial \psi_u}{\partial \theta_j}. \end{aligned}$$

► Further ($j, l = 1, \dots, k$),

$$\begin{aligned}
 J_{j,l}(\theta) &= \mathbb{E}_{p(\mathbf{y}|\theta)} \left[\frac{\partial \log\{p(\mathbf{Y}|\theta)\}}{\partial \theta_j} \cdot \frac{\partial \log\{p(\mathbf{Y}|\theta)\}}{\partial \theta_l} \right] \\
 &= \mathbb{E}_{p(\mathbf{y}|\theta)} \left[\sum_{u=1}^k \sum_{v=1}^k \frac{\partial \psi_u}{\partial \theta_j} \cdot \frac{\partial \log\{p^*(\mathbf{Y}|\psi)\}}{\partial \psi_u} \cdot \frac{\partial \log\{p^*(\mathbf{Y}|\psi)\}}{\partial \psi_v} \cdot \frac{\partial \psi_v}{\partial \theta_l} \right] \\
 &= \sum_{u=1}^k \sum_{v=1}^k \frac{\partial \psi_u}{\partial \theta_j} \cdot \mathbb{E}_{p(\mathbf{y}|\theta)} \left[\frac{\partial \log\{p^*(\mathbf{Y}|\psi)\}}{\partial \psi_u} \cdot \frac{\partial \log\{p^*(\mathbf{Y}|\psi)\}}{\partial \psi_v} \right] \cdot \frac{\partial \psi_v}{\partial \theta_l} \\
 &= \sum_{u=1}^k \sum_{v=1}^k \frac{\partial \psi_u}{\partial \theta_j} \cdot J_{u,v}^*(\psi) \cdot \frac{\partial \psi_v}{\partial \theta_l}.
 \end{aligned}$$

- ▶ That is, $\mathbb{J}(\boldsymbol{\theta}) = \mathbb{D}^\top \mathbb{J}^*(\boldsymbol{\psi}(\boldsymbol{\theta})) \mathbb{D}$, where $\mathbb{D} = \left(\frac{\partial \boldsymbol{\psi}}{\partial \theta_1}, \dots, \frac{\partial \boldsymbol{\psi}}{\partial \theta_k} \right)$.
- ▶ Due to the fact that H is regular injective and $\mathbb{J}(\boldsymbol{\theta}) > 0$ (for any $\boldsymbol{\theta} \in \Theta$) this also shows that $\mathbb{J}^*(\boldsymbol{\psi}) > 0$ (for any $\boldsymbol{\psi} \in \Psi$).
- ▶ Now for $\mathbf{y} \in \mathbb{R}^{N^*}$ and while using the substitution formula for integrals:

$$\begin{aligned} c(\mathbf{y}) &= \int_{\Theta} \rho(\mathbf{y} | \boldsymbol{\theta}) \left[\det\{\mathbb{J}(\boldsymbol{\theta})\} \right]^{1/2} d\boldsymbol{\theta} \\ &= \int_{\Psi} \rho^*(\mathbf{y} | \boldsymbol{\psi}) \left[\det\{\mathbb{D}^\top \mathbb{J}^*(\boldsymbol{\psi}) \mathbb{D}\} \right]^{1/2} [\det(\mathbb{D})]^{-1} d\boldsymbol{\psi} \\ &= \int_{\Psi} \rho^*(\mathbf{y} | \boldsymbol{\psi}) \left[\det\{\mathbb{J}^*(\boldsymbol{\psi})\} \right]^{1/2} d\boldsymbol{\psi}. \end{aligned}$$

- Analogously for any $B \subseteq \Theta$, $B \in \mathcal{B}^k$ and any $\mathbf{y} \in \mathbb{R}^{N^*}$ such that $c(\mathbf{y}) > 0$:

$$\begin{aligned} & \frac{1}{c(\mathbf{y})} \int_B p(\mathbf{y} | \boldsymbol{\theta}) \left[\det\{\mathbb{J}(\boldsymbol{\theta})\} \right]^{1/2} d\boldsymbol{\theta} \\ &= \frac{1}{c(\mathbf{y})} \int_{H(B)} p^*(\mathbf{y} | \boldsymbol{\psi}) \left[\det\{\mathbb{D}^\top \mathbb{J}^*(\boldsymbol{\psi}) \mathbb{D}\} \right]^{1/2} [\det(\mathbb{D})]^{-1} d\boldsymbol{\psi} \\ &= \frac{1}{c(\mathbf{y})} \int_{H(B)} p^*(\mathbf{y} | \boldsymbol{\psi}) \left[\det\{\mathbb{J}^*(\boldsymbol{\psi})\} \right]^{1/2} d\boldsymbol{\psi}. \end{aligned}$$



Model	$p(\mathbf{y} \boldsymbol{\theta})$	$p^*(\mathbf{y} \boldsymbol{\psi})$
Prior	$p(\boldsymbol{\theta}) \propto \left[\det\{\mathbb{J}(\boldsymbol{\theta})\} \right]^{1/2}$	$p(\boldsymbol{\psi}) \propto \left[\det\{\mathbb{J}^*(\boldsymbol{\psi})\} \right]^{1/2}$
Posterior	$p(\boldsymbol{\theta} \mathbf{y})$ $\propto p(\mathbf{y} \boldsymbol{\theta}) \left[\det\{\mathbb{J}(\boldsymbol{\theta})\} \right]^{1/2}$	$p^*(\boldsymbol{\psi} \mathbf{y})$ $\propto p^*(\mathbf{y} \boldsymbol{\psi}) \left[\det\{\mathbb{J}^*(\boldsymbol{\psi})\} \right]^{1/2}$

Jeffreys

$$P(\boldsymbol{\theta} \in B | \mathbf{Y} = \mathbf{y}) = P(\boldsymbol{\eta} \in H(B) | \mathbf{Y} = \mathbf{y})$$

Example 2.5 (Bernoulli (alternative) i.i.d. sample).

Data: $Y_1, \dots, Y_N \stackrel{i.i.d.}{\sim} \text{Alt}(\theta)$, $0 < \theta < 1$.

$$p(\mathbf{y} | \theta) = \theta^{\sum_{i=1}^N y_i} (1 - \theta)^{N - \sum_{i=1}^N y_i},$$

$$\ell(\theta; \mathbf{y}) := \log\{p(\mathbf{y} | \theta)\} = \sum_{i=1}^N y_i \log(\theta) + (N - \sum_{i=1}^N y_i) \log(1 - \theta),$$

$$\mathbf{U}(\theta; \mathbf{y}) = \frac{\partial}{\partial \theta} \ell(\theta) = \frac{\sum_{i=1}^N y_i}{\theta} - \frac{N - \sum_{i=1}^N y_i}{1 - \theta},$$

$$\mathbb{I}(\theta; \mathbf{y}) = -\frac{\partial^2}{\partial \theta^2} \ell(\theta) = \frac{\sum_{i=1}^N y_i}{\theta^2} + \frac{N - \sum_{i=1}^N y_i}{(1 - \theta)^2},$$

$$\mathbb{J}(\theta) = \mathbb{E}_{p(\mathbf{y} | \theta)} \mathbb{I}(\theta; \mathbf{Y}) = \frac{N}{\theta(1 - \theta)}.$$

Example 2.5 (Bernoulli (alternative) i.i.d. sample).

Jeffreys prior

$$p(\theta) \propto \theta^{-1/2} (1 - \theta)^{-1/2}, \quad 0 < \theta < 1 \quad \equiv \text{Be}(1/2, 1/2).$$

Let $\psi = \theta/(1 - \theta)$ (odds).

Jeffreys prior

$$p(\psi) \propto \psi^{-1/2} (1 + \psi)^{-1}, \quad 0 < \psi < \infty.$$

Example 2.6 (Poisson i.i.d. sample).

Data: $Y_1, \dots, Y_N \stackrel{i.i.d.}{\sim} \text{Po}(\theta), \theta \in (0, \infty)$.

$$p(\mathbf{y} | \theta) = \frac{1}{\prod_{i=1}^N y_i!} \exp(-N\theta) \theta^{\sum_{i=1}^N y_i},$$

$$\ell(\theta; \mathbf{y}) := \log\{p(\mathbf{y} | \theta)\} = -\sum_{i=1}^N \log(y_i!) - N\theta + \sum_{i=1}^N y_i \log(\theta),$$

$$\mathbf{U}(\theta; \mathbf{y}) = \frac{\partial}{\partial \theta} \ell(\theta) = -N + \frac{\sum_{i=1}^N y_i}{\theta},$$

$$\mathbb{I}(\theta; \mathbf{y}) = -\frac{\partial^2}{\partial \theta^2} \ell(\theta) = \frac{\sum_{i=1}^N y_i}{\theta^2},$$

$$\mathbb{J}(\theta) = \mathbb{E}_{p(\mathbf{y} | \theta)} \mathbb{I}(\theta; \mathbf{Y}) = \frac{N}{\theta}.$$

Example 2.6 (Poisson i.i.d. sample).

Jeffreys prior: $p(\theta) \propto \frac{1}{\theta^{1/2}} \equiv \text{Ga}(0, 1/2)$.

It is *improper*.

Section **2.6**

Exercises

Exercise 2.1 (The Bayesian Choice, Example 1.2.4 (Laplace, 1786)).

Considering male and female births in Paris, Laplace wants to test whether the probability π of a male birth is above $1/2$. For 251 527 male and 241 945 female births, assuming that π has a uniform prior distribution on $(0, 1)$, Laplace obtains

$$P(\pi \leq 1/2 \mid (251\,527; 241\,945)) = 1.15 \times 10^{-42}.$$

He then deduces that this probability π is more likely to be above 50%.

How does he obtain the above formula?

Exercise 2.2 (Double exponential (Laplace) distribution).

Data and model: $\mathbf{Y} \equiv Y_1, \dots, Y_N \stackrel{i.i.d.}{\sim} \text{DE}(\theta)$, $\theta > 0$ parameter.

That is, each Y_i ($i = 1, \dots, N$) has a density

$$f(y; \theta) = \frac{\theta}{2} \exp(-\theta |y|), \quad y \in \mathbb{R}.$$

Remark: $\mathbb{E}(Y_i) = 0$, $\text{var}(Y_i) = \frac{2}{\theta^2}$.

3

Bayesian statistical inference

Section **3.1**

Inference

Bayesian statistical inference

Everything based on the **posterior** distribution

$$p(\boldsymbol{\theta} | \mathbf{y}) = \frac{p(\mathbf{y} | \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int_{\Theta} p(\mathbf{y} | \boldsymbol{\theta}^*) p(\boldsymbol{\theta}^*) d\boldsymbol{\theta}^*} \propto p(\mathbf{y} | \boldsymbol{\theta}) p(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta.$$

In the **univariate** case ($\theta \in \Theta \subseteq \mathbb{R}$):

Posterior cdf:

$$G(\theta; \mathbf{y}) := \int_{-\infty}^{\theta} p(\theta^* | \mathbf{y}) d\theta^*, \quad \theta \in \mathbb{R}.$$

Quantiles of the posterior distribution:

$$G^{-1}(\alpha; \mathbf{y}) := \inf\{\theta : G(\theta; \mathbf{y}) \geq \alpha\}, \quad 0 < \alpha < 1.$$

Note: In the multivariate case, we are usually (also) interested in the univariate characteristics related to the **margins** of the **joint** posterior distribution $p(\boldsymbol{\theta} | \mathbf{y})$.

Suitable **location** characteristic of the posterior distribution

1. **Posterior mean** (if it exists):

$$\bar{\theta} := \mathbb{E}_{p(\theta | \mathbf{y})} \theta = \int_{\Theta} \theta p(\theta | \mathbf{y}) d\theta.$$

2. **Posterior median** (in the univariate case):

$$\tilde{\theta} := G^{-1}(0.5; \mathbf{y}).$$

“Uncertainty” (in the univariate case)

≡ “spread” of the posterior distribution

1. **Posterior standard deviation** (if it exists):

$$\text{SD}_{p(\theta|\mathbf{y})}(\theta) = \sqrt{\int_{\Theta} (\theta - \bar{\theta})^2 p(\theta|\mathbf{y}) d\theta}.$$

→ counterpart of the **standard error** from the frequentist statistics.

2. **Posterior quartiles**:

$$Q_1(\theta|\mathbf{y}) := G^{-1}(0.25; \mathbf{y}), \quad Q_3(\theta|\mathbf{y}) := G^{-1}(0.75; \mathbf{y}).$$

Definition 3.1 Credible set/region.

We say that the Borel set $C_\alpha(\mathbf{y}) \subset \Theta$ ($0 < \alpha < 1$) is the $100(1 - \alpha)\%$ credible set/region (*věrohodnostní množina/oblast*) for the parameter θ if

$$P(\theta \in C_\alpha(\mathbf{y}) \mid \mathbf{Y} = \mathbf{y}) = \int_{C_\alpha(\mathbf{y})} p(\theta \mid \mathbf{y}) d\theta = 1 - \alpha.$$

The credible region is (indeed) not uniquely specified by Definition 3.1. The number $1 - \alpha$ is called the *credible level* (*věrohodnost*) (compare with *coverage* (*pokrytí*)) in the frequentist statistics.

Highest posterior density credible region

Definition 3.2 Highest posterior density credible region.

The credible region $C_\alpha(\mathbf{y})$ is called the **highest posterior density (HPD)** credible region if it satisfies

$$C_\alpha(\mathbf{y}) = \{\boldsymbol{\theta} \in \Theta : p(\boldsymbol{\theta} | \mathbf{y}) \geq k_\alpha\},$$

where k_α is the highest constant such that

$$\int_{C_\alpha(\mathbf{y})} p(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta} = 1 - \alpha.$$

If $C_\alpha(\mathbf{y})$ is the HPD credible region then for any other credible region $C_\alpha^*(\mathbf{y})$ with the same credible level

$$\int_{C_\alpha(\mathbf{y})} d\boldsymbol{\theta} \leq \int_{C_\alpha^*(\mathbf{y})} d\boldsymbol{\theta},$$

i.e., the HPD credible region has the **lowest volume** (in \mathbb{R}^k) among all credible regions with the same credible level.

Univariate case, $\theta \in \Theta \subseteq \mathbb{R}$

If $C_\alpha(\mathbf{y}) = (L_\alpha(\mathbf{y}), U_\alpha(\mathbf{y}))$ we call it **100(1 - α)% credible interval**.

Equal-tail (ET) credible interval

$$L_\alpha(\mathbf{y}) = G^{-1}(\alpha/2; \mathbf{y}), \quad U_\alpha(\mathbf{y}) = G^{-1}(1 - \alpha/2; \mathbf{y}).$$

HPD credible interval

\equiv credible interval which is also the HPD credible region.

- o It is the shortest credible interval.
- o If the posterior distribution is **symmetric** and **unimodal** then HPD = ET.

Hypothesis testing

There exist more (different) approaches to hypothesis testing in a Bayesian setting.

Here: “What is the posterior evidence against a given point $\theta_0 \in \Theta$ based on the credible region?”

G.E.P. Box and G. Tiao, 1973, *Bayesian Inference in Statistical Analysis*;

L. Held, 2004, Simultaneous posterior probability statements from Monte Carlo output, *J. of Comp. and Graph. Stat.*

Analogy to testing $H_0: \theta = \theta_0$ against $H_1: \theta \neq \theta_0$ for chosen $\theta_0 \in \Theta$ while using the duality between testing and confidence intervals/regions.

⚡ However, don't look for frequentist interpretations like type I error, its probability etc.!

Definition 3.3 Bayesian P-value.

For a given point θ_0 and a given approach to construct the credible regions $C_\alpha(\mathbf{y})$, $0 < \alpha < 1$, with $C_0(\mathbf{y}) := \Theta$ and $C_1(\mathbf{y}) := \emptyset$, the **Bayesian P-value** $p(\theta_0; \mathbf{y})$ is defined as

$$p(\theta_0; \mathbf{y}) = \inf\{\alpha : \theta_0 \notin C_\alpha(\mathbf{y})\}.$$

It is **not** the P-value in a frequentist sense!

But it has a similar interpretation:

- $p(\theta_0; \mathbf{y}) \rightarrow 0$: small posterior evidence for θ_0
(high posterior evidence against θ_0);
- $p(\theta_0; \mathbf{y}) \rightarrow 1$: high posterior evidence for θ_0
(small posterior evidence against θ_0).

→ It is popular among practitioners who use it as the classical P-value,
why not. . . .

Not easy calculation when based on the HPD credible regions.

Easily calculated in a **univariate** case ($\theta \in \Theta \subseteq \mathbb{R}$) when the P-value is based on the **equal-tail credible interval**

$$\rightarrow \quad p(\theta_0; \mathbf{y}) = 2 \min\{G(\theta_0; \mathbf{y}), 1 - G(\theta_0; \mathbf{y})\}.$$

“One-sided” variants could be defined as well.

In the literature, different “Bayesian P-values” (with a different meaning) can be found.

Bayesian prediction

Consider **independent** observations and the model

$$\mathbf{Y} \equiv \mathbf{Y}_1, \dots, \mathbf{Y}_N, \quad \mathbf{Y}_i \sim f_i(\mathbf{y}_i; \boldsymbol{\theta}), \quad i = 1, \dots, N.$$

f_i has the same functional form for all i 's and depends on i only through known factors (e.g., explanatory variables in the regression context).

→
$$p(\mathbf{y} | \boldsymbol{\theta}) = \prod_{i=1}^N f_i(\mathbf{y}_i; \boldsymbol{\theta}).$$

Consider the prior distribution $p(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$ which leads to the posterior

$$p(\boldsymbol{\theta} | \mathbf{y}) \propto p(\mathbf{y} | \boldsymbol{\theta}) p(\boldsymbol{\theta}) = \left\{ \prod_{i=1}^N f_i(\mathbf{y}_i; \boldsymbol{\theta}) \right\} p(\boldsymbol{\theta}).$$

Let \mathbf{Y}_{new} be an d_{new} -dimensional random vector being

- **independent**, given the stochastic model, of $\mathbf{Y} \equiv \mathbf{Y}_1, \dots, \mathbf{Y}_N$;
- generated by the **same probabilistic mechanism** as \mathbf{Y} , i.e.,

$$\mathbf{Y}_{new} \sim f_{new}(\mathbf{y}_{new}; \boldsymbol{\theta})$$

with known additional factors (explanatory variables, ...).

Definition 3.4 Posterior predictive distribution.

The **posterior predictive distribution** of \mathbf{Y}_{new} is the distribution whose density at $\mathbf{y}_{new} \in \mathbb{R}^{d_{new}}$ is given as

$$f_{pred}(\mathbf{y}_{new}) := \mathbb{E}_{p(\theta | \mathbf{y})} f_{new}(\mathbf{y}_{new}; \theta) = \int_{\Theta} f_{new}(\mathbf{y}_{new}; \theta) p(\theta | \mathbf{y}) d\theta.$$

Notes

- The density $f_{pred}(\mathbf{y}_{new}) = \int_{\Theta} f_{new}(\mathbf{y}_{new}; \theta) p(\theta | \mathbf{y}) d\theta$ is called the **posterior predictive density**
 \equiv (marginal) distribution of \mathbf{Y}_{new} after the information on θ included in data \mathbf{Y} is taken into account.

- Compare it with the marginal/integrated likelihood of the new observation:

$$\begin{aligned} p(\mathbf{y}_{new}) &= \int_{\Theta} p(\mathbf{y}_{new}, \theta) d\theta = \int_{\Theta} p(\mathbf{y}_{new} | \theta) p(\theta) d\theta \\ &= \int_{\Theta} f_{new}(\mathbf{y}_{new}; \theta) p(\theta) d\theta. \end{aligned}$$

- \equiv (marginal) distribution of \mathbf{Y}_{new} when only the prior information on θ is taken into account.

Point prediction

Suitable characteristic of the posterior predictive distribution, e.g.,

$$\hat{\mathbf{Y}}_{new} := \mathbb{E}_{f_{pred}(\cdot)} \mathbf{Y}_{new} = \int_{\mathbb{R}^{d_{new}}} \mathbf{y}^* f_{pred}(\mathbf{y}^*) d\nu_{new}(\mathbf{y}^*) \quad (\text{if it exists}),$$

ν_{new} : the appropriate measure related to the density of \mathbf{Y}_{new} .

Interval prediction in the univariate case ($d_{new} = 1$)

→ credible interval (HPD, ET, ...)

based on the posterior predictive density f_{pred} .

Section **3.2**

Exercise: Normal linear model

Exercise 3.1 (Normal linear model).

Data and model: $\mathbf{Y} \equiv Y_1, \dots, Y_N$ independent, $Y_i \sim \mathcal{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$, where

- $\mathbf{x}_1, \dots, \mathbf{x}_k \in \mathbb{R}^k$ are known constants;
- $\boldsymbol{\beta} \in \mathbb{R}^k$, $0 < \sigma^2 < \infty$ are unknown parameters,
 $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2)^\top \in \Theta$, $\Theta = \mathbb{R}^k \times (0, \infty)$;
- It is assumed that the $N \times k$ matrix

$$\mathbb{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_N^\top \end{pmatrix}$$

has a full column rank, $\text{rank}(\mathbb{X}) = k < N$.

Example: Weighing of light objects

Example 3.1 (Weighing of light objects).

From Box and Tiao, 1973, Bayesian Inference in Statistical Analysis.

We want to find a weight of two (very) light objects (A and B).

β_1 : weight of the object A, β_2 : weight of the object B.

Experiment (obtained values, μg):

- o 2-times: object A on the weights \rightarrow 109, 85.*
- o 9-times: object B on the weights
 \rightarrow 114, 121, 140, 122, 125, 129, 98, 134, 133.*
- o 7-times: both objects A and B on the weights
 \rightarrow 217, 203, 243, 229, 233, 221, 221.*

\rightarrow *Data $\mathbf{Y} = (Y_1, \dots, Y_N)^\top$, $N = 18$.*

\rightarrow *(Bayesian) inference on β_1 and β_2 through the linear model.*

Exercise: Normal linear model

Data and model: $\mathbf{Y} \equiv Y_1, \dots, Y_N$ independent, $Y_i \sim \mathcal{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$.

Further notation:

$$\text{SS}(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^N (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 = (\mathbf{y} - \mathbb{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbb{X}\boldsymbol{\beta}),$$

$$\mathbf{b} = \mathbf{b}(\mathbf{y}) = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbf{y} = (98.89, 124.42)^\top,$$

$$\text{SS}_e = \text{SS}_e(\mathbf{y}) = \text{SS}(\mathbf{b}; \mathbf{y}) = 2525.7.$$

The model implies:

$$\begin{aligned} p(\mathbf{y} | \boldsymbol{\theta}) &= (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2} \text{SS}(\boldsymbol{\beta}; \mathbf{y})\right\} \\ &= (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{\text{SS}_e}{2\sigma^2} - \frac{1}{2\sigma^2} (\boldsymbol{\beta} - \mathbf{b})^\top \mathbb{X}^\top \mathbb{X} (\boldsymbol{\beta} - \mathbf{b})\right\}, \quad \mathbf{y} \in \mathbb{R}^N. \end{aligned}$$

$$\ell(\boldsymbol{\theta}; \mathbf{y}) = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{N}{2} \log(\sigma^2) - \frac{\text{SS}(\boldsymbol{\beta}; \mathbf{y})}{2\sigma^2},$$

$$\mathbf{U}(\boldsymbol{\theta}; \mathbf{y}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{y}) = \begin{pmatrix} \frac{1}{\sigma^2} (\mathbb{X}^\top \mathbf{y} - \mathbb{X}^\top \mathbb{X} \boldsymbol{\beta}) \\ -\frac{N}{2\sigma^2} + \frac{\text{SS}(\boldsymbol{\beta}; \mathbf{y})}{2\sigma^4} \end{pmatrix},$$

$$\mathbb{I}(\boldsymbol{\theta}; \mathbf{y}) = -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \ell(\boldsymbol{\theta}; \mathbf{y}) = \begin{pmatrix} \frac{1}{\sigma^2} \mathbb{X}^\top \mathbb{X} & \frac{1}{\sigma^4} (\mathbb{X}^\top \mathbf{y} - \mathbb{X}^\top \mathbb{X} \boldsymbol{\beta}) \\ * & \frac{\text{SS}(\boldsymbol{\beta}; \mathbf{y})}{\sigma^6} - \frac{N}{2\sigma^4} \end{pmatrix},$$

$$\mathbf{J}(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{y}|\boldsymbol{\theta})} \mathbb{I}(\boldsymbol{\theta}; \mathbf{y}) = \begin{pmatrix} \frac{1}{\sigma^2} \mathbb{X}^\top \mathbb{X} & \mathbf{0} \\ \mathbf{0}^\top & \frac{N}{2\sigma^4} \end{pmatrix}.$$

Exercise: Normal linear model

Special case: $x_i = 1, i = 1, \dots, N, Y_1, \dots, Y_N \sim \mathcal{N}(\beta, \sigma^2)$.

$\rightarrow \mathbb{X}^\top \mathbb{X} = N,$

$$\mathbb{J}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{N}{2\sigma^4} \end{pmatrix}.$$

Jeffreys prior

1. For $\beta \in \mathbb{R}^k$ when $0 < \sigma^2 < \infty$ known.
2. For $0 < \sigma^2 < \infty$ when $\beta \in \mathbb{R}^k$ known.
3. For $\boldsymbol{\theta} = (\beta, \sigma^2)^\top \in \mathbb{R} \times (0, \infty)$ in the special case.

$$\begin{aligned} p(\mathbf{y} | \boldsymbol{\theta}) &= (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2} \text{SS}(\boldsymbol{\beta}; \mathbf{y})\right\} \\ &= (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{\text{SS}_e}{2\sigma^2} - \frac{1}{2\sigma^2} (\boldsymbol{\beta} - \mathbf{b})^\top \mathbb{X}^\top \mathbb{X} (\boldsymbol{\beta} - \mathbf{b})\right\}, \quad \mathbf{y} \in \mathbb{R}^N. \end{aligned}$$

Conjugate system

$$\tau := \sigma^{-2}, \boldsymbol{\theta} := (\boldsymbol{\beta}^\top, \tau)^\top \in \mathbb{R}^k \times (0, \infty) = \Theta.$$

$$p(\boldsymbol{\beta}, \tau) = p(\boldsymbol{\beta} | \tau) p(\tau), \quad p(\boldsymbol{\beta} | \tau) \sim \mathcal{N}(\boldsymbol{\beta}_0, \tau^{-1} \boldsymbol{\Sigma}_0),$$

$$p(\tau) \sim \text{Ga}(c_0, d_0).$$

$$\sim \mathcal{N}\text{-Ga}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0, c_0, d_0),$$

$\boldsymbol{\beta}_0 \in \mathbb{R}^k, \boldsymbol{\Sigma}_0 > \mathbf{0}, c_0 > 0, d_0 > 0$: hyperparameters.

Semiconjugate system

$$p(\beta, \tau) = p(\beta) p(\tau) \quad (\text{independence of } \beta \text{ and } \tau \text{ in the prior}),$$

$$p(\beta) \sim \mathcal{N}(\beta_0, \Sigma_0),$$

$$p(\tau) \sim \text{Ga}(c_0, d_0),$$

$\beta_0 \in \mathbb{R}^k$, $\Sigma_0 > 0$, $c_0 > 0$, $d_0 > 0$: hyperparameters.

Then $p(\beta, \tau | \mathbf{y}) \sim \mathcal{N}\text{-Ga}(\cdot, \cdot, \cdot, \cdot)$.

Jeffreys motivated improper prior

$$p(\beta) \propto 1 \equiv \mathcal{N}_k(\mathbf{0}, \mathbb{O}^{-1}),$$

$$p(\tau) \propto \frac{1}{\tau} \equiv \text{Ga}(0, 0), \quad p(\log \sigma) \propto 1.$$

Does it lead to proper posterior?

Exercise: Normal linear model

Posterior distribution, $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \tau)^\top \in \mathbb{R}^k \times (0, \infty)$

$$\begin{aligned} p(\mathbf{y} | \boldsymbol{\theta}) &= (2\pi)^{-N/2} \tau^{N/2} \exp\left\{-\frac{\tau}{2} \text{SS}(\boldsymbol{\beta}; \mathbf{y})\right\} \\ &= (2\pi)^{-N/2} \tau^{N/2} \exp\left\{-\tau \frac{\text{SS}_e}{2} - \frac{\tau}{2} (\boldsymbol{\beta} - \mathbf{b})^\top \mathbb{X}^\top \mathbb{X} (\boldsymbol{\beta} - \mathbf{b})\right\}, \quad \mathbf{y} \in \mathbb{R}^N, \end{aligned}$$

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

$$p(\boldsymbol{\beta}, \tau | \mathbf{y}) = p(\boldsymbol{\beta} | \tau, \mathbf{y}) p(\tau | \mathbf{y}),$$

$$p(\tau | \mathbf{y}) \sim \text{Ga}\left(\frac{N-k}{2}, \frac{\text{SS}_e}{2}\right),$$

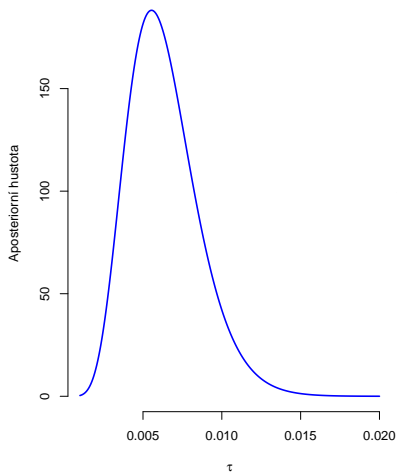
$$p(\boldsymbol{\beta} | \tau, \mathbf{y}) \sim \mathcal{N}_k(\mathbf{b}, \tau^{-1} (\mathbb{X}^\top \mathbb{X})^{-1}),$$

$$p(\boldsymbol{\beta} | \mathbf{y}) \sim \mathbf{b} + \text{mvt}_k\left(N-k, \frac{\text{SS}_e}{N-k} (\mathbb{X}^\top \mathbb{X})^{-1}\right).$$

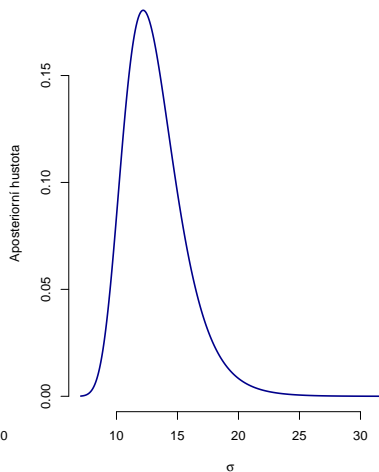
Example: Weighing of light objects

$$p(\tau | \mathbf{y}) \sim \text{Ga}\left(\frac{N-k}{2}, \frac{SS_e}{2}\right) = \text{Ga}(8, 1262.8).$$

Reziduální p esnost



Reziduální sm rodatná odchylka



Example: Weighing of light objects

$$p(\tau | \mathbf{y}) \sim \text{Ga}\left(\frac{N-k}{2}, \frac{\text{SS}_e}{2}\right) = \text{Ga}(8, 1\,262.8).$$

$$\mathbb{E}(\tau | \mathbf{Y} = \mathbf{y}) = \frac{n-k}{\text{SS}_e}$$

$$\text{var}(\tau | \mathbf{Y} = \mathbf{y}) = \frac{2(n-k)}{\text{SS}_e^2}$$

$$\mathbb{E}(\sigma^2 | \mathbf{Y} = \mathbf{y}) = \frac{\text{SS}_e}{n-k-2}, \quad \text{for } n-k > 2$$

$$\text{var}(\sigma^2 | \mathbf{Y} = \mathbf{y}) = \frac{2\text{SS}_e^2}{(n-k-2)^2(n-k-4)}, \quad \text{for } n-k > 4$$

$$\mathbb{E}(\sigma | \mathbf{Y} = \mathbf{y}) = ???$$

$$\text{var}(\sigma | \mathbf{Y} = \mathbf{y}) = ???$$

Example: Weighing of light objects

$$p(\tau | \mathbf{y}) \sim \text{Ga}\left(\frac{N-k}{2}, \frac{\text{SS}_e}{2}\right) = \text{Ga}(8, 1262.8).$$

Possible point estimates

$$\hat{\tau}_1 = \mathbb{E}(\tau | \mathbf{Y} = \mathbf{y}) = 0.00633 \quad \hat{\sigma}_1 = \mathbb{E}(\sigma | \mathbf{Y} = \mathbf{y}) = ???$$

$$\hat{\tau}_2 = \text{med}(\tau | \mathbf{Y} = \mathbf{y}) = 0.00607 \quad \hat{\sigma}_2 = \text{med}(\sigma | \mathbf{Y} = \mathbf{y}) = 12.83$$

95% credible intervals

$$\text{ET cred. interval} \quad \tau : (0.00273, 0.01142) \quad \sigma : (9.36, 19.12)$$

$$\text{HPD cred. interval} \quad \tau : (0.00235, 0.01079) \quad \sigma : (8.87, 18.22)$$

Multivariate t-distribution

$\mathbf{T} \sim \text{mvt}_k(\nu, \boldsymbol{\Sigma})$, if $\mathbf{T} = \mathbf{U} \sqrt{\frac{\nu}{V}}$, where

- $\boldsymbol{\Sigma}$: positive definite matrix (the scale matrix),
- $\mathbf{U} \sim \mathcal{N}_k(\mathbf{0}, \boldsymbol{\Sigma})$,
- $V \sim \chi_\nu^2$,
- \mathbf{U} and V independent.

It has a density

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

→ Can be used to define the mvt distribution for non-integer $\nu \in (0, \infty)$.

- $\mathbb{E}\mathbf{T} = \mathbf{0}$, if $\nu > 1$,
- $\text{var}\mathbf{T} = \frac{\nu}{\nu-2} \boldsymbol{\Sigma}$, if $\nu > 2$,
- $\text{modus}\mathbf{T} = \mathbf{0}$.

Shifted multivariate t-distribution

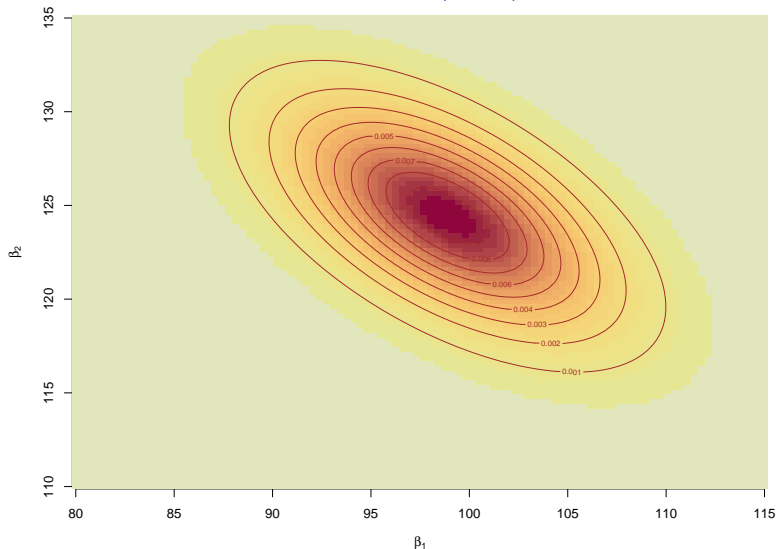
For $\boldsymbol{\mu} \in \mathbb{R}^k$, the random vector $\mathbf{Z} = \boldsymbol{\mu} + \mathbf{T}$, where $\mathbf{T} \sim \text{mvt}_k(\nu, \boldsymbol{\Sigma})$ has a density

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

- $\mathbb{E}\mathbf{Z} = \boldsymbol{\mu}$, if $\nu > 1$,
- $\text{var}\mathbf{Z} = \frac{\nu}{\nu-2} \boldsymbol{\Sigma}$, if $\nu > 2$,
- $\text{modus}\mathbf{Z} = \boldsymbol{\mu}$.

Example: Weighing of light objects

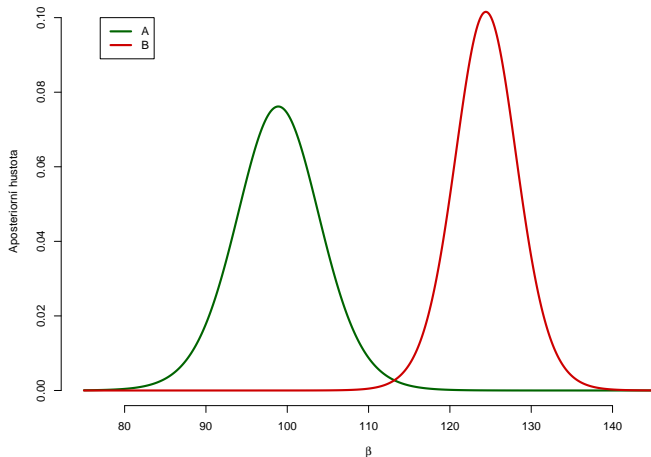
$$p(\beta | \mathbf{y}) \sim \mathbf{b} + \text{mvt}_k\left(N-k, \frac{\text{SS}_e}{N-k} (\mathbb{X}^T \mathbb{X})^{-1}\right) = \begin{pmatrix} 98.89 \\ 124.42 \end{pmatrix} + \text{mvt}_2\left(16, \frac{2525.7}{16} (\mathbb{X}^T \mathbb{X})^{-1}\right).$$



Example: Weighing of light objects

$$p(\beta | \mathbf{y}) \sim \mathbf{b} + \text{mvt}_k \left(N-k, \frac{\text{SS}_e}{N-k} (\mathbb{X}^\top \mathbb{X})^{-1} \right) = \begin{pmatrix} 98.89 \\ 124.42 \end{pmatrix} + \text{mvt}_2 \left(16, \frac{2525.7}{16} (\mathbb{X}^\top \mathbb{X})^{-1} \right).$$

→ For each $j = 1, \dots, k$ $\frac{\beta_j - b_j}{\sqrt{\frac{\text{SS}_e}{N-k} v_{j,j}}} \Big| \mathbf{Y} = \mathbf{y} \sim t_{N-k}$, $v_{j,j} = (j, j)$ diag. elem. of $(\mathbb{X}^\top \mathbb{X})^{-1}$.



Example: Weighing of light objects

$$p(\boldsymbol{\beta} | \mathbf{y}) \sim \mathbf{b} + \text{mvt}_k\left(N-k, \frac{\text{SS}_e}{N-k} (\mathbb{X}^\top \mathbb{X})^{-1}\right) = \begin{pmatrix} 98.89 \\ 124.42 \end{pmatrix} + \text{mvt}_2\left(16, \frac{2525.7}{16} (\mathbb{X}^\top \mathbb{X})^{-1}\right).$$

Possible point estimates

$$\hat{\beta}_1 = \mathbb{E}(\beta_1 | \mathbf{Y} = \mathbf{y}) = \text{med}(\beta_1 | \mathbf{Y} = \mathbf{y}) = 98.89$$

$$\hat{\beta}_2 = \mathbb{E}(\beta_2 | \mathbf{Y} = \mathbf{y}) = \text{med}(\beta_2 | \mathbf{Y} = \mathbf{y}) = 124.42$$

95% credible intervals

ET as well as HPD cred. interval $\beta_1 : (87.96, 109.83)$

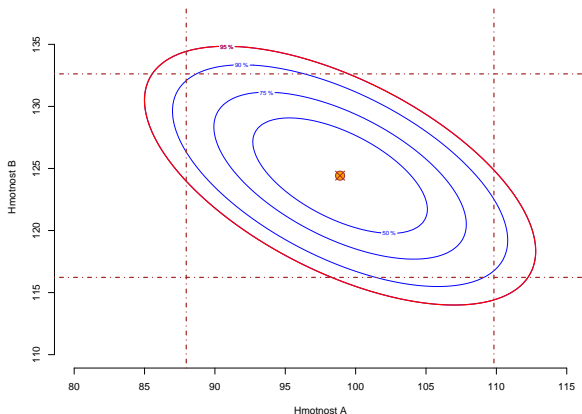
$\beta_2 : (116.22, 132.62)$

Example: Weighing of light objects

$$p(\beta | \mathbf{y}) \sim \mathbf{b} + \text{mvt}_k\left(N-k, \frac{\text{SS}_e}{N-k} (\mathbb{X}^T \mathbb{X})^{-1}\right) = \begin{pmatrix} 98.89 \\ 124.42 \end{pmatrix} + \text{mvt}_2\left(16, \frac{2525.7}{16} (\mathbb{X}^T \mathbb{X})^{-1}\right).$$

$$\rightarrow \frac{1}{k} (\beta - \mathbf{b})^T \left(\frac{N-k}{\text{SS}_e} \mathbb{X}^T \mathbb{X}\right) (\beta - \mathbf{b}) \mid \mathbf{Y} = \mathbf{y} \sim F_{k, N-k}.$$

→ 100(1 - α)% HPD credible region for β:



Exercise: Normal linear model

Proper semiconjugate prior

$$p(\boldsymbol{\beta}, \tau) = p(\boldsymbol{\beta}) p(\tau) \quad (\text{independence of } \boldsymbol{\beta} \text{ and } \tau \text{ in the prior}),$$

$$p(\boldsymbol{\beta}) \sim \mathcal{N}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0),$$

$$p(\tau) \sim \text{Ga}(c_0, d_0),$$

$\boldsymbol{\beta}_0 \in \mathbb{R}^k$, $\boldsymbol{\Sigma}_0 > 0$, $c_0 > 0$, $d_0 > 0$: hyperparameters.

$$p(\boldsymbol{\beta}, \tau | \mathbf{y}) = p(\boldsymbol{\beta} | \tau, \mathbf{y}) p(\tau | \mathbf{y}),$$

$$p(\tau | \mathbf{y}) \sim \text{Ga}\left(c_0 + \frac{N-k}{2}, d_0 + \frac{\text{SS}_e}{2}\right),$$

$$p(\boldsymbol{\beta} | \tau, \mathbf{y}) \sim \mathcal{N}_k(\mathbb{Q}^{-1} \boldsymbol{\mu}_{\text{canon}}, \mathbb{Q}^{-1}),$$

where $\mathbb{Q} = \boldsymbol{\Sigma}_0^{-1} + \tau \mathbb{X}^T \mathbb{X}$ precision matrix,

$\boldsymbol{\mu}_{\text{canon}} = \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\beta}_0 + \tau \mathbb{X}^T \mathbb{X} \mathbf{b}$ canonical mean.

4

Monte Carlo posterior inference

Section 4.1

Monte Carlo integration in statistics

Bayesian statistical inference

Everything based on the **posterior** distribution for $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^\top \in \Theta \subseteq \mathbb{R}^k$:

$$p(\boldsymbol{\theta} | \mathbf{y}) = \frac{p(\mathbf{y} | \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int_{\Theta} p(\mathbf{y} | \boldsymbol{\theta}^*) p(\boldsymbol{\theta}^*) d\boldsymbol{\theta}^*} \propto p(\mathbf{y} | \boldsymbol{\theta}) p(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta.$$

We need

- For a measurable function $t : \Theta \rightarrow \mathbb{R}$

$$\bar{t}_{\boldsymbol{\theta}} := \mathbb{E}_{p(\boldsymbol{\theta} | \mathbf{y})} t(\boldsymbol{\theta}) = \int_{\Theta} t(\boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta} \quad (\text{if it exists}).$$

- In the univariate case ($\theta \in \Theta \subseteq \mathbb{R}$):

$$G(\theta; \mathbf{y}) := \int_{-\infty}^{\theta} p(\theta^* | \mathbf{y}) d\theta^*, \quad \theta \in \mathbb{R},$$

$$G^{-1}(\alpha; \mathbf{y}) := \inf\{\theta : G(\theta; \mathbf{y}) \geq \alpha\}, \quad 0 < \alpha < 1.$$

→ credible intervals, ...

All that needs to calculate the integral $\int_{\Theta} p(\mathbf{y} | \boldsymbol{\theta}^*) p(\boldsymbol{\theta}^*) d\boldsymbol{\theta}^*$.

Monte Carlo integration in statistics

Task: For a measurable function $t : \Theta \rightarrow \mathbb{R}$, to calculate **numerically**

$$\bar{t}_\theta := \mathbb{E}_{p(\theta | \mathbf{y})} t(\theta) = \int_{\Theta} t(\theta) p(\theta | \mathbf{y}) d\theta \quad (\text{if it exists}).$$

Assumption:

$$\int_{\Theta} |t(\theta)| p(\theta | \mathbf{y}) d\theta < \infty.$$

Monte Carlo principle:

- Let $S_M = \{\theta^{(1)}, \dots, \theta^{(M)}\}$, $\theta^{(1)}, \dots, \theta^{(M)} \stackrel{\text{i.i.d.}}{\sim} p(\theta | \mathbf{y})$.
- Then (*law of large numbers*)

$$\hat{t}_M := \frac{1}{M} \sum_{m=1}^M t(\theta^{(m)}) \xrightarrow{P} \mathbb{E}_{p(\theta | \mathbf{y})} t(\theta) = \bar{t}_\theta \quad \text{as } M \rightarrow \infty.$$

- $\hat{t}_M =$ **Monte Carlo (MC) approximation/estimate** of \bar{t}_θ .

Precision of the MC integration:

Let $\int_{\Theta} \{t(\theta)\}^2 p(\theta | \mathbf{y}) d\theta < \infty$ and

$$\sigma_t^2 := \int_{\Theta} \{t(\theta) - \bar{t}\}^2 p(\theta | \mathbf{y}) d\theta.$$

Then

$$v_M := \text{var}(\hat{t}_M) = \text{var}\left\{\frac{1}{M} \sum_{m=1}^M t(\theta^{(m)})\right\} = \frac{\sigma_t^2}{M}.$$

Law of large numbers:

$$\hat{\sigma}_M^2 := \frac{1}{M-1} \sum_{m=1}^M \{t(\theta^{(m)}) - \hat{t}_M\}^2 \xrightarrow{P} \sigma_t^2.$$

Monte Carlo Error

$$\text{MCE}(\hat{t}_M) := \sqrt{\frac{1}{M(M-1)} \sum_{m=1}^M \{t(\theta^{(m)}) - \hat{t}_M\}^2} = \sqrt{\frac{\hat{\sigma}_M^2}{M}}.$$

Central limit theorem & Cramér-Sluckij:

$$\frac{\hat{t}_M - \bar{t}_\theta}{\text{MCE}(\hat{t}_M)} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) \quad \text{as } M \rightarrow \infty.$$

→ 100(1 - α)%, 0 < α < 1, confidence bounds for the approximation:

$$\hat{t}_M \pm \text{MCE}(\hat{t}_M) u_{1-\alpha/2},$$

$u_{1-\alpha/2}$: quantiles of $\mathcal{N}(0, 1)$.

Practice:

- $\mathcal{S}_M = \{\theta^{(1)}, \dots, \theta^{(M)}\}$ generated by a computer.
- The “sample size” M can be arbitrarily high.
- The Monte Carlo error can be arbitrarily low.

Really?

Section **4.2**
Special cases

$$S_M \equiv \boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(M)} \stackrel{\text{i.i.d.}}{\sim} p(\boldsymbol{\theta} | \mathbf{y})$$

Take $j \in \{1, \dots, k\}$, $x \in \mathbb{R}$.

$$t(\boldsymbol{\theta}) = \theta_j: \quad \frac{1}{M} \sum_{m=1}^M \theta_j^{(m)} \approx \mathbb{E}_{p(\boldsymbol{\theta} | \mathbf{y})} \theta_j;$$

$$t(\boldsymbol{\theta}) = \mathbb{I}_{[\theta_j \leq x]}(\boldsymbol{\theta}): \quad \frac{\#[\theta_j^{(m)} \leq x]}{M} \approx \mathbb{P}(\theta_j \leq x | \mathbf{Y} = \mathbf{y})$$

- o density estimate of $p(\theta_j | \mathbf{y})$ (histogram, kernel, ...)
→ HPD credible intervals for θ_j ;
- o quantiles of $p(\theta_j | \mathbf{y})$ → ET credible intervals for θ_j .

Special cases

Take $r : \mathbb{R}^k \rightarrow \mathbb{R}$ a measurable function, $x \in \mathbb{R}$.

$$t(\theta) = r(\theta): \quad \frac{1}{M} \sum_{m=1}^M r(\theta^{(m)}) \approx \mathbb{E}_{p(\theta | \mathbf{y})} r(\theta);$$

$$t(\theta) = \mathbb{I}_{[r(\theta) \leq x]}(\theta): \quad \frac{\# [r(\theta^{(m)}) \leq x]}{M} \approx \mathbb{P}(r(\theta) \leq x \mid \mathbf{Y} = \mathbf{y})$$

- o density estimate of $p(r(\theta) \mid \mathbf{y})$ (histogram, kernel, ...)
→ HPD credible intervals for $r(\theta)$;
- o quantiles of $p(r(\theta) \mid \mathbf{y})$ → ET credible intervals for $r(\theta)$.

Function r can be relatively complex. Nevertheless, no additional integration is required to calculate the MC estimates of $\mathbb{E}_{p(\theta | \mathbf{y})} r(\theta)$, $\mathbb{P}(r(\theta) \leq x \mid \mathbf{Y} = \mathbf{y})$, $p(r(\theta) \mid \mathbf{y})$, ...

How to generate

$$S_M \equiv \theta^{(1)}, \dots, \theta^{(M)} \stackrel{\text{i.i.d.}}{\sim} p(\theta | \mathbf{y})$$

???

Section 4.3

Random numbers generation

Random numbers generation

TASK

Generate $\theta^{(1)}, \dots, \theta^{(M)} \stackrel{\text{i.i.d.}}{\sim} F(\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^k$, F : cumulative distribution function (cdf);

$f(\cdot)$: density;

In a univariate case: $F^{-1}(\alpha) := \inf\{\theta : F(\theta) \geq \alpha\}$, $0 < \alpha < 1$;

Assume: Θ is the **support** of the distribution, i.e.,

$$P(\theta \in \Theta) = 1, \quad \forall \tilde{\Theta} \subset \mathbb{R}^k P(\theta \in \Theta \setminus \tilde{\Theta}) > 0 \implies P(\theta \in \tilde{\Theta}) < 1.$$

Inverse transform sampling

1. Generate $U \sim \text{Unif}(0, 1)$.
2. $\theta := F^{-1}(U)$.

Note:

If F is **continuous** and **strictly increasing** on Θ , it is easily seen that for any $x \in \mathbb{R}$

$$P(\theta \leq x) = P(F^{-1}(U) \leq x) = F(x).$$

The useful method if F^{-1} can be easily/efficiently calculated.

Sampling based on transformations

Based on transformations of distributions from which we (computer) are able to sample from.

Example 1: $U_1, U_2 \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(0, 1)$, $\mu \in \mathbb{R}$, $0 < \sigma < \infty$, then for

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

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

$$\theta_1, \theta_2 \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2).$$

Example 2: $U \sim \text{Unif}(0, 1)$, $0 < \lambda < \infty$, then for

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

$$\theta \sim \text{Exp}(\lambda).$$

Consecutive sampling from conditional distributions

Assume $\theta = (\theta_1^\top, \theta_2^\top, \dots, \theta_k^\top)^\top$ and

$$\begin{aligned} f(\theta) &= f(\theta_1, \dots, \theta_k) \\ &= f(\theta_1 \mid \theta_2, \dots, \theta_k) f(\theta_2 \mid \theta_3, \dots, \theta_k) \cdots f(\theta_{k-1} \mid \theta_k) f(\theta_k), \end{aligned}$$

where it is possible to generate **easily** from all (conditional) distributions on the RHS of the decomposition.

Step 1 Generate $\tilde{\theta}_k \sim f(\theta_k)$.

Step 2 Generate $\tilde{\theta}_{k-1} \sim f(\theta_{k-1} \mid \theta_k = \tilde{\theta}_k)$.

⋮

Step $k - 1$ Generate $\tilde{\theta}_2 \sim f(\theta_2 \mid \theta_3 = \tilde{\theta}_3, \dots, \theta_k = \tilde{\theta}_k)$.

Step k Generate $\tilde{\theta}_1 \sim f(\theta_1 \mid \theta_2 = \tilde{\theta}_2, \theta_3 = \tilde{\theta}_3, \dots, \theta_k = \tilde{\theta}_k)$.

$$\rightarrow \quad \tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_k)^\top \sim f(\theta).$$

Random numbers generation

More


NMST535 *Simulation Methods* (summer term, sometimes).

Luc Devroye. *Non-Uniform Random Variate Generation*. New York: Springer-Verlag, 1986.

Christian P. Robert, George Casella. *Monte Carlo Statistical Methods, 2nd Ed.* New York: Springer-Verlag.

Practical applications

Statistical packages have methods implemented to generate (pseudo)random numbers from most common (even multivariate) distributions.

 functions `runif`, `rnorm`, `rexp`, ...

For most (even slightly complex) Bayesian models, the posterior distribution is **multivariate** and **not common**...



Section 4.4

Exercise: Normal linear model

Exercise 4.1 (Normal linear model).

Data and model: $\mathbf{Y} \equiv Y_1, \dots, Y_N$ independent, $Y_i \sim \mathcal{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, \tau^{-1})$, where

- $\mathbf{x}_1, \dots, \mathbf{x}_k \in \mathbb{R}^k$ are known constants;
- $\boldsymbol{\beta} \in \mathbb{R}^k$, $0 < \tau < \infty$ are unknown parameters,
 $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \tau)^\top \in \Theta$, $\Theta = \mathbb{R}^k \times (0, \infty)$;
- It is assumed that the $N \times k$ matrix

$$\mathbb{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_N^\top \end{pmatrix}$$

has a full column rank, $\text{rank}(\mathbb{X}) = k < N$.

Example: Weighing of light objects

Example 4.1 (Weighing of light objects).

From Box and Tiao, 1973, Bayesian Inference in Statistical Analysis.

We want to find a weight of two (very) light objects (A and B).

β_1 : weight of the object A, β_2 : weight of the object B.

Experiment (obtained values, μg):

- o 2-times: object A on the weights \rightarrow 109, 85.*
- o 9-times: object B on the weights
 \rightarrow 114, 121, 140, 122, 125, 129, 98, 134, 133.*
- o 7-times: both objects A and B on the weights
 \rightarrow 217, 203, 243, 229, 233, 221, 221.*

\rightarrow *Data $\mathbf{Y} = (Y_1, \dots, Y_N)^\top$, $N = 18$.*

\rightarrow *(Bayesian) inference on β_1 and β_2 through the linear model.*

Exercise: Normal linear model

Data and model: $\mathbf{Y} \equiv Y_1, \dots, Y_N$ independent, $Y_i \sim \mathcal{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, \tau^{-1})$.

Further notation:

$$\text{SS}(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^N (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 = (\mathbf{y} - \mathbb{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbb{X}\boldsymbol{\beta}),$$

$$\mathbf{b} = \mathbf{b}(\mathbf{y}) = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbf{y} = (98.89, 124.42)^\top,$$

$$\text{SS}_e = \text{SS}_e(\mathbf{y}) = \text{SS}(\mathbf{b}; \mathbf{y}) = 2525.7.$$

The model implies:

$$\begin{aligned} p(\mathbf{y} | \boldsymbol{\theta}) &= (2\pi)^{-N/2} \tau^{N/2} \exp\left\{-\frac{\tau}{2} \text{SS}(\boldsymbol{\beta}; \mathbf{y})\right\} \\ &= (2\pi)^{-N/2} \tau^{N/2} \exp\left\{-\tau \frac{\text{SS}_e}{2} - \frac{\tau}{2} (\boldsymbol{\beta} - \mathbf{b})^\top \mathbb{X}^\top \mathbb{X} (\boldsymbol{\beta} - \mathbf{b})\right\}, \quad \mathbf{y} \in \mathbb{R}^N. \end{aligned}$$

Jeffreys motivated improper prior

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

$$p(\boldsymbol{\beta}) \propto 1 \equiv \mathcal{N}_k(\mathbf{0}, \mathbb{O}^{-1}),$$

$$p(\tau) \propto \frac{1}{\tau} \equiv \text{Ga}(0, 0), \quad p(\log \sigma) \propto 1.$$

Posterior distribution

$$p(\boldsymbol{\beta}, \tau | \mathbf{y}) = p(\boldsymbol{\beta} | \tau, \mathbf{y}) p(\tau | \mathbf{y}),$$

$$p(\tau | \mathbf{y}) \sim \text{Ga}\left(\frac{N-k}{2}, \frac{\text{SS}_e}{2}\right),$$

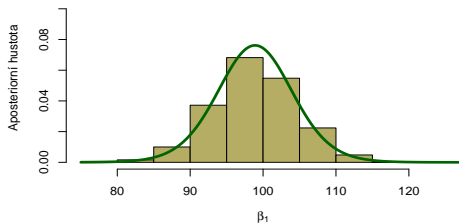
$$p(\boldsymbol{\beta} | \tau, \mathbf{y}) \sim \mathcal{N}_k(\mathbf{b}, \tau^{-1} (\mathbb{X}^\top \mathbb{X})^{-1}),$$

$$p(\boldsymbol{\beta} | \mathbf{y}) \sim \mathbf{b} + \text{mvt}_k\left(N-k, \frac{\text{SS}_e}{N-k} (\mathbb{X}^\top \mathbb{X})^{-1}\right).$$

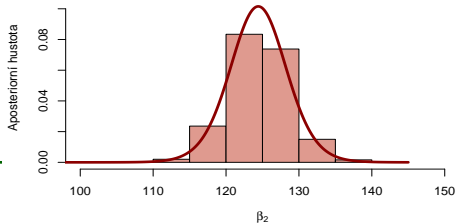
Example: Weighing of light objects

Marginal posterior densities (M=1 000)

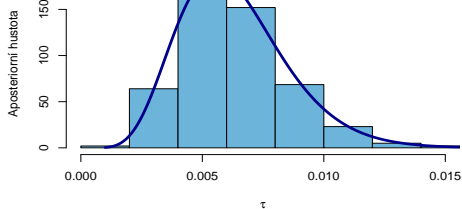
Hmotnost A



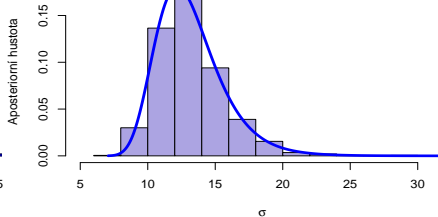
Hmotnost B



Inverzní rozptyl



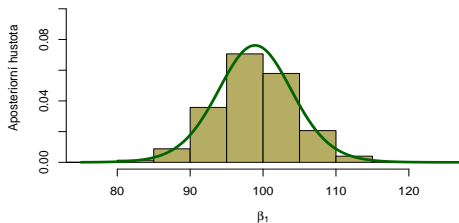
Sm rodatná odchylka



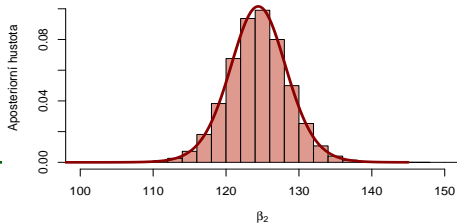
Example: Weighing of light objects

Marginal posterior densities (M=1 000 000)

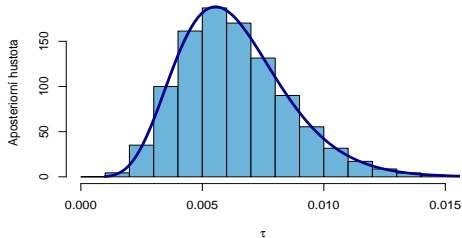
Hmotnost A



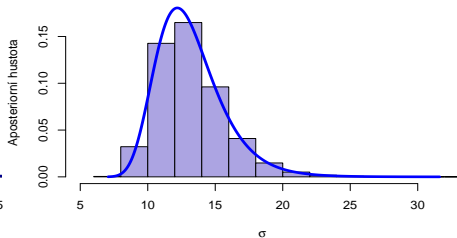
Hmotnost B



Inverzní rozptyl

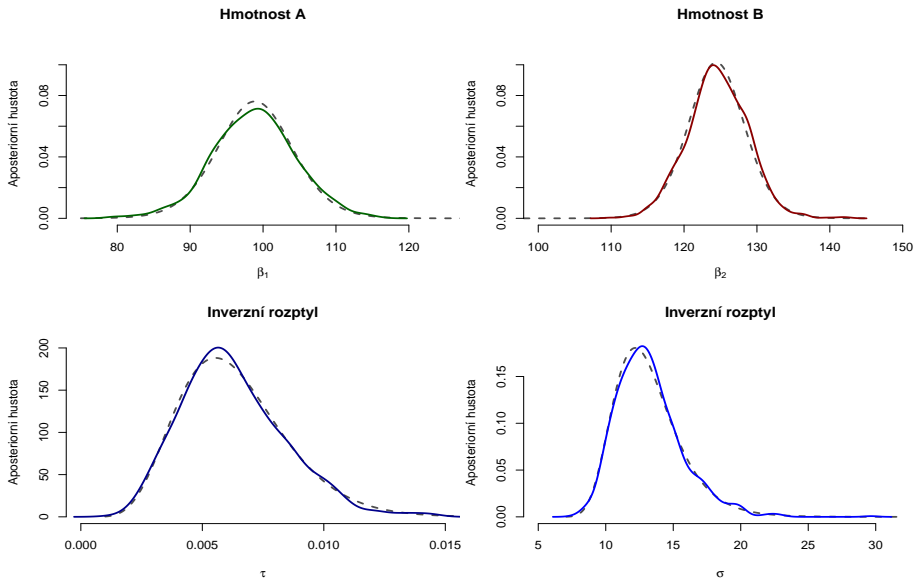


Sm rodatná odchylka



Example: Weighing of light objects

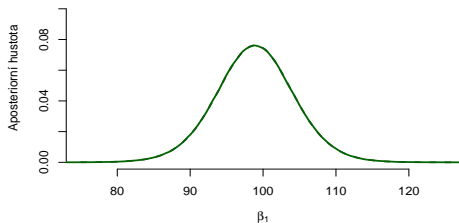
Marginal posterior densities (M=1 000)



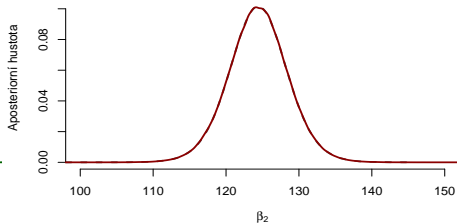
Example: Weighing of light objects

Marginální aposteriorní hustoty ($M=1\ 000\ 000$)

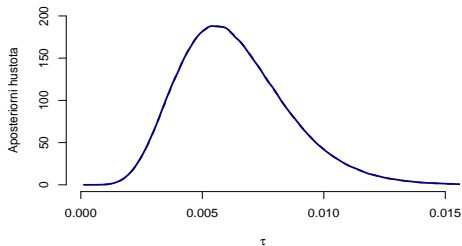
Hmotnost A



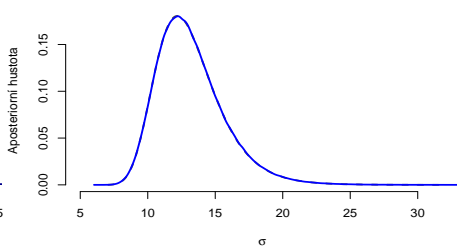
Hmotnost B



Inverzní rozptyl

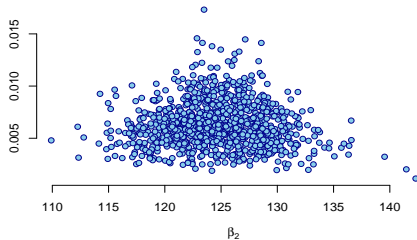
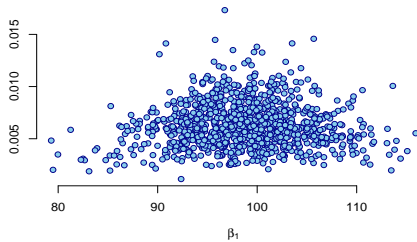
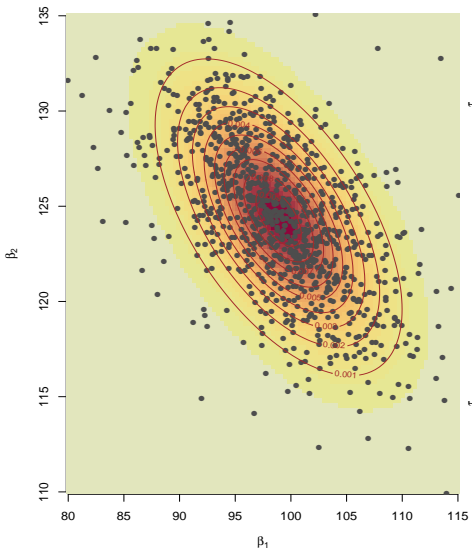


Inverzní rozptyl



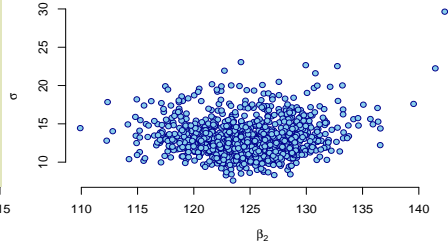
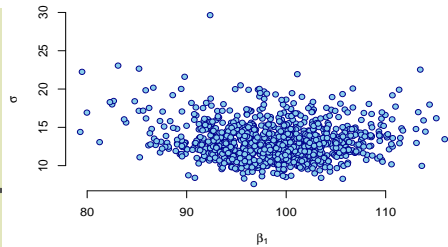
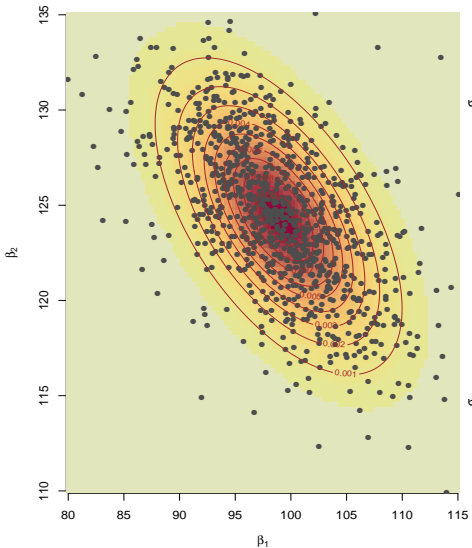
Example: Weighing of light objects

Joint samples from the posterior distribution (M=1 000)



Example: Weighing of light objects

Joint samples from the posterior distribution ($M=1\ 000$)



Example: Weighing of light objects

Posterior inference for β ($M=1\ 000$)

	β_1	β_2
Posterior mean	98.8947	124.4211
MC estimate	98.7944	124.6197
MC error	0.1804	0.1312
Posterior median	98.8947	124.4211
MC estimate	98.7813	124.5673
95% ET cred. interval	(87.9641; 109.8253)	(116.2231; 132.6190)
MC estimate	(86.9761; 110.0815)	(116.7594; 132.6802)
95% HPD cred. interval	(87.9641; 109.8253)	(116.2231; 132.6190)
MC estimate	(87.9245; 110.7076)	(116.4892; 132.2210)

Example: Weighing of light objects

Posterior inference for β ($M=1\,000\,000$)

	β_1	β_2
Posterior mean	98.8947	124.4211
MC estimate	98.8874	124.4192
MC error	0.0055	0.0041
Posterior median	98.8947	124.4211
MC estimate	98.8849	124.4196
95% ET cred. interval	(87.9641; 109.8253)	(116.2231; 132.6190)
MC estimate	(87.9680; 109.8184)	(116.2239; 132.6113)
95% HPD cred. interval	(87.9641; 109.8253)	(116.2231; 132.6190)
MC estim	(88.0765; 109.9202)	(116.1496; 132.5329)

Example: Weighing of light objects

Posterior inference for τ a σ (M=1 000)

	τ	σ
Posterior mean	0.00633	?
MC estimate	0.00629	13.207
MC error	0.0000689	0.0776
Posterior median	0.00607	?
MC estimate	0.00601	12.904
95% ET cred. interval	(0.00273; 0.01142)	?
MC estimate	(0.00272; 0.01097)	(9.547; 19.183)
95% HPD cred. interval	?	?
MC estimate	(0.00248; 0.01039)	(8.987; 18.186)

Example: Weighing of light objects

Posterior inference for τ a σ (M=1 000 000)

	τ	σ
Posterior mean	0.00633	?
MC estimate	0.00633	13.198
MC error	0.0000022	0.0025
Posterior median	0.00607	?
MC estimate	0.00607	12.838
95% ET cred. interval	(0.00273; 0.01142)	?
MC estimate	(0.00274; 0.01142)	(9.356; 19.116)
95% HPD cred. interval	?	?
MC estimate	(0.00237; 0.01081)	(8.872; 18.224)

5

Markov chain Monte Carlo (MCMC) methods

Section **5.1**

Introduction

Alternative to generation of a random sample.

Construct the **Markov chain**

$$\theta^{(0)} \rightarrow \theta^{(1)} \rightarrow \dots \rightarrow \theta^{(B)} \longrightarrow \theta^{(B+1)} \rightarrow \theta^{(B+2)} \rightarrow \dots \rightarrow \theta^{(B+M)}$$

whose **stationary/limiting** distribution is the **requested** distribution.

→ From certain point B , the **Markov chain** behaves **almost** as the random sample and can be used in **almost** the same way as with the **Monte Carlo** methods to calculate integrals etc.

→ **Markov chain Monte Carlo (MCMC)**.

Markov chain Monte Carlo

NMSA334 *Random Processes 1*: Markov chains with a **countable** state space.

Here: $\theta \in \Theta \subseteq \mathbb{R}^k$, need a (random) sample from $p(\theta | \mathbf{y}) =: f(\theta)$

\equiv distribution which is continuous w.t.to the Lebesgue measure on $(\Theta, \mathcal{B}(\Theta))$.

\rightarrow **Uncountable** state space.

Sound theory: NMTP539 *Markov Chain Monte Carlo Methods*

\rightarrow Here: Scetch of the theory needed to understand the principles.

Many books exist. . .

- o Dani Gamerman, Hedibert F. Lopes (2006).
Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, 2nd Edition. Boca Raton: Chapman & Hall/CRC.
- o Steve Brooks, Andrew Gelman, Galin L. Jones, Xiao-Li Meng (2011).
Handbook of Markov Chain Monte Carlo. Boca Raton: Chapman & Hall/CRC.

Section **5.2**

Markov chains with a general state space

Markov chains

$\Theta \subseteq \mathbb{R}^k$: the **state space**;

$f(\theta)$: distribution/density which is continuous w.r.to the Lebesgue measure
on (Θ, \mathcal{T}) , $\mathcal{T} = \mathcal{B}(\Theta)$

→ **target** distribution/density,

in a Bayesian analysis, $f(\theta) \equiv p(\theta | \mathbf{y})$.

Definition 5.1 Markov kernel.

The measurable mapping $P : \Theta \times \mathcal{T} \rightarrow [0, 1]$ is called the **Markov kernel** (*markovské jádro*) on (Θ, \mathcal{T}) if

1. For each $T \in \mathcal{T}$, $P(\cdot, T)$ is a non-negative measurable function on Θ .
2. For each $\theta \in \Theta$, $P(\theta, \cdot)$ is a probability measure on \mathcal{T} .

- Generalization of **transition probabilities**.
- $P(\theta, T)$: probability of transition from the state θ to a state in the set T .

Transition density

For each $\theta \in \Theta$, there exist a **density** of the distribution $P(\theta, \cdot)$
(w.r.to the Lebesgue measure)

\equiv **transition density** (*přechodová hustota*)

→ Notation: $p(\theta, \psi)$, $\psi \in \Theta$,

i.e., for $T \in \mathcal{T}$

$$P(\theta, T) = \int_T p(\theta, \psi) d\psi.$$

The transition density could also be viewed as a conditional density

$$p(\theta, \psi) \equiv p(\psi | \theta), \quad \theta \in \Theta, \psi \in \Theta.$$

Definition 5.2 Homogeneous Markov chain.

The random process $\{\theta^{(m)} : m = 0, 1, \dots\}$ is the **homogeneous Markov chain** (**homogenní markovský řetězec**) with the **transition kernel** (**s přechodovým jádrem**) P and the **initial distribution** (**počátečním rozdělením**) $f_0(\theta)$ if for each $m \in \mathbb{N}_0$ its finite-dimensional distributions satisfy for any $T_0, \dots, T_m \in \mathcal{T}$ the (markovian) condition

$$\begin{aligned} & P(\theta^{(0)} \in T_0, \theta^{(1)} \in T_1, \dots, \theta^{(m)} \in T_m) \\ &= \int_{T_0} \int_{T_1} \cdots \int_{T_{m-1}} P(\theta^{(m-1)}, T_m) p(\theta^{(m-2)}, \theta^{(m-1)}) d\theta^{(m-1)} \\ & \quad \cdots p(\theta^{(0)}, \theta^{(1)}) d\theta^{(1)} f_0(\theta^{(0)}) d\theta^0 \\ &= \int_{T_0} \int_{T_1} \cdots \int_{T_{m-1}} \int_{T_m} p(\theta^{(m-1)}, \theta^{(m)}) d\theta^{(m)} p(\theta^{(m-2)}, \theta^{(m-1)}) d\theta^{(m-1)} \\ & \quad \cdots p(\theta^{(0)}, \theta^{(1)}) d\theta^{(1)} f_0(\theta^{(0)}) d\theta^0. \end{aligned}$$

Properties

For any measurable and bounded function h on (Θ, \mathcal{T}) and any $m \in \mathbb{N}_0$:

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

For any $T \in \mathcal{T}$, with $h = \mathbb{I}_T$, we get

$$\mathbb{P}(\boldsymbol{\theta}^{(m+1)} \in T \mid \boldsymbol{\theta}^{(m)}, \dots, \boldsymbol{\theta}^{(0)}) = \mathbb{P}(\boldsymbol{\theta}^{(m+1)} \in T \mid \boldsymbol{\theta}^{(m)}),$$

markovian property.

Homogeneous Markov chains

Distribution of the Markov chain at time $m + 1$

$\pi(\theta)$: distribution/density which is continuous w.r.to the Lebesgue measure on (Θ, \mathcal{T}) ,
→ distribution/density of a state θ at a certain time point m .

Notation for $T \in \mathcal{T}$:

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

While using the transition density, we can also write

$$\pi P(T) = \int_{\Theta} \left(\int_{\mathcal{T}} p(\theta, \psi) d\psi \right) \pi(\theta) d\theta = \int_{\Theta} \int_{\mathcal{T}} p(\theta, \psi) \pi(\theta) d\psi d\theta.$$

Homogeneous Markov chains

Distribution of the Markov chain at time $m + 1$

Fubini theorem (for any $T \in \mathcal{T}$):

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

→ $\pi P(\cdot)$ is again the **probability distribution** on (Θ, \mathcal{T}) with the density

$$\int_{\Theta} p(\theta, \psi) \pi(\theta) d\theta \quad \text{w.r.to the Lebesgue measure.}$$

If $\pi(\cdot)$ is the distribution of the state of the Markov chain at time m then $\pi P(\cdot)$ is the distribution of the state at time $m + 1$.

Homogeneous Markov chains

Stationary distribution

Definition 5.3 Stationary distribution.

The probability distribution $\pi(\cdot)$ is called the **stationary distribution** (*stacionární rozdělení*) of the homogeneous Markov chain with the transition kernel P if $\pi(\cdot) = \pi P(\cdot)$, that is if for any $T \in \mathcal{T}$

$$\pi P(T) = \pi(T),$$

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

Definition 5.4 Reversibility.

The homogeneous Markov chain with the transition kernel P is called **reversible** (*reversibilní*) with respect to the probability distribution π if for any $T, S \in \mathcal{T}$

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

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

Homogeneous Markov chains

Reversibility

$\int_{\mathcal{T}} P(\theta, S) \pi(\theta) d\theta$ can be viewed as a **joint** probability distribution

on $(\Theta \times \Theta, \mathcal{T} \otimes \mathcal{T})$ with a joint probability measure of sets $T, S \in \mathcal{T}$

$$Q_1(T, S) = \int_{\mathcal{T}} P(\theta, S) \pi(\theta) d\theta = \int_{\mathcal{T}} \int_S p(\theta, \psi) \pi(\theta) d\psi d\theta,$$

and a **joint** density (w.r.to the Lebesgue measure)

$$q_1(\theta, \psi) = p(\theta, \psi) \pi(\theta), \quad \theta, \psi \in \Theta.$$

$\int_{\mathcal{S}} P(\psi, T) \pi(\psi) d\psi$ can be viewed as a **joint** probability distribution

on $(\Theta \times \Theta, \mathcal{T} \otimes \mathcal{T})$ with a joint probability measure of sets $T, S \in \mathcal{T}$

$$Q_2(S, T) = \int_{\mathcal{S}} P(\psi, T) \pi(\psi) d\psi = \int_{\mathcal{S}} \int_T p(\psi, \theta) \pi(\psi) d\theta d\psi,$$

and a **joint** density (w.r.to the Lebesgue measure)

$$q_2(\psi, \theta) = p(\psi, \theta) \pi(\psi), \quad \psi, \theta \in \Theta.$$

Homogeneous Markov chains

Reversibility

Reversibility with respect to the distribution π

For any $T, S \in \mathcal{T}$

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

\equiv Equality of the probability measures.

→ The **joint** distribution of states at times m and $m + 1$ is the same as the **joint** distribution of states at times $m + 1$ and m .

Homogeneous Markov chains

Reversibility

Detailed balance condition (*detailní podmínka rovnováhy*)

for a transition density:

$$p(\theta, \psi) \pi(\theta) = p(\psi, \theta) \pi(\psi) \quad \text{for almost all } \theta, \psi \in \Theta.$$

It (clearly) **implies** reversibility:

$$\int_T \int_S p(\theta, \psi) \pi(\theta) d\psi d\theta = \int_S \int_T p(\psi, \theta) \pi(\psi) d\theta d\psi \quad \text{for any } T, S \in \mathcal{T}.$$

Reversibility implies stationarity

Take $S = \Theta$ in the definition of reversibility, for any $T \in \mathcal{T}$:

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

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

$$\pi(T) = \pi P(T).$$

detailed balance condition \implies reversibility \implies stationarity
with respect to the distribution π

Homogeneous Markov chains

m -step transition probability kernel

For $\theta \in \Theta$ and $T \in \mathcal{T}$, denote $P^0(\theta, T) := \delta_\theta(T)$ (Dirac at θ).

Definition 5.5 m -step transition probability kernel.

The m -step transition probability kernel (*přechodové jádro m -tého řádu*) of the homogeneous Markov chain with the transition kernel \mathbf{P} is given inductively as

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

Chapman-Kolmogorov equality

For any $n \leq m$

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

Homogeneous Markov chains

Limiting distribution

Definition 5.6 Limiting distribution.

The probability distribution π on (Θ, \mathcal{T}) is called **the limiting distribution** (*limitní rozdělení*) of the Markov chain $\{\theta^{(m)} : m = 0, 1, \dots\}$ generated by the transition kernel P if

$$\lim_{m \rightarrow \infty} P^m(\theta, T) = \pi(T) \quad \text{for almost all } \theta \in \Theta \text{ and for all } T \in \mathcal{T}.$$

Note. If π is the limiting distribution then for **arbitrary** initial distribution f_0 and for any $T \in \mathcal{T}$

$$P(\theta^{(m)} \in T) = \int_{\Theta} P^m(\theta, T) f_0(\theta) d\theta \xrightarrow{m \rightarrow \infty} \int_{\Theta} \pi(T) f_0(\theta) d\theta = \pi(T).$$

Limiting distribution must be stationary

If π is the **limiting** distribution of the homogeneous Markov chain then it is also the **stationary** distribution of this chain.

We have for any $T \in \mathcal{T}$ and for almost all $\theta \in \Theta$

$$\begin{aligned}\pi(T) &= \lim_{m \rightarrow \infty} P^m(\theta, T) = \lim_{m \rightarrow \infty} \int_{\Theta} P(\psi, T) P^{m-1}(\theta, \psi) d\psi \\ &= \int_{\Theta} P(\psi, T) \pi(\psi) d\psi = \pi P(T).\end{aligned}$$

TO REMEMBER

detailed balance condition \implies reversibility \implies stationarity
with respect to the distribution π

π is limiting \implies π is stationary

There is no guarantee that the limiting distribution exists.

Section **5.3**

MCMC principles

Připomenutí, čím se zabýváme

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

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

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

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

▣ Monte Carlo integrace

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

$$p(\theta, \psi) f(\theta) = p(\psi, \theta) f(\psi) \quad \text{pro s.v. } \theta, \psi \in \Theta$$

a máme potřebný markovský řetězec.

▮▶ Toto není nikterak obtížné, jak bude záhy ukázáno.

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

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

$$\hat{t}_M = \frac{1}{M} \sum_{m=1}^M t(\theta^{(B+m)}) \xrightarrow{P} \int_{\Theta} t(\theta) f(\theta) d\theta \quad \text{pro } M \rightarrow \infty.$$

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

$$v_M = \text{var}_{f(\theta)}(\hat{t}_M)$$

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

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

- 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 $f(\theta)$.
 - 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!

Druhá největší obtíž při praktické aplikaci MCMC

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

$$v_M = \text{var}_{f(\theta)}(\hat{t}_M) = \frac{\text{var}_{f(\theta)}(t(\theta))}{M}.$$

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

$$v_M = \text{var}_{f(\theta)}(\hat{t}_M) \geq \frac{\text{var}_{f(\theta)}(t(\theta))}{M}.$$

- 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 .

- 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 $v_M = \frac{1}{M} \text{var}_{f(\theta)}(t(\theta))$.
- Konstrukce markovského řetězce s nízkou autokorelovaností snadná není a obtížnost této snahy závisí na konkrétní aplikaci.

Section **5.4**
Gibbs algorithm

- 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:

- $\Theta = \prod_{i=1}^k \Theta_i$, $\theta = (\theta_1^\top, \dots, \theta_k^\top)^\top$
- Cílové (stacionární) rozdělení má hustotu $f(\theta)$ vzhledem k součinné míře $\lambda_1 \otimes \dots \otimes \lambda_k$, přičemž λ_i je σ -konečná míra s $\lambda_i(\Theta_i) > 0$ ($i = 1, \dots, k$).
ZDE: Lebesgueova míra na $(\mathbb{R}^d, \mathcal{B}^d)$.
- $\Theta = \{\theta : f(\theta) > 0\}$.
- Jsme schopni (snadno) generovat z **plně podmíněných** (*full conditional*) rozdělení

$$f(\theta_i | \theta_{-i}) = f(\theta_i | \theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_k).$$

Algoritmus:

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

2. (i) generuj $\boldsymbol{\theta}_1^{(m+1)}$ z podmíněného rozdělení

$$f(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2^{(m)}, \dots, \boldsymbol{\theta}_k^{(m)}).$$

(ii) generuj $\boldsymbol{\theta}_2^{(m+1)}$ z podmíněného rozdělení

$$f(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1^{(m+1)}, \boldsymbol{\theta}_3^{(m)}, \dots, \boldsymbol{\theta}_k^{(m)}).$$

(iii) generuj $\boldsymbol{\theta}_3^{(m+1)}$ z podmíněného rozdělení

$$f(\boldsymbol{\theta}_3 \mid \boldsymbol{\theta}_1^{(m+1)}, \boldsymbol{\theta}_2^{(m+1)}, \boldsymbol{\theta}_4^{(m)}, \dots, \boldsymbol{\theta}_k^{(m)}).$$

⋮

(k) generuj $\boldsymbol{\theta}_k^{(m+1)}$ z podmíněného rozdělení

$$f(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_1^{(m+1)}, \dots, \boldsymbol{\theta}_{k-1}^{(m+1)}).$$

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

Přechodová hustota

$$p(\theta, \psi) = \prod_{i=1}^k f(\psi_i | \psi_1, \dots, \psi_{i-1}, \theta_{i+1}, \dots, \theta_k).$$

- Odpovídá přechodovému jádru

$$P(\theta, T) = \int_T p(\theta, \psi) d\psi.$$

Theorem 5.1

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

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

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
 - $\Theta = \prod_{i=1}^k \Theta_i$, $\theta = (\theta_1^\top, \dots, \theta_k^\top)^\top$.
 - Cílové (stacionární) rozdělení má hustotu $f(\theta)$ vzhledem k součinnové míře $\lambda_1 \otimes \dots \otimes \lambda_k$, přičemž λ_i je σ -konečná míra s $\lambda_i(\Theta_i) > 0$ ($i = 1, \dots, k$).
 - $\Theta = \{\theta : f(\theta) > 0\}$.
- 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!

- Lze ukázat, že markovský řetězec generovaný Gibbsovým algoritmem **ne-splňuje** 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.

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

Plně podmíněná rozdělení

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

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

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 $f(\theta_i | \theta_{-i})$ 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 $f(\theta)$ nezávislí),
 - o “rodiče” θ_i .

Lineární model

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

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

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

Příklad: Lineární model s neinformativním apriorním rozdělením

Věrohodnost

- Označme: $\mathbf{b} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbf{y}$,
 $SS_e = \|\mathbf{y} - \mathbb{X}\mathbf{b}\|^2$.
- Věrohodnost:

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

Příklad: Lineární model s neinformativním apriorním rozdělením

Aposteriorní rozdělení

- Bylo odvozeno:

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

kde $p(\tau | \mathbf{y}) \sim \mathcal{G}\left(\frac{n-k}{2}, \frac{SS_e}{2}\right)$,

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

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

Příklad: Lineární model s neinformativním apriorním rozdělením

Plně podmíněná rozdělení

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

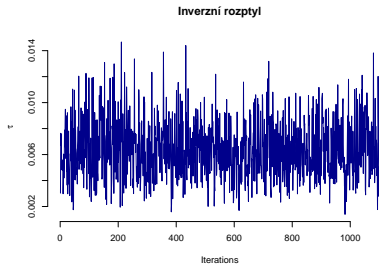
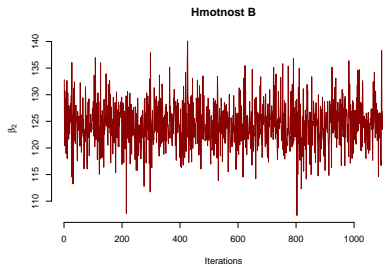
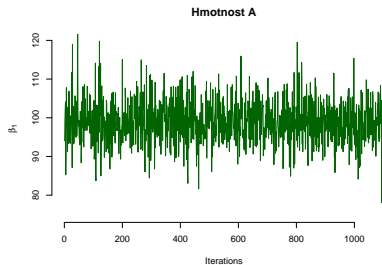
$$p(\boldsymbol{\beta} | \dots) = p(\boldsymbol{\beta} | \tau, \mathbf{y}) \sim \mathcal{N}_k(\mathbf{b}, \tau^{-1} (\mathbb{X}^\top \mathbb{X})^{-1}),$$

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

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

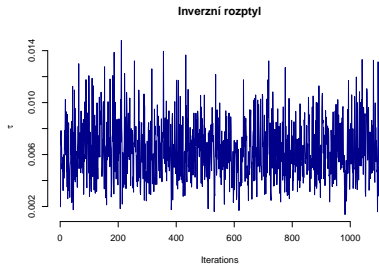
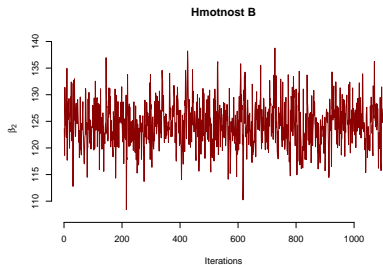
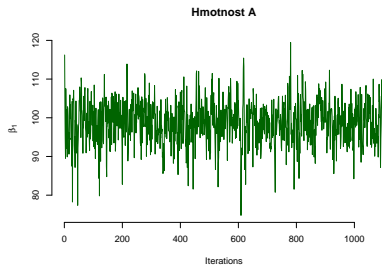
Příklad: Vážení lehkých objektů

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



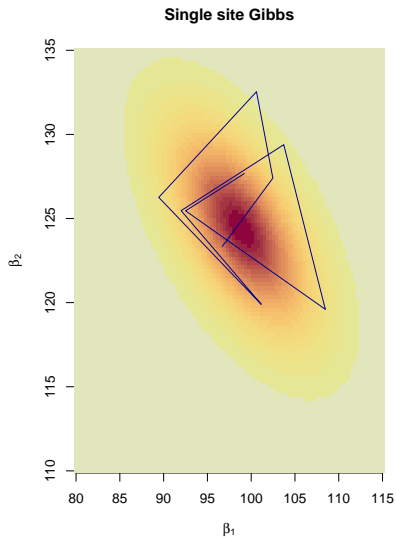
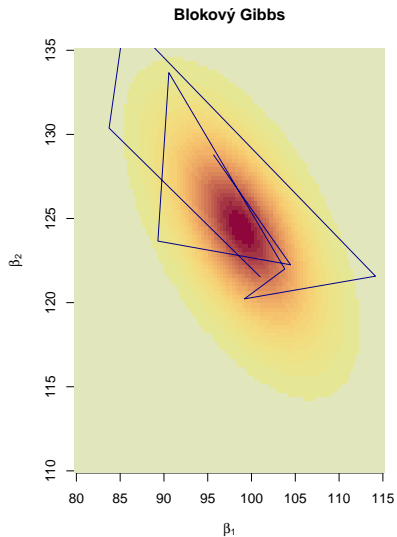
Příklad: Vážení lehkých objektů

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



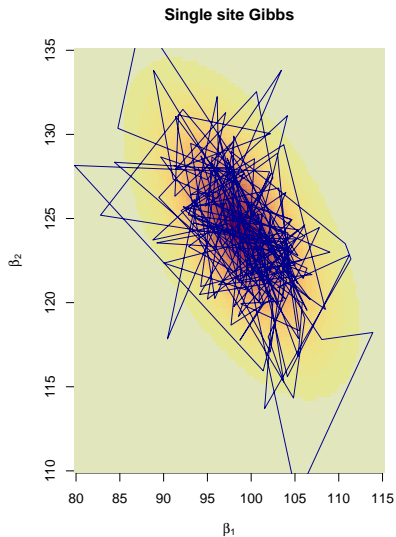
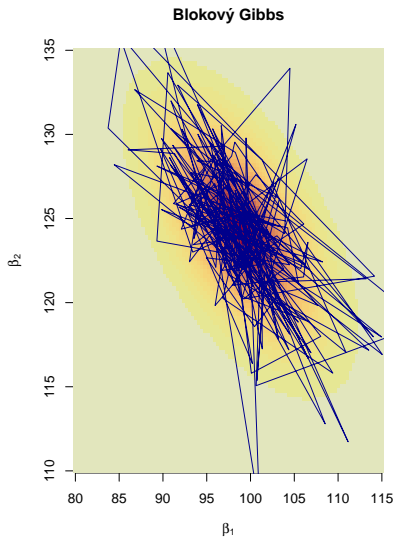
Příklad: Vážení lehkých objektů

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



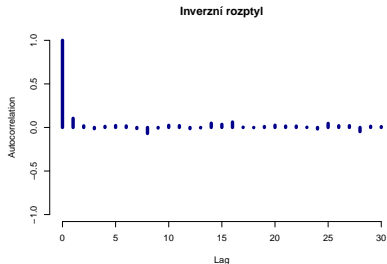
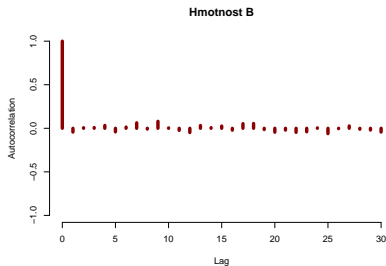
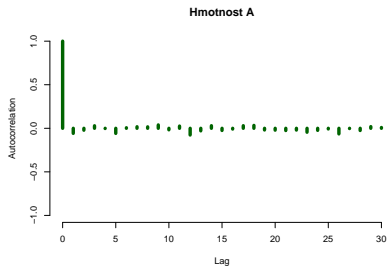
Příklad: Vážení lehkých objektů

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



Příklad: Vážení lehkých objektů

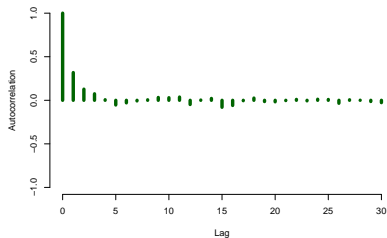
Blokový Gibbsův algoritmus: Odhady autokorelačních funkcí ($B=100$, $M=1\ 000$)



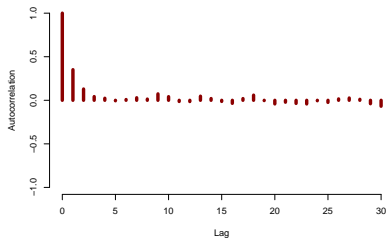
Příklad: Vážení lehkých objektů

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

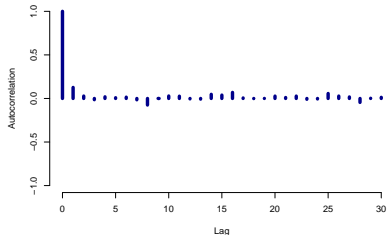
Hmotnost A



Hmotnost B



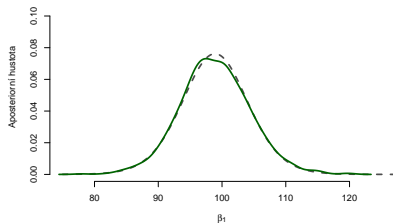
Inverzní rozptyl



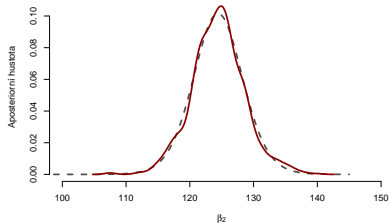
Příklad: Vážení lehkých objektů

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

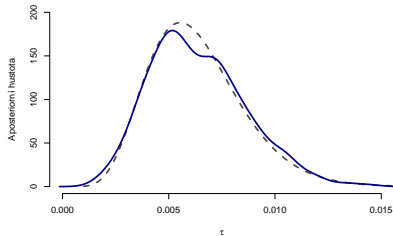
Hmotnost A



Hmotnost B



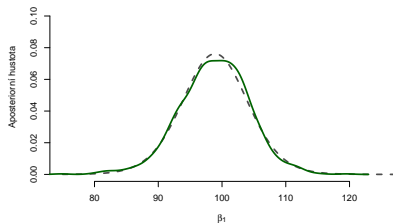
Inverzní rozptyl



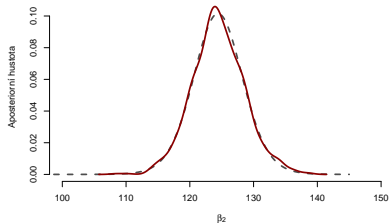
Příklad: Vážení lehkých objektů

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

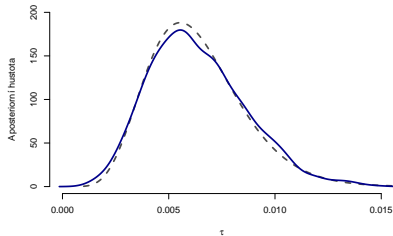
Hmotnost A



Hmotnost B



Inverzní rozptyl



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 inferenze 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)

Section **5.5**

Metropolis-Hastings algorithm

- 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 má hustotu $f(\theta)$ vzhledem k σ -konečné míře λ s $\lambda(\Theta) > 0$.
ZDE: Lebesgueova míra na $(\mathbb{R}^d, \mathcal{B}^d)$, resp. na $(\Theta, \mathcal{B}(\Theta))$.
- $\Theta = \{\theta : f(\theta) > 0\}$

Algoritmus:

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

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

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

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

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

Poznámky

Poznámky:

- Pro aplikaci MH algoritmu není potřeba znát normující konstantu cílové hustoty $f(\theta)$.
 - Ideální pro použití v bayesovské statistice.
- Návrhová hustota $q(\theta, \psi)$ může být **libovolná**.
 - Nevhodná volba $q(\theta, \psi)$ však vede k vysoké autokorelaci a s tím spojené neefektivitě.
 - Příliš “ambiciózní” $q(\theta, \psi)$ vede k malým pravděpodobnostem přijetí návrhu a řetězec pak dlouho setrvává v jednom stavu
 - ▣ vysoká autokorelace.
 - Příliš “opatrné” $q(\theta, \psi)$ 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. $q(\theta, \psi) = q(\psi, \theta) \quad \forall \theta, \psi \in \Theta$
 - ▣▶ **Metropolisův** algoritmus.
- Hlavní část pravděpodobnosti přijetí

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

obvykle počítáme v logaritmickém měřítku, tj.

$$\begin{aligned} \log\{\alpha^*(\theta^{(m)}, \psi)\} &= \log\{f(\psi)\} + \log\{q(\psi, \theta^{(m)})\} \\ &\quad - \log\{f(\theta^{(m)})\} - \log\{q(\theta^{(m)}, \psi)\} \end{aligned}$$

▣▶ 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*)

$$q(\theta, \psi) = q_0(\psi) \quad \forall \theta \in \Theta.$$

- 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: $q_0(\psi) = f(\psi)$
 - ▣ generujeme přímo náhodný výběr z cílového rozdělení $f(\theta)$.

Metropolisův-Hastingsův algoritmus

Návrhové rozdělení

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

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

$$q(\theta, \psi) = q_0(\psi - \theta) \quad \forall \theta \in \Theta.$$

- q_0 : nějaká hustota vzhledem k σ -konečné míře λ s nosičem na Θ .
- Návrh: $\psi = \theta + \mathbf{Z}$, kde \mathbf{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. $q_0(\mathbf{z}) = q_0(-\mathbf{z})$), potom $q(\theta, \psi) = q(\psi, \theta)$ a při počítání pravděpodobnosti přijetí nemusíme vůbec počítat hodnoty hustoty q_0 (resp. návrhové hustoty q).

Theorem 5.2 .

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

Proof. Viz přednáška NMTP539 Metody Markov Chain Monte Carlo.



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

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

$$q(\theta, \psi) = q_0(\psi - \theta), \quad q_0(\mathbf{z}) = q_0(-\mathbf{z})$$

(symetrická náhodná procházka)

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

- Podrobnosti viz přednáška NMTP539.

Některé další metody pro generování z plně podmíněných rozdělání

- **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 rozdělání z exponenciální třídy rozdělání.

- **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**.

Lineární model

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

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

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

Příklad: Lineární model s neinformativním apriorním rozdělením

Věrohodnost

- Označme: $\mathbf{b} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbf{y}$,
 $SS_e = \|\mathbf{y} - \mathbb{X}\mathbf{b}\|^2$.
- Věrohodnost:

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

Příklad: Lineární model s neinformativním apriorním rozdělením

Aposteriorní rozdělení

- Bylo odvozeno:

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

kde $p(\tau | \mathbf{y}) \sim \mathcal{G}\left(\frac{n-k}{2}, \frac{SS_e}{2}\right)$,

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

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

Příklad: Lineární model s neinformativním apriorním rozdělením

Plně podmíněná rozdělení

$$p(\boldsymbol{\beta} | \dots) = p(\boldsymbol{\beta} | \tau, \mathbf{y}) \sim \mathcal{N}_k(\mathbf{b}, \tau^{-1}(\mathbb{X}^T \mathbb{X})^{-1}),$$

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

Příklad: Lineární model s neinformativním apriorním rozdělením

Metropolis within Gibbs algoritmus

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

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

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

- V kroku $m + 1$ algoritmu navrhujeme β_{prop} vygenerované z rozdělení $\mathcal{N}_k(\beta^{(m)}, \mathbb{D}_{prop})$.
- Hlavní část pravděpodobnosti přijetí návrhu je

$$\begin{aligned} \alpha^*(\beta^{(m)}, \beta_{prop}) &= \frac{p(\beta_{prop} | \dots) q(\beta_{prop}, \beta^{(m)})}{p(\beta^{(m)} | \dots) q(\beta^{(m)}, \beta_{prop})} = \frac{p(\beta_{prop} | \dots)}{p(\beta^{(m)} | \dots)} \\ &= \exp \left[-\frac{\tau^{(m)}}{2} \left\{ (\beta_{prop} - \mathbf{b})^\top \mathbb{X}^\top \mathbb{X} (\beta_{prop} - \mathbf{b}) - (\beta^{(m)} - \mathbf{b})^\top \mathbb{X}^\top \mathbb{X} (\beta^{(m)} - \mathbf{b}) \right\} \right]. \end{aligned}$$

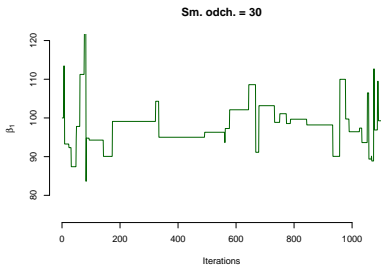
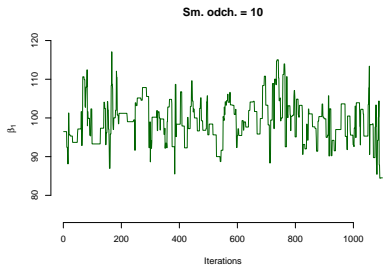
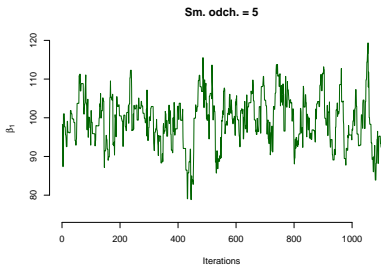
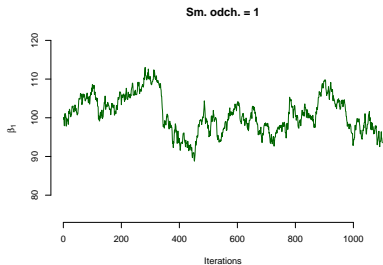
Příklad: Vážení lehkých objektů

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

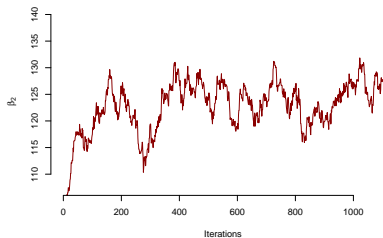
Metropolis within Gibbs algorithm: Generované hodnoty β_1 (B=100, M=1 000)



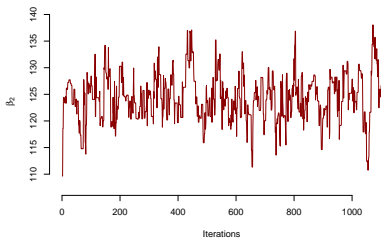
Příklad: Vážení lehkých objektů

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

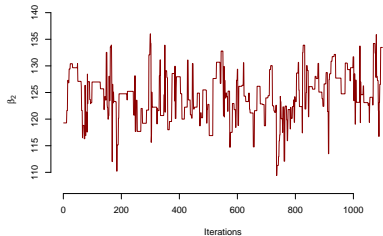
Sm. odch. = 1



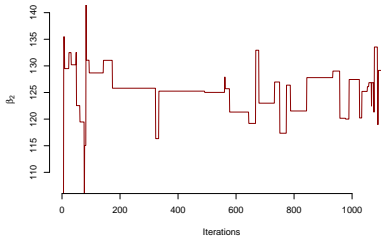
Sm. odch. = 5



Sm. odch. = 10

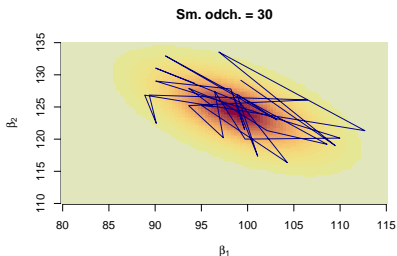
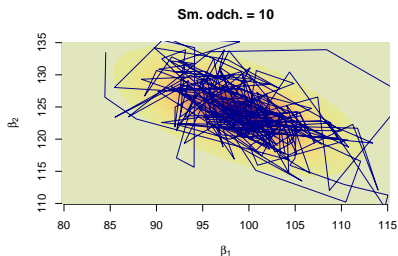
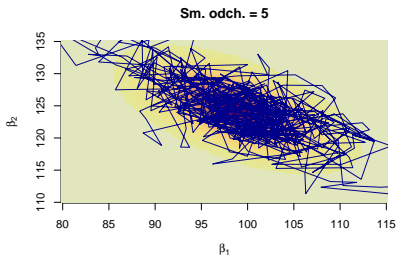
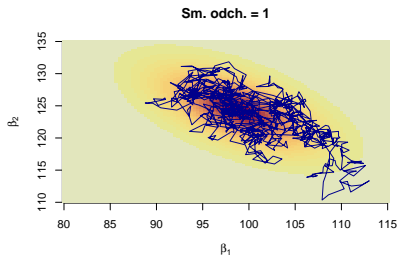


Sm. odch. = 30



Příklad: Vážení lehkých objektů

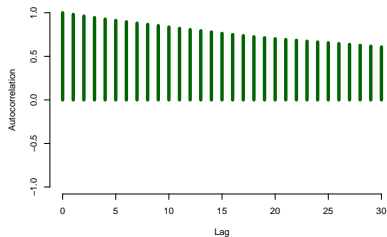
Metropolis within Gibbs algorithm: Generované hodnoty β (iterace 101 – 1 100)



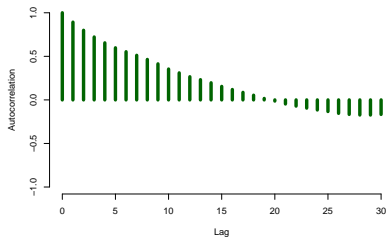
Příklad: Vážení lehkých objektů

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

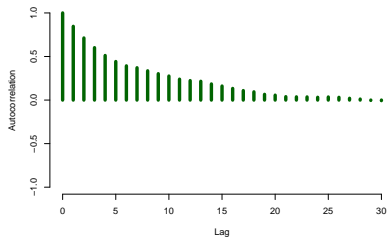
Sm. odch. = 1



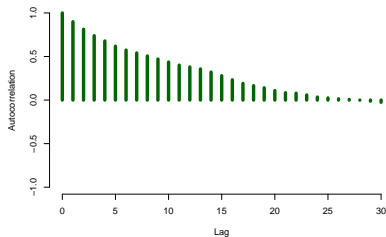
Sm. odch. = 5



Sm. odch. = 10



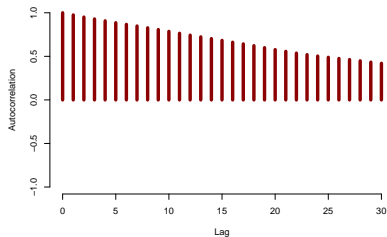
Sm. odch. = 30



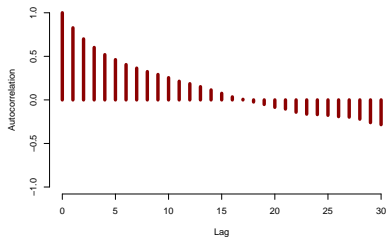
Příklad: Vážení lehkých objektů

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

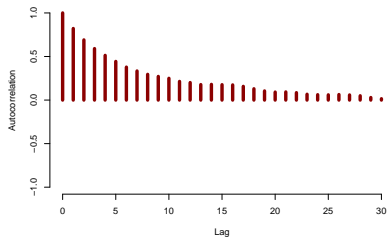
Sm. odch. = 1



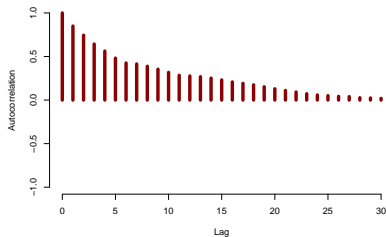
Sm. odch. = 5



Sm. odch. = 10



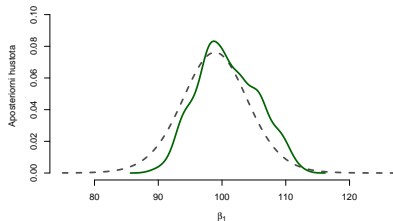
Sm. odch. = 30



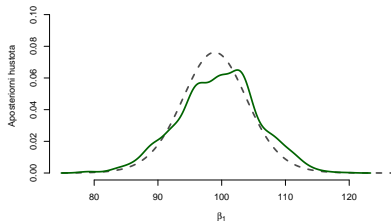
Příklad: Vážení lehkých objektů

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

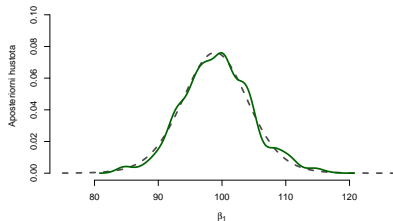
Smer. odch. = 1



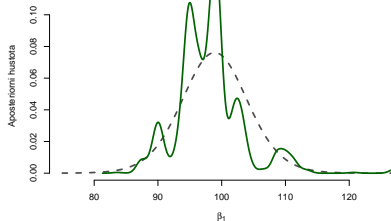
Smer. odch. = 5



Smer. odch. = 10



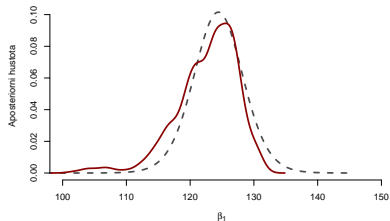
Smer. odch. = 30



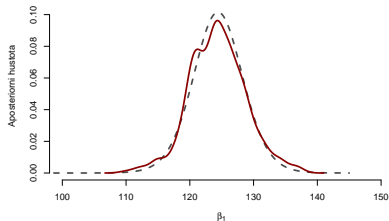
Příklad: Vážení lehkých objektů

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

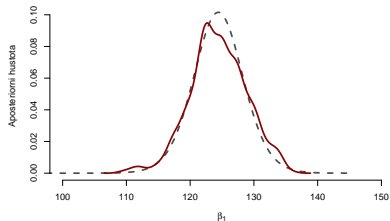
Smer. odch. = 1



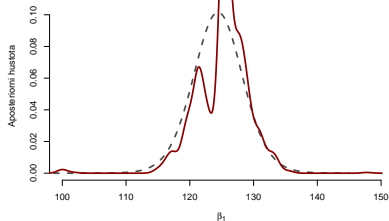
Smer. odch. = 5



Smer. odch. = 10



Smer. odch. = 30



6

Hierarchical models

Section **6.1**

Hierarchical prior

- Choice of a prior influences (possibly a lot) the **posterior**.
- Bayesian statistics can easily be misused.
- Most “bayesian” applications from past 30 years:
 - not motivated by an attempt to utilize any prior information
 - main motivation: the proposed model cannot be (not even numerically) estimated in a frequentist way (typically using the **maximum likelihood**), nevertheless, it is possible to estimate it in a bayesian way using computer **simulations (MCMC, ...)**
 - mostly, there is no real prior information available.

- The prior information is only rarely rich enough to be able to consider the chosen prior distribution to be precisely and **without any error** defined.
- It is necessary to express **uncertainty** in a choice of the prior distribution.

▶ Bayesian model with a hierarchically specified prior distribution

- decomposition of the prior distribution into several levels of conditional distributions,
- uncertainty on each level is expressed by a prior distribution on the next level.

Definition 6.1 Bayesian model with a hierarchically specified prior distribution.

Bayesian model with a hierarchically specified prior distribution is a statistical model with the likelihood $L(\boldsymbol{\psi}) = p(\mathbf{y} | \boldsymbol{\psi})$ and a prior distribution $p(\boldsymbol{\psi})$ where $p(\boldsymbol{\psi})$ is factorized into conditional distributions

$$p_0(\boldsymbol{\psi} | \zeta_1), p_1(\zeta_1 | \zeta_2), \dots, p_{m-1}(\zeta_{m-1} | \zeta_m)$$

and a marginal distribution $p_m(\zeta_m)$ such that

$$p(\boldsymbol{\psi}) =$$

$$\int p_0(\boldsymbol{\psi} | \zeta_1) p_1(\zeta_1 | \zeta_2) \cdots p_{m-1}(\zeta_{m-1} | \zeta_m) p_m(\zeta_m) d\zeta_1 \cdots d\zeta_m.$$

Parameters contained in ζ_i are called **hyperparameters** of the i -th level ($1 \leq i \leq m$).

Model

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

- **Parameters:** $\boldsymbol{\psi} = (\boldsymbol{\beta}^\top, \tau)^\top$, kde $\tau = \sigma^{-2} > 0$
- **Likelihood:** $L(\boldsymbol{\psi}) = p(\mathbf{y} | \boldsymbol{\psi}) \equiv \mathcal{N}_n(\mathbb{X}\boldsymbol{\beta}, \tau^{-1} \mathbf{I}_n)$
- **Conjugate prior distribution:**

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

$$p(\boldsymbol{\beta} | \tau) \equiv \mathcal{N}_k(\boldsymbol{\beta}_0, \tau^{-1} \boldsymbol{\Sigma}_0)$$

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

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

Example: Linear model

- Do not consider c_0 and/or d_0 as fixed constants but allow for a randomness in their selection.

- ▮ **hierarchical prior**

- for instance:

$$p(\tau | d_0) \equiv \mathcal{G}(c_0, d_0)$$

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

- c_0 : 1st level fixed (hyper)parameter
- d_0 : 1st level random hyperparameter
- g_0, h_0 : 2nd level fixed (hyper)parameters

Section **6.2**

Hierarchical likelihood

Hierarchically specified likelihood

- Hierarchical specification (also in a non-bayesian context) is often a natural way on how to construct a **realistic** stochastic model for a particular real situation/application/data
- Benchmark application areas
 - Data obtained by a **stratified** sample or otherwise *grouped* data.
 - Repeated measurements (per subject).
 - **Longitudinal** data (biostatistics), **panel** data (econometrics).

Example: NTP TER84073 trial on mice

- Data from *National Toxicology Program*
- 94 pregnant mice were given, in pre-specified moments, certain amount of ethylene glycol (EG).
- On a day 17 of pregnancy, the mice were killed and the fetuses weights have been recorded.

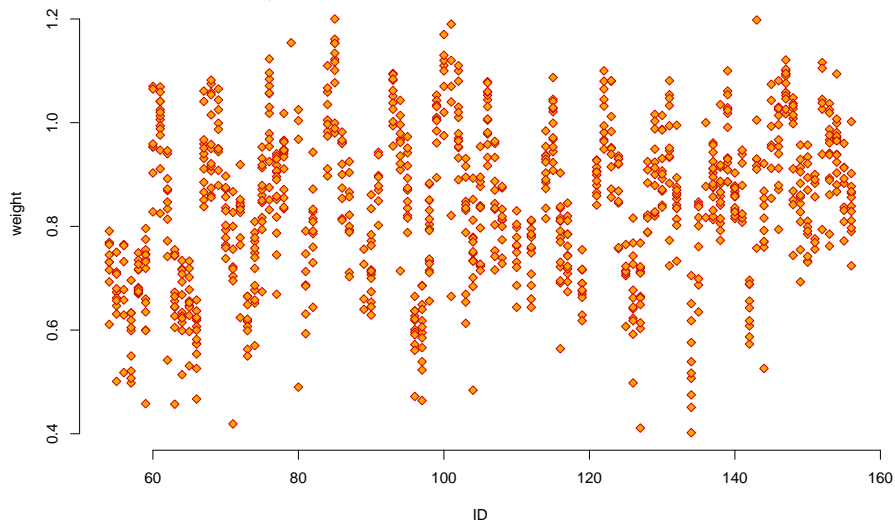
- **Stochastic representation of data:**

$Y_{i,j}$ = weight of the j -th fetus of i -th mouse, $i = 1, \dots, N$, $j = 1, \dots, n_i$

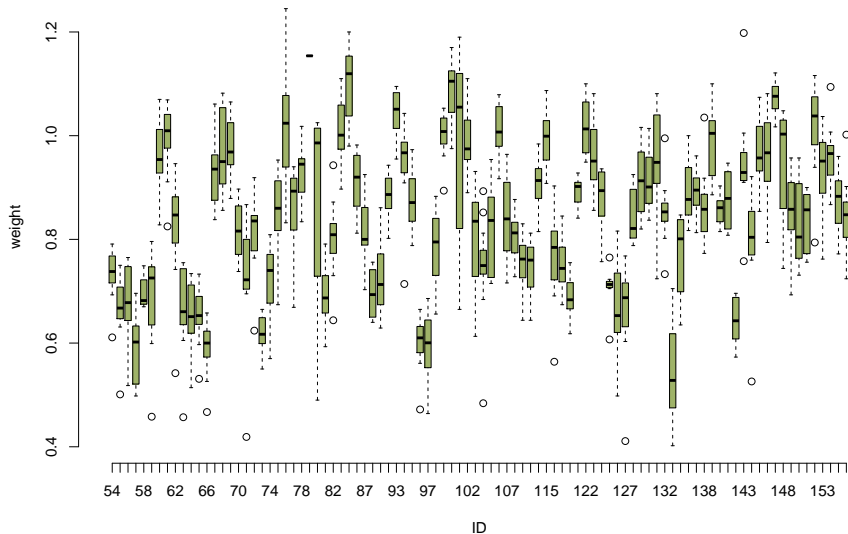
- **Primary aim:**

Estimation and inference for $\mu = \mathbb{E}Y_{i,j}$

Example: NTP TER84073 trial on mice



Example: NTP TER84073 trial on mice



Example: NTP TER84073 trial on mice

- Possible stochastic model:

$$Y_{i,j} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2) \quad \text{for } i = 1, \dots, N, j = 1, \dots, n_i.$$


-
- Can we justify independence of $Y_{1,1}, \dots, Y_{N,n_N}$?
 - Any idea on how to model association between $Y_{i,1}, \dots, Y_{i,n_i}$ for given i ?

Example: NTP TER84073 trial on mice


- More realistic model (being perhaps wrong but useful):

$$Y_{i,j} | b_i \sim \mathcal{N}(b_i, \sigma^2), \quad i = 1, \dots, N, j = 1, \dots, n_i,$$
$$b_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, d^2), \quad i = 1, \dots, N.$$

- While assuming, for each i , (conditional) independence $Y_{i,1} | b_i, \dots, Y_{i,n_i} | b_i$
- and independence of b_1, \dots, b_N .

- b_i : mean fetus weight for mouse i ;
- σ^2 : variability between weights of fetuses **within** each mouse;
  **within** (group) variability;
- μ : mean fetus weight in the whole population

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

- d^2 : variability of “levels” of individual mice
  **between** (groups) variability.

Example: NTP TER84073 trial on mice

- More realistic model (being perhaps wrong but useful):

$$Y_{i,j} | b_i \sim \mathcal{N}(b_i, \sigma^2), \quad i = 1, \dots, N, j = 1, \dots, n_i,$$
$$b_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, d^2), \quad i = 1, \dots, N.$$

- It models in a **certain** way also correlation between weights of fetuses of each individual mouse:

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

$$\implies \text{cor}(Y_{i,j}, Y_{i,k}) = \frac{d^2}{\sigma^2 + d^2}$$

\implies **intraclass correlation** (*vnitroskupinová korelace*)

Some additional talking

Hierarchical models

General note

- Also in other types of model, we can distinguish three types of parameters (in a bayesian sense):
 - latent/hidden data
 - ▣ in the following usually denoted by ξ
 - “genuine” parameters \equiv parameters also in a frequentist sense
 - ▣ in the following usually denoted by ψ
 - random hyperparameters
 - ▣ in the following usually denoted by ζ
- Parameters for a bayesian model are then $\theta = (\xi^\top, \psi^\top, \zeta^\top)^\top$.

Hierarchical models

General note, cont'd

- **Joint** prior distribution is then given by a factorization

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

- $p(\xi | \psi)$: “**structural**” part of the prior distribution

▮▮▮▮ follows from a considered stochastic model used in a particular situation/for a particular data structure

- $p(\psi | \zeta) p(\zeta)$: “**genuine**” prior distribution

“ANOVA” hierarchical model

Example of a prior distribution

- E.g., **conjugate prior** distribution with additional random hyperparameters to make the prior **flatter/weaker**:

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

- b_0, d_0 : **random** hyperparameters, tj. $\zeta = (b_0, d_0)^\top$
- $\mu_0, k_0, b_0, p_0, r_0, d_0, g_0, h_0$: fixed hyperparameters
- Graphical representation: **DAG** (*directed acyclic graph*)
 - Circles: random nodes
 - Squares: non-random nodes uzly
 - Expression of conditional (in)dependencies

“ANOVA” hierarchical model

Example of a prior distribution

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

Non-random (fixed) (hyper)parameters:

- μ_0, k_0
- a_0, p_0, r_0
- c_0, g_0, h_0

How to choose them?

“ANOVA” hierarchical model

Choice of fixed (hyper)parameters

- If no reasonable prior information is available, the non-random (hyper)parameters are usually chosen such that the resulting priori distribution is so called (*weakly informative*).
 - $p(\boldsymbol{\psi} | \mathbf{y}) \propto L_F(\boldsymbol{\psi}) p(\boldsymbol{\psi})$,
 - $p(\boldsymbol{\psi})$ represents “new”/“old” observations in the likelihood.
- We try to choose the prior such that the influence of such “new”/“old” observations on the likelihood was as weak as possible.
 - We want to satisfy, as much as possible, that $p(\boldsymbol{\psi}) \propto 1$.
 - It is sufficient if this holds relatively to $L_F(\boldsymbol{\psi})$.
- ▣▶ A particular choice of fixed hyperparameters is often (at least partly) motivated by observed data.

“ANOVA” hierarchical model

(Partly) data motivated choice of fixed hyperparameters

- μ_0 : prior mean for $\mathbb{E}Y_{i,j} = \mathbb{E}b_i = \mu$
 - $\mu_0 \approx \bar{y}$
- k_0 : influences prior inverse variance (precision) of μ
 - k_0 close to 0
- a_0, c_0, ρ_0, g_0 : “degrees of freedom”/“shape” of the gamma distribution
 - usually between 0 and 1
- r_0, h_0 : “rate” parameter of the gamma distribution in the last hierarchical level
 - usually chosen close to 0

“ANOVA” hierarchical model

Posterior distribution

- $p(\theta | \mathbf{y})$ can be derived in a standard way
 - ▮ How?
- In this particular case, when selecting the conjugate system, it can still be derived analytically.
- For many other prior choices, analytic derivation is not possible (integral in the denominator of the Bayes theorem).
- The posterior inference is usually based on a computer simulation (Monte Carlo methods).

7

(Generalized) linear mixed models

Section 7.1

Linear mixed model

Normal linear mixed model (almost as in the NMST422 course)

$$\mathbf{Y}_i = \mathbb{X}_i \boldsymbol{\beta} + \mathbb{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \dots, N$$

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

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

- ▶ $\mathbf{b}_1, \dots, \mathbf{b}_N, \boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_N$ mutually independent

Normal linear mixed model, hierarchically specified

$$\mathbf{Y}_i | \mathbf{b}_i \sim \mathcal{N}_{n_i}(\mathbb{X}_i \boldsymbol{\beta} + \mathbb{Z}_i \mathbf{b}_i, \boldsymbol{\Sigma}_i), \quad i = 1, \dots, N$$

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

▶ $\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{b}_1 \end{pmatrix}, \dots, \begin{pmatrix} \mathbf{Y}_N \\ \mathbf{b}_N \end{pmatrix}$ mutually independent

Example 1: NTP TER84073 trial on mice

Normal linear mixed model

- ▶ Stochastic representation of data:

$Y_{i,j}$ = weight of the j -th fetus of i -th mouse, $i = 1, \dots, N$, $j = 1, \dots, n_i$

- ▶ Normal LMM:

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

$$\mathbf{Y}_i | b_i \sim \mathcal{N}_{n_i}(b_i \mathbf{1}, \boldsymbol{\Sigma}_i)$$

$$b_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, d^2)$$

- ▶ In a general notation

$$\mathbb{X}_i \text{ none, } \mathbb{Z}_i = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}_{n_i}, \quad \mathbb{D} = d^2$$

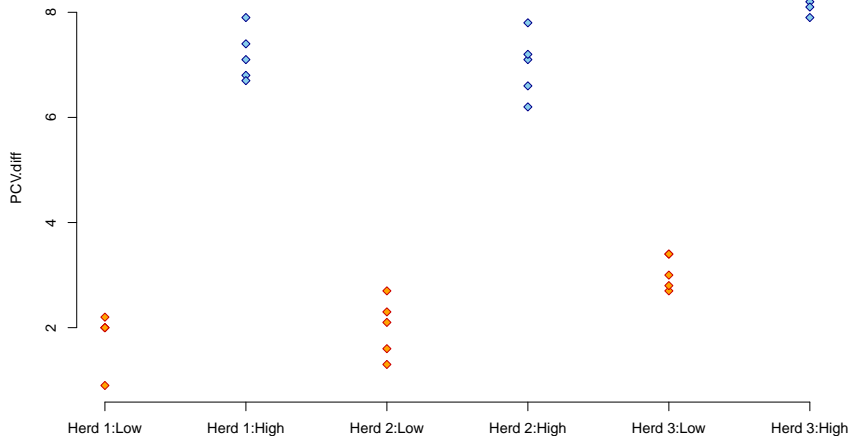
Example 2: Influence of the Berenil dose on a treatment of trypanosomosis

Split-plot design

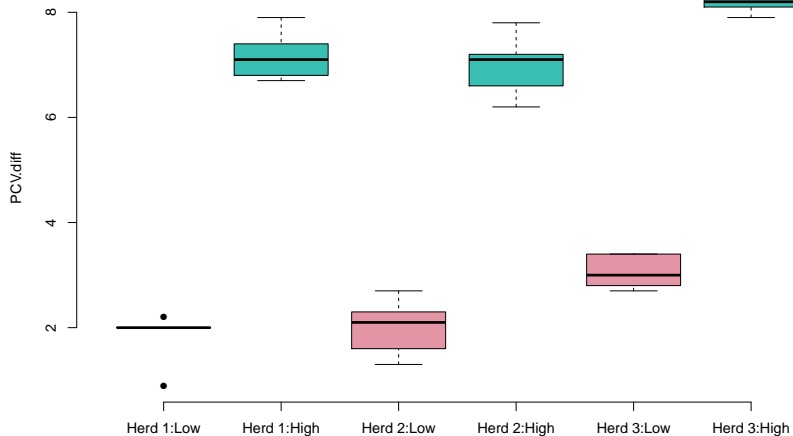
- An experiment to evaluate the effect of the Berenil dose on the success of treatment of trypanosomosis in cows
- The success of the treatment is measured by the increase in PCV (packed cell volume) after a certain period of drug administration
- Is there a difference in the treatment success between **low** and **high** Berenil dose
- Cows from **three** herds (from different herds)
- Each herd was randomly divided into two (approximately equal) parts, with cows from one part of the herd being treated with a low dose of Berenil, cows from the other part of the herd treated with a high dose of Berenil
- **Stochastic representation of data:**

$$Y_{i,j} = \text{PCV increment at } j\text{th cow of the } i\text{th herd}$$

Example 2: Influence of the Berenil dosis on a treatment of trypanosomosis



Example 2: Influence of the Berenil dosis on a treatment of trypanosomosis



Example 2: Influence of the Berenil dosis on a treatment of trypanosomosis

Linear mixed model

1. Dosis having **the same** effect in all herds.

$$Y_{i,j} = b_i + \mathbb{I}[\text{dosis}_{i,j} = \text{high}]\beta + \varepsilon_{i,j}$$

2. Dosis having **not necessarily the same** effect across herds.

$$Y_{i,j} = b_{i,1} + \mathbb{I}[\text{dosis}_{i,j} = \text{high}] b_{i,2} + \varepsilon_{i,j}$$

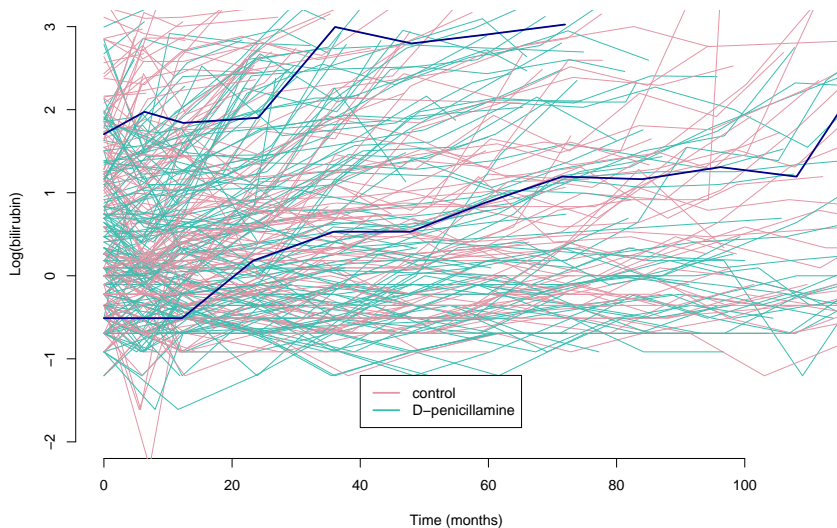
Example 3: Evolution of the bilirubin level a patients with PBC

Longitudinal study

- Study conducted in 1974–1984 at Mayo Clinic
- 312 patients with PBC (primary biliary cholangitis/cirrhosis)
- 158 patients treated by D-penicillamin
- 154 patients treated by a standard treatment
- One of aims of the study: groups comparison (D-penicillamin vs. standard) with respect to the bilirubin evolution (one of signs of severity of PBC)
- Patients were examined in (irregular) intervals
- Median follow-up time: 6,3 years (IQR 3,7 – 8,9 years)
- **Stochastic representation of data:**

$Y_{i,j}$ = logarithm of the bilirubin level of the i th patient at time $t_{i,j}$

Example 3: Evolution of the bilirubin level a patients with PBC



Example 3: Evolution of the bilirubin level a patients with PBC

Possible linear mixed model

- Possible model for one patient: **line in time**

1. The same **mean evolution** in log-bilirubin in both groups:

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

2. **Different slopes** in the two groups:

$$Y_{i,j} = b_{i,1} + b_{i,2} t_{i,j} + \beta \mathbb{I}[\text{treatment}_i = \text{D-penicillamin}] t_{i,j} + \varepsilon_{i,j}$$

Hierarchically specified model for data

$$\mathbf{Y}_i | \mathbf{b}_i \sim \mathcal{N}_{n_i}(\mathbb{X}_i \boldsymbol{\beta} + \mathbb{Z}_i \mathbf{b}_i, \boldsymbol{\Sigma}_i), \quad i = 1, \dots, N$$

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

- $\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{b}_1 \end{pmatrix}, \dots, \begin{pmatrix} \mathbf{Y}_N \\ \mathbf{b}_N \end{pmatrix}$ mutually independent


“Genuine” parameters (parameters for a frequentist analysis)

- β : fixed effects
 - (population) influence of regressors included in the matrix \mathbb{X}
- $\mu = \mathbb{E}\mathbf{b}_i$ ($i = 1, \dots, N$) : expectations of random effects
 - (population) influence of regressors included in the matrix \mathbb{Z}
- $\Sigma_i = \text{var}(\mathbf{Y}_i | \mathbf{b}_i)$ ($i = 1, \dots, N$) : “intraclass/within” covariance matrix
 - often assumed $\Sigma_i = \sigma^2 \mathbf{I}_{n_i}$
 - conditional independence
 - other association structures can be modelled (AR(d), ...)
- $\mathbb{D} = \text{var}\mathbf{b}_i$ ($i = 1, \dots, N$) : “between” group covariance matrix
 - usually no special structure assumed for \mathbb{D} , it is only requested that $\mathbb{D} > 0$ (positive definite matrix)

- **Genuine parameters:** $\psi = (\beta^\top, \mu^\top, \underbrace{\text{par}(\Sigma_1, \dots, \Sigma_N)}_{\text{často pouze } \sigma^2}, \text{par}(\mathbb{D}))^\top$
- **Likelihood for frequentist estimation:**

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

where $\mathbb{V}_i = \mathbb{Z}_i \mathbb{D} \mathbb{Z}_i^\top + \Sigma_i$

- When using the **maximum likelihood**, it is necessary to maximize $L_F(\psi)$ under the constraint $\Sigma_i > 0$ (for all $i = 1, \dots, N$) and $\mathbb{D} > 0$
 -  packages lme4, nlme
 - SAS procedure MIXED

Linear mixed model, bayesian specification

Latent data and the likelihood of the bayesian model

- Additional stochastic components of the model:
 - ≡ random effect vectors $\mathbf{b}_1, \dots, \mathbf{b}_N$
 - ▮▶ Additional “parameters” with the bayesian approach
 - ≡ latent data
- **Latent data:** $\boldsymbol{\xi} = (\mathbf{b}_1^\top, \dots, \mathbf{b}_N^\top)^\top$
- **Likelihood of the bayesian model:**

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

Section 7.2

Generalized linear mixed model

Example: NTP TER84073 trial on mice

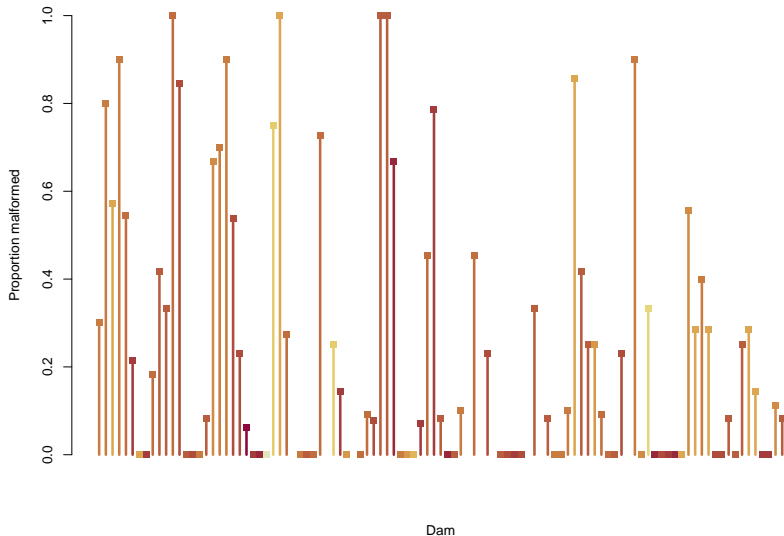
- Data from *National Toxicology Program*
- 94 pregnant mice were given, in pre-specified moments, certain amount of ethylene glycol (EG).
- On a day 17 of pregnancy, the mice were killed and
 - (a) the number of fetuses ($n_i, i = 1, \dots, 94$) was recorded;
 - (b) indication on whether a developmental defect was present on the fetus.
- **Stochastic representation of data:**

$$Y_{i,j} = \begin{cases} 0, & \text{if } j\text{th fetus of } i\text{th mouse without any defect,} \\ 1, & \text{if } j\text{th fetus of } i\text{th mouse with a defect.} \end{cases}$$

- **Primary aim:**
Estimation and inference for $\pi = \mathbb{E}Y_{i,j} = P(Y_{i,j} = 1)$.

Example: NTP TER84073 trial on mice

Observed proportions of fetuses with developmental defects



Example: NTP TER84073 trial on mice

Possible model ($j = 1, \dots, n_i$ for each $i = 1, \dots, N$)

- $Y_{i,j} | \pi_i \sim \mathcal{A}(\pi_i)$.

- $\pi_i \stackrel{\text{i.i.d.}}{\sim}$ from a certain distribution.

▸ It represents the fact that each mouse has different (genetic, ...) dispositions for their embryos to develop developmental defects.

- The mice included in the study are a random sample from the mouse population and it is therefore reasonable to assume that π_i ($i = 1, \dots, N$) are random as well.

- Suitable reparameterization of π_i 's can be used to avoid constraints $0 < \pi_i < 1$ ($i = 1, \dots, N$), e.g.,

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

where $b_i \stackrel{\text{i.i.d.}}{\sim}$ certain distribution.

▸ **Logistic regression with random effects.**

Example: NTP TER84073 trial on mice

Logistic regression with normally distributed random effects

Possible model ($j = 1, \dots, n_i$ for each $i = 1, \dots, N$)

- $Y_{i,j} | b_i$ independent with the $\mathcal{A}(\pi_i)$ distribution, where $\pi_i = \frac{e^{b_i}}{1+e^{b_i}}$.
- b_i i.i.d. with the $\mathcal{N}(\mu, d^2)$ distribution.

Parameters (genuine)

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

Likelihood (for a frequentist approach)

$$\begin{aligned} L_F(\psi) &= p(\mathbf{y} | \psi) = \prod_{i=1}^N p(\mathbf{y}_i | \psi) = \prod_{i=1}^N \int_{\mathbb{R}} \prod_{j=1}^{n_i} p(y_{i,j} | b_i, \psi) p(b_i | \psi) db_i \\ &= \prod_{i=1}^N \int_{\mathbb{R}} \pi_i^{\sum_{j=1}^{n_i} y_{i,j}} (1 - \pi_i)^{n_i - \sum_{j=1}^{n_i} y_{i,j}} \varphi(b_i | \mu, d^2) db_i \\ &= \prod_{i=1}^N \int_{\mathbb{R}} \frac{e^{b_i \sum_{j=1}^{n_i} y_{i,j}}}{(1 + e^{b_i})^{n_i}} \varphi(b_i | \mu, d^2) db_i. \end{aligned}$$

Section **7.3**

Prior distribution

Normal (generalized) linear mixed model

Prior distribution

- We use a decomposition which follows the hierarchical structure of the model

$$p(\xi, \psi) = p(\xi | \psi) p(\psi)$$

- Part I

$$p(\xi | \psi) = p(\mathbf{b}_1, \dots, \mathbf{b}_N | \beta, \mu, \Sigma_1, \dots, \Sigma_N, \mathbb{D}) =$$

$$p(\mathbf{b}_1, \dots, \mathbf{b}_N | \mu, \mathbb{D}) = \dots = \prod_{i=1}^N \varphi(\mathbf{b}_i | \mu, \mathbb{D})$$

- Part II: $p(\psi)$
 - “standard” prior distribution for “genuine” parameters
 - possibly including (random) hyperparameters

Normal (generalized) linear mixed model

Prior distribution

- **Parameters of the bayesian model** $\theta = (\xi^\top, \psi^\top, \zeta^\top)^\top$

- $\xi = (\mathbf{b}_1^\top, \dots, \mathbf{b}_N^\top)^\top$
- $\psi = (\beta^\top, \mu^\top, \text{par}(\Sigma_1, \dots, \Sigma_N), \text{par}(\mathbb{D}))^\top$
- ζ : possibly additional hyperparameters

- **Decomposition of the prior distribution:**

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

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

Normal (generalized) linear mixed model

Prior distribution

- We need to specify $p(\psi)$
- Mostly hierarchically with additional hyperparameters

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

that is $p(\psi, \zeta) = p(\psi | \zeta) p(\zeta)$

- $\psi = (\beta^\top, \mu^\top, \text{par}(\Sigma_1, \dots, \Sigma_N), \text{par}(\mathbb{D}))^\top$
- In the following, only situations with $\Sigma_j = \sigma^2 \mathbf{I}_{n_j}$, i.e., $\text{par}(\Sigma_1, \dots, \Sigma_N) = \sigma^2$
- Notation:

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

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

$$\Rightarrow \psi = (\beta^\top, \mu^\top, \tau, \text{par}(\mathbb{Q}))^\top$$

Normal (generalized) linear mixed model

Possible choices of the prior distribution

- Usually prior independence for sets of “related” parameters is assumed, e.g.,

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

or

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

where $\zeta = (\zeta^{(1)\top}, \zeta^{(2)\top}, \zeta^{(3)\top}, \zeta^{(4)\top})^\top$

Normal (generalized) linear mixed model

Possible choices of the prior distribution

β , μ : interpretation of expected values

- Reasonable prior distribution:

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

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

- Other reasonable prior distribution:

$$p(\beta) \sim \mathcal{N}(\beta_0, \Sigma_0^\beta)$$

$$p(\mu) \sim \mathcal{N}(\mu_0, \Sigma_0^\mu),$$

- β_0 , Σ_0^β , μ_0 , Σ_0^μ : fixed/random hyperparameters
- frequent reasonable choice:
 - $\beta_0 = \mathbf{0}$ (not for the intercept term)
 - $\mu_0 = \mathbf{0}$ (not for the intercept term)
 - Σ_0^β , Σ_0^μ : diagonal matrix with large numbers on a diagonal (what large means?)

Normal (generalized) linear mixed model

Possible choices of the prior distribution

τ : inverted variance of deviations of subject i from the mean value (given by regressors) of this subject

- Reasonable prior distribution:

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

- c_τ, d_τ : fixed/random hyperparameters
- ensures (among the other things) that $P(\tau > 0 | \mathbf{Y}) = 1$
- frequent reasonable choice of hyperparameters:
 - c_τ : prior “degrees of freedom”, i.e., $c_\tau \in (0, 1]$ leads to weakly informative distribution
 - to remind: $\mathbb{E}_\tau = \frac{c_\tau}{d_\tau}$, $\text{var}_\tau = \frac{c_\tau}{d_\tau^2}$
 - ▮ d_τ : “precision” of the prior gamma distribution
 - d_τ close to 0 may lead to a weakly informative distribution
 - d_τ : often is taken to be random with another gamma distribution (with fixed hyperparameters) as a prior

Normal (generalized) linear mixed model

Possible choices of the prior distribution

Q: inverted covariance matrix of “levels” of individual subjects

- **Multivariate** prior distribution is needed which generates, with a probability of 1, a positive definite matrix

▣▶ **Wishart** distribution

$$\mathbf{Q} \sim \mathcal{W}_p(\nu, \mathbf{\Xi}), \text{ if}$$

$$\mathbf{Q} = \sum_{i=1}^{\nu} \mathbf{z}_i \mathbf{z}_i^{\top},$$

- where
- $\mathbf{z}_1, \dots, \mathbf{z}_{\nu} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mathbf{0}, \mathbf{\Xi})$
 - $\mathbf{\Xi}$ is a positive definite scale matrix
 - $\nu > p - 1$ are degrees of freedom
 - Clearly: $P(\mathbf{Q} > \mathbf{0}) = 1$
 - It is a multivariate generalization of the χ_{ν}^2 distribution

$\mathbb{Q} \sim \mathcal{W}_p(\nu, \Xi)$ has a density

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

- $\nu > p - 1$: also non-integer “degrees of freedom” can be considered
 ▸ generalization of the basic Wishart distribution
- $\mathbb{E}\mathbb{Q} = \nu\Xi$
- $\mathcal{W}_1(\nu, 1) \equiv \mathcal{G}\left(\frac{\nu}{2}, \frac{1}{2}\right) \equiv \chi_\nu^2$
- $\mathcal{W}_1(\nu, \Xi) \equiv \mathcal{G}\left(\frac{\nu}{2}, \frac{\Xi^{-1}}{2}\right)$

Normal (generalized) linear mixed model

Possible choices of the prior distribution

\mathbb{Q} : inverted covariance matrix of “levels” of individual subjects

- Reasonable prior distribution:

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

- Ξ_Q, ν_Q : fixed/random hyperparameters
- frequent reasonable choice of hyperparameters:
 - ν_Q : prior “degrees of freedom”, i.e., $\nu_Q \in (p - 1, p]$ leads to weakly informative prior
 - Ξ_Q : usually chosen as a **diagonal** matrix, e.g., $\Xi_Q = \text{diag}(\gamma_{Q,1}^{-1}, \dots, \gamma_{Q,p}^{-1})$
 - inverted scale matrix (Ξ_Q^{-1}) is the “precision” matrix of the Wishart distribution
 - ▮ diagonal values of Ξ_Q^{-1} (i.e., $\gamma_{Q,1}, \dots, \gamma_{Q,p}$) close to 0 may lead to weakly informative prior distribution

Normal (generalized) linear mixed model

Summary

- **“Genuine” parameters:** $\psi = (\beta^\top, \mu^\top, \tau, \text{par}(\mathbb{Q}))^\top$
- **Latent data:** $\xi = (\mathbf{b}_1^\top, \dots, \mathbf{b}_N^\top)^\top$
- **Likelihood of the bayesian model:**

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

- Decomposition of the **prior distribution:**

$$p(\xi, \psi) = p(\xi | \psi) p(\psi) = \prod_{i=1}^N \varphi(\mathbf{b}_i | \mu, \mathbb{Q}^{-1}) p(\beta) p(\mu) p(\tau) p(\mathbb{Q})$$

Section 7.4

Posterior distribution

Normal (generalized) linear mixed model

Posterior distribution

- If no random hyperparameters:

$$p(\xi, \psi | \mathbf{y}) \propto L(\xi, \psi) p(\xi | \psi) p(\psi)$$

- If some hyperparameters (included in ζ) random:

$$p(\xi, \psi, \zeta | \mathbf{y}) \propto L(\xi, \psi) p(\xi | \psi) p(\psi | \zeta) p(\zeta)$$

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

Normal (generalized) linear mixed model

Posterior distribution of “genuine” parameters

- Marginal posterior distribution of “genuine” parameters:

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

-
- Owing to the hierarchical structure, we also have

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

$$\text{since } p(\boldsymbol{\psi}) = \int p(\boldsymbol{\psi} | \boldsymbol{\zeta}) p(\boldsymbol{\zeta}) d\boldsymbol{\zeta}$$

$$L_F(\boldsymbol{\psi}) = \int L(\boldsymbol{\xi}, \boldsymbol{\psi}) p(\boldsymbol{\xi} | \boldsymbol{\psi}) d\boldsymbol{\xi}$$

and the Fubini theorem holds

- **Sdružené** aposteriorní rozdělení pro $\theta = (\psi^\top, \xi^\top, \zeta^\top)^\top$ vyjádříme snadno až na **multiplikativní konstantu**
- Primárně nás však zajímají hlavně **marginální** aposteriorní rozdělení pro sady parametrů nebo dokonce jednotlivé parametry
 - $p(\beta | \mathbf{y}), p(\beta_j | \mathbf{y}), p(\mu | \mathbf{y}), p(\tau | \mathbf{y}), \dots$
 - 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

- Často nás zajímá též aposteriorní rozdělení pro nějakou měřitelnou funkci t
 - transformaci původních parametrů
- Příklad:
 - $Y_{i,j} | b_i \sim \mathcal{N}(b_i, \sigma^2)$, $b_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, d^2)$, $i = 1, \dots, N$, $j = 1, \dots, n_i$
 - Bylo: $\text{cor}(Y_{i,j}, Y_{i,k}) = d^2 / (\sigma^2 + d^2) =: \varrho$
 - ▮ Z $p(\mu, \sigma^2, d^2 | \mathbf{y})$ potřeba pomocí věty o transformaci a integrováním odvodit $p(\varrho | \mathbf{y})$
 - Vzpomeňte si na svoje úspěchy na tomto poli (obvykle v poměrně jednoduchých situacích) ze 3. ročníku...
- Odvození $p(t(\theta) | \mathbf{y})$ z $p(\theta | \mathbf{y})$ (které již samo o sobě obvykle známe až na multiplikační konstantu) je často analyticky neproveditelné

- 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** počítačové vybavení
 - do cca začátku 90. let 20. století jenom omezeně dostupné/nedostupné (nejenom v ČSSR)

8

Bayesian data augmentation

Section **8.1**

Introduction

Data augmentation (rozšiřování dat)

- Typically, we need to generate from the posterior distribution with a density

$$p(\psi | \mathbf{y}) = \frac{L_{obs}(\psi) p(\psi)}{\int_{\Theta} L_{obs}(\psi) p(\psi) d\lambda(\psi)} \propto L_{obs}(\psi) p(\psi)$$

with respect to some σ -finite measure λ , where

- $L_{obs}(\psi) = p(\mathbf{y} | \psi)$: likelihood (of observed data)
 - $p(\psi)$: prior distribution
- To use the MCMC, usually:
 - no need to know the normalizing constant $\int_{\Theta} L_{obs}(\psi) p(\psi) d\lambda(\psi)$;
 - it is useful if the quantity $L_{obs}(\psi) p(\psi)$ can be easily calculated/evaluated for any $\psi \in \Theta$.
- Often, the likelihood $L_{obs}(\psi)$ is not easily calculated
 - the most frequent complication: an integral must be calculated to get the value of $L_{obs}(\psi)$.

Example 1: Linear mixed model

Model (for $i = 1, \dots, N$)

- $\mathbf{Y}_i | \mathbf{b}_i$ independent with the $\mathcal{N}_{n_i}(\mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\mathbf{b}_i, \sigma^2\mathbf{I}_{n_i})$ distribution
- \mathbf{b}_i i.i.d. with the $\mathcal{N}_q(\boldsymbol{\mu}, \mathbb{D})$ distribution

Parameters

- $\boldsymbol{\psi} = (\boldsymbol{\beta}^\top, \boldsymbol{\mu}^\top, \sigma^2, \text{vec}(\mathbb{D}))^\top$

Likelihood (of observed data)

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

Example 2: Logistic regress. with normally distrib. random effects

Model ($j = 1, \dots, n_i$ for each $i = 1, \dots, N$)

- $Y_{i,j} | b_i$ independent with the $\mathcal{A}(\pi_i)$ distribution, where $\pi_i = \frac{e^{b_i}}{1+e^{b_i}}$.
- b_i i.i.d. with the $\mathcal{N}(\mu, d^2)$ distribution.

Parameters

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

Likelihood (of observed data)

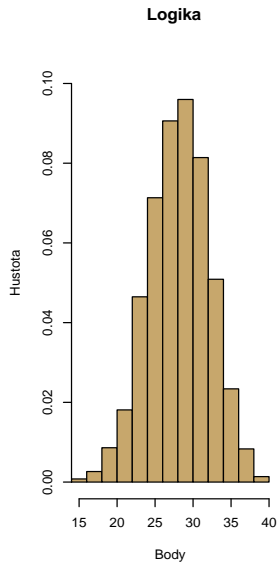
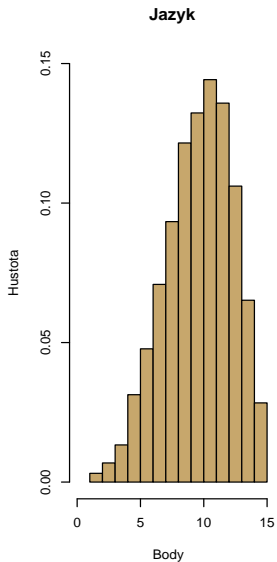
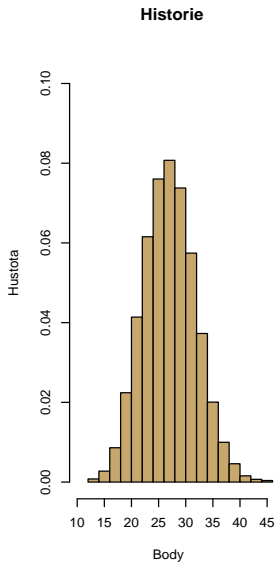
$$\begin{aligned} L_{obs}(\psi) &= p(\mathbf{y} | \psi) = \prod_{i=1}^N p(\mathbf{y}_i | \psi) = \prod_{i=1}^N \int_{\mathbb{R}} \prod_{j=1}^{n_i} p(y_{i,j} | b_i, \psi) p(b_i | \psi) db_i \\ &= \prod_{i=1}^N \int_{\mathbb{R}} \pi_i^{\sum_{j=1}^{n_i} y_{i,j}} (1 - \pi_i)^{n_i - \sum_{j=1}^{n_i} y_{i,j}} \varphi(b_i | \mu, d^2) db_i \\ &= \prod_{i=1}^N \int_{\mathbb{R}} \frac{e^{b_i \sum_{j=1}^{n_i} y_{i,j}}}{(1 + e^{b_i})^{n_i}} \varphi(b_i | \mu, d^2) db_i. \end{aligned}$$

Example 3: Entrance exam results

- In 1999, certain faculty (of law) of a certain Czech university (not in Pilsen) organized 11 regular terms and one extra term (Term #12) of the entrance exam.
- Someone has noticed that the proportion of admitted students was much higher during Term #12 as compared to all regular terms.
 - The smart students made mass excuses for the regular terms and showed up only at the extra term?
 - Significantly more stimulating classroom atmosphere during Term #12?
 - Miracle?
- It showed that the questions for Term #12 of the entrance examination mysteriously escaped (from a locked vault) and were (in certain circles) available for purchase prior to this term.
- The whole event was, according to both Rector of the university as well as Dean of the faculty, the work of the *gangster mafia* standing outside the faculty. Some more details still at <http://www.cibulka.com/nnoviny/nn2000/nn1900/obsah/05.htm> through the WayBack Machine (2019/05/28).

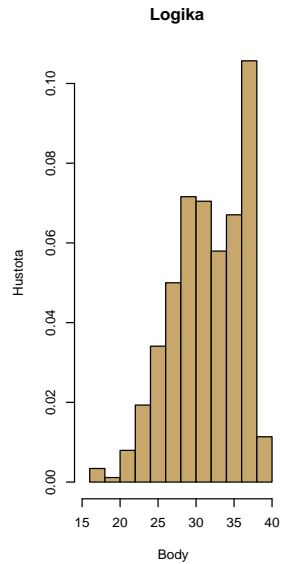
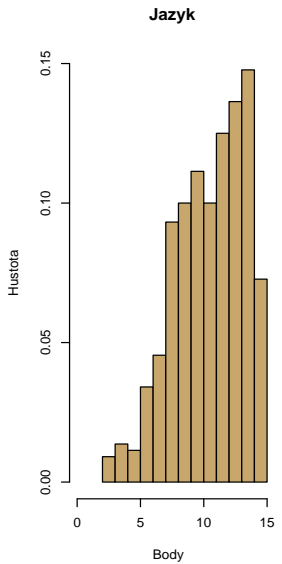
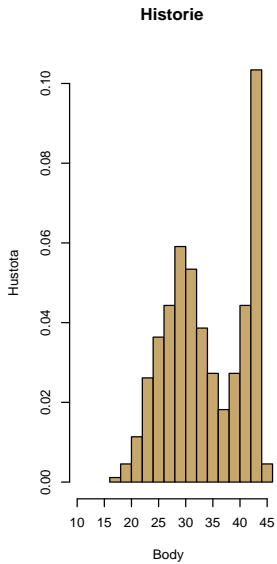
Example 3: Entrance exam results

Terms #1-11



Example 3: Entrance exam results

Term #12



Example 3: Entrance exam results

Possible model for results of the Term #12

Possible model (for $i = 1, \dots, N$)

- $\mathbf{Y}_i = (Y_{i,1}, Y_{i,2}, Y_{i,3})^\top$: points of the i th student at different parts of the exam.
- Students belong to **two** populations:
 1. usual students (proportion w_1);
 2. students connected to the gangster mafia (proportion w_2 , $w_1 + w_2 = 1$).
- We will **assume** that the (joint) distribution of points is multivariate normal in **in each** population, with means μ_1 and μ_2 and the covariance matrices Σ_1 and Σ_2 .

⇒ Distribution of \mathbf{Y}_i : **mixture** of normal distributions with a density

$$p(\mathbf{y}_i | \psi) = w_1 \varphi(\mathbf{y}_i | \mu_1, \Sigma_1) + w_2 \varphi(\mathbf{y}_i | \mu_2, \Sigma_2)$$

Example 3: Entrance exam results

Possible model for results of the Term #12

- Must be estimated:
 - weights (proportions) w_1, w_2 ,
 - means μ_1, μ_2 ,
 - covariance matrices Σ_1, Σ_2 .

$$\Rightarrow \psi = (w_1, w_2, \mu_1^\top, \mu_2^\top, \text{vec}(\Sigma_1), \text{vec}(\Sigma_2))^\top$$

Likelihood (of observed data)

$$L_{obs}(\psi) = p(\mathbf{y} | \psi) = \prod_{i=1}^N p(\mathbf{y}_i | \psi) = \prod_{i=1}^N \left\{ \sum_{k=1}^2 w_k \varphi(\mathbf{y}_i | \mu_k, \Sigma_k) \right\}.$$

Section **8.2**
Principles

- We have seen that the likelihood of observed data $L_{obs}(\psi) = p(\mathbf{y} | \psi)$ (which is a basis of the posterior density $p(\psi | \mathbf{y}) \propto L_{obs}(\psi) p(\psi)$) cannot always be expressed as a **product** of “nice” functions.
- Sometimes, $L_{obs}(\psi)$ cannot even be expressed analytically:
 - logistic regression with random effects,
 - generalized linear mixed models (GLMM).
- Nevertheless, the “likelihood” can often be simplified if we start considering more parameters:
 - let us denote them ξ ;
 - “likelihood” is then $L_{augm}(\xi, \psi) = p(\mathbf{y} | \xi, \psi)$;
 - $L_{augm}(\xi, \psi)$ will be called **augmented** likelihood (*rozšířená věrohodnost*).

Data augmentation (rozšiřování dat)

Principles

- Parameters ξ often takes a meaning of **unobservable** (or only **indirectly observable** data).
 - hence the term **data augmentation** (*rozšiřování dat*);
 - the term comes from the paper 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.
- \mathbf{y} : observed data (*pozorovaná/pozorovatelná data*).
- (\mathbf{y}, ξ) : complete data (*úplná/rozšířená data*).

Data augmentation (rozšiřování dat)

Principles

- Primarily we are interested in $p(\boldsymbol{\psi} | \mathbf{y}) \propto \underbrace{p(\mathbf{y} | \boldsymbol{\psi})}_{L_{obs}(\boldsymbol{\psi})} p(\boldsymbol{\psi})$,

where $p(\mathbf{y} | \boldsymbol{\psi}) = L_{obs}(\boldsymbol{\psi})$ follows from the assumed (not necessarily hierarchical) model.

- Suppose that for a suitable $\boldsymbol{\xi}$ with a joint density

$$p(\boldsymbol{\xi}, \boldsymbol{\psi} | \mathbf{y}) \propto p(\mathbf{y} | \boldsymbol{\xi}, \boldsymbol{\psi}) p(\boldsymbol{\xi}, \boldsymbol{\psi}) = \underbrace{p(\mathbf{y} | \boldsymbol{\xi}, \boldsymbol{\psi})}_{L_{augm}(\boldsymbol{\xi}, \boldsymbol{\psi})} p(\boldsymbol{\xi} | \boldsymbol{\psi}) p(\boldsymbol{\psi})$$

we can work much more easily.

- $L_{augm}(\boldsymbol{\xi}, \boldsymbol{\psi})$: model (likelihood) for observable data if the augmented data are considered as additional model parameters.
- $p(\boldsymbol{\xi} | \boldsymbol{\psi})$: model (likelihood) for augmented data.

Principles

- Suppose that we ensure that the following holds

$$p(\psi | \mathbf{y}) = \int p(\xi, \psi | \mathbf{y}) d\lambda(\xi)$$

that is, the term $p(\psi | \mathbf{y})$ is a marginal distribution of $\psi | \mathbf{Y} = \mathbf{y}$ which corresponds to the joint distribution $(\xi, \psi) | \mathbf{Y} = \mathbf{y}$.

Principles

- If we perform the posterior inference based on a simulation, we generate a random sample/Markov chain

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

with a stationary distribution with a density $p(\xi, \psi | \mathbf{y})$.

- If $p(\psi | \mathbf{y})$ is a marginal density corresponding to the joint density $p(\xi, \psi | \mathbf{y})$ then

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

is a random sample/Markov chain with a stationary distribution with a density $p(\psi | \mathbf{y})$.

Data augmentation (rozšiřování dat)

Principles

- We have: $p(\psi | \mathbf{y}) \propto L_{obs}(\psi) p(\psi)$,
 $p(\xi, \psi | \mathbf{y}) \propto L_{augm}(\xi, \psi) p(\xi | \psi) p(\psi)$.
 - $L_{obs}(\psi)$ follows from the assumed model for observed data.
 - $L_{augm}(\xi, \psi) p(\xi | \psi)$ follows from the considered data augmentation.
 - $L_{augm}(\xi, \psi)$: model (likelihood) for observed data if the augmented data are considered as additional model parameters.
 - $p(\xi | \psi)$: model (likelihood) for augmented data.
- If is necessary to perform the augmentation such that $p(\psi | \mathbf{y})$ is a marginal distribution which corresponds to the joint distribution $p(\xi, \psi | \mathbf{y})$.
- If is necessary to perform the augmentation such that

$$L_{obs}(\psi) \propto \int L_{augm}(\xi, \psi) p(\xi | \psi) d\lambda(\xi).$$

Data augmentation (rozšiřování dat)

Principles

- To satisfy that $p(\psi | \mathbf{y})$ is a marginal distribution which corresponds to the joint distribution $p(\xi, \psi | \mathbf{y})$, it is sufficient to ensure

$$L_{obs}(\psi) = p(\mathbf{y} | \psi) \\ \propto \int p(\mathbf{y} | \xi, \psi) p(\xi | \psi) d\lambda(\xi) = \int L_{augm}(\xi, \psi) p(\xi | \psi) d\lambda(\xi).$$

- The term

$$p(\mathbf{y} | \xi, \psi) p(\xi | \psi) = L_{augm}(\xi, \psi) p(\xi | \psi)$$

is equal to $p(\mathbf{y}, \xi | \psi)$ and can be interpreted as a likelihood if we had observed **complete** data.

- $L_{compl}(\psi) := L_{augm}(\xi, \psi) p(\xi | \psi)$
= likelihood of **complete data**.

Principles

- To satisfy that $p(\psi | \mathbf{y})$ is a marginal distribution which corresponds to the joint distribution $p(\xi, \psi | \mathbf{y})$, it is sufficient to specify the (augmented) model which includes unobserved data ξ such that the involved likelihoods correspond to each other.

► This is directly ensured in case of **hierarchical** models where the augmented data ξ have a clear role in the description of a stochastic mechanism which generates the observed data \mathbf{y} .

Specification of a hierarchical model:

1. $L_{augm}(\xi, \psi) = p(\mathbf{y} | \xi, \psi)$: 1st hierarchical level
(model for observed data given the unobserved data).
2. $p(\xi | \psi)$: 2nd hierarchical level (model for unobserved data).
3. $L_{obs}(\psi)$ (marginal model for observed data)

is then “calculated” as $L_{obs}(\psi) = \int L_{augm}(\xi, \psi) p(\xi | \psi) d\lambda(\xi)$.

Data augmentation (rozšiřování dat)

Summary of terminology

Data, parameters

- \mathbf{y} : observed data;
- ψ : (genuine) parameters
 - statistical inference on them is our primary aim;
- ξ : unobserved (or only indirectly observed) data, additional parameters.

Likelihoods

- $L_{obs}(\psi) = p(\mathbf{y} | \psi)$: likelihood of observed data;
- $L_{augm}(\xi, \psi) = p(\mathbf{y} | \xi, \psi)$: augmented likelihood of observed data;
- $p(\xi | \psi)$: likelihood of augmented data;
- $L_{compl}(\psi) = p(\xi, \mathbf{y} | \psi) = L_{augm}(\xi, \psi) p(\xi | \psi)$:
likelihood of complete data.

- The single terms must be specified such that

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

Section **8.3**

Examples

Example 1: Linear mixed model

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

- $Y_i \mid \mathbf{b}_i$ independent with the $\mathcal{N}_{n_i}(\mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\mathbf{b}_i, \sigma^2\mathbf{I}_{n_i})$ distribution,
- \mathbf{b}_i i.i.d. with the $\mathcal{N}_q(\boldsymbol{\mu}, \mathbb{D})$ distribution.

Parameters

- $\boldsymbol{\psi} = (\boldsymbol{\beta}^\top, \boldsymbol{\mu}^\top, \sigma^2, \text{vec}(\mathbb{D}))^\top$.

Indirectly observable (augmented) data

- $\boldsymbol{\xi} = (\mathbf{b}_1^\top, \dots, \mathbf{b}_N^\top)^\top$.

Example 1: Linear mixed model

Likelihood of **observed data**

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

Augmented likelihood

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

Likelihood of **augmented data**

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

Example 2: Logistic regress. with normally distrib. random effects

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

- $Y_{i,j} | b_i$ independent with the distribution $\mathcal{A}(\pi_i)$, where $\pi_i = \frac{e^{b_i}}{1+e^{b_i}}$,
- b_i i.i.d. with the $\mathcal{N}(\mu, d^2)$ distribution.

Parameters

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

Indirectly observable (augmented) data

- $\xi = (b_1, \dots, b_N)^\top$.

Example 2: Logistic regress. with normally distrib. random effects

Likelihood of **observed data**

$$\begin{aligned}L_{obs}(\psi) &= \prod_{i=1}^N \int_{\mathbb{R}} \frac{e^{b_i \sum_{j=1}^{n_i} y_{i,j}}}{(1 + e^{b_i})^{n_i}} \varphi(b_i | \mu, d^2) db_i \\ &= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \left\{ \prod_{i=1}^N \frac{e^{b_i \sum_{j=1}^{n_i} y_{i,j}}}{(1 + e^{b_i})^{n_i}} \right\} \left\{ \prod_{i=1}^N \varphi(b_i | \mu, d^2) \right\} db_1 \cdots db_N.\end{aligned}$$

Augmented likelihood

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

Likelihood of **augmented data**

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

Example 3: Normal mixture model ($K > 1$ groups)

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

- Y_i i.i.d. with the mixture distribution with a density

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

Parameters

- $\psi \equiv \mathbf{w}, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K$

$$\mathbf{w} = (w_1, \dots, w_K)^\top, 0 < w_k < 1, \sum_{k=1}^K w_k = 1$$

Indirectly observable (augmented) data

- ???

Example 3: Normal mixture model ($K > 1$ groups)

Likelihood of **observed data**

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

Augmented likelihood

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

Likelihood of **augmented data**

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

Example 3: Normal mixture model ($K > 1$ groups)

- It holds also now:

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

- λ is now the product count measure on $\{1, \dots, K\}^N$ and hence

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

Section **8.4**

Notes

Notes

- Even Bayesian Data Augmentation is not for free. . .
- We are increasing (often quite considerably) **dimension** of the parametric space.
 - If we generate from the posterior distribution using the MCMC methods, it can be quite complicated to construct the Markov chain with a low autocorrelation allowing to explore the target (posterior) distribution within a reasonable number of iterations.
- When using the BDA, we primarily work with the posterior distribution

$$p(\xi, \psi | \mathbf{y}) \propto L_{augm}(\xi, \psi) p(\xi | \psi) p(\psi).$$

- Also here the prior distribution $p(\psi)$ can be (often is) specified hierarchically using random hyperparameters ζ with a prior distribution $p(\zeta)$.
- In reality, we then work with the posterior distribution

$$p(\xi, \psi, \zeta | \mathbf{y}) \propto L_{augm}(\xi, \psi) p(\xi | \psi) p(\psi | \zeta) p(\zeta).$$

Data augmentation (rozšiřování dat)

Other areas of use

- Models for **censored** data:
 - not only right-censored but also more general **interval**-censored data;
 - not only noninformative but also **informative** censoring.
- And many others. . .

9

Bayesian model selection

Section **9.1**
Bayes factor

Bayesian model

Data: \mathbf{Y} ,

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

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

Definition 9.1 Integrated (marginal) likelihood.

Marginal density of \mathbf{Y} following from the joint distribution of (\mathbf{Y}, θ) is called the **integrated (marginal) likelihood**, i.e.,

$$p(\mathbf{y}) = \int_{\Theta} p(\mathbf{y}, \theta) d\theta = \int_{\Theta} \underbrace{p(\mathbf{y} | \theta)}_{L(\theta)} p(\theta) d\theta.$$

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**.

- Interest in selecting a model from a set of candidate models M_1, \dots, M_r .
- Model M_k , $k = 1, \dots, r$:
 - Likelihood: $p_k(\mathbf{y} | \boldsymbol{\theta}_k) = L(\boldsymbol{\theta}_k), \quad \boldsymbol{\theta}_k \in \Theta_k \subset \mathbb{R}^{p_k},$
 - Prior distribution: $p_k(\boldsymbol{\theta}_k),$
 - Integrated likelihood: $p_k(\mathbf{y}) = \int_{\Theta} p_k(\mathbf{y} | \boldsymbol{\theta}_k) p_k(\boldsymbol{\theta}_k) d\boldsymbol{\theta}_k.$
- Integrated likelihood $p_k(\mathbf{y})$ can also be interpreted as distribution of data under validity of model M_k :

$$p_k(\mathbf{y}) = p(\mathbf{y} | M_k), \quad k = 1, \dots, r.$$

Model selection

- Let $P(M_1), \dots, P(M_r)$ be the **prior** probabilities of models M_1, \dots, M_r :

$$0 < P(M_k) < 1, \quad k = 1, \dots, r \quad \sum_{k=1}^r P(M_k) = 1.$$

- For example (but not necessarily): $P(M_k) = \frac{1}{r}, k = 1, \dots, r.$
- Model selection in Bayesian context can be based on **posterior** probabilities of models M_1, \dots, M_r :

$$P(M_k | \mathbf{y}) = \frac{p(\mathbf{y} | M_k) P(M_k)}{\sum_{l=1}^r p(\mathbf{y} | M_l) P(M_l)}, \quad k = 1, \dots, r.$$

- Choose model with the maximal posterior probability.
- “Small” complication: Integrated likelihood $p_k(\mathbf{y}) = p(\mathbf{y} | M_k)$ must be calculated for each model which requires calculation of (usually complicated/intractable) integral.

Definition 9.2 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.,

$$\text{BF}(M_k, M_j) = \frac{p(\mathbf{y} | M_k)}{p(\mathbf{y} | M_j)} = \frac{p_k(\mathbf{y})}{p_j(\mathbf{y})}.$$

Remarks

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

$$= \frac{P(M_k | \mathbf{y})}{P(M_j | \mathbf{y})} = \frac{p(\mathbf{y} | M_k) P(M_k)}{p(\mathbf{y} | M_j) P(M_j)} = \text{BF}(M_k, M_j) \underbrace{\frac{P(M_k)}{P(M_j)}}_{\text{prior odds}(M_k, M_j)}.$$

- With the uniform prior distribution for the competing models:

$$\text{posterior odds}(M_k, M_j) = \text{BF}(M_k, M_j).$$

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 $p_k(\mathbf{y})$ 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).
- $p_k(\mathbf{y})$ is not well defined when the prior distribution $p_k(\theta_k)$ 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.

Section **9.2**

Posterior predictive distribution

Bayesian model

Data: \mathbf{Y} ,

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

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

Posterior predictive distribution

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

Definition 9.3 Posterior predictive distribution.

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

We have

$$\begin{aligned} p(\mathbf{y}_{new} | \mathbf{y}) &= \int_{\Theta} p(\mathbf{y}_{new}, \theta | \mathbf{y}) d\theta = \int_{\Theta} p(\mathbf{y}_{new} | \theta, \mathbf{y}) p(\theta | \mathbf{y}) d\theta \\ &= \int_{\Theta} \underbrace{p(\mathbf{y}_{new} | \theta)}_{L_{new}(\theta)} p(\theta | \mathbf{y}) d\theta. \end{aligned}$$

Integrated likelihood

$$p(\mathbf{y}) = \int_{\Theta} L(\theta) p(\theta) d\theta$$

- ≡ 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

$$p(\mathbf{y}_{new} | \mathbf{y}) = \int_{\Theta} L_{new}(\theta) p(\theta | \mathbf{y}) d\theta$$

- ≡ Distribution of (new) data when the unknown parameters are averaged over their **posterior** distribution.
- ➡ Evidence of the model **after** using the data \mathbf{Y} for inference on unknown θ .

Section **9.3**

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.**

Definition 9.4 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

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

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

In context of statistical modelling

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

Then

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

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

Definition 9.5 Deviance of a model.

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

$$D(\theta; \mathbf{y}) = -2 \log\{p(\mathbf{y} | \theta)\} = -2 \log\{L(\theta)\}$$

is called the **deviance** of the model.

Remarks

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

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

- 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).

Typically

Data: $\mathbf{Y} = (\mathbf{Y}_1^\top, \dots, \mathbf{Y}_n^\top)^\top,$

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

$\mathbf{Y}_1, \dots, \mathbf{Y}_n$ (conditionally) independent given θ .

Deviance

$$\begin{aligned} D(\theta; \mathbf{y}) &= -2 \log \left\{ \prod_{i=1}^n p_i(\mathbf{y}_i | \theta) \right\} = -2 \sum_{i=1}^n \log \{ p_i(\mathbf{y}_i | \theta) \} \\ &= \sum_{i=1}^n \underbrace{\left[-2 \log \{ p_i(\mathbf{y}_i | \theta) \} \right]}_{D_i(\theta; \mathbf{y}_i)}. \end{aligned}$$

Section **9.4**

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: $\mathbf{Y} = (\mathbf{Y}_1^\top, \dots, \mathbf{Y}_n^\top)^\top$.

(New, not yet observed) data: $\mathbf{Y}_{new} = (\mathbf{Y}_{new,1}^\top, \dots, \mathbf{Y}_{new,n}^\top)^\top$.

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

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

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

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

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

Bayesian inference

Prior distribution: $p(\theta)$.

Integrated likelihood: $p(\mathbf{y}) = \int_{\Theta} p(\mathbf{y} | \theta) p(\theta) d\theta$

Evidence of the model before unknown parameters being estimated.

Inference on unknown θ based on the posterior distribution:

$$p(\theta | \mathbf{y}) = \frac{p(\mathbf{y} | \theta) p(\theta)}{p(\mathbf{y})}.$$

Posterior predictive distribution: $p(\mathbf{y}_{new} | \mathbf{y}) = \int_{\Theta} p(\mathbf{y}_{new} | \theta) p(\theta | \mathbf{y}) d\theta$

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

Posterior predictive deviance

Posterior predictive distribution:

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

Definition 9.6 Posterior predictive deviance.

Quantity

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

will be called the **posterior predictive deviance**.

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

Posterior predictive deviance

We have

$$\begin{aligned}\bar{D}_{pred} &= \mathbb{E}_{p(\theta | \mathbf{y})} D(\theta; \mathbf{y}_{new}) = \mathbb{E}_{p(\theta | \mathbf{y})} \left\{ \sum_{i=1}^n D_i(\theta; \mathbf{y}_{new,i}) \right\} \\ &= \sum_{i=1}^n \underbrace{\mathbb{E}_{p(\theta | \mathbf{y})} D_i(\theta; \mathbf{y}_{new,i})}_{\bar{D}_{pred,i}} = \sum_{i=1}^n \int_{\Theta} \underbrace{\left[-2 \log \{ p_i(\mathbf{y}_{new,i} | \theta) \} \right]}_{D_i(\theta, \mathbf{y}_{new,i})} p(\theta | \mathbf{y}) d\theta.\end{aligned}$$

- To calculate \bar{D}_{pred} in practice (to be able to use it for model selection), we need the value of “new” data \mathbf{Y}_{new} .

Posterior predictive deviance

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

≡ Values of new data $\mathbf{Y}_{new} = \mathbf{y}_{new}$ needed.

▮ Measure of the **loss of prediction**.

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

$$\bar{D}_{pred,i} \approx \bar{D}_{pred,i}^{CV} = \mathbb{E}_{p(\theta | \mathbf{y}_{-i})} D_i(\theta, \mathbf{y}_i) = \int_{\Theta} D_i(\theta, \mathbf{y}_i) p(\theta | \mathbf{y}_{-i}) d\theta.$$

Cross-validated posterior predictive deviance

Definition 9.7 Cross-validated posterior predictive deviance.

Quantity

$$\begin{aligned}\bar{D}_{pred}^{CV} &= \sum_{i=1}^n \underbrace{\mathbb{E}_{p(\theta | \mathbf{y}_{-i})} D_i(\theta; \mathbf{y}_i)}_{\bar{D}_{pred,i}^{CV}} \\ &= \sum_{i=1}^n \int_{\Theta} \underbrace{\left[-2 \log\{p_i(\mathbf{y}_i | \theta)\}\right]}_{D_i(\theta, \mathbf{y}_i)} p(\theta | \mathbf{y}_{-i}) d\theta.\end{aligned}$$

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).

Definition 9.8 Posterior expected deviance.

Quantity

$$\bar{D} = \mathbb{E}_{p(\theta | \mathbf{y})} D(\theta; \mathbf{y}) = \int_{\Theta} \underbrace{\left[-2 \log \{ p(\mathbf{y} | \theta) \} \right]}_{D(\theta; \mathbf{y})} p(\theta | \mathbf{y}) d\theta$$

will be called the **posterior expected deviance**.

We have

$$\begin{aligned} \bar{D} &= \mathbb{E}_{p(\theta | \mathbf{y})} D(\theta; \mathbf{y}) = \mathbb{E}_{p(\theta | \mathbf{y})} \sum_{i=1}^n D_i(\theta; \mathbf{y}_i) \\ &= \sum_{i=1}^n \underbrace{\mathbb{E}_{p(\theta | \mathbf{y})} D_i(\theta; \mathbf{y}_i)}_{\bar{D}_i} = \sum_{i=1}^n \int_{\Theta} \underbrace{\left[-2 \log \{ p_i(\mathbf{y}_i | \theta) \} \right]}_{D_i(\theta, \mathbf{y}_i)} p(\theta | \mathbf{y}) d\theta. \end{aligned}$$

Posterior expected deviance

$$\bar{D} = \sum_{i=1}^n \bar{D}_i = \sum_{i=1}^n \mathbb{E}_{p(\theta | \mathbf{y})} D_i(\theta, \mathbf{y}_i).$$

≡ Only the observed data $\mathbf{Y} = \mathbf{y}$ needed.

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

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

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

Theorem 9.1

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

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

Reminder

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

$$\bar{D}_i = \mathbb{E}_{p(\theta | \mathbf{y})} D_i(\theta, \mathbf{y}_i) = -2 \mathbb{E}_{p(\theta | \mathbf{y})} \log p_i(\mathbf{y}_i | \theta).$$

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

$$\begin{aligned} \text{KL}_2 &:= \text{KL}\left(p(\boldsymbol{\theta} | \mathbf{y}_{-i}), p(\boldsymbol{\theta} | \mathbf{y})\right) \\ &= \int_{\Theta} p(\boldsymbol{\theta} | \mathbf{y}) \log \left\{ \frac{p(\boldsymbol{\theta} | \mathbf{y})}{p(\boldsymbol{\theta} | \mathbf{y}_{-i})} \right\} d\boldsymbol{\theta} \\ &= \int_{\Theta} p(\boldsymbol{\theta} | \mathbf{y}) \log \left\{ \frac{p(\mathbf{y} | \boldsymbol{\theta}) p(\boldsymbol{\theta}) p(\mathbf{y}_{-i})}{p(\mathbf{y}_{-i} | \boldsymbol{\theta}) p(\boldsymbol{\theta}) p(\mathbf{y})} \right\} d\boldsymbol{\theta} \\ &= \int_{\Theta} p(\boldsymbol{\theta} | \mathbf{y}) \log \left\{ \frac{p_i(\mathbf{y}_i | \boldsymbol{\theta}) p(\mathbf{y}_{-i})}{p(\mathbf{y})} \right\} d\boldsymbol{\theta} \\ &= \int_{\Theta} p(\boldsymbol{\theta} | \mathbf{y}) \log \{p_i(\mathbf{y}_i | \boldsymbol{\theta})\} d\boldsymbol{\theta} - \log \left\{ \frac{p(\mathbf{y})}{p(\mathbf{y}_{-i})} \right\} \\ &= -\frac{1}{2} \mathbb{E}_{p(\boldsymbol{\theta} | \mathbf{y})} D_i(\boldsymbol{\theta}, \mathbf{y}_i) - \log \left\{ \frac{p(\mathbf{y})}{p(\mathbf{y}_{-i})} \right\} \\ &= -\frac{1}{2} \bar{D}_i - \log \left\{ \frac{p(\mathbf{y})}{p(\mathbf{y}_{-i})} \right\}. \end{aligned}$$

That is,

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

Definition 9.9 Expected optimism.

Quantity

$$\rho_{opt,i} = \mathbb{E}\left(\bar{D}_{pred,i}^{CV} - \bar{D}_i \mid \mathbf{Y}_{-i}\right), \quad i = 1, \dots, n$$

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

Penalized expected deviance

Definition 9.10 Penalized expected deviance (PED).

Quantity

$$\text{PED} = \underbrace{\sum_{i=1}^n \bar{D}_i}_{\bar{D}} + \underbrace{\sum_{i=1}^n \rho_{opt,i}}_{\rho_{opt}} = \sum_{i=1}^n \underbrace{(\bar{D}_i + \rho_{opt,i})}_{\text{PED}_i}$$

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

Quantity

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

will be called the **overall optimism**, quantity

$$\text{PED}_i = \bar{D}_i + \rho_{opt,i}, \quad i = 1, \dots, n,$$

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

Penalized expected deviance and cross-validated posterior predictive deviance

Theorem 9.2 Penalized expected deviance and cross-validated posterior predictive deviance.

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

$$\mathbb{E}(\text{PED}_i \mid \mathbf{Y}_{-i}) = \mathbb{E}\left(\overline{D}_{pred,i}^{CV} \mid \mathbf{Y}_{-i}\right).$$

With respect to cross-validation

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

is equivalent to

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

Penalized expected deviance and cross-validated posterior predictive deviance

Proof.

$$\begin{aligned}\mathbb{E}(\text{PED}_i \mid \mathbf{Y}_{-i}) &= \mathbb{E}\left\{\bar{D}_i + \underbrace{\mathbb{E}\left(\bar{D}_{pred,i}^{CV} - \bar{D}_i \mid \mathbf{Y}_{-i}\right)}_{\rho_{opt,i}} \mid \mathbf{Y}_{-i}\right\} \\ &= \mathbb{E}\left(\bar{D}_i \mid \mathbf{Y}_{-i}\right) + \mathbb{E}\left(\bar{D}_{pred,i}^{CV} \mid \mathbf{Y}_{-i}\right) - \mathbb{E}\left(\bar{D}_i \mid \mathbf{Y}_{-i}\right) \\ &= \mathbb{E}\left(\bar{D}_{pred,i}^{CV} \mid \mathbf{Y}_{-i}\right).\end{aligned}$$



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

$$p_{opt,i} = \mathbb{E}\left(\bar{D}_{pred,i}^{CV} - \bar{D}_i \mid \mathbf{Y}_{-i}\right), \quad i = 1, \dots, n.$$

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

Deviance information criterion

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

$$\rho_{opt} = \sum_{i=1}^n \rho_{opt,i} \approx \rho_D = \bar{D} - D(\hat{\theta}(\mathbf{y}); \mathbf{y}),$$

where $\hat{\theta}(\mathbf{y}) = \mathbb{E}_{p(\theta | \mathbf{y})} \theta$ (posterior mean of θ).

- Terminology: $D(\hat{\theta}(\mathbf{y}); \mathbf{y})$: **plug-in** deviance;
 ρ_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 ρ_D used in place of ρ_{opt}
▣ **Deviance information criterion (DIC).**

Deviance information criterion (DIC)

$$\begin{aligned} \text{DIC} &= \bar{D} + p_D \\ &= \bar{D} + \left\{ \bar{D} - D(\hat{\theta}(\mathbf{y}); \mathbf{y}) \right\} \\ &= 2\bar{D} - D(\hat{\theta}(\mathbf{y}); \mathbf{y}). \end{aligned}$$

- DIC \equiv approximation to \bar{D}_{pred}^{CV} 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.