Dept. of Probability and Mathematical Statistics

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

doc. RNDr. Arnošt Komárek, Ph.D.

NMST431 Bayesian Methods

Spring term 2022–23

1

Introduction

Section **1.1**

History

Where it began

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, Nonconformist minister and mathematician)

Bayes theorem (for events)

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

where $A_1, \dots, A_m$ events such that

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

Conditional probability

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

In this context:

$\mathrm{P}(A_k)$: prior (*apriorní*) probability of event A_k,

$\mathrm{P}\big(A_k \,\big|\, B\big)$: posterior (*aposteriorní*) probability of event A_k, given the fact that the event B occured.

Bayes theorem (for events)

- Derived by Thomas Bayes for $\mathrm{P}(A_k) = \frac{1}{m}$ for all $k = 1, \ldots, m$.
- 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

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

$$\mathrm{P}\big(\mathrm{INFECT} \,\big|\, \mathrm{TEST}+\big) = \frac{\mathrm{sens} \cdot \mathrm{prev}}{\mathrm{sens} \cdot \mathrm{prev} + (1 - \mathrm{spec}) \cdot (1 - \mathrm{prev})}$$

$$\mathrm{P}\big(\mathrm{INFECT} \,\big|\, \mathrm{TEST}-\big) = \frac{(1 - \mathrm{sens}) \cdot \mathrm{prev}}{(1 - \mathrm{sens}) \cdot \mathrm{prev} + \mathrm{spec} \cdot (1 - \mathrm{prev})}$$

depends on prevalence

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

The (testing) business must go on

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

Bayesian methods since 18th century

- 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 Finneti – 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 London.

Bayesian methods in past $\approx$ 30 years

- 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...

Pouze do povolaných rukou. . .

NALS | COVID-19 Science

Science Current Issue First release papers Archive About Submit manuscript

HOME > SCIENCE > VOL. 371, NO. 6531 > INFERRING THE EFFECTIVENESS OF GOVERNMENT INTERVENTIONS AGAINST COVID-19

RESEARCH ARTICLE

Inferring the effectiveness of government interventions against COVID-19

JAN M. BRAUNER, SÖREN MINDERMANN, MRINANK SHARMA, DAVID JOHNSTON, JOHN SALVATIER, TOMÁŠ GAVENČIAK, ANNA B. STEPHENSON, GAVIN LEECH, GEORGE ALTMAN, [...], AND JAN KULVEIT +9 authors Authors Info & Affiliations

SCIENCE · 15 Dec 2020 · Vol 371, Issue 6531 · DOI: 10.1126/science.abd9338

How to hold down transmission

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.

Science, this issue p. eabd9338

Structured Abstract

Section **1.2**

Basics

Setting for (simpler) statistical problems

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

Model (obvious and frequent part): $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N \overset{\text{i.i.d.}}{\sim} \boldsymbol{Y}$.

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

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

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

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

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

➔ Likelihood (given observed data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$)

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

Frequentist (classical) approach

$\boldsymbol{\theta} \in \Theta$ is unknown constant.

Principal tasks: point/interval estimation, hypothesis testing,

e.g., $\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} L(\boldsymbol{\theta})$ (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 $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$) under the same stochastic conditions (model) as the original data (e.g., unbiasedness);

(b) we add more data ($\boldsymbol{Y}_{N+1}, \boldsymbol{Y}_{N+2}, \ldots$) 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 $N << \infty$) that rely on asymptotics ($N = \infty$)?

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

Statistical inference

Bayesian approach
We postulate that some (stochastic) information is available about $\boldsymbol{\theta} \in \Theta$ **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 $\boldsymbol{\theta}$

given prior information and given data at hand.

No repeated sampling of data behind.

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

➔ not really needed for practical problems.

Bayesian approach

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

➔ 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);

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

$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top$: 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;

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

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

Bayes theorem

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.

$p(\boldsymbol{\theta})$: prior distribution for model parameters.

$\boldsymbol{Y} \equiv \left(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N\right) \equiv$ Data, $N^\star = N \cdot d$.

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

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

Theorem 1.1 Bayes.

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

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \begin{cases} \dfrac{p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\displaystyle\int_{\Theta} p(\boldsymbol{y} \mid \boldsymbol{\theta}^*)\, p(\boldsymbol{\theta}^*)\,\mathrm{d}\boldsymbol{\theta}^*}, & \displaystyle\int_{\Theta} p(\boldsymbol{y} \mid \boldsymbol{\theta}^*)\, p(\boldsymbol{\theta}^*)\,\mathrm{d}\boldsymbol{\theta}^* > 0, \\ 0, & \textit{otherwise.} \end{cases}$$

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

Bayes theorem

Bayes theorem, more important part: $p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$,

$\propto$ constants w.r.t. $\boldsymbol{\theta}$.

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

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

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

➔ meaning?

This is also the lunch price...

Basic theoretical problems

- Choice of a prior distribution ($p(\boldsymbol{\theta})$),

  see next part.

- Point and interval estimation, hypothesis testing, ...

  based on the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$.

Point and interval estimation

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

If it exists, perhaps

$$\widehat{\boldsymbol{\theta}} := \int_\Theta \boldsymbol{\theta}\, p(\boldsymbol{\theta} \,|\, \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta} = \mathbb{E}\big(\boldsymbol{\theta} \,\big|\, \boldsymbol{Y} = \boldsymbol{y}\big)?$$

Extension of the lunch price?

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

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

then

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

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)

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

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

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

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

Role of the prior distribution

Posterior distribution (typically)

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

$$= \left\{ \prod_{i=1}^{N} p(\boldsymbol{y}_i \mid \boldsymbol{\theta}) \right\} p(\boldsymbol{\theta})$$

$$= \left\{ \prod_{i=1}^{N} L_i(\boldsymbol{\theta}) \right\} L_{prior}(\boldsymbol{\theta}).$$

Prior distribution

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

Prior distribution

Knowledge on possible values of unknown parameters $\boldsymbol{\theta}$ 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 $p(\boldsymbol{\theta})$ concentrated on a confidence/credible set on $\boldsymbol{\theta}$ 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"

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{y}) \propto p(\boldsymbol{y} \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta}), \qquad \boldsymbol{\theta} \in \Theta.$$

- Support of $p(\boldsymbol{\theta} \,|\, \boldsymbol{y}) \subseteq$ support of $p(\boldsymbol{\theta})$.
- Client does not like negative values of θ.

  Náš zákazník, náš pán. $p(\theta) \sim \mathrm{Ga}(1, 1)$

  $\Rightarrow \mathrm{P}\big(\theta > 0 \,\big|\, \boldsymbol{Y} = \boldsymbol{y}\big) = 1$ (for any data $\boldsymbol{y}$).
- Client wishes $\widehat{\theta} \in (1, 2)$. *Služebníček.* $p(\theta) \sim \mathrm{Unif}(1, 2)$

  $\Rightarrow \mathrm{P}\big(\theta \in (1, 2) \,\big|\, \boldsymbol{Y} = \boldsymbol{y}\big) = 1$ (for any data $\boldsymbol{y}$).
- Client desires $\widehat{\theta} = 1$. *Nic není nemožné.* $p(\theta) \sim \mathrm{Dirac}(1)$

  $\Rightarrow \mathrm{P}\big(\theta = 1 \,\big|\, \boldsymbol{Y} = \boldsymbol{y}\big) = 1$ (for any data $\boldsymbol{y}$).

Prior distribution, client oriented approach

or  ?

Prior distribution

For majority of problems, nothing/not much is known in advance on $\boldsymbol{\theta}$.

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

- for given model $p(\boldsymbol{y} \,|\, \boldsymbol{\theta})$, $p(\boldsymbol{\theta})$ leads to "more tractable" calculation of the marginal likelihood $p(\boldsymbol{y}) = \int p(\boldsymbol{y} \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$;
- allow for specification of $p(\boldsymbol{\theta})$ 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 $p(\boldsymbol{y} \,|\, \boldsymbol{\theta})$ if and only if for any $\boldsymbol{\eta} \in \mathcal{H}$ and any $\boldsymbol{y} \in \mathbb{R}^d$ that satisfy

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

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

- Condition (2.1), its part > 0, means that we only care about data $\boldsymbol{Y} = \boldsymbol{y}$ 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 $p(\boldsymbol{y} \,|\, \boldsymbol{\theta})$ then the respective posterior $p(\boldsymbol{\theta} \,|\, \boldsymbol{y}) = q(\boldsymbol{\theta};\, \boldsymbol{\eta}_{\text{poster}}(\boldsymbol{y}))$ for some $\boldsymbol{\eta}_{\text{poster}}(\boldsymbol{y}) \in \mathcal{H}$.

Example 2.1 (Normal model with known variance).

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

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

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

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

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

Conjugate system of prior distributions

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

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

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

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

$$p(\boldsymbol{y} \,|\, \boldsymbol{\theta}) \;=\; g_1\big(\boldsymbol{T}(\boldsymbol{y});\, \boldsymbol{\theta}\big)\, g_2(\boldsymbol{y}), \qquad \boldsymbol{\theta} \in \Theta,$$

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

$$\mathcal{H} \;=\; \Big\{\boldsymbol{\eta} \in \mathbb{R}^r :\; 0 < \int_\Theta g_1\big(\boldsymbol{\eta};\, \boldsymbol{\theta}\big)\mathrm{d}\boldsymbol{\theta} < \infty\Big\}.$$

Further, let

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

The system

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

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

Conjugate system of prior distributions

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

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

$$\begin{aligned} p(\boldsymbol{y} \,|\, \theta) &= \theta^{\sum_{i=1}^N y_i} (1-\theta)^{N - \sum_{i=1}^N y_i}, \\ q(\theta;\, a,\, b) &\propto \theta^{a-1} (1-\theta)^{b-1} \\ &\sim \mathrm{Be}(a,\, b), \qquad \boldsymbol{\eta} \equiv a > 0,\ b > 0, \\ p(\theta \,|\, \boldsymbol{y};\, a,\, b) &\sim \mathrm{Be}\Big(a + \sum_{i=1}^N y_i,\, b + N - \sum_{i=1}^N y_i\Big). \end{aligned}$$

Conjugate system of prior distributions

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

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

$$
\begin{aligned}
p(\boldsymbol{y} \,|\, \theta) &= \exp(-\theta\, N) \frac{\theta^{\sum_{i=1}^N y_i}}{\prod_{i=1}^N y_i!}, \\
q(\theta;\, a,\, b) &\propto \theta^{a-1} \exp(-\theta\, b) \\
&\sim \text{Ga}(a,\, b), \qquad \boldsymbol{\eta} \equiv a > 0,\ b > 0, \\
p(\theta \,|\, \boldsymbol{y};\, a,\, b) &\sim \text{Ga}\Big(a + \sum_{i=1}^N y_i,\, b + N\Big).
\end{aligned}
$$

Conjugate system of prior distributions

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

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

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

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

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

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

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

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

$$q(\boldsymbol{\theta};\, \boldsymbol{\eta}) \;\propto\; ???$$

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

☞ 

Conjugate system of prior distributions

In a particular situation (analysis of a given dataset): value $\boldsymbol{\eta} \in \mathcal{H}$ 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 $\boldsymbol{Y}_{old}$ the model $p\big(\boldsymbol{y}_{old} \,\big|\, \boldsymbol{\theta}\big), \boldsymbol{\theta} \in \Theta$.

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

➔ The integrated/marginal likelihood od historical data is

$$p(\boldsymbol{y}_{old};\, \boldsymbol{\eta}) = \int_{\Theta} p\big(\boldsymbol{y}_{old} \,\big|\, \boldsymbol{\theta}\big)\, q(\boldsymbol{\theta};\, \boldsymbol{\eta})\, \mathrm{d}\boldsymbol{\theta}$$

➔ Likelihood of historical data which depends on unknown (hyper)parameter $\boldsymbol{\eta}$.

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

➔ $\widehat{\boldsymbol{\eta}}$.

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

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

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

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

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

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

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

Prior for the analyzis with the "new" data is then $p(\theta) \ \sim \ \mathcal{N}(\widehat{\mu}_q, \widehat{\sigma}_q^2)$.

Hierarchical prior

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

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

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

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

Data and the model: $\boldsymbol{Y} \equiv Y_1, \ldots, Y_N \overset{i.i.d.}{\sim} \mathcal{N}(\theta, \sigma_0^2)$,
$\theta \in \mathbb{R}$: *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 analyzis:

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

Hierarchical prior

Both $\boldsymbol{\theta}$ (primary parameters) as well as $\boldsymbol{\eta}$ (hyperparameters) can be viewed as just parameters of the Bayesian model having the (joint) prior

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

where the model for data does not depend on $\boldsymbol{\eta}$, i.e., where

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

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

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

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

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

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 $p(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \boldsymbol{y})$ rather than to work directly with $p(\boldsymbol{\theta} \mid \boldsymbol{y})$.

Multilevel hierarchical prior

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

M levels of hyperparameters: $\boldsymbol{\eta}_1 \in \mathcal{H}_1, \ldots, \boldsymbol{\eta}_M \in \mathcal{H}_M$.

Hierarchically specified joint prior

$$
\begin{aligned}
p(\boldsymbol{\theta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_M) \\
&= p(\boldsymbol{\theta} \,|\, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_M) \cdot p(\boldsymbol{\eta}_1 \,|\, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_M) \cdots p(\boldsymbol{\eta}_{M-1} \,|\, \boldsymbol{\eta}_M) \cdot p(\boldsymbol{\eta}_M) \\
&\overset{\text{assume}}{=} p(\boldsymbol{\theta} \,|\, \boldsymbol{\eta}_1) \cdot p(\boldsymbol{\eta}_1 \,|\, \boldsymbol{\eta}_2) \cdots p(\boldsymbol{\eta}_{M-1} \,|\, \boldsymbol{\eta}_M) \cdot p(\boldsymbol{\eta}_M).
\end{aligned}
$$

Prior for the primary parameters $\boldsymbol{\theta}$:

$$
p(\boldsymbol{\theta}) = \int_{\mathcal{H}_1} \cdots \int_{\mathcal{H}_{M-1}} \int_{\mathcal{H}_M} p(\boldsymbol{\theta} \,|\, \boldsymbol{\eta}_1) \cdot p(\boldsymbol{\eta}_1 \,|\, \boldsymbol{\eta}_2) \cdots p(\boldsymbol{\eta}_{M-1} \,|\, \boldsymbol{\eta}_M) \cdot p(\boldsymbol{\eta}_M) \, \mathrm{d}\boldsymbol{\eta}_M \, \mathrm{d}\boldsymbol{\eta}_{M-1} \cdots \mathrm{d}\boldsymbol{\eta}_1.
$$

Section **2.4**

Noninformative prior

Noninformative prior (princip neurčitosti)

Uniform distribution over the whole parameter space Θ

$$q(\boldsymbol{\theta}) \propto 1, \qquad \boldsymbol{\theta} \in \Theta.$$

It can be

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

➔ improper density (*nevlastní hustota*).

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

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

leading to *proper* posterior

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

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

Example 2.4 (Noniformative 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 $\psi = \log(\sigma), \quad \psi \in \mathbb{R}$*.*

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

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

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

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

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 $p(\boldsymbol{y} \mid \boldsymbol{\theta})$, $\boldsymbol{y} \in N^*$ (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 an non-empty open set in $\mathbb{R}^k$.

(ii) The set $M = \{\boldsymbol{y} : p(\boldsymbol{y} \mid \boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta}$.

(iii) For almost all $\boldsymbol{y} \in \mathbb{R}^{N^*}$, for all $\boldsymbol{\theta} \in \Theta$ and for each $j = 1, \ldots, k$ there exist a finite partial derivative $p_j'(\boldsymbol{y} \mid \boldsymbol{\theta}) := \frac{\partial}{\partial \theta_j} p(\boldsymbol{y} \mid \boldsymbol{\theta})$.

(iv) For each $\boldsymbol{\theta} \in \Theta$ and each $j = 1, \ldots, k$ $\int_M p_j'(\boldsymbol{y} \mid \boldsymbol{\theta}) \mathrm{d}\nu(\boldsymbol{y}) = 0$.

(v) For each $\boldsymbol{\theta} \in \Theta$ and each pair (j, l) there exist a finite integral

$$J_{j,l}(\boldsymbol{\theta}) = \int_M \frac{p_j'(\boldsymbol{y} \mid \boldsymbol{\theta})\, p_l'(\boldsymbol{y} \mid \boldsymbol{\theta})}{p^2(\boldsymbol{y} \mid \boldsymbol{\theta})}\, p(\boldsymbol{y} \mid \boldsymbol{\theta})\, \mathrm{d}\nu(\boldsymbol{y}).$$

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

Notes

- Matrix $\mathbb{J}(\boldsymbol{\theta})$ is called the Fisher information matrix.
- In case of i.i.d. data $\boldsymbol{Y} \equiv \big(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N\big)$, $\boldsymbol{Y}_i \overset{\text{i.i.d.}}{\sim} p_0\big(\boldsymbol{y}_i \,\big|\, \boldsymbol{\theta}\big)$, $\boldsymbol{y}_i \in \mathbb{R}^d$, $i = 1, \ldots, N$, $N^* = N\, d$, 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 $p_0\big(\cdot \,\big|\, \boldsymbol{\theta}\big)$.
- For each $\boldsymbol{\theta} \in \Theta$ and each $j = 1, \ldots, k$

  $$\frac{p_j'\big(\boldsymbol{y} \,\big|\, \boldsymbol{\theta}\big)}{p\big(\boldsymbol{y} \,\big|\, \boldsymbol{\theta}\big)} = \frac{\partial}{\partial \theta_j} \log\Big\{p\big(\boldsymbol{y} \,\big|\, \boldsymbol{\theta}\big)\Big\} =: U_j(\boldsymbol{\theta};\, \boldsymbol{y}).$$

- Vector

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

  is called the score vector.

Maximum-likelihood theory in a nutshell

Notes, cont'd

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

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

- Under regularity conditions also

$$\mathbb{J}(\boldsymbol{\theta}) \;=\; \mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{\theta})} \Bigl[-\frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top} \log\Bigl\{ p\bigl(\boldsymbol{Y} \bigm| \boldsymbol{\theta}\bigr) \Bigr\} \Bigr].$$

- Matrix

$$\mathbb{I}(\boldsymbol{\theta};\ \boldsymbol{y}) := -\frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top} \log\Bigl\{ p\bigl(\boldsymbol{y} \bigm| \boldsymbol{\theta}\bigr) \Bigr\}$$

  is called the observed information matrix.

- Asymptotic normality of the MLE $\widehat{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \log\Bigl\{ p\bigl(\boldsymbol{y} \bigm| \boldsymbol{\theta}\bigr) \Bigr\}$ related to above matrices.

Theorem 2.1 Jeffreys.

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

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

Let H *be a regular injective function (prosté zobrazení)* $\Theta \longrightarrow \Psi \in \mathcal{B}^k$. *Let* $\psi = H(\boldsymbol{\theta})$ *and* $p^*(\mathbf{y} \,|\, \psi) = p\big(\mathbf{y} \,|\, H^{-1}(\psi)\big)$. *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* $c(\mathbf{y}) > 0$

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

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

Jeffreys $$\mathrm{P}(\boldsymbol{\theta} \in B \mid \boldsymbol{Y} = \boldsymbol{y}) = \mathrm{P}(\boldsymbol{\eta} \in H(B) \mid \boldsymbol{Y} = \boldsymbol{y})$$

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

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

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

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

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

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

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

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

Jeffreys prior

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

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

Jeffreys prior

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

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

Data: $Y_1, \ldots, Y_N \overset{i.i.d.}{\sim} \mathrm{Po}(\theta)$, $\theta \in (0, \infty)$.

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

$$\ell(\theta;\, \boldsymbol{y}) := \log\Bigl\{p\bigl(\boldsymbol{y} \,\big|\, \theta\bigr)\Bigr\} \;=\; -\sum_{i=1}^N \log(y_i!) - N\theta + \sum_{i=1}^N y_i \log(\theta),$$

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

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

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

Improper Jeffreys prior for Poisson sample

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

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

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

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

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: $\boldsymbol{Y} \equiv Y_1, \ldots, Y_N \overset{i.i.d.}{\sim} \mathrm{DE}(\theta)$, $\theta > 0$ *parameter.*

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

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

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

3

Bayesian statistical inference

Section **3.1**

Inference

Bayesian statistical inference

Everything based on the posterior distribution

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int_{\Theta} p(\boldsymbol{y} \mid \boldsymbol{\theta}^*)\, p(\boldsymbol{\theta}^*)\, \mathrm{d}\boldsymbol{\theta}^*} \propto p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}), \qquad \boldsymbol{\theta} \in \Theta.$$

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

Posterior cdf:

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

Quantiles of the posterior distribution:

$$G^{-1}(\alpha;\, \boldsymbol{y}) := \inf\bigl\{\theta :\; G(\theta;\, \boldsymbol{y}) \ge \alpha\bigr\}, \quad 0 < \alpha < 1.$$

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

Point estimate

Suitable location characteristic of the posterior distribution

1. Posterior mean (if it exists):

$$\overline{\boldsymbol{\theta}} := \mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y})} \boldsymbol{\theta} = \int_{\Theta} \boldsymbol{\theta} \, p(\boldsymbol{\theta} \mid \boldsymbol{y}) \, \mathrm{d}\boldsymbol{\theta}.$$

2. Posterior median (in the univariate case):

$$\widetilde{\theta} := G^{-1}(0.5; \, \boldsymbol{y}).$$

"Uncertainty" (in the univariate case)

$\equiv$ "spread" of the posterior distribution

1. Posterior standard deviation (if it exists):

$$\mathrm{SD}_{p(\theta \mid \boldsymbol{y})}(\theta) \;=\; \sqrt{\int_{\Theta} \left(\theta - \overline{\theta}\right)^2 p(\theta \mid \boldsymbol{y})\,\mathrm{d}\theta}.$$

➔ counterpart of the standard error from the frequentist statistics.

2. Posterior quartiles:

$$Q_1\left(\theta \,\middle|\, \boldsymbol{y}\right) \;:=\; G^{-1}(0.25;\, \boldsymbol{y}), \qquad Q_3\left(\theta \,\middle|\, \boldsymbol{y}\right) \;:=\; G^{-1}(0.75;\, \boldsymbol{y}).$$

Credible set/region

Definition 3.1 Credible set/region.

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

$$\mathsf{P}\bigl(\boldsymbol{\theta} \in C_\alpha(\boldsymbol{y}) \bigm| \boldsymbol{Y} = \boldsymbol{y}\bigr) \;=\; \int_{C_\alpha(\boldsymbol{y})} p\bigl(\boldsymbol{\theta} \bigm| \boldsymbol{y}\bigr)\,\mathrm{d}\boldsymbol{\theta} \;=\; 1 - \alpha.$$

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.

Definition 3.2 Highest posterior density credible region.

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

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

where k_α is the highest constant such that

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

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

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

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

Credible interval

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

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

Equal-tail (ET) credible interval

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

HPD credible interval

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

- It is the shortest credible interval.
- 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 $\boldsymbol{\theta}_0 \in \Theta$ 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: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ against H_1: $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ for chosen $\boldsymbol{\theta}_0 \in \Theta$ while using the duality between testing and confidence intervals/regions.

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

Definition 3.3 Bayesian P-value.

For a given point $\boldsymbol{\theta}_0$ and a given approach to construct the credible regions $C_\alpha(\boldsymbol{y})$, $0 < \alpha < 1$, with $C_0(\boldsymbol{y}) := \Theta$ and $C_1(\boldsymbol{y}) := \emptyset$, the Bayesian P-value $p(\boldsymbol{\theta}_0;\ \boldsymbol{y})$ is defined as

$$p(\boldsymbol{\theta}_0;\ \boldsymbol{y}) = \inf\bigl\{\alpha :\ \boldsymbol{\theta}_0 \notin C_\alpha(\boldsymbol{y})\bigr\}.$$

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

But it has a similar interpretation:

- $p(\boldsymbol{\theta}_0;\ \boldsymbol{y}) \to 0$: small posterior evidence for $\boldsymbol{\theta}_0$
  (high posterior evidence against $\boldsymbol{\theta}_0$);
- $p(\boldsymbol{\theta}_0;\ \boldsymbol{y}) \to 1$: high posterior evidence for $\boldsymbol{\theta}_0$
  (small posterior evidence against $\boldsymbol{\theta}_0$).

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

Bayesian P-value

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

$$\rightarrow \qquad p(\theta_0;\, \boldsymbol{y}) = 2\,\min\bigl\{G(\theta_0;\, \boldsymbol{y}),\; 1 - G(\theta_0;\, \boldsymbol{y})\bigr\}.$$

“One-sided” variants cound 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

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

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).

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

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

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

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

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

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

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

Definition 3.4 Posterior predictive distribution.

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

$$f_{pred}(\boldsymbol{y}_{new}) := \mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y})} f_{new}(\boldsymbol{y}_{new}; \boldsymbol{\theta}) = \int_{\Theta} f_{new}(\boldsymbol{y}_{new}; \boldsymbol{\theta}) \, p(\boldsymbol{\theta} \mid \boldsymbol{y}) \, \mathrm{d}\boldsymbol{\theta}.$$

Notes

- The density $f_{pred}(\boldsymbol{y}_{new}) = \int_{\Theta} f_{new}(\boldsymbol{y}_{new};\, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \,|\, \boldsymbol{y})\, \mathrm{d}\boldsymbol{\theta}$ is called the posterior predictive density

  $\equiv$ (marginal) distribution of $\boldsymbol{Y}_{new}$ after the information on $\boldsymbol{\theta}$ included in data $\boldsymbol{Y}$ is taken into account.

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

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

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

  $\equiv$ (marginal) distribution of $\boldsymbol{Y}_{new}$ when only the prior information on $\boldsymbol{\theta}$ is taken into account.

Point prediction

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

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

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

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: $\boldsymbol{Y} \equiv Y_1, \ldots, Y_N$ *independent,* $Y_i \sim \mathcal{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2)$, *where*

- $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathbb{R}^k$ *are known constants;*
- $\boldsymbol{\beta} \in \mathbb{R}^k$, $0 < \sigma^2 < \infty$ *are unknown parameters,*

  $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top, \sigma^2\right)^\top \in \Theta, \quad \Theta = \mathbb{R}^k \times (0, \infty);$
- *It is assumed that the* $N \times k$ *matrix*

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

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

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, μ*g):*

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

➔ *Data* $\boldsymbol{Y} = \left(Y_1, \ldots, Y_N\right)^\top$, $N = 18$.

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

Exercise: Normal linear model

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

Further notation:

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

$$\boldsymbol{b} = \boldsymbol{b}(\boldsymbol{y}) \;=\; \bigl(\mathbb{X}^\top \mathbb{X}\bigr)^{-1} \mathbb{X}^\top \boldsymbol{y} \qquad = \bigl(98.89,\, 124.42\bigr)^\top,$$

$$\mathrm{SS}_e = \mathrm{SS}_e(\boldsymbol{y}) \;=\; \mathrm{SS}(\boldsymbol{b};\, \boldsymbol{y}) \qquad = 2\,525.7.$$

The model implies:

$$p(\boldsymbol{y} \,|\, \boldsymbol{\theta}) \;=\; \bigl(2\,\pi\,\sigma^2\bigr)^{-N/2} \exp\Bigl\{-\frac{1}{2\sigma^2}\,\mathrm{SS}(\boldsymbol{\beta};\, \boldsymbol{y})\Bigr\}$$

$$=\; \bigl(2\,\pi\,\sigma^2\bigr)^{-N/2} \exp\Bigl\{-\frac{\mathrm{SS}_e}{2\sigma^2} \;-\; \frac{1}{2\sigma^2}\,\bigl(\boldsymbol{\beta} - \boldsymbol{b}\bigr)^\top \mathbb{X}^\top \mathbb{X}\,\bigl(\boldsymbol{\beta} - \boldsymbol{b}\bigr)\Bigr\}, \qquad \boldsymbol{y} \in \mathbb{R}^N.$$

Exercise: Normal linear model

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

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

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

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

Exercise: Normal linear model

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

➔ $\mathbb{X}^\top \mathbb{X} = N$,

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

Jeffreys prior

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

Exercise: Normal linear model

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

Conjugate system

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

$$
\begin{aligned}
p(\boldsymbol{\beta},\, \tau) = p(\boldsymbol{\beta} \,|\, \tau)\, p(\tau), \qquad & p(\boldsymbol{\beta} \,|\, \tau) \sim \mathcal{N}(\boldsymbol{\beta}_0,\, \tau^{-1}\,\boldsymbol{\Sigma}_0), \\
& p(\tau) \sim \mathrm{Ga}(c_0,\, d_0). \\
\sim \mathcal{N}\text{-Ga}(\boldsymbol{\beta}_0,\, \boldsymbol{\Sigma}_0,\, c_0,\, d_0), &
\end{aligned}
$$

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

Exercise: Normal linear model

Semiconjugate system

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

$$p(\boldsymbol{\beta}) \sim \mathcal{N}(\beta_0, \boldsymbol{\Sigma}_0),$$

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

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

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

Jeffreys motivated improper prior

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

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

Does it lead to proper posterior?

Exercise: Normal linear model

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

$$p(\boldsymbol{y} \,|\, \boldsymbol{\theta}) \;=\; (2\,\pi)^{-N/2}\,\tau^{N/2} \exp\Bigl\{-\frac{\tau}{2}\,\mathrm{SS}(\boldsymbol{\beta};\, \boldsymbol{y})\Bigr\}$$

$$=\; (2\,\pi)^{-N/2}\,\tau^{N/2} \exp\Bigl\{-\tau\,\frac{\mathrm{SS}_e}{2} \;-\; \frac{\tau}{2}\,(\boldsymbol{\beta} - \boldsymbol{b})^\top\,\mathbb{X}^\top\mathbb{X}\,(\boldsymbol{\beta} - \boldsymbol{b})\Bigr\}, \quad \boldsymbol{y} \in \mathbb{R}^N,$$

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

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

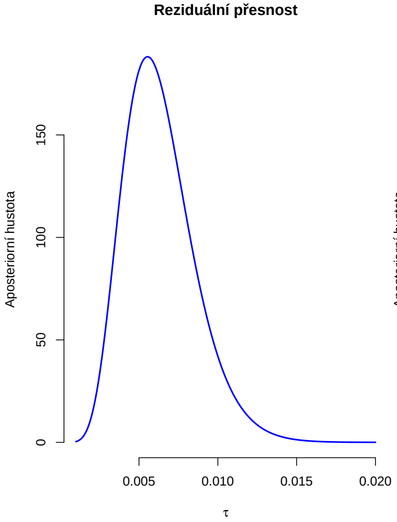
$$p(\tau \,|\, \boldsymbol{y}) \;\sim\; \mathrm{Ga}\Bigl(\frac{N-k}{2},\, \frac{\mathrm{SS}_e}{2}\Bigr),$$

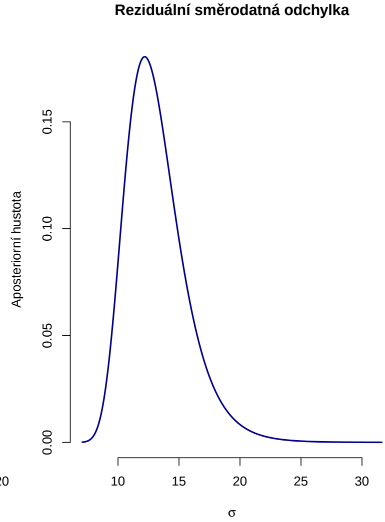
$$p(\boldsymbol{\beta} \,|\, \tau,\, \boldsymbol{y}) \;\sim\; \mathcal{N}_k\bigl(\boldsymbol{b},\, \tau^{-1}\,(\mathbb{X}^\top\mathbb{X})^{-1}\bigr),$$

$$p(\boldsymbol{\beta} \,|\, \boldsymbol{y}) \;\sim\; \boldsymbol{b} + \mathrm{mvt}_k\Bigl(N-k,\, \frac{\mathrm{SS}_e}{N-k}\,(\mathbb{X}^\top\mathbb{X})^{-1}\Bigr).$$

Example: Weighing of light objects

$$p(\tau \mid \boldsymbol{y}) \sim \mathsf{Ga}\Big(\frac{N-k}{2}, \frac{\mathsf{SS}_e}{2}\Big) = \mathsf{Ga}(8,\ 1\,262.8).$$





Example: Weighing of light objects

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

$$\mathbb{E}(\tau \mid \boldsymbol{Y} = \boldsymbol{y}) = \frac{n-k}{\mathrm{SS}_e}$$

$$\mathrm{var}(\tau \mid \boldsymbol{Y} = \boldsymbol{y}) = \frac{2(n-k)}{\mathrm{SS}_e^2}$$

$$\mathbb{E}(\sigma^2 \mid \boldsymbol{Y} = \boldsymbol{y}) = \frac{\mathrm{SS}_e}{n-k-2}, \qquad \text{for } n-k > 2$$

$$\mathrm{var}(\sigma^2 \mid \boldsymbol{Y} = \boldsymbol{y}) = \frac{2\,\mathrm{SS}_e^2}{(n-k-2)^2\,(n-k-4)}, \qquad \text{for } n-k > 4$$

$$\mathbb{E}(\sigma \mid \boldsymbol{Y} = \boldsymbol{y}) = ???$$

$$\mathrm{var}(\sigma \mid \boldsymbol{Y} = \boldsymbol{y}) = ???$$

Example: Weighing of light objects

$$p(\tau \mid \boldsymbol{y}) \sim \mathrm{Ga}\Big(\frac{N-k}{2}, \frac{\mathrm{SS}_e}{2}\Big) = \mathrm{Ga}(8,\ 1\,262.8).$$

Possible **point estimates**

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

$\widehat{\tau}_2 = \mathrm{med}\big(\tau \mid \boldsymbol{Y} = \boldsymbol{y}\big) = 0.00607$ $\qquad$ $\widehat{\sigma}_2 = \mathrm{med}\big(\sigma \mid \boldsymbol{Y} = \boldsymbol{y}\big) = 12.83$

95% credible intervals

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

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

Multivariate t-distribution

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

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

It has a density

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

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

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

Shifted multivariate t-distribution

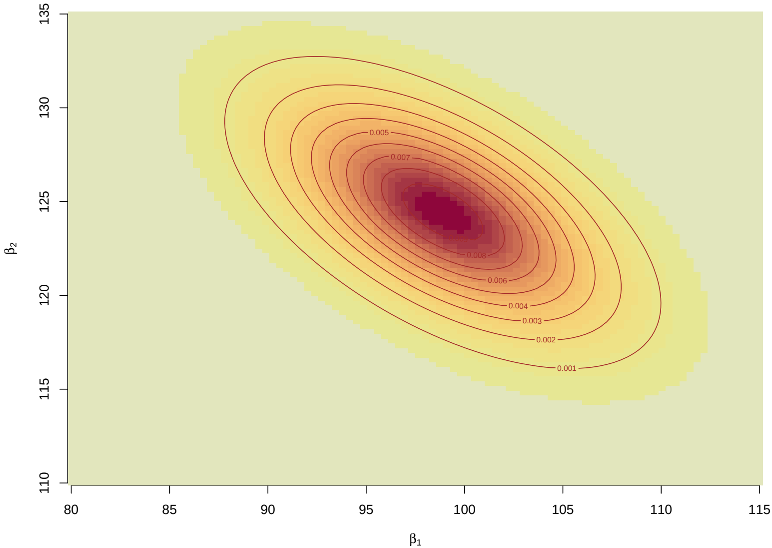
For $\boldsymbol{\mu} \in \mathbb{R}^k$, the random vector $\boldsymbol{Z} = \boldsymbol{\mu} + \boldsymbol{T}$, where $\boldsymbol{T} \sim \mathrm{mvt}_k(\nu, \boldsymbol{\Sigma})$ has a density

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

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

Example: Weighing of light objects

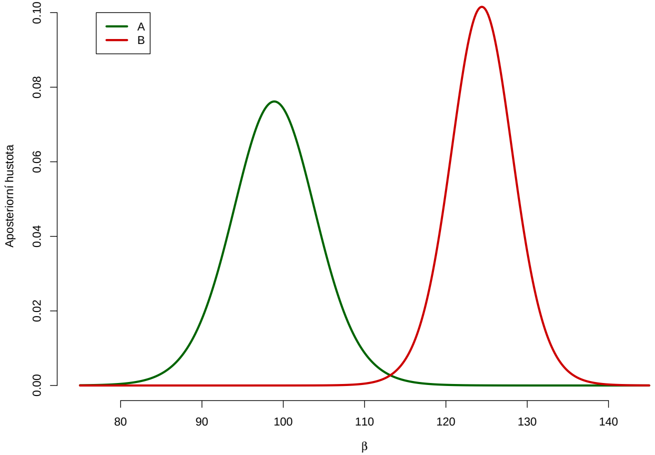
$$p(\beta \mid \boldsymbol{y}) \sim \boldsymbol{b} + \mathrm{mvt}_k\Big(N-k,\ \frac{SS_e}{N-k}\,(\mathbb{X}^\top\mathbb{X})^{-1}\Big) = \begin{pmatrix} 98.89 \\ 124.42 \end{pmatrix} + \mathrm{mvt}_2\Big(16,\ \frac{2\,525.7}{16}\,(\mathbb{X}^\top\mathbb{X})^{-1}\Big).$$



Example: Weighing of light objects

$$p(\beta \,|\, \boldsymbol{y}) \;\sim\; \boldsymbol{b} + \mathrm{mvt}_k\Bigl(N-k,\ \frac{SS_e}{N-k}\,(\mathbb{X}^\top\mathbb{X})^{-1}\Bigr) = \begin{pmatrix} 98.89 \\ 124.42 \end{pmatrix} + \mathrm{mvt}_2\Bigl(16,\ \frac{2\,525.7}{16}\,(\mathbb{X}^\top\mathbb{X})^{-1}\Bigr).$$

➔ For each $j = 1, \ldots, k$ $\quad \dfrac{\beta_j - b_j}{\sqrt{\frac{SS_e}{N-k}\, v_{j,j}}} \,\Bigg|\, \boldsymbol{Y} = \boldsymbol{y} \;\sim\; \mathrm{t}_{N-k}, \quad v_{j,j} = (j,j)$ diag. elem. of $(\mathbb{X}^\top\mathbb{X})^{-1}$.



Example: Weighing of light objects

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) \;\sim\; \boldsymbol{b} + \mathrm{mvt}_k\Big(N-k,\, \frac{\mathrm{SS}_e}{N-k}\,(\mathbb{X}^\top\mathbb{X})^{-1}\Big) = \begin{pmatrix} 98.89 \\ 124.42 \end{pmatrix} + \mathrm{mvt}_2\Big(16,\, \frac{2\,525.7}{16}\,(\mathbb{X}^\top\mathbb{X})^{-1}\Big).$$

Possible **point estimates**

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

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

95% credible intervals

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

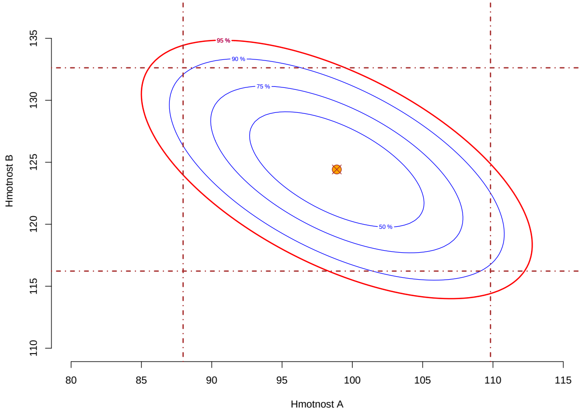
$\beta_2 : (116.22,\; 132.62)$

Example: Weighing of light objects

$$p(\boldsymbol{\beta} \,|\, \boldsymbol{y}) \;\sim\; \boldsymbol{b} + \mathrm{mvt}_k\Big(N-k,\; \frac{SS_e}{N-k}\,(\mathbb{X}^\top\mathbb{X})^{-1}\Big) = \begin{pmatrix} 98.89 \\ 124.42 \end{pmatrix} + \mathrm{mvt}_2\Big(16,\; \frac{2\,525.7}{16}\,(\mathbb{X}^\top\mathbb{X})^{-1}\Big).$$

$$\rightarrow \qquad \frac{1}{k}\,(\boldsymbol{\beta}-\boldsymbol{b})^\top\Big(\frac{N-k}{SS_e}\,\mathbb{X}^\top\mathbb{X}\Big)(\boldsymbol{\beta}-\boldsymbol{b})\;\Big|\;\boldsymbol{Y}=\boldsymbol{y} \;\sim\; \mathrm{F}_{k,N-k}.$$

→ $100\,(1-\alpha)\%$ HPD credible region for $\boldsymbol{\beta}$:



Exercise: Normal linear model

Proper semiconjugate prior

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

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

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

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

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

$$p(\tau \,|\, \boldsymbol{y}) \sim \mathrm{Ga}\Big(c_0 + \frac{N-k}{2},\, d_0 + \frac{\mathrm{SS}_e}{2}\Big),$$

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

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

$\boldsymbol{\mu}_{canon} = \boldsymbol{\Sigma}_0^{-1}\, \boldsymbol{\beta}_0 + \tau\, \mathbb{X}^\top \mathbb{X}\, \boldsymbol{b}$ 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 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top \in \Theta \subseteq \mathbb{R}^k$:

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

We need

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

$$\overline{t_{\boldsymbol{\theta}}} := \mathbb{E}_{p(\boldsymbol{\theta} \,|\, \boldsymbol{y})} t(\boldsymbol{\theta}) = \int_\Theta t(\boldsymbol{\theta})\, p(\boldsymbol{\theta} \,|\, \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta} \qquad \text{(if it exists).}$$

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

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

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

  ➔ credible intervals, ...

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

Monte Carlo integration in statistics

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

$$\overline{t_{\boldsymbol{\theta}}} := \mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y})} t(\boldsymbol{\theta}) = \int_{\Theta} t(\boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta} \qquad \text{(if it exists).}$$

Assumption:

$$\int_{\Theta} \big|t(\boldsymbol{\theta})\big|\, p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta} \;<\; \infty.$$

Monte Carlo **principle**:

- Let $\mathcal{S}_M = \{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}\}, \quad \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{\theta} \mid \boldsymbol{y})$.
- Then (*law of large numbers*)

$$\widehat{t_M} := \frac{1}{M} \sum_{m=1}^{M} t\big(\boldsymbol{\theta}^{(m)}\big) \xrightarrow{\text{P}} \mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{y})} t(\boldsymbol{\theta}) = \overline{t_{\boldsymbol{\theta}}} \qquad \text{as } M \to \infty.$$

- $\widehat{t_M}$ = Monte Carlo (MC) approximation/estimate of $\overline{t_{\boldsymbol{\theta}}}$.

Precision of the MC integration:

Let $\int_\Theta \left\{t(\boldsymbol{\theta})\right\}^2 p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta} < \infty$ and

$$\sigma_t^2 := \int_\Theta \left\{t(\boldsymbol{\theta}) - \overline{t_\theta}\right\}^2 p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}.$$

Then

$$v_M := \mathrm{var}\big(\widehat{t_M}\big) = \mathrm{var}\left\{\frac{1}{M}\sum_{m=1}^{M} t\big(\boldsymbol{\theta}^{(m)}\big)\right\} = \frac{\sigma_t^2}{M}.$$

Law of large numbers:

$$\widehat{\sigma}_M^2 := \frac{1}{M-1}\sum_{m=1}^{M}\left\{t\big(\boldsymbol{\theta}^{(m)}\big) - \widehat{t_M}\right\} \xrightarrow{\mathrm{P}} \sigma_t^2.$$

Monte Carlo Error

$$\mathrm{MCE}\big(\widehat{t_M}\big) := \sqrt{\frac{1}{M\,(M-1)}\sum_{m=1}^{M}\left\{t\big(\boldsymbol{\theta}^{(m)}\big) - \widehat{t_M}\right\}} = \sqrt{\frac{\widehat{\sigma}_M^2}{M}}.$$

Monte Carlo integration in statistics

Central limit theorem & Cramér-Sluckij:

$$\frac{\widehat{t_M} - \overline{t_\theta}}{\mathrm{MCE}(\widehat{t_M})} \xrightarrow{\mathcal{D}} \mathcal{N}(0,\,1) \qquad \text{as } M \to \infty.$$

➔ $100\,(1-\alpha)\%$, $0 < \alpha < 1$, confidence bounds for the approximation:

$$\widehat{t_M} \pm \mathrm{MCE}(\widehat{t_M})\, u_{1-\alpha/2},$$

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

Practice:

- $\mathcal{S}_M = \{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}\}$ 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

Special cases

$$\mathcal{S}_M \equiv \boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(M)} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{\theta} \mid \boldsymbol{y})$$

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

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

$t(\boldsymbol{\theta}) = \mathbb{I}_{[\theta_j \le x]}(\boldsymbol{\theta})$: $$\frac{\#[\theta_j^{(m)} \le x]}{M} \approx \mathrm{P}\left(\theta_j \le x \mid \boldsymbol{Y} = \boldsymbol{y}\right)$$

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

Special cases

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

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

$t(\theta) = \mathbb{I}_{[r(\theta) \le x]}(\theta)$: $$\frac{\#\left[r\left(\theta^{(m)}\right) \le x\right]}{M} \approx \mathrm{P}\left(r(\theta) \le x \,\middle|\, Y = y\right)$$

- density estimate of $p\left(r(\theta) \,\middle|\, y\right)$ (histogram, kernel, ...)
  - ➔ HPD credible intervals for $r(\theta)$;
- quantiles of $p\left(r(\theta) \,\middle|\, y\right)$ ➔ ET credible intervals for $r(\theta)$.

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

No free lunch…

How to generate

$$\mathcal{S}_M \equiv \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{\theta} \mid \boldsymbol{y})$$

???

Section **4.3**

Random numbers generation

TASK

Generate $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)} \overset{\text{i.i.d.}}{\sim} F(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$, 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.,

$$\mathsf{P}(\boldsymbol{\theta} \in \Theta) = 1, \qquad \forall \widetilde{\Theta} \subset \mathbb{R}^k \mathsf{P}(\boldsymbol{\theta} \in \Theta \setminus \widetilde{\Theta}) > 0 \quad \Longrightarrow \quad \mathsf{P}(\boldsymbol{\theta} \in \widetilde{\Theta}) < 1.$$

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 $x \in \mathbb{R}$

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

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

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

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

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

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 $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \ldots, \boldsymbol{\theta}_k^\top\right)^\top$ and

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

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

Step 1	Generate $\widetilde{\boldsymbol{\theta}}_k \sim f(\boldsymbol{\theta}_k)$.
Step 2	Generate $\widetilde{\boldsymbol{\theta}}_{k-1} \sim f\left(\boldsymbol{\theta}_{k-1} \,\middle|\, \boldsymbol{\theta}_k = \widetilde{\boldsymbol{\theta}}_k\right)$.
$\vdots$	
Step $k-1$	Generate $\widetilde{\boldsymbol{\theta}}_2 \sim f\left(\boldsymbol{\theta}_2 \,\middle|\, \boldsymbol{\theta}_3 = \widetilde{\boldsymbol{\theta}}_3, \ldots, \boldsymbol{\theta}_k = \widetilde{\boldsymbol{\theta}}_k\right)$.
Step k	Generate $\widetilde{\boldsymbol{\theta}}_1 \sim f\left(\boldsymbol{\theta}_1 \,\middle|\, \boldsymbol{\theta}_2 = \widetilde{\boldsymbol{\theta}}_2, \boldsymbol{\theta}_3 = \widetilde{\boldsymbol{\theta}}_3, \ldots, \boldsymbol{\theta}_k = \widetilde{\boldsymbol{\theta}}_k\right)$.

$$\rightarrow \qquad \widetilde{\boldsymbol{\theta}} = \left(\widetilde{\boldsymbol{\theta}}_1, \widetilde{\boldsymbol{\theta}}_2, \ldots, \widetilde{\boldsymbol{\theta}}_k\right)^\top \sim f(\boldsymbol{\theta}).$$

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.

R 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: $\boldsymbol{Y} \equiv Y_1, \ldots, Y_N$ *independent,* $Y_i \sim \mathcal{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta},\, \tau^{-1})$, *where*

- $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathbb{R}^k$ *are known constants;*
- $\boldsymbol{\beta} \in \mathbb{R}^k$, $0 < \tau < \infty$ *are unknown parameters,*

  $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top,\, \tau\right)^\top \in \Theta, \quad \Theta = \mathbb{R}^k \times (0,\, \infty);$
- *It is assumed that the* $N \times k$ *matrix*

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

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

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, μ*g):*

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

➔ *Data* $\boldsymbol{Y} = \left(Y_1, \ldots, Y_N\right)^\top$, $N = 18$.

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

Exercise: Normal linear model

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

Further notation:

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

$$\boldsymbol{b} = \boldsymbol{b}(\boldsymbol{y}) \;=\; \bigl(\mathbb{X}^\top \mathbb{X}\bigr)^{-1} \mathbb{X}^\top \boldsymbol{y} \qquad = \bigl(98.89,\, 124.42\bigr)^\top,$$

$$\mathrm{SS}_e = \mathrm{SS}_e(\boldsymbol{y}) \;=\; \mathrm{SS}(\boldsymbol{b};\, \boldsymbol{y}) \qquad = 2\,525.7.$$

The model implies:

$$p\bigl(\boldsymbol{y} \,\big|\, \boldsymbol{\theta}\bigr) \;=\; (2\,\pi)^{-N/2}\, \tau^{N/2} \exp\Bigl\{-\frac{\tau}{2}\, \mathrm{SS}(\boldsymbol{\beta};\, \boldsymbol{y})\Bigr\}$$

$$=\; (2\,\pi)^{-N/2}\, \tau^{N/2} \exp\Bigl\{-\tau\,\frac{\mathrm{SS}_e}{2} \;-\; \frac{\tau}{2}\,\bigl(\boldsymbol{\beta} - \boldsymbol{b}\bigr)^\top \mathbb{X}^\top \mathbb{X}\,\bigl(\boldsymbol{\beta} - \boldsymbol{b}\bigr)\Bigr\}, \qquad \boldsymbol{y} \in \mathbb{R}^N.$$

Exercise: Normal linear model

Jeffreys motivated improper prior

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

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

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

Posterior distribution

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

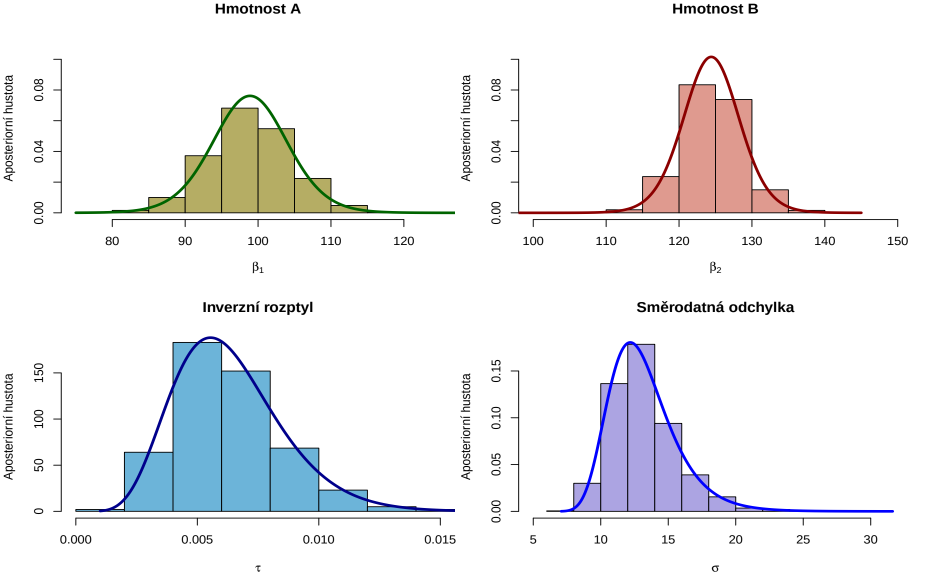
$$p(\tau \,|\, \boldsymbol{y}) \sim \mathrm{Ga}\Big(\frac{N-k}{2}, \frac{\mathrm{SS}_e}{2}\Big),$$

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

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

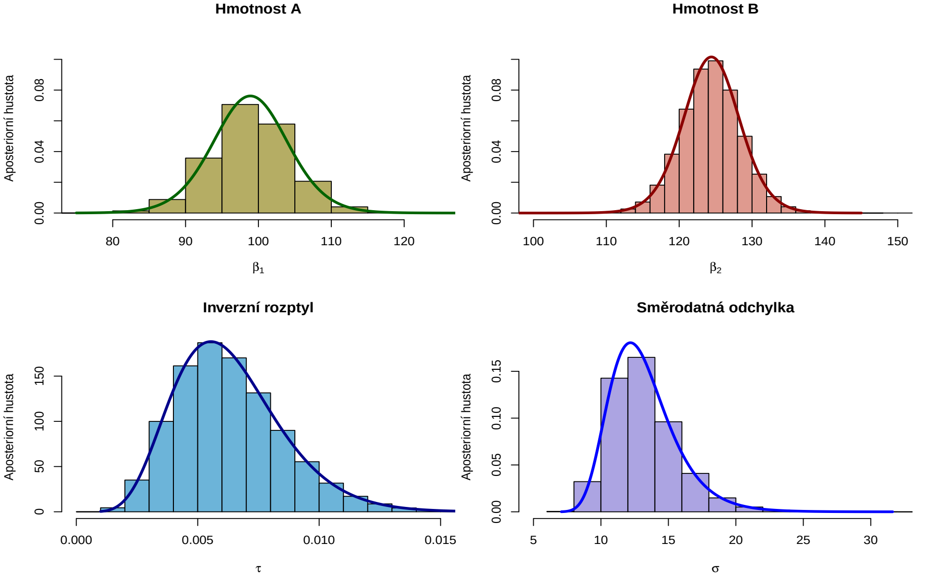
Example: Weighing of light objects

Marginal posterior densities (M=1 000)



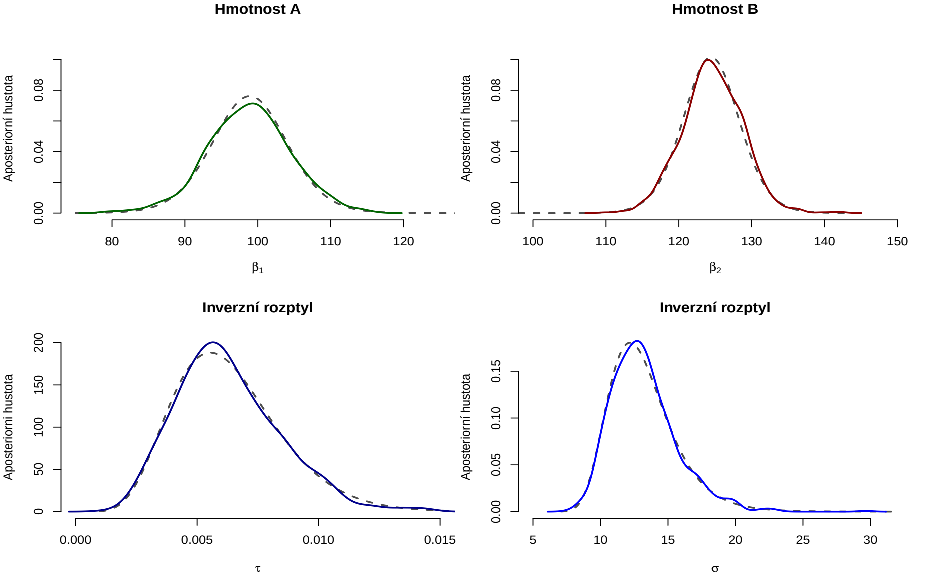
Example: Weighing of light objects

Marginal posterior densities (M=1 000 000)



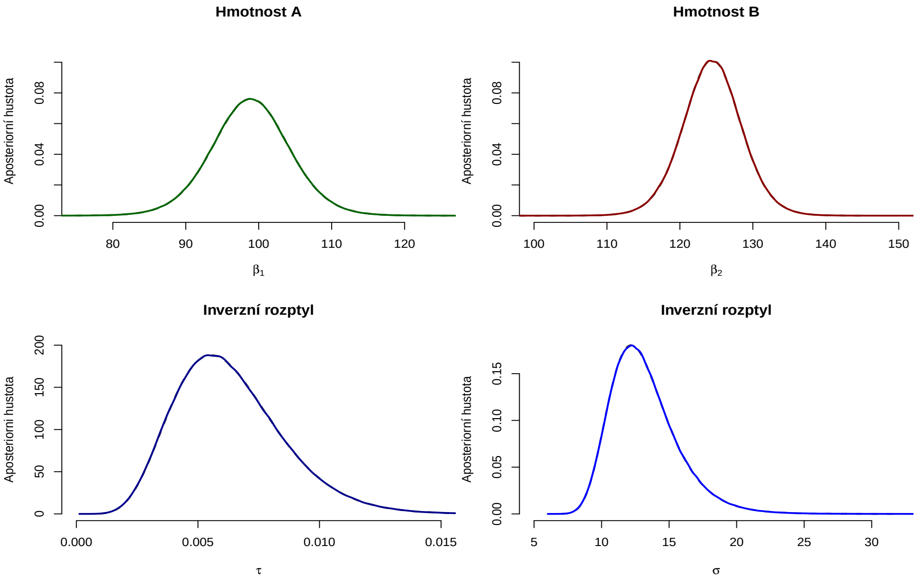
Example: Weighing of light objects

Marginal posterior densities (M=1 000)



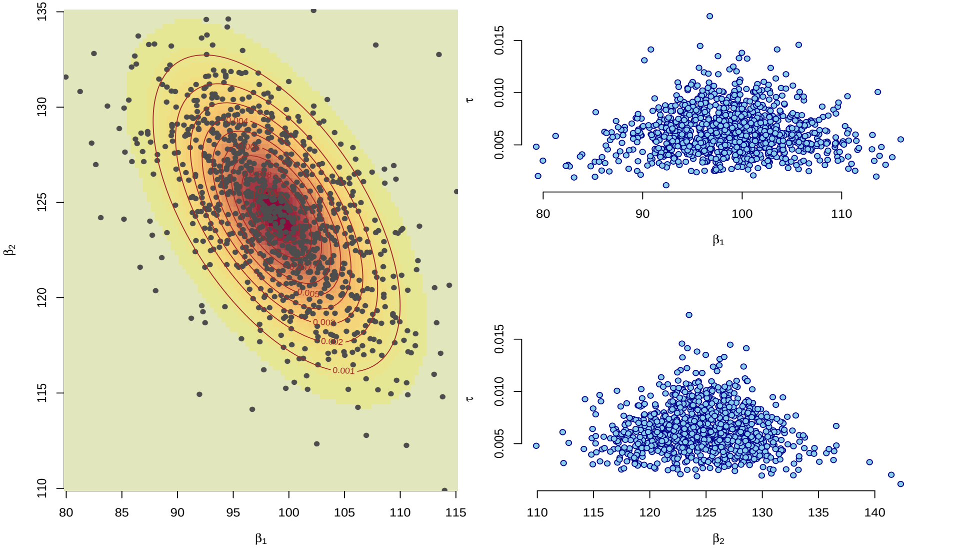
Example: Weighing of light objects

Marginální aposteriorní hustoty (M=1 000 000)



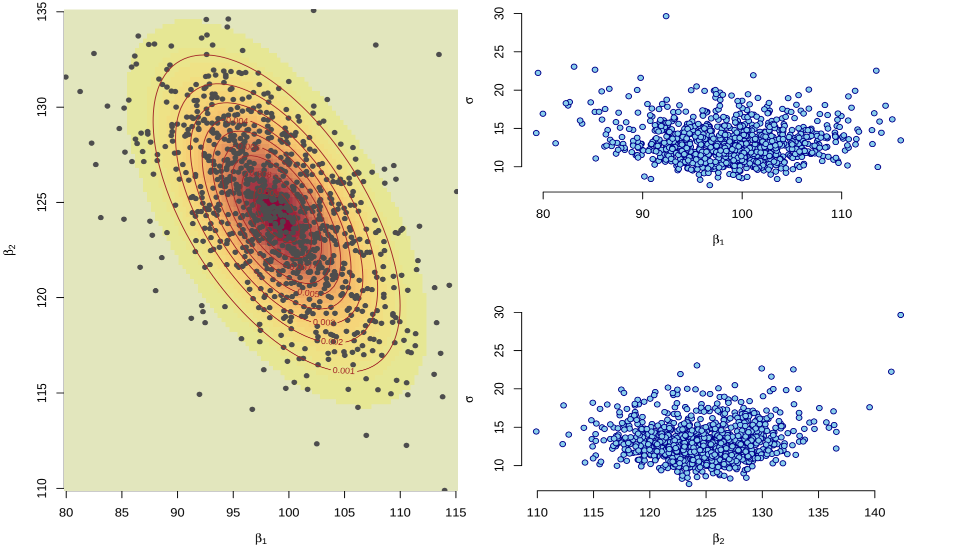
Example: Weighing of light objects

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

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

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

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

➔ Uncountable state space.

Sound theory: NMTP539 *Markov Chain Monte Carlo Methods*

➔ Here: Scetch of the theory needed to understand the principles.

Many books exist. . .

- Dani Gamerman, Hedibert F. Lopes (2006).
  Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, 2nd Edition. Boca Raton: Chapman & Hall/CRC.
- 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;

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

➔ target distribution/density,

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

Definition 5.1 Markov kernel.

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

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

- Generalization of transition probabilities.
- $\mathsf{P}(\boldsymbol{\theta}, T)$: probability of transition from the state $\boldsymbol{\theta}$ to a state in the set T.

Transition density

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

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

➔ Notation: $p(\boldsymbol{\theta}, \boldsymbol{\psi})$, $\boldsymbol{\psi} \in \Theta$,

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

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

The transition density could also be viewed as a conditional density

$$p(\boldsymbol{\theta}, \boldsymbol{\psi}) \equiv p\big(\boldsymbol{\psi} \,\big|\, \boldsymbol{\theta}\big), \qquad \boldsymbol{\theta} \in \Theta, \ \boldsymbol{\psi} \in \Theta.$$

Definition 5.2 Homogeneous Markov chain.

The random process $\{\boldsymbol{\theta}^{(m)} : m = 0, 1, \dots\}$ is the homogeneous Markov chain (homogenní markovský řetězec) with the transition kernel (*s přechodovým jádrem*) P and the initial distribution (*počátečním rozdělením*) $f_0(\boldsymbol{\theta})$ 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

$$\mathsf{P}\big(\boldsymbol{\theta}^{(0)} \in T_0, \boldsymbol{\theta}^{(1)} \in T_1, \dots, \boldsymbol{\theta}^{(m)} \in T_m\big)$$

$$= \int_{T_0} \int_{T_1} \cdots \int_{T_{m-1}} \mathsf{P}(\boldsymbol{\theta}^{(m-1)}, T_m)\, p(\boldsymbol{\theta}^{(m-2)}, \boldsymbol{\theta}^{(m-1)}) \mathrm{d}\boldsymbol{\theta}^{(m-1)} \cdots p(\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}) \mathrm{d}\boldsymbol{\theta}^{(1)}\, f_0(\boldsymbol{\theta}^{(0)}) \mathrm{d}\boldsymbol{\theta}^{0}$$

$$= \int_{T_0} \int_{T_1} \cdots \int_{T_{m-1}} \int_{T_m} p(\boldsymbol{\theta}^{(m-1)}, \boldsymbol{\theta}^{(m)}) \mathrm{d}\boldsymbol{\theta}^{(m)}\, p(\boldsymbol{\theta}^{(m-2)}, \boldsymbol{\theta}^{(m-1)}) \mathrm{d}\boldsymbol{\theta}^{(m-1)} \cdots p(\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}) \mathrm{d}\boldsymbol{\theta}^{(1)}\, f_0(\boldsymbol{\theta}^{(0)}) \mathrm{d}\boldsymbol{\theta}^{0}.$$

Properties

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

$$\mathbb{E}\Big[h\big(\boldsymbol{\theta}^{(m+1)}\big)\,\Big|\,\boldsymbol{\theta}^{(m)}, \ldots, \boldsymbol{\theta}^{(0)}\Big] \;=\; \mathbb{E}\Big[h\big(\boldsymbol{\theta}^{(m+1)}\big)\,\Big|\,\boldsymbol{\theta}^{(m)}\Big].$$

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

$$\mathsf{P}\big(\boldsymbol{\theta}^{(m+1)} \in T \,\big|\, \boldsymbol{\theta}^{(m)}, \ldots, \boldsymbol{\theta}^{(0)}\big) \;=\; \mathsf{P}\big(\boldsymbol{\theta}^{(m+1)} \in T \,\big|\, \boldsymbol{\theta}^{(m)}\big),$$

markovian property.

Homogeneous Markov chains

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

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

➔ distribution/density of a state $\boldsymbol{\theta}$ at a certain time point m.

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

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

While using the transition density, we can also write

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

Homogeneous Markov chains

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

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

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

$$= \int_T \int_\Theta p(\boldsymbol{\theta}, \boldsymbol{\psi})\pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \, \mathrm{d}\boldsymbol{\psi}.$$

➔ $\pi \mathrm{P}(\cdot)$ is again the probability distribution on $(\Theta, \mathcal{T})$ with the density

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

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

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 $\pi(\cdot) = \pi\mathsf{P}(\cdot)$, that is if for any $T \in \mathcal{T}$

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

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

$$\int_T \mathsf{P}(\boldsymbol{\theta}, S)\, \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} = \int_S \mathsf{P}(\boldsymbol{\psi}, T)\, \pi(\boldsymbol{\psi})\, \mathrm{d}\boldsymbol{\psi}$$

$$\int_T \int_S p(\boldsymbol{\theta}, \boldsymbol{\psi})\, \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\psi}\, \mathrm{d}\boldsymbol{\theta} = \int_S \int_T p(\boldsymbol{\psi}, \boldsymbol{\theta})\, \pi(\boldsymbol{\psi})\, \mathrm{d}\boldsymbol{\theta}\, \mathrm{d}\boldsymbol{\psi}.$$

Homogeneous Markov chains

Reversibility

$\displaystyle\int_T \mathrm{P}(\boldsymbol{\theta},\, S)\, \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$ can be viewed as a joint probability distribution

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

$$Q_1(T,\, S) \;=\; \int_T \mathrm{P}(\boldsymbol{\theta},\, S)\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \;=\; \int_T \int_S p(\boldsymbol{\theta},\, \boldsymbol{\psi})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\psi}\mathrm{d}\boldsymbol{\theta},$$

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

$$q_1(\boldsymbol{\theta},\, \boldsymbol{\psi}) = p(\boldsymbol{\theta},\, \boldsymbol{\psi})\pi(\boldsymbol{\theta}), \qquad \boldsymbol{\theta},\, \boldsymbol{\psi} \in \Theta.$$

$\displaystyle\int_S \mathrm{P}(\boldsymbol{\psi},\, T)\, \pi(\boldsymbol{\psi})\, \mathrm{d}\boldsymbol{\psi}$ can be viewed as a joint probability distribution

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

$$Q_2(S,\, T) \;=\; \int_S \mathrm{P}(\boldsymbol{\psi},\, T)\pi(\boldsymbol{\psi})\mathrm{d}\boldsymbol{\psi} \;=\; \int_S \int_T p(\boldsymbol{\psi},\, \boldsymbol{\theta})\pi(\boldsymbol{\psi})\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\psi},$$

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

$$q_2(\boldsymbol{\psi},\, \boldsymbol{\theta}) = p(\boldsymbol{\psi},\, \boldsymbol{\theta})\pi(\boldsymbol{\psi}), \qquad \boldsymbol{\psi},\, \boldsymbol{\theta} \in \Theta.$$

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.

Reversibility

Detailed balance condition (*detailní podmínka rovnováhy*)

for a transition density:

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

It (clearly) implies reversibility:

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

Reversibility implies **stationarity**

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

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

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

$$\pi(T) \;=\; \pi\mathrm{P}(T).$$

detailed balance condition ➠ reversibility ➠ stationarity

with respect to the distribution π

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

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 P is given inductively as

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

Chapman-Kolmogorov equality

For any $n \le m$

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

Definition 5.6 Limitting distribution.

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

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

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

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

Limitting distribution must be stationary

If π is the limitting 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 $\boldsymbol{\theta} \in \Theta$

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

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

TO REMEBER

detailed balance condition ➠ reversibility ➠ stationarity

with respect to the distribution π

π is limitting ➠ π is stationary

There is no guarantee that the limitting distribution exists.

Section **5.3**

MCMC principles

Principy MCMC

Připomenutí, čím se zabýváme

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

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

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

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

➠ Monte Carlo integrace

Principy MCMC

- Nechť $\{\boldsymbol{\theta}^{(m)} : m = 0, \ldots\}$ je homogenní markovský řetězec se stacionárním rozdělením $f(\boldsymbol{\theta})$.
  - Víme: reversibilita vzhledem k $f(\boldsymbol{\theta})$ implikuje stacionaritu vzhledem k $f(\boldsymbol{\theta})$.

- Stačí tedy zvolit přechodové jádro markovského řetězce tak, aby přechodová hustota $p(\boldsymbol{\theta}, \boldsymbol{\psi})$ splňovala detailní podmínku rovnováhy vzhledem k $f(\boldsymbol{\theta})$.

- Stačí tedy volit přechodovou hustotu tak, aby splňovala

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

a máme potřebný markovský řetězec.

➠ Toto není nikterak obtížné, jak bude záhy ukázáno.

Principy MCMC

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

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

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

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

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

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

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

Principy MCMC

Ergodická věta

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

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

Největší obtíž při praktické aplikaci MCMC

- Zjistit, od kterého okamžiku již lze (s přiměřeně malou chybou) považovat rozdělení stavů vygenerovaného markovského řetězce za limitní $=$ stacionární $f(\theta)$.
  - 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 $\boldsymbol{\theta}^{(B+1)}, \ldots, \boldsymbol{\theta}^{(B+M)}$ nejsou obecně nezávislé a tudíž (i při předpokladu konvergence k limitnímu rozdělení) není nutně pravda

$$v_M = \mathrm{var}_{f(\boldsymbol{\theta})}\big(\widehat{t}_M\big) = \frac{\mathrm{var}_{f(\boldsymbol{\theta})}\big(t(\boldsymbol{\theta})\big)}{M}.$$

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

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

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

Největší obtíže

- Snaha konstruovat markovský řetězec tak, aby autokorelace byla co nejnižší.
  - Nulová autokorelace $\equiv$ co do přesnosti se markovský řetězec chová stejně jako náhodný výběr, kde $v_M = \dfrac{1}{M}\,\mathrm{var}_{f(\boldsymbol{\theta})}\big(t(\boldsymbol{\theta})\big)$.
- 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:

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

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

Algoritmus:

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

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

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

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

Přechodová hustota

$$p(\boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^{k} f(\psi_i \,|\, \psi_1, \ldots, \psi_{i-1}, \boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_k).$$

- Odpovídá přechodovému jádru

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

Stacionární rozdělení

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
  - $\Theta = \prod_{i=1}^{k} \Theta_i$, $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_k^\top\right)^\top$.
  - Cílové (stacionární) rozdělení má hustotu $f(\boldsymbol{\theta})$ vzhledem k součinové míře $\lambda_1 \otimes \cdots \otimes \lambda_k$, přičemž λ_i je σ-konečná míra s $\lambda_i(\Theta_i) > 0$ ($i = 1, \ldots, k$).
  - $\Theta = \{\boldsymbol{\theta} : f(\boldsymbol{\theta}) > 0\}$.
- 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 nesplňuje detailní podmínku rovnováhy, tj. řetězec není reversibilní vzhledem k rozdělení f.
- Reversibility lze dosáhnout několika způsoby:
  - Generujeme střídavě odpředu a odzadu.
  - Pořadí vybíráme náhodně.
    - V každém podkroku Gibbsova algoritmu generujeme i-tý podvektor s pravděpodobností p_i ($0 < p_i < 1, \sum_{i=1}^{k} p_i = 1$).
    - Častá volba je $p_i = 1/k$ (rovnoměrné rozdělení).
    - *Random scan* Gibbsův algoritmus.

- V principu lze generovat ze všech jednorozměrných podmíněných rozdělení.
  - V případě, že složky $\boldsymbol{\theta}$ jsou v cílovém rozdělení $f(\boldsymbol{\theta})$ významně korelovány, vede generování z jednorozměrných podmíněných rozdělení k markovskému řetězci s velkou autokorelací.
  - Ideální situace je stav, kdy podvektory $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ jsou v cílovém rozdělení $f(\boldsymbol{\theta})$ co možná nejméně korelovány.

Gibbsův algoritmus

Plně podmíněná rozdělení

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

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

  přičemž $\propto$ nyní znamená, že vše, co neobsahuje $\boldsymbol{\theta}_i$ je konstantou.

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

Příklad: Lineární model s neinformativním apriorním rozdělením

Lineární model

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

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

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

Příklad: Lineární model s neinformativním apriorním rozdělením

Věrohodnost

- Označme: $\boldsymbol{b} = \bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\mathbb{X}^\top\boldsymbol{y}$,
  $\mathrm{SS}_e = \bigl\|\boldsymbol{y} - \mathbb{X}\boldsymbol{b}\bigr\|^2$.
- Věrohodnost:

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

Příklad: Lineární model s neinformativním apriorním rozdělením

Aposteriorní rozdělení

- Bylo odvozeno:

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

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

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

- Dále bylo odvozeno: $p(\boldsymbol{\beta} \mid \boldsymbol{y}) \sim \text{MVT}_{k,n-k}\left(\boldsymbol{b}, \frac{SS_e}{n-k}(\mathbb{X}^\top\mathbb{X})^{-1}\right)$.

- Pomocí Gibbsova algoritmu sestrojíme markovský řetězec, který bude mít rozdělení $p(\boldsymbol{\beta}, \tau \mid \boldsymbol{y})$ jako stacionární i limitní.

Příklad: Lineární model s neinformativním apriorním rozdělením

Plně podmíněná rozdělení

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

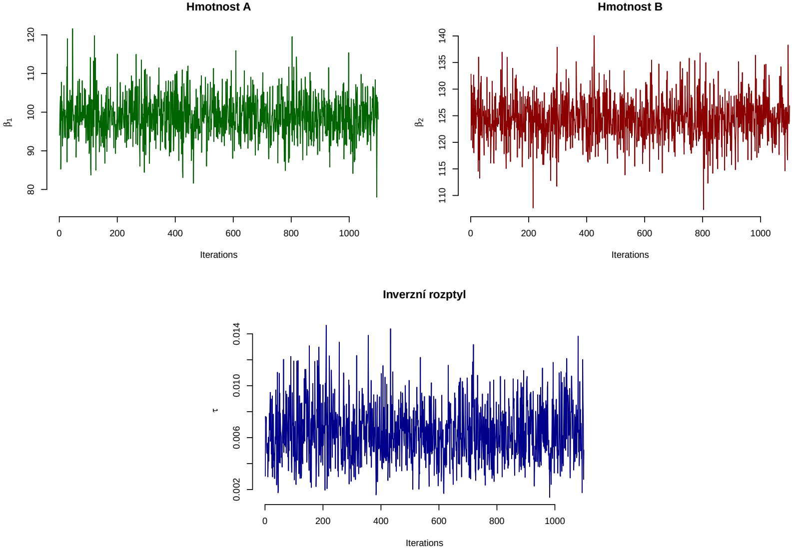
$$p(\boldsymbol{\beta} \,|\, \cdots) = p(\boldsymbol{\beta} \,|\, \tau,\, \boldsymbol{y}) \sim \mathcal{N}_k\Bigl(\boldsymbol{b},\, \tau^{-1}\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\Bigr),$$

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

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

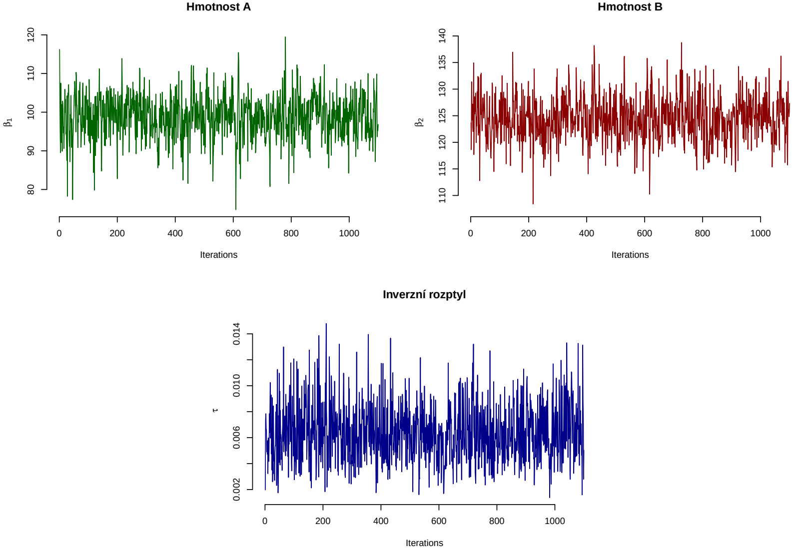
Příklad: Vážení lehkých objektů

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



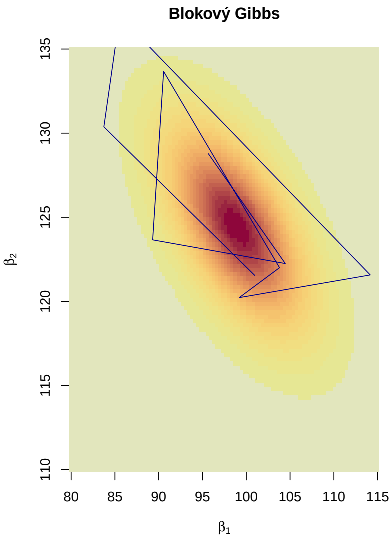
Příklad: Vážení lehkých objektů

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



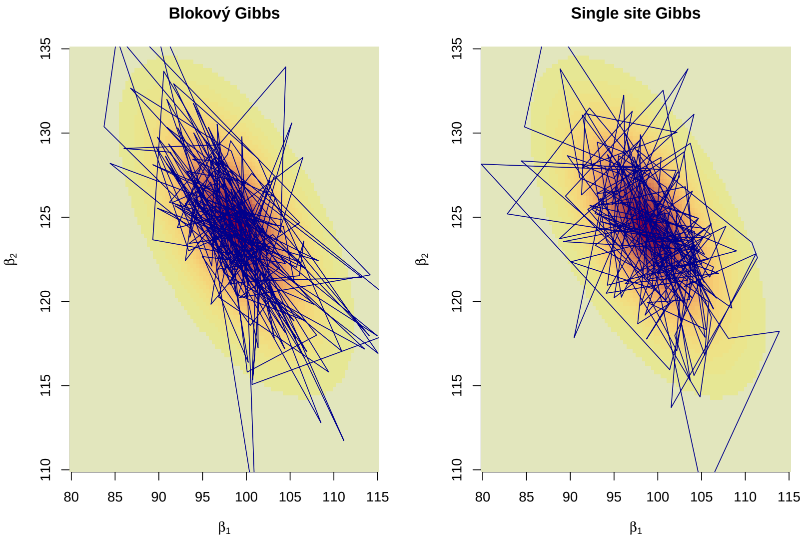
Příklad: Vážení lehkých objektů

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



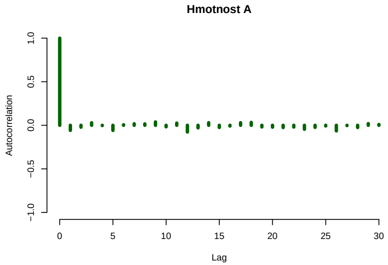
Příklad: Vážení lehkých objektů

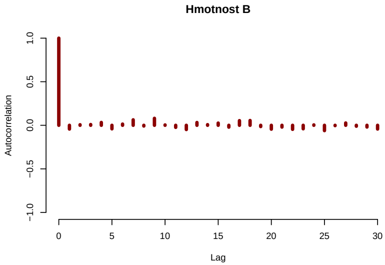
Gibbsův algoritmus: Generované hodnoty β (iterace 101 – 300)

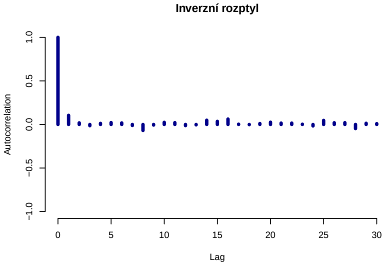


Příklad: Vážení lehkých objektů

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

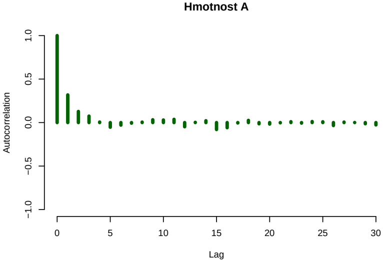


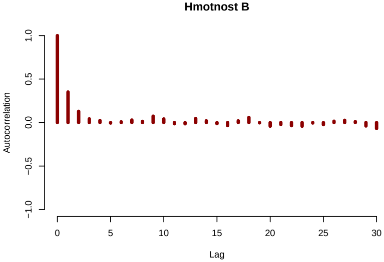


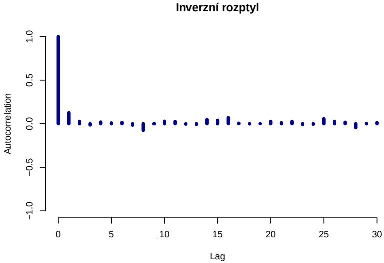


Příklad: Vážení lehkých objektů

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

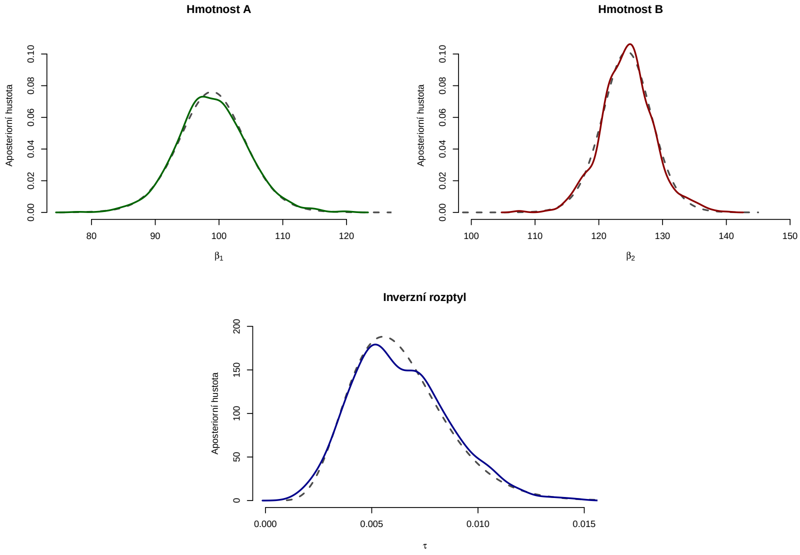




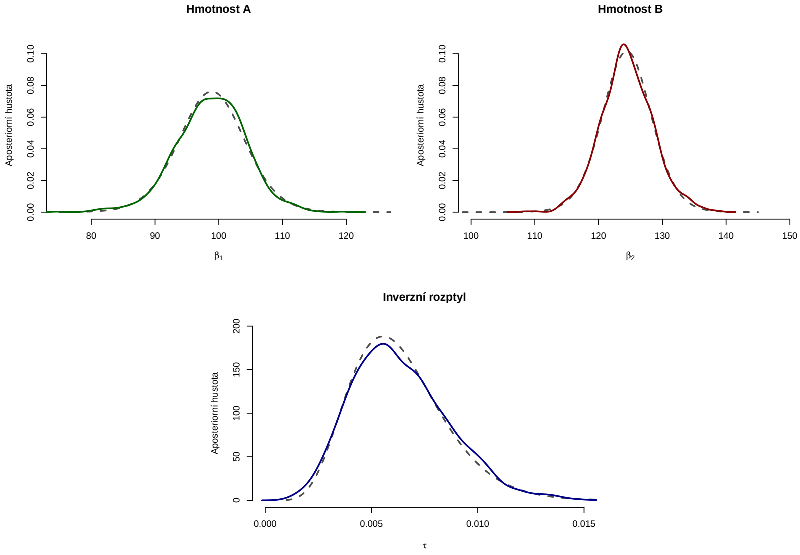


Příklad: Vážení lehkých objektů

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



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



Příklad: Vážení lehkých objektů

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

Blokový Gibbsův algoritmus

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

Příklad: Vážení lehkých objektů

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

Single site Gibbsův algoritmus

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

Section **5.5**

Metropolis-Hastings algorithm

Metropolisův-Hastingsův algoritmus

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

Předpoklady:

- Parametrický prostor Θ

- Cílové (stacionární) rozdělení má hustotu $f(\boldsymbol{\theta})$ vzhledem k σ-konečné míře λ s $\lambda(\Theta) > 0$.

  ZDE: Lebesgueova míra na $\left(\mathbb{R}^d, \mathcal{B}^d\right)$, resp. na $\left(\Theta, \mathcal{B}(\Theta)\right)$.

- $\Theta = \left\{\boldsymbol{\theta} : f(\boldsymbol{\theta}) > 0\right\}$

Metropolisův-Hastingsův algoritmus

Algoritmus:

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

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

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

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

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

Poznámky:

- Pro aplikaci MH algoritmu není potřeba znát normující konstantu cílové hustoty $f(\boldsymbol{\theta})$.
    - Ideální pro použití v bayesovské statistice.

- Návrhová hustota $q(\boldsymbol{\theta}, \boldsymbol{\psi})$ může být libovolná.
    - Nevhodná volba $q(\boldsymbol{\theta}, \boldsymbol{\psi})$ však vede k vysoké autokorelaci a s tím spojené neefektivitě.
    - Příliš “ambiciózní” $q(\boldsymbol{\theta}, \boldsymbol{\psi})$ vede k malým pravděpodobnostem přijetí návrhu a řetězec pak dlouho setrvává v jednom stavu

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

      ➡ vysoká autokorelace.

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

Metropolisův-Hastingsův algoritmus

Poznámky

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

- Hlavní část pravděpodobnosti přijetí

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

obvykle počítáme v logaritmickém měřítku, tj.

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

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

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

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

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

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

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

Návrhové rozdělení

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

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

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

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

➠ Potřeba vhodně zvolit rozptyly.

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

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

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

  (symetrická náhodná procházka)

  a $q_0(\boldsymbol{z}) > 0$ pro všechna $\mathbb{R}^d$ (např. vícerozměrné normální nebo t-rozdělení).
- Podrobnosti viz přednáška NMTP539.

Hybridní algoritmy

Některé další metody pro generování z plně podmíněných rozdělení

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

Příklad: Lineární model s neinformativním apriorním rozdělením

Lineární model

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

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

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

Příklad: Lineární model s neinformativním apriorním rozdělením

Věrohodnost

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

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

- Věrohodnost:

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

Příklad: Lineární model s neinformativním apriorním rozdělením

Aposteriorní rozdělení

- Bylo odvozeno:

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

kde $\quad p(\tau\,|\,\boldsymbol{y}) \sim \mathcal{G}\Bigl(\dfrac{n-k}{2},\,\dfrac{\mathsf{SS}_e}{2}\Bigr),$

$\quad p(\beta\,|\,\tau,\,\boldsymbol{y}) \sim \mathcal{N}_k\Bigl(\boldsymbol{b},\,\tau^{-1}\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\Bigr).$

- Dále bylo odvozeno: $p(\beta\,|\,\boldsymbol{y}) \sim \mathsf{MVT}_{k,\,n-k}\Bigl(\boldsymbol{b},\,\dfrac{\mathsf{SS}_e}{n-k}\bigl(\mathbb{X}^\top\mathbb{X}\bigr)^{-1}\Bigr)$

- Pomocí algoritmu Metropolis within Gibbs algoritmu sestrojíme markovský řetězec, který bude mít rozdělení $p(\beta,\,\tau\,|\,\boldsymbol{y})$ jako stacionární i limitní.

Příklad: Lineární model s neinformativním apriorním rozdělením

Plně podmíněná rozdělení

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

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

Příklad: Lineární model s neinformativním apriorním rozdělením

Metropolis within Gibbs algoritmus

- Regresní parametry $\boldsymbol{\beta}$ budeme generovat pomocí symetrické náhodné procházky s návrhovou hustotou
$$q(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = q_0(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1),$$
$$\text{kde } q_0 \sim \mathcal{N}_k(\mathbf{0}, \mathbb{D}_{prop}), \qquad \mathbb{D}_{prop} = \mathrm{diag}(d^2_{1,prop}, \ldots, d^2_{p,prop}).$$
  - V kroku $m+1$ algoritmu navrhneme $\boldsymbol{\beta}_{prop}$ vygenerované z rozdělení $\mathcal{N}_k(\boldsymbol{\beta}^{(m)}, \mathbb{D}_{prop})$.
- Hlavní část pravděpodobnosti přijetetí návrhu je

$$\alpha^*(\boldsymbol{\beta}^{(m)}, \boldsymbol{\beta}_{prop}) = \frac{p(\boldsymbol{\beta}_{prop} \mid \cdots)\, q(\boldsymbol{\beta}_{prop}, \boldsymbol{\beta}^{(m)})}{p(\boldsymbol{\beta}^{(m)} \mid \cdots)\, q(\boldsymbol{\beta}^{(m)}, \boldsymbol{\beta}_{prop})} = \frac{p(\boldsymbol{\beta}_{prop} \mid \cdots)}{p(\boldsymbol{\beta}^{(m)} \mid \cdots)}$$
$$= \exp\left[-\frac{\tau^{(m)}}{2}\left\{(\boldsymbol{\beta}_{prop} - \boldsymbol{b})^\top \mathbb{X}^\top \mathbb{X}(\boldsymbol{\beta}_{prop} - \boldsymbol{b}) - (\boldsymbol{\beta}^{(m)} - \boldsymbol{b})^\top \mathbb{X}^\top \mathbb{X}(\boldsymbol{\beta}^{(m)} - \boldsymbol{b})\right\}\right].$$

Příklad: Vážení lehkých objektů

Metropolis within Gibbs algoritmus

- Budou čtyři ukázky vygenerované při použití různých varianční matic v návrhové normální hustotě.
  1. $\mathbb{D}_{prop} = \text{diag}(1, 1)$

     ➠ proporce přijatých návrhů 0,87.
  2. $\mathbb{D}_{prop} = \text{diag}(5^2, 5^2)$

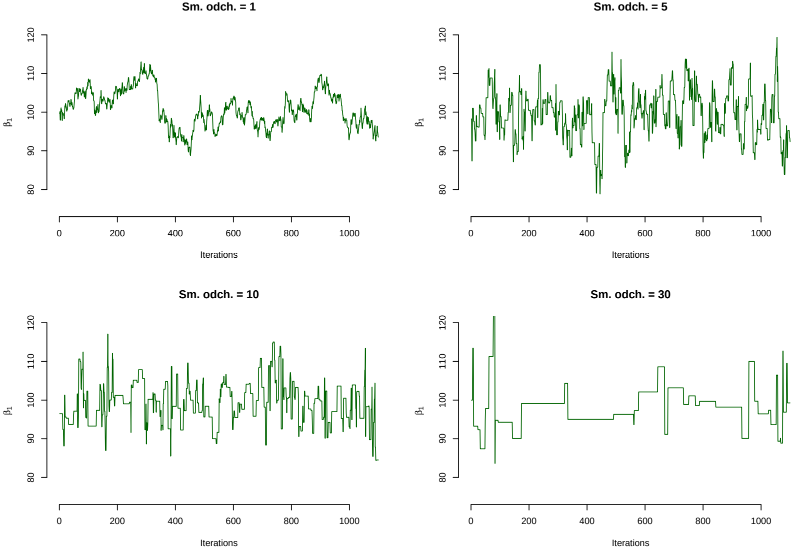
     ➠ proporce přijatých návrhů 0,47.
  3. $\mathbb{D}_{prop} = \text{diag}(10^2, 10^2)$

     ➠ proporce přijatých návrhů 0,22.
  4. $\mathbb{D}_{prop} = \text{diag}(30^2, 30^2)$

     ➠ proporce přijatých návrhů 0,04.

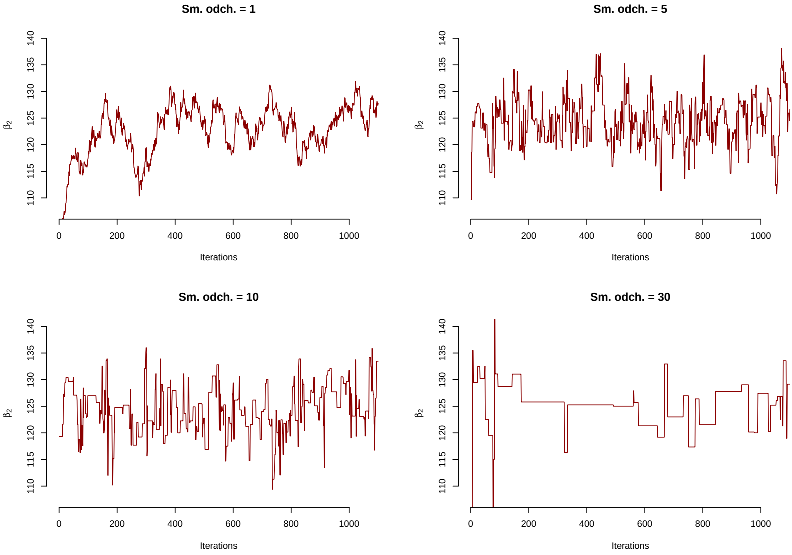
Příklad: Vážení lehkých objektů

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



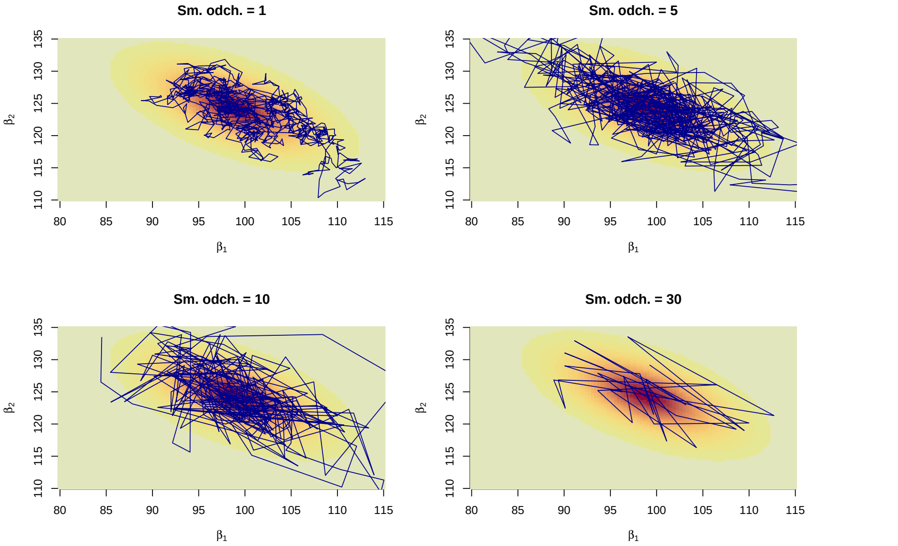
Příklad: Vážení lehkých objektů

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



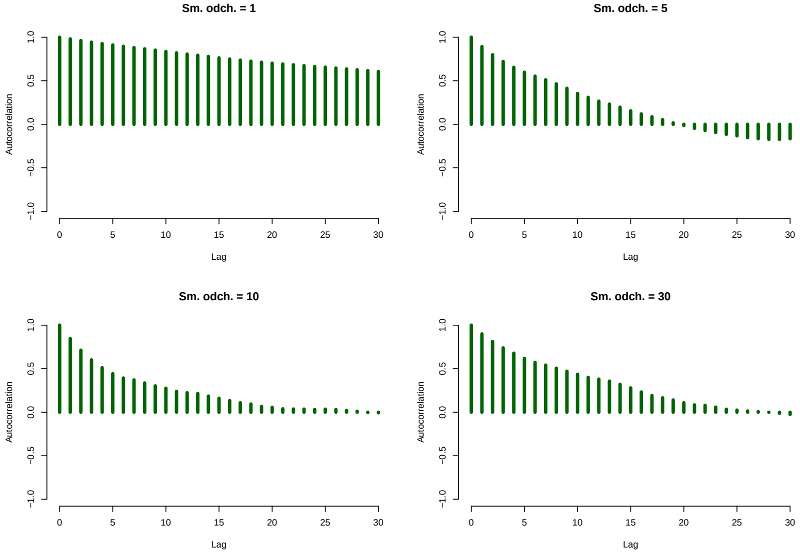
Příklad: Vážení lehkých objektů

Metropolis within Gibbs algoritmus: Generované hodnoty $\boldsymbol{\beta}$ (iterace 101 – 1 100)



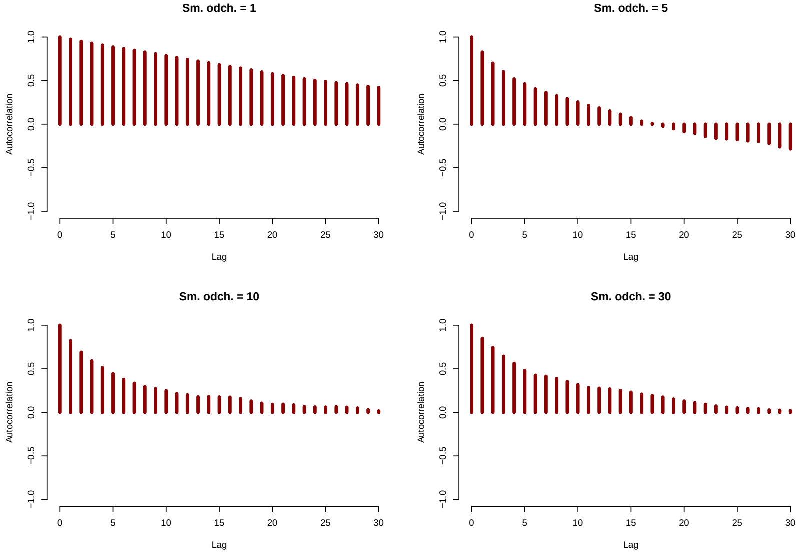
Příklad: Vážení lehkých objektů

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



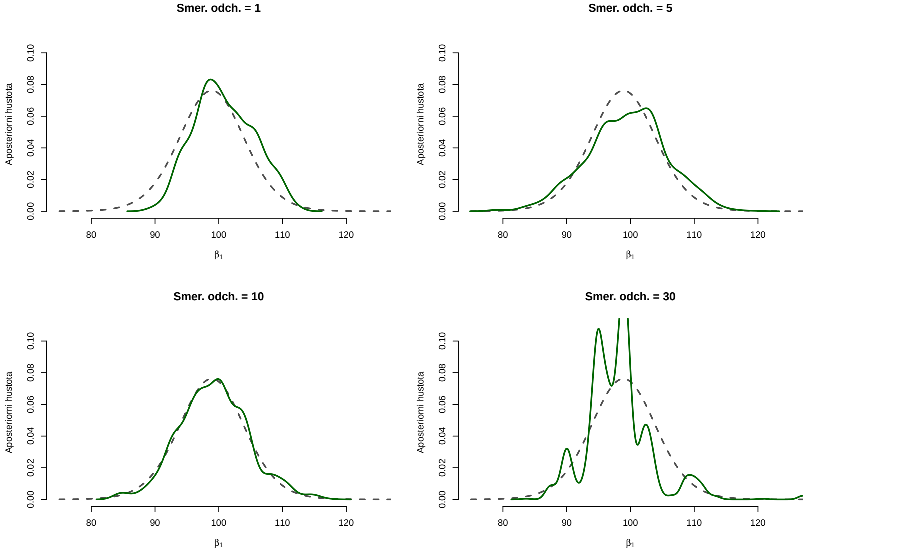
Příklad: Vážení lehkých objektů

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



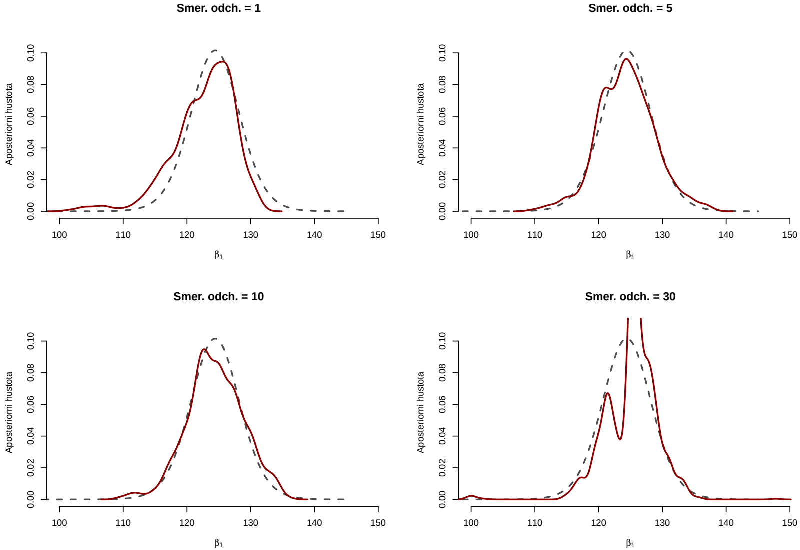
Příklad: Vážení lehkých objektů

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



Příklad: Vážení lehkých objektů

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



6

Hierarchial models

Section **6.1**

Hierarchical prior

Apriorní rozdělení

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

Apriorní rozdělení

- Pouze málokdy je apriorní informace dostatečně bohatá na to, abychom mohli zvolené apriorní rozdělení považovat za přesně a bez jakékoliv chyby definované.

- Potřeba vhodným způsobem vyjádřit nejistotu při volbě apriorního rozdělení.

➠ Bayesovský model s hierarchicky specifikovaným apriorním rozdělením

  - rozklad apriorního rozdělení do několika úrovní podmíněných rozdělení,

  - nejistota na libovolné úrovni je vyjádřena apriorním rozdělením v další úrovni.

Bayesovský model s hierarchicky specifikovaným apriorním rozdělením

Definition 6.1 Bayesovský model s hierarchicky specifikovaným apriorním rozdělením.

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

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

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

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

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

Příklad: Lineární model

Model

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

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

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

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

Příklad: Lineární model

- Volba parametrů gama apriorního rozdělení pro τ (resp. samotná volba gama rozdělení) může mít významný vliv na výsledné aposteriorní rozdělení.

- Nepovažujme c_0 a/nebo d_0 za pevné konstanty, ale umožněme náhodnost při jejich výběru.

  ➡ hierarchický model

  - například:

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

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

Section **6.2**

Hierarchical likelihood

Hierarchicky specifikovaná věrohodnost

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

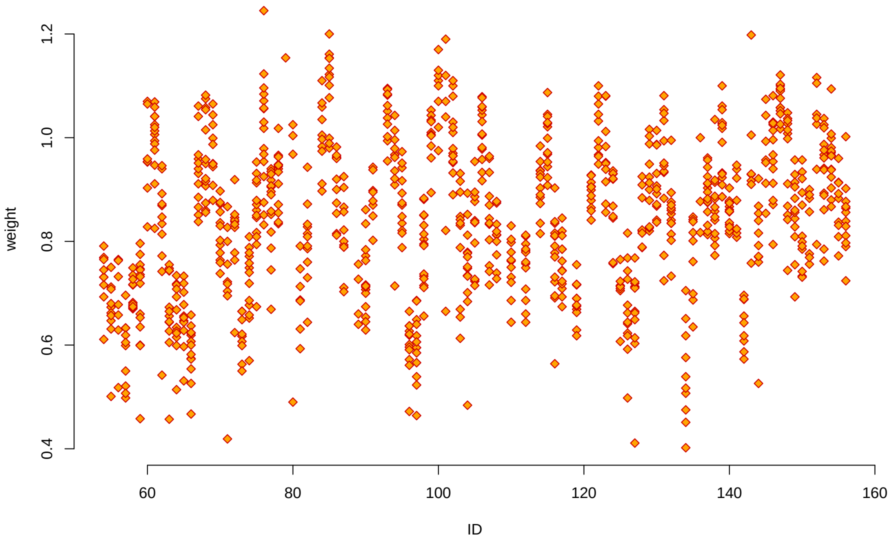
Příklad: NTP TER84073 pokus na myších

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

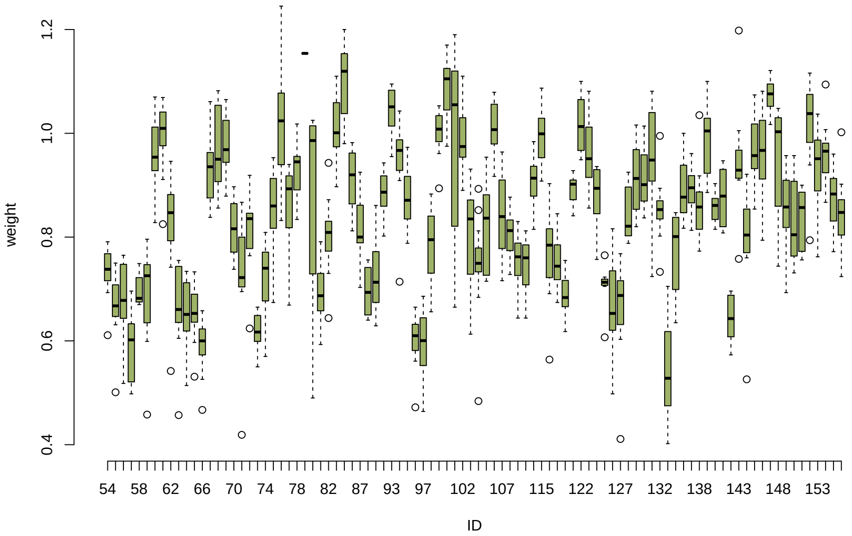
  $Y_{i,j} =$ hmotnost j-tého zárodku i-té myši, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$
- Primární cíl:

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

Příklad: NTP TER84073 pokus na myších



Příklad: NTP TER84073 pokus na myších



Příklad: NTP TER84073 pokus na myších

- Možný pravděpodobnostní model:

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

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

Příklad: NTP TER84073 pokus na myších

- Realističtější pravděpodobnostní model:

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

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

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

Příklad: NTP TER84073 pokus na myších

- Realističtější pravděpodobnostní model:

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

- Modelujeme též jistým způsobem korelaci mezi zárodky jedné myši:

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

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

➠ vnitroskupinová korelace (*intraclass correlation*)

Delší další povídání

Hierarchické modely

Obecná poznámka

- Též v jiných modelech lze často rozlišit tři typy parametrů (v bayesovském smyslu):
  - skrytá data

    ➠ v dalším budeme obvykle značit $\boldsymbol{\xi}$
  - "čisté" parametry $\equiv$ parametry též ve frekventistickém pojetí

    ➠ v dalším budeme obvykle značit $\boldsymbol{\psi}$
  - náhodné hyperparametry

    ➠ v dalším budeme obvykle značit $\boldsymbol{\zeta}$

- Parametry pro bayesovský model jsou potom $\boldsymbol{\theta} = \left(\boldsymbol{\xi}^\top, \boldsymbol{\psi}^\top, \boldsymbol{\zeta}^\top\right)^\top$.

Obecná poznámka, pokračování

- Sdružené apriorní rozdělení je zadáno rozkladem

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

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

    ➡ plyne z uvažovaného pravděpdobnostního modelu použitého pro popis situace

  - $p(\psi \mid \zeta)\, p(\zeta)$: “skutečné” apriorní rozdělení

"ANOVA" hierarchický model

Konkrétní apriorní rozdělení (jedna z možností)

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

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

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

"ANOVA" hierarchický model

Apriorní rozdělení (jedna z možností)

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

Nenáhodné (pevné) (hyper)parametry:

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

➠ Jak je volit?

“ANOVA” hierarchický model

Volba nenáhodných (hyper)parametrů

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

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

“ANOVA” hierarchický model

(Částečně) datově motivovaná volba pevných hyperparametrů

- μ_0 : apriorní střední hodnota pro $\mathbb{E}Y_{i,j} = \mathbb{E}b_i = \mu$
  - $\mu_0 \approx \overline{y}$

- k_0 : ovlivňuje apriorní inverzní rozptyl (přesnost) pro μ
  - k_0 blízké 0

- a_0, c_0, p_0, g_0 : “stupně volnosti” gama rozdělení
  - obvykle mezi 0 a 1

- r_0, h_0 : “rate” parametr gama rozdělení v poslední hierarchické úrovni
  - obvykle se volí blízké 0

“ANOVA” hierarchický model

Aposteriorní rozdělení

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

  ➠ Jak?

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

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

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

7

(Generalized) linear mixed models

Section **7.1**

Linear mixed model

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

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

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

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

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

Normal linear mixed model, hierarchically specified

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

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

- $\begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{b}_1 \end{pmatrix}, \ldots, \begin{pmatrix} \boldsymbol{Y}_N \\ \boldsymbol{b}_N \end{pmatrix}$ mutually independent

Příklad 1: NTP TER84073 pokus na myších

Normální lineární smíšený model

- Pravděpodobnostní reprezentace dat:

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

- Normální LMM:

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

- V obecném značení

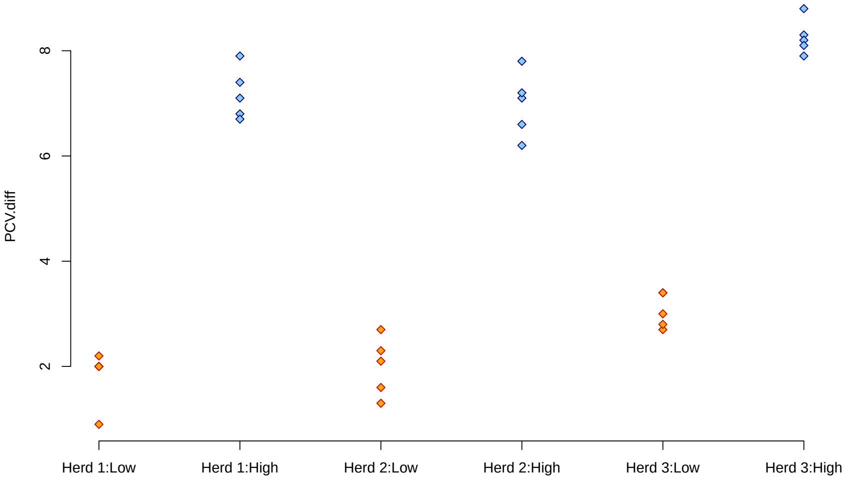
$$\mathbb{X}_i \text{ není}, \quad \mathbb{Z}_i = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}_i = \sigma^2\, \mathbf{I}_{n_i}, \quad \mathbb{D} = d^2$$

Příklad 2: Vliv dávky Berenilu na léčbu trypanosomosis

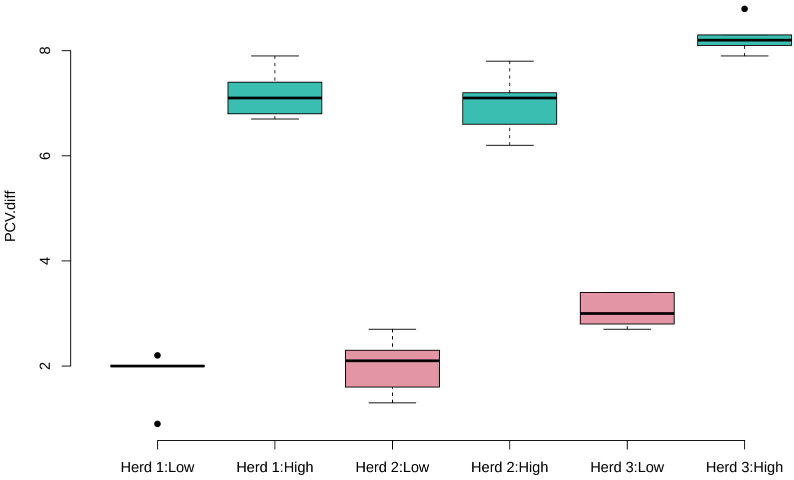
- Experiment mající za úkol vyhodnotit vliv dávky léku Berenil na úspěšnost léčby trypanosomosis u krav
- Úspěšnost léčby je měřena pomocí přírůstku PCV (packed cell volume) po určité době podávání léku
- Zajímá nás, zda existuje rozdíl v úspěšnosti léčby nízkou či vysokou dávkou Berenilu
- Experiment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

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

Příklad 2: Vliv dávky Berenilu na léčbu trypanosomosis



Příklad 2: Vliv dávky Berenilu na léčbu trypanosomosis



Příklad 2: Vliv dávky Berenilu na léčbu trypanosomosis

Lineární smíšený model

1. Dávka má stejný účinek ve všech stádech.

$$Y_{i,j} \;=\; b_i + \mathbb{I}[\text{dávka}_{i,j} = \text{vysoká}]\beta + \varepsilon_{i,j}$$

2. Dávka nemá nutně stejný účinek ve všech stádech.

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

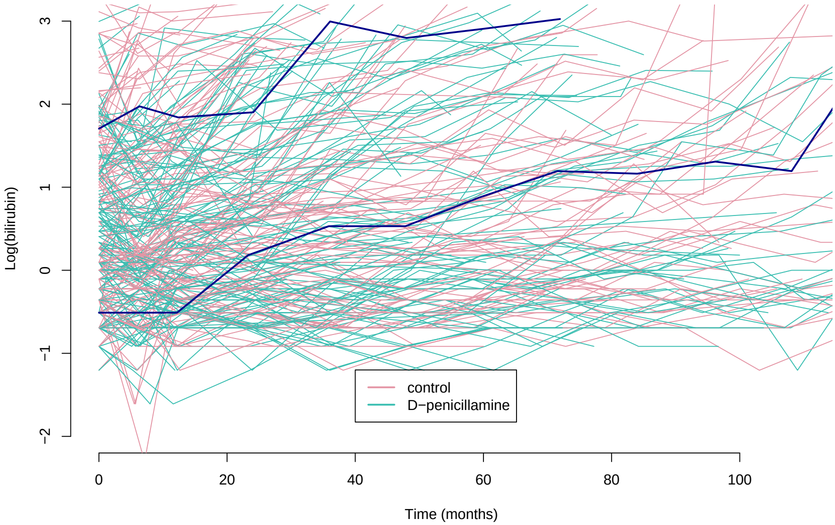
Příklad 3: Vývoj hladiny bilirubinu u pacientů s PBC

Longitudinální studie

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

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

Příklad 3: Vývoj hladiny bilirubinu u pacientů s PBC



Příklad 3: Vývoj hladiny bilirubinu u pacientů s PBC

Možný lineární smíšený model

- Možný model pro jednoho pacienta: přímka v čase

1. Stejný střední vývoj log-bilirubinu v obou skupinách:

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

2. Různá střední směrnice v jednotlivých skupinách:

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

Hierarchically specified model for data

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

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

- $\begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{b}_1 \end{pmatrix}, \ldots, \begin{pmatrix} \boldsymbol{Y}_N \\ \boldsymbol{b}_N \end{pmatrix}$ mutually independent

Linear mixed model, bayesian specification

“Čisté” parametry (parametry ve frekventistickém smyslu)

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

Linear mixed model, bayesian specification

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

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

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

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

Skrytá data a věrohodnost bayesovského modelu

- Další náhodné složky modelu:

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

  ➠ Další "parametry" při bayesovském přístupu

  $\equiv$ skrytá data

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

- **Věrohodnost bayesovského modelu:**

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

Section **7.2**

Generalized linear mixed model

Příklad: NTP TER84073 pokus na myších

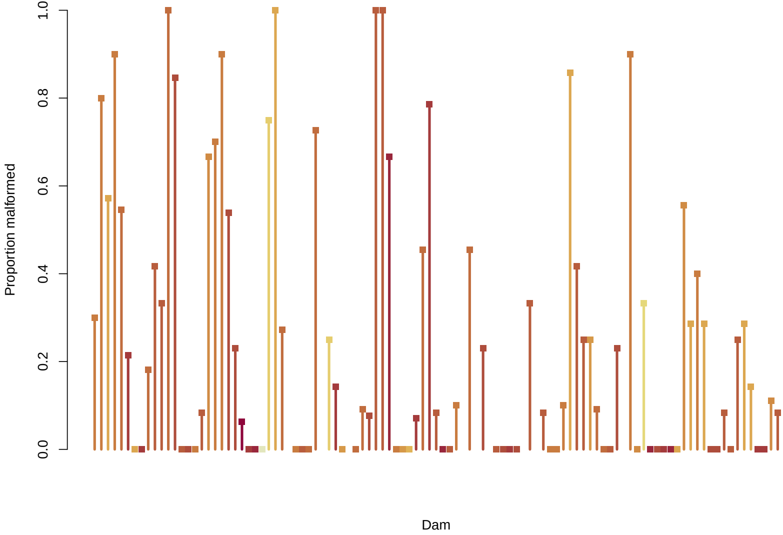
- Data z *National Toxicology Program*.
- 94 těhotným myším byla podána v předem určených momentech určená množství etylenglykolu (EG).
- V 17. dnu těhotenství byly myši usmrceny a následně byl zaznamenán počet zárodků ($n_i, i = 1, \ldots, 94$) a indikace, zda se u jednotlivých zárodků vyskytovala vývojová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

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

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

Příklad: NTP TER84073 pokus na myších

Pozorované proporce zárodků s vývojovými vadami



Příklad: NTP TER84073 pokus na myších

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

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

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

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

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

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

➠ Logistická regrese s náhodnými efekty.

Příklad: NTP TER84073 pokus na myších

Logistická regrese s normálně rozdělenými náhodnými efekty

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

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

Parametry (klasické)

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

Věrohodnost (frekventistická)

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

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

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

Section **7.3**

Prior distribution

Normální (zobecněný) lineární smíšený model

Apriorní rozdělení

- Využijeme rozklad, který sleduje hierarchickou strukturu modelu

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

- První část

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

- Druhá část $p(\psi)$
  - “standardní” apriorní rozdělení pro “čisté” parametry
  - obvykle se při jeho specifikaci zavádějí na principu obecných hiearchických modelů další náhodné hyperparametry ζ

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

- **Rozklad apriorního rozdělení:**

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

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

Normální (zobecněný) lineární smíšený model

Možné volby pro nestrukturální část apriorního rozdělení

- Potřeba zvolit $p(\psi)$, které často specifikujeme hierarchicky pomocí dalších hyperparametrů jako
$$p(\psi) = \int p(\psi \,|\, \zeta)\, p(\zeta)\, d\zeta,$$
to jest $p(\psi, \zeta) = p(\psi \,|\, \zeta)\, p(\zeta)$
- $\boldsymbol{\psi} = \left(\boldsymbol{\beta}^\top, \boldsymbol{\mu}^\top, \mathrm{par}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N), \mathrm{par}(\mathbb{D})\right)^\top$
- V dalším se budeme zabývat pouze situací, kdy $\boldsymbol{\Sigma}_i = \sigma^2\, \mathbf{I}_{n_i}$, tj. $\mathrm{par}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N) = \sigma^2$
- Označíme dále
$$\tau = \sigma^{-2}$$
$$\mathbb{Q} = \mathbb{D}^{-1}$$

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

Normální (zobecněný) lineární smíšený model

Možné volby pro nestrukturální část apriorního rozdělení

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

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

respektive

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

kde $\zeta = \bigl({\zeta^{(1)}}^\top, {\zeta^{(2)}}^\top, {\zeta^{(3)}}^\top, {\zeta^{(4)}}^\top\bigr)^\top$

Normální (zobecněný) lineární smíšený model

Možné volby pro nestrukturální část apriorního rozdělení

$\boldsymbol{\beta}$, $\boldsymbol{\mu}$ mají interpretaci středních hodnot

- Smysluplné apriorní rozdělení:

$$p(\boldsymbol{\beta}) \propto 1$$
$$p(\boldsymbol{\mu}) \propto 1$$

- Jiné smysluplné apriorní rozdělení:

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

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

Normální (zobecněný) lineární smíšený model

Možné volby pro nestrukturální část apriorního rozdělení

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

- Smysluplné apriorní rozdělení:

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

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

      ➠ d_τ : "přesnost" apriorního gama rozdělení
    - d_τ blízké 0 může vést k slabě informativnímu rozdělení

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

Normální (zobecněný) lineární smíšený model

Možné volby pro nestrukturální část apriorního rozdělení

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

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

➠ Wishartovo rozdělení

Wishartovo rozdělení

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

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

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

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

Wishartovo rozdělení

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

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

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

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

Normální (zobecněný) lineární smíšený model

Možné volby pro nestrukturální část apriorního rozdělení

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

- Smysluplné apriorní rozdělení:

$$p(\mathbb{Q}) \sim \mathcal{W}(\nu_Q, \boldsymbol{\Xi}_Q)$$

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

      ➠ diagonální hodnoty $\boldsymbol{\Xi}^{-1}$ (tj. $\gamma_{Q,1}, \ldots, \gamma_{Q,p}$) blízké 0 mohou vést k slabě informativnímu apriornímu rozdělení

      aposteriorního rozdělení

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

Normální (zobecněný) lineární smíšený model

Shrnutí

- **"čisté" parametry:** $\psi = \left(\boldsymbol{\beta}^\top, \boldsymbol{\mu}^\top, \tau, \mathrm{par}(\mathbb{Q})\right)^\top$
- **Skrytá data:** $\xi = \left(\boldsymbol{b}_1^\top, \ldots, \boldsymbol{b}_N^\top\right)^\top$
- **Věrohodnost bayesovského modelu:**

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

- Dekompozice **apriorního rozdělení:**

$$p(\xi,\,\psi) = p(\xi\,|\,\psi)\,p(\psi) = \prod_{i=1}^{N} \varphi\bigl(\boldsymbol{b}_i\,|\,\boldsymbol{\mu},\,\mathbb{Q}^{-1}\bigr)\,p(\boldsymbol{\beta})\,p(\boldsymbol{\mu})\,p(\tau)\,p(\mathbb{Q})$$

Section **7.4**

Posterior distribution

- Nejsou-li žádné náhodné hyperparametry:

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

- Jsou-li některé hyperparametry (označené jako ζ) náhodné:

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

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

Normální (zobecněný) lineární smíšený model

Aposteriorní rozdělení "čistých" parametrů

- Marginální aposteriorní rozdělení "čistých" parametrů:

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

- Díky hierarchické struktuře též platí

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

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

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

a platí Fubiniova věta

Inference založená na aposteriorním rozdělení

- Sdružené aposteriorní rozdělení pro $\boldsymbol{\theta} = \left(\boldsymbol{\psi}^\top, \boldsymbol{\xi}^\top, \boldsymbol{\zeta}^\top\right)^\top$ vyjádříme snadno až na multiplikativní konstantu

- Primárně nás však zajímají hlavně marginální aposteriorní rozdělení pro sady parametrů nebo dokonce jednotlivé parametry
  - $p(\boldsymbol{\beta} \,|\, \boldsymbol{y})$, $p(\beta_j \,|\, \boldsymbol{y})$, $p(\boldsymbol{\mu} \,|\, \boldsymbol{y})$, $p(\tau \,|\, \boldsymbol{y})$, ...
  - z nich odvozujeme aposteriorní střední hodnotu, věrohodnostní množiny atp.

- Při výpočtu marginálních aposteriorních rozdělení se již nelze vyhnout integrování
  - analyticky proveditelné pro jednodušší modely s apriorními rozděleními vykazujícími alespoň nějaký stupeň konjugovanosti
  - analyticky pracné pro složitější modely i s apriorními rozděleními vykazujícími alespoň nějaký stupeň konjugovanosti
  - analyticky často neproveditelné

Inference založená na aposteriorním rozdělení

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

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

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

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

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

Inference založená na aposteriorním rozdělení

- Vše výše řečené je důvodem toho, že bayesovská statistika byla až cca do začátku 90. let 20. století relativně málo prakticky využívána, a pokud ano, tak jenom v souvislosti s poměrně jednoduchými modely

- Řešení problému s nemožností odvozovat potřebné výrazy analyticky:

  ➠ aposteriorní inference založená na simulaci

  - k jejímu efektivnímu použití potřeba přiměřeně výkonné počítadlo
  - do cca začátku 90. let 20. století jenom omezeně dostupné/nedostupné (nejenom v ČSSR)

8

Bayesian data augmentation

Section **8.1**

Introduction

Rozšiřování dat (data augmentation)

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

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

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

Příklad 1: Lineární smíšený model

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

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

Parametry

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

Věrohodnost (pozorovaných dat)

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

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

Příklad 2: Logistická regrese s normálně rozdělenými náhodnými efekty

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

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

Parametry

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

Věrohodnost (pozorovaných dat)

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

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

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

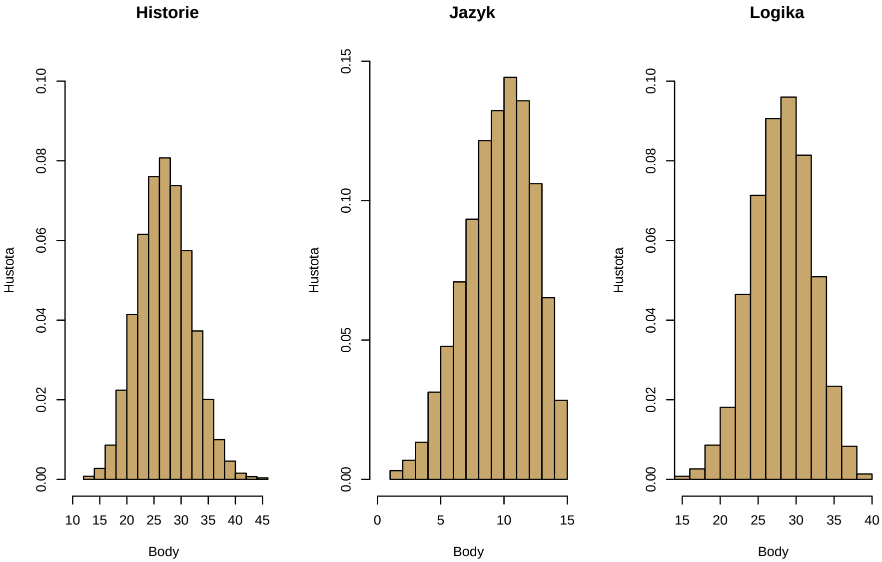
Příklad 3: Výsledky přijímacích zkoušek

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

  Viz http://www.cibulka.com/nnoviny/nn2000/nn1900/obsah/05.htm.

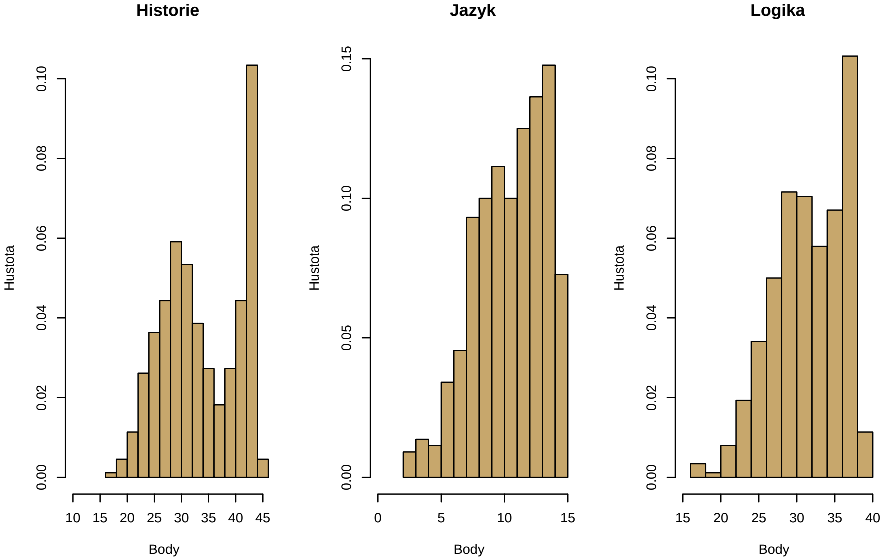
Příklad 3: Výsledky přijímacích zkoušek

Termíny č. 1–11



Příklad 3: Výsledky přijímacích zkoušek

Termín č. 12



Příklad 3: Výsledky přijímacích zkoušek

Možný model pro výsledky termínu č. 12

Možný model (pro $i = 1, \ldots, N$)

- $\boldsymbol{Y}_i = \bigl(Y_{i,1},\ Y_{i,2},\ Y_{i,3}\bigr)^\top$: bodové zisky $i$tého studenta u jednotlivých částí zkoušky.
- Studenti pocházejí ze dvou populací:
  1. běžní studenti (proporce w_1);
  2. studenti napojení na gangsterskou mafii (proporce w_2, $w_1 + w_2 = 1$).
- Budeme předpokládat, že (sdružené) rozdělení bodových zisků je v každé populaci normální se střední hodnotou $\boldsymbol{\mu}_1$, resp. $\boldsymbol{\mu}_2$ a varianční maticí $\boldsymbol{\Sigma}_1$, resp. $\boldsymbol{\Sigma}_2$.

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

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

Příklad 3: Výsledky přijímacích zkoušek

Možný model pro výsledky termínu č. 12

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

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

Věrohodnost (pozorovaných dat)

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

Section **8.2**

Principles

Rozšiřování dat (data augmentation)

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

    (augmented likelihood).

Rozšiřování dat (data augmentation)

Principy

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

Rozšiřování dat (data augmentation)

Principy

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

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

- Řekněme, že pro vhodné $\boldsymbol{\psi}$ se se sdruženou hustotou

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

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

- Řekněme, že zajistíme, aby platilo

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

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

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

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

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

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

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

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

Rozšiřování dat (data augmentation)

Principy

- Máme: $$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto L_{obs}(\boldsymbol{\theta})\, p(\boldsymbol{\theta}),$$
  $$p(\boldsymbol{\psi}, \boldsymbol{\theta} \mid \boldsymbol{y}) \propto L_{augm}(\boldsymbol{\psi}, \boldsymbol{\theta})\, p(\boldsymbol{\psi} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}).$$
  - $L_{obs}(\boldsymbol{\theta})$ plyne z předpokládaného modelu pro pozorovaná data.
  - $L_{augm}(\boldsymbol{\psi}, \boldsymbol{\theta})\, p(\boldsymbol{\psi} \mid \boldsymbol{\theta})$ plyne z uvažovaného rozšiřování dat.
    - $L_{augm}(\boldsymbol{\psi}, \boldsymbol{\theta})$: model (věrohodnost) pro pozorovaná data, jestliže na doplněná data pohlížíme jako na další parametry modelu.
    - $p(\boldsymbol{\psi} \mid \boldsymbol{\theta})$: model (věrohodnost) pro doplněná data.
- Rozšiřování je potřeba udělat tak, aby $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ bylo marginální hustotou odpovídající sdružené hustotě $p(\boldsymbol{\psi}, \boldsymbol{\theta} \mid \boldsymbol{y})$.
- Rozšiřování je tedy potřeba udělat tak, aby

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

Rozšiřování dat (data augmentation)

Principy

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

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

- Výraz

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

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

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

$=$ věrohodnost úplných dat.

Rozšiřování dat (data augmentation)

Principy

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

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

Specifikace hierarchického modelu:

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

   (model pro pozorovaná data za podmínky nepozorovaných dat).

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

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

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

Shrnutí terminologie

Data, parametry

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

Věrohodnosti

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

  věrohodnost úplných dat.

- Jednotlivé věrohodnosti potřeba specifikovat tak, aby

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

Section **8.3**

Examples

Příklad 1: Lineární smíšený model

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

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

Parametry

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

Nepřímo pozorovatelná (doplněná) data

- $\boldsymbol{\psi} = \bigl(\boldsymbol{b}_1^\top, \dots, \boldsymbol{b}_N^\top\bigr)^\top$.

Příklad 1: Lineární smíšený model

Věrohodnost **pozorovaných dat**

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

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

Rozšířená věrohodnost

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

Věrohodnost **doplněných dat**

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

Příklad 2: Logistická regrese s normálně rozdělenými náhodnými efekty

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

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

Parametry

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

Nepřímo pozorovatelná (doplněná) data

- $\boldsymbol{\psi} = \bigl(b_1, \ldots,\, b_N\bigr)^\top$.

Příklad 2: Logistická regrese s normálně rozdělenými náhodnými efekty

Věrohodnost **pozorovaných dat**

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

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

Rozšířená věrohodnost

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

Věrohodnost **doplněných dat**

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

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

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

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

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

Parametry

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

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

Nepřímo pozorovatelná (doplněná) data

- ???

Příklad 3: Normální směsový model

Věrohodnost **pozorovaných dat**

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

Rozšířená věrohodnost

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

Věrohodnost **doplněných dat**

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

Příklad 3: Normální směsový model

- Též nyní platí, že

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

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

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

Section **8.4**

Notes

Rozšiřování dat (data augmentation)

Poznámky

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

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

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

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

Rozšiřování dat (data augmentation)

Další oblasti využití

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

9

Bayesian model selection

Section **9.1**

Bayes factor

Bayesian model

Data: $\boldsymbol{Y}$,

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

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

Definition 9.1 Integrated (marginal) likelihood.

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

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

Remarks

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

Model selection

- Interest in selecting a model from a set of candidate models $M_1, \ldots, M_r$.
- Model M_k, $k = 1, \ldots, r$:

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

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

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

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

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

Model selection

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

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

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

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

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

  - Choose model with the maximal posterior probability.

- “Small” complication: Integrated likelihood $p_k(\boldsymbol{y}) = p(\boldsymbol{y} \,|\, \mathsf{M}_k)$ 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.,

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

Remarks

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

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

- With the uniform prior distribution for the competing models:

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

Jeffreys' scale of evidence for Bayes factor

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

Problems with Bayes factor

- The integrated likelihoods $p_k(\boldsymbol{y})$ which enter the Bayes factor are, in fact, the means (expected values) of the likelihood (under model M_k) with respect to the prior distribution (under model M_k).

- $p_k(\boldsymbol{y})$ is not well defined when the prior distribution $p_k(\boldsymbol{\theta}_k)$ is improper.

- Bayes factor is numerically unstable when proper but diffuse (weakly informative) prior distributions used.

- There exist numerous approaches that were suggested in the literature to overcome above problems.

Further reading

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

Section **9.2**

Posterior predictive distribution

Bayesian model

Data: $\boldsymbol{Y}$,

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

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

Posterior predictive distribution

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

Definition 9.3 Posterior predictive distribution.

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

We have

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

Posterior predictive distribution versus integrated likelihood

Integrated likelihood

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

≡ Distribution of data when the unknown parameters are averaged over their prior distribution.

➠ Evidence of the model before unknown parameters being estimated.

Posterior predictive distribution

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

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

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

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

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

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

In context of statistical modelling

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

Then

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

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

Deviance of a model

Definition 9.5 Deviance of a model.

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

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

is called the deviance of the model.

Remarks

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

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

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

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

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

Deviance

$$D(\boldsymbol{\theta};\, \boldsymbol{y}) = -2\log\Bigl\{\prod_{i=1}^{n} p_i(\boldsymbol{y}_i \,|\, \boldsymbol{\theta})\Bigr\} = -2\sum_{i=1}^{n}\log\bigl\{p_i(\boldsymbol{y}_i \,|\, \boldsymbol{\theta})\bigr\}$$

$$= \sum_{i=1}^{n} \underbrace{\Bigl[-2\,\log\bigl\{p_i(\boldsymbol{y}_i \,|\, \boldsymbol{\theta})\bigr\}\Bigr]}_{D_i(\boldsymbol{\theta};\, \boldsymbol{y}_i)}.$$

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

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

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

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

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

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

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

Bayesian inference

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

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

Evidence of the model before unknown parameters being estimated.

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

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

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

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

Posterior predictive deviance

Posterior predictive distribution:

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

Definition 9.6 Posterior predictive deviance.

Quantity

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

will be called the posterior predictive deviance.

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

Posterior predictive deviance

We have

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

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

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

Posterior predictive deviance

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

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

➠ Measure of the loss of prediction.

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

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

Definition 9.7 Cross-validated posterior predictive deviance.

Quantity

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

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

will be called the cross-validated posterior predictive deviance.

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

Definition 9.8 Posterior expected deviance.

Quantity

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

will be called the posterior expected deviance.

We have

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

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

Cross-validation and posterior expected deviance

Posterior expected deviance

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

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

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

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

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

Theorem 9.1 .

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

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

Reminder

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

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

Cross-validation and posterior expected deviance

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

Cross-validation and posterior expected deviance

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

Cross-validation and posterior expected deviance

That is,

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

Definition 9.9 Expected optimism.

Quantity

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

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

Definition 9.10 Penalized expected deviance (PED).

Quantity

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

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

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

will be called the overall optimism, quantity

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

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

Theorem 9.2 Penalized expected deviance and cross-validated posterior predictive deviance.

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

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

- With respect to cross-validation

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

is equivalent to

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

Penalized expected deviance and cross-validated posterior predictive deviance

Proof.

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

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

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

❑

Practical use of PED

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

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

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

Deviance information criterion

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

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

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

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

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

- PED with p_D used in place of p_{opt}
  ➠ **Deviance information criterion (DIC)**.

Deviance information criterion (DIC)

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

- DIC $\equiv$ approximation to $\overline{D}_{pred}^{CV}$ 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.