Introduction to Bayesian Statistics

Helga Wagner

IFAS, JKU Linz

Summerschool Ligist 2019

Part 1:
Introduction to Bayesian Statistics

- Why Bayes?
- Statistical Inference
- The Bayesian Approach to Statistical Inference
- Conjugate Bayesian Analysis
  - Probability of a rare disease
  - The Normal model

Why Bayes?

Why Bayes?

Reverend Thomas Bayes

(*1702 in London, †1761)

In 1763 Richard Price read *An Essay Towards solving a Problem in the Doctrine of Chances* where Bayes solves the problem of "inverse" probabilities

$\Longrightarrow$ Bayes rule and Bayesian inference has its birth

Bayes rule

Theorem

For two events A and B with $P(B) > 0$ the probability of A conditioning on B is given as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}$$

Example: Screening-Test

Available information

- prevalence of the disease in the population $P(A)$
- properties of the screening test
  - sensitivity: $P(T+|A)$
  - specificity: $P(T-|A^C)$

What is the probability to have the disease given a positive test result ?

$\Longrightarrow$ positive predictive value $P(A|T+)$?

Example: Screening-Test

- prevalence $P(A) = 0.002$
- sensitivity: 0.98; specifity: 0.935

	$P(\sim \mid A)$	$P(\sim \mid A^C)$
$T+$	0.980	0.065
$T-$	0.020	0.935
Σ	1.000	1.000

$\Longrightarrow$

	$P(A \mid \sim)$	$P(A^C \mid \sim)$	Σ
$T+$	0.02933	0.97067	1.000
$T-$	0.00004	0.99996	1.000
Σ	0.002	0.998	

$$P(A|T+) = 0.02933 \approx 15P(A) \qquad P(A|T-) = 0.00004 \approx \frac{1}{50}P(A)$$

$\Rightarrow$ Bayesian learning, i.e. updating prior beliefs

Topics

goals: you

- will be familiar with the Bayesian approach to statistical learning
- know basic concepts of Bayesian statistcs
- can perform a conjugate Bayesian analysis
- know how to perfom Bayesian inference using Monte Carlo methods

Literature:
Hoff, Peter D. (2009). A first Course in Bayesian Statistics. Springer
Material: `https://pdhoff.github.io/book/`

Statistical Inference

Statistical Inference

- statistical induction
  use a data sample to infer population characteristics
  - observed data **y** quantify the outcome of a survey or experiment
  - parameter θ quantifies unknown population characteristic
- information (uncertainty)
  - pre-experimentally (before the experiment) **y** and θ are unknown
  - post-experimentally (after the experiment) **y** is known; θ is unknown
- goal: inference on the parameter θ after observing **y**

Example: Probability of a Rare Disease

- interest is in prevalence θ of an infectious disease in a population
  - a random sample of $n = 20$ persons is checked for infection
  - y is the number of infected persons in the sample

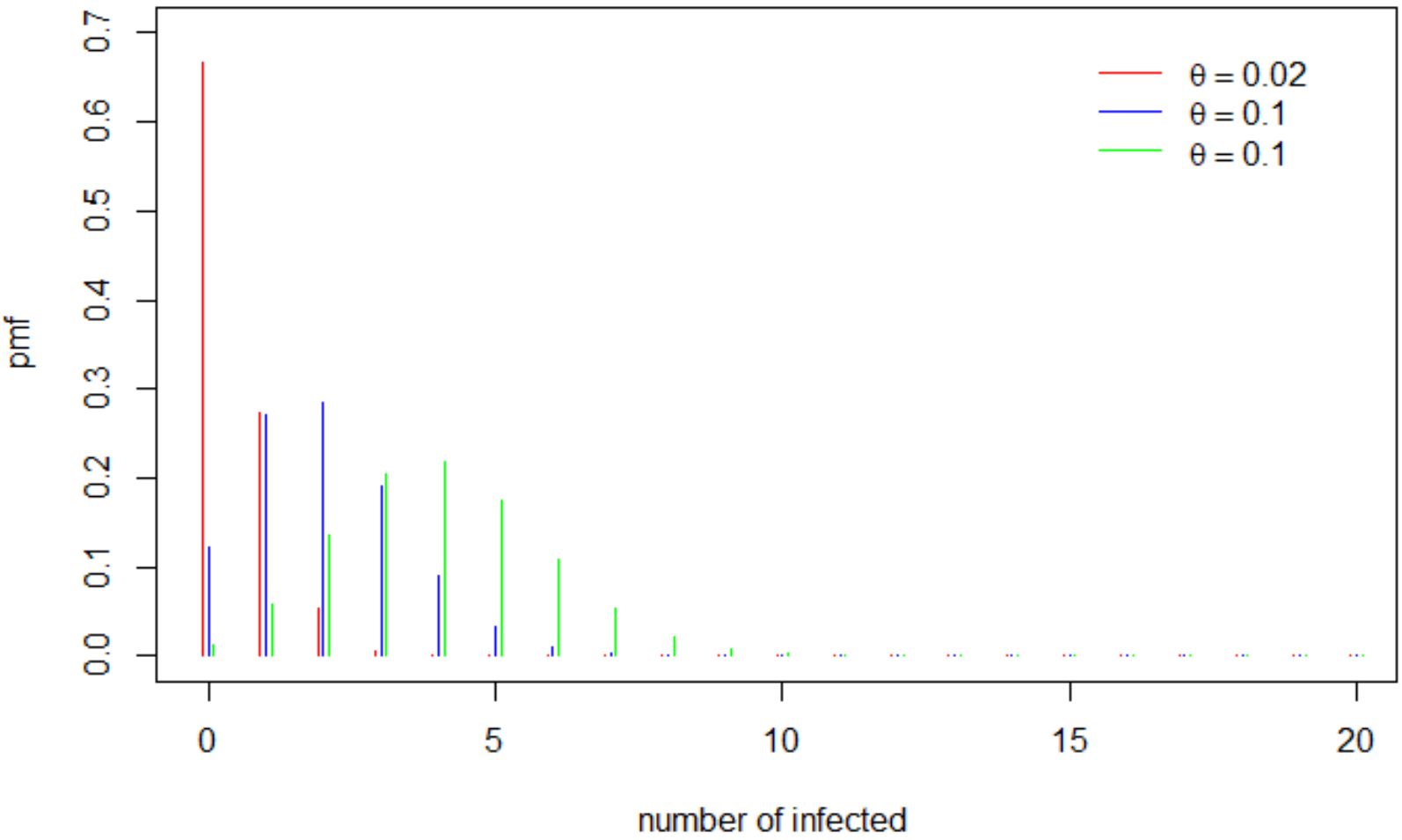
  what is the information on θ?
- statistical model
  - parameter space $\Theta = [0, 1]$
  - sample space $\mathcal{Y} = \{0, 1, ..., 20\}$

  For a large population a reasonable sampling model for Y is

$$Y|\theta \sim \text{BiNom}(20, \theta)$$

Binomial distributon

$$Y \sim \text{BiNom}(n, \theta) \Longrightarrow \quad p(Y = k) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$



Probability of a Rare Disease: Statistical Analysis

- sampling model: $y \sim \text{BiNom}(20, \theta)$
- data: $n = 20, y = 0$

Statistical Tasks:

- point estimation: $\hat{\theta}$
- interval estimation: determine $[\theta_l = l(Y), \theta_u = u(Y)]$ such that
$$P(l(Y) < \theta < u(Y)) = 1 - \alpha$$
- hypothesis tests: e.g. $\mathbf{H}_0 : \theta \geq 0.1$ vs. $\mathbf{H}_1 : \theta < 0.1$
- prediction: number of infected $\tilde{Y}$ in a new sample of size $\tilde{n}$

Probability of a Rare Disease: Frequentist Analysis

- point estimation: ML- estimation
  - ML-estimate $\hat{\theta}$ is the value of θ maximizing the probability to observe y in Binomial sampling of size n
  - likelihood: $P(Y = 0|\theta) = (1 - \theta)^n$
  - MLE for θ

$$\hat{\theta} = \overline{y} = \frac{y}{n} = 0$$

- interval estimation:
  - approximative: Wald confidence interval

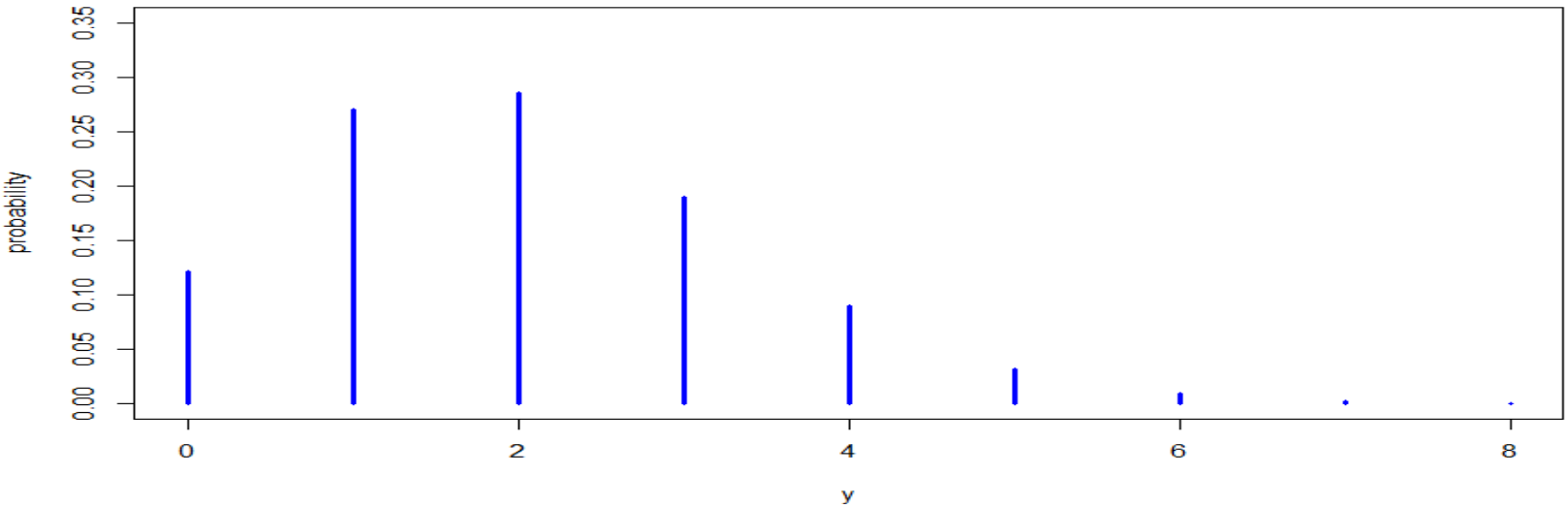
$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\hat{\theta}(1-\hat{\theta})/n}$$

has length 0 if $\hat{\theta} = 0$ for any significance level
(frequentist coverage is asymptotically correct but not for small n)

Probability of a Rare Event: Frequentist Analysis

- hypothesis testing: the p-value is supremum of the probability to observe y (or an even more extreme value) under $\mathbf{H}_0$

$$p = \sup_{\theta \geq 0.1} P(Y \leq 0|\theta)$$



distribution of Y for $\theta = 0.1$

therefore $p = P(Y = 0|\theta = 0.1) = 0.1215$

Probability of a Rare Event: Frequentist Analysis

- prediction: conditional on $\hat{\theta}$

$$\tilde{Y}|\hat{\theta} \sim \mathbf{B}_{\tilde{n},\hat{\theta}}$$

$\Longrightarrow P(\tilde{Y} = 0|\hat{\theta} = 0) = 1$ for any $\tilde{n}$

The Bayesian Approach to Statistical Inference

Bayesian Approach

- quantify beliefs (information) on unknown quantities/events
  $\Longrightarrow$ beliefs can be expressed via probability distributions
- update beliefs in light of new information
  $\Longrightarrow$ information update (inductive learning) via Bayes rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}$$

Information Update

Bayes rule allows information update

- A and B are independent

$$P(A \cap B) = P(A)P(B) \quad \Longrightarrow \quad P(A|B) = P(A)$$

Information that B has occurred provides no information on the probability of A

- A and B are dependent

$$P(A \cap B) > P(A)P(B) \quad \Longrightarrow \quad P(A|B) > P(A)$$

or

$$P(A \cap B) < P(A)P(B) \quad \Longrightarrow \quad P(A|B) < P(A)$$

Information that B has occurred provides information on the probability of A.

Bayes Approach: Basics Summary

- (subjective) uncertainty or prior knowledge is quantified through probability distributions
- before data are collected there is uncertainty on
  - the observables (data) $\mathbf{y}$
  - the unobservables, i.e. the unknown parameter (vector) θ

$\Longrightarrow$ specification of a joint stochastic model for $(\mathbf{y}, \theta)$

Bayesian analysis

- parameter and sample space
  - the parameter space Θ is the set of all possible values for θ
  - the sample space $\mathcal{Y}$ is the set of all possible data sets
- Bayesian modelling (quantification of information)
  - prior distribution $p(\theta)$
    quantifies the belief that θ is the value of the population parameter for all $\theta \in \Theta$ before observing the data
  - sample model $p(\mathbf{y}|\theta)$
    quantifies the belief that $\mathbf{y}$ will be observed for all $\theta \in \Theta$ and $\mathbf{y} \in \mathcal{Y}$
- Bayesian inference
  - based only on posterior distribution $p(\theta|\mathbf{y})$
  - the posterior distribution combines information on θ from prior and data

Bayes Theorem

Bayes Theorem (normalized)

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

where $p(\sim)$ denotes

- the probability density for continuous random variables
- probability function for discrete random variables

Bayes theorem

- describes how uncertainty on the parameter is changed by the information in the data
- holds for discrete and continuous random variables $\mathbf{y}, \theta$

Bayes Theorem

Bayes Theorem (non-normalized)

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

posterior $\propto$ likelihood $\times$ prior

The normalizing constant is given as

$$p(\mathbf{y}) = \sum_{\boldsymbol{\theta}\in\Theta} p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \quad \boldsymbol{\theta} \ldots \text{discrete} \tag{1}$$

$$p(\mathbf{y}) = \int_{\Theta} p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \quad \boldsymbol{\theta} \ldots \text{continuous} \tag{2}$$

Conjugate Bayesian Analysis

Probability of a rare disease

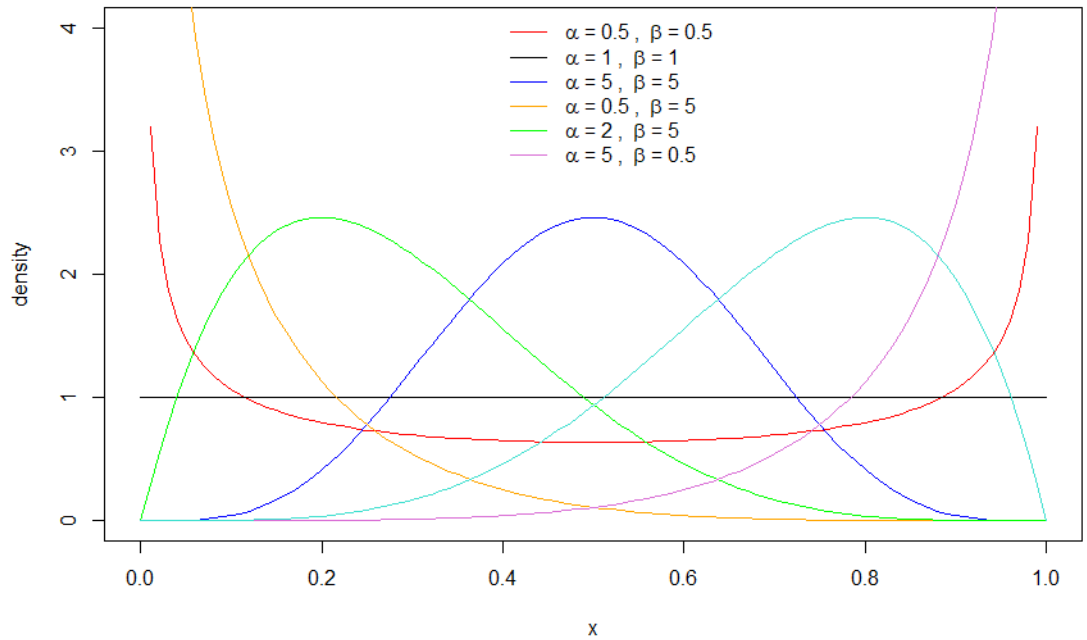
Probability of a Rare Disease: Bayesian Analysis

- data: $n = 20, y = 0$
- sampling model: $y \sim \text{BiNom}\,(20, \theta)$
  likelihood: $P(Y = 0|\theta) = (1 - \theta)^n$
- of interest: $P(\theta|\, Y = 0)$
- a reasonable prior distribution for $\theta \in \mathcal{Y} = [0, 1]$ is a Beta-distribution distribution $\mathcal{B}\,(a, b)$

Beta-distribution

- parameters: $a, b > 0$
- probability density function (pdf)

$$p(x|\mathcal{B}(a,b)) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} x^{a-1}(1-x)^{b-1} \quad \text{for } x \in [0,1]$$



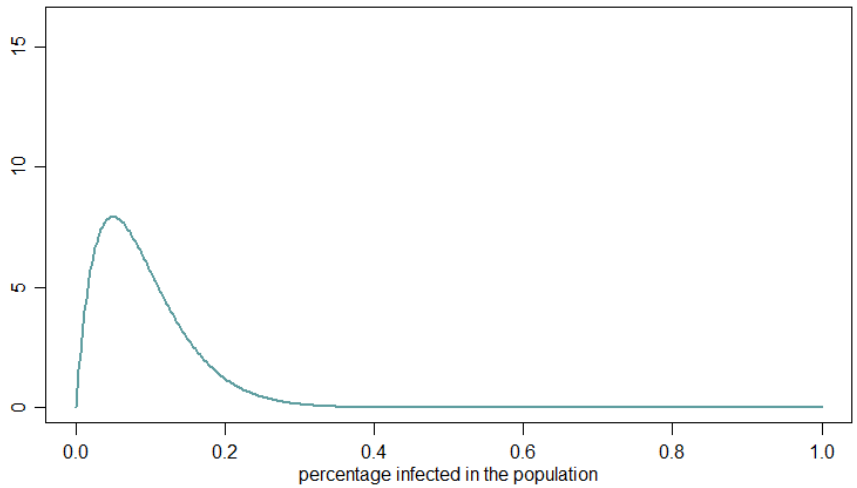
- moments

$$\mathrm{E}(X) = \frac{a}{a+b} \qquad \mathrm{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

Probability of a Rare Disease: Prior Distribution

- infection rates from 0.05 to 0.2 , average prevalence of 0.1
  $\Longrightarrow$ assign substantial probability to $[0.05, 0.2]$
- specify $p(\theta)$ as a Beta-distribution $\mathcal{B}(a, b)$,

$$\theta \sim \mathcal{B}(2, 20)$$



$$\mathrm{E}(\theta) = 0.09 \qquad \text{mode}(\theta) = 0.05$$
$$P(\theta < 0.1) = 0.64 \quad P(0.05 < \theta < 0.2) = 0.66$$

Probability of a Rare Disease: Derivation of the Posterior Distribution

- data: y of the n tested persons were infected
- the posterior $p(\theta|y)$ is derived by Bayes theorem

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\ &= \frac{1}{p(y)} \left(\begin{array}{c} n \\ y \end{array} \right) \theta^y (1-\theta)^{n-y} \cdot \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \theta^{a-1}(1-\theta)^{b-1} \\ &= c(n, y, a, b) \cdot \theta^{a+y-1}(1-\theta)^{b+n-y-1} \end{aligned}$$

- this is the kernel of the $\mathcal{B}(a+y, b+n-y)$ distribution

$$\theta|y \sim \mathcal{B}(a+y, b+n-y)$$

Conjugacy

- the binomial sampling model
- combined with Beta prior for θ

yields a Beta posterior

Definition (Conjugacy)

A class of prior distributions $\mathcal{P}$ for a parameter θ is called conjugate for the sampling model $p(y|\theta)$ if

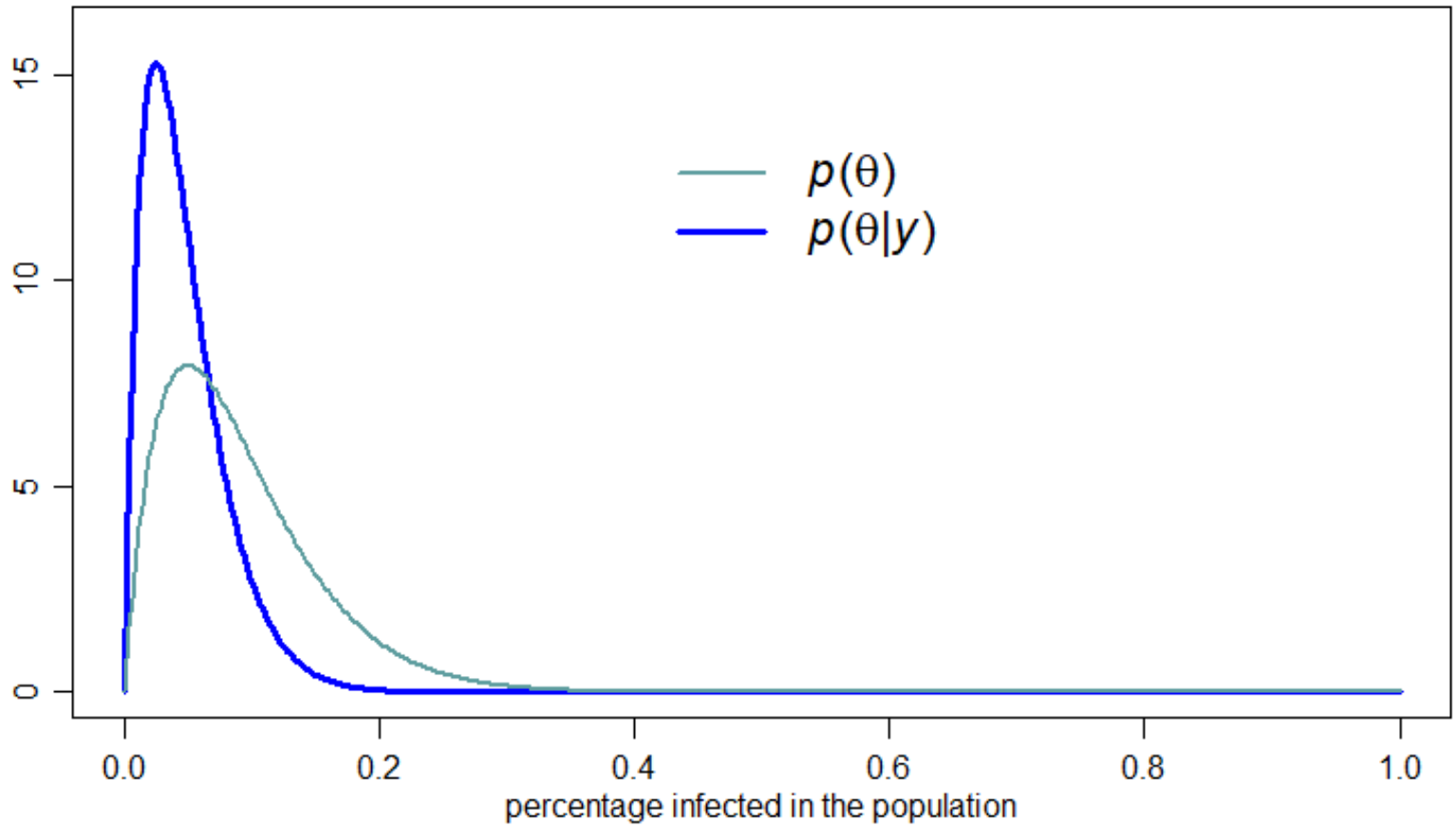
$$p(\theta) \in \mathcal{P} \Longrightarrow p(\theta|y) \in \mathcal{P}$$

$\Longrightarrow$ the class of Beta prior distributions is conjugate for the Binomial sampling model

Probability of a Rare Disease: Posterior Distribution

for $n = 20, y = 0, a = 2$ and $b = 20$ we get

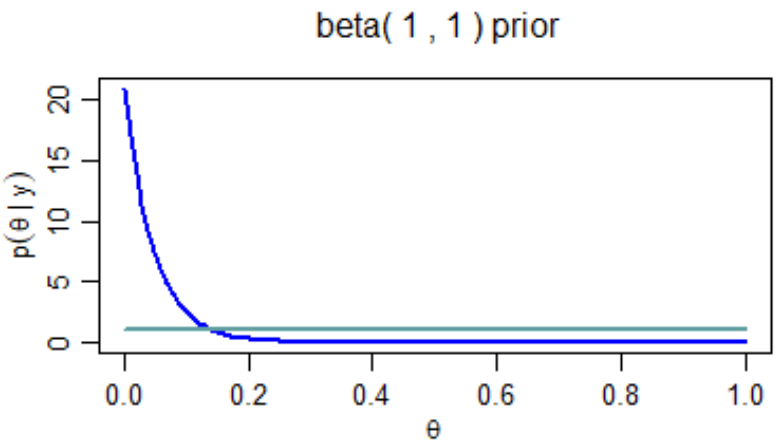
$$\theta|y \sim \mathcal{B}(2, 40)$$

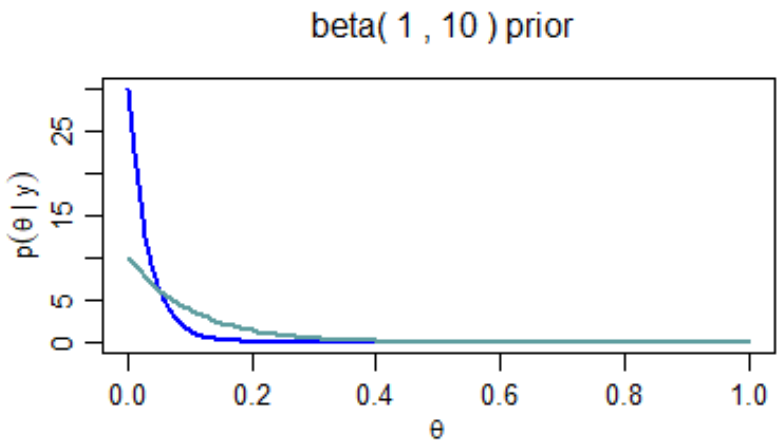


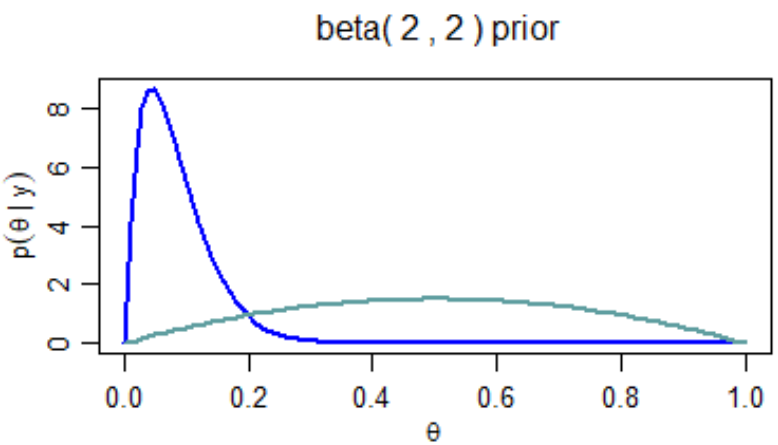
Probability of a Rare Disease: Sensitivity Analysis

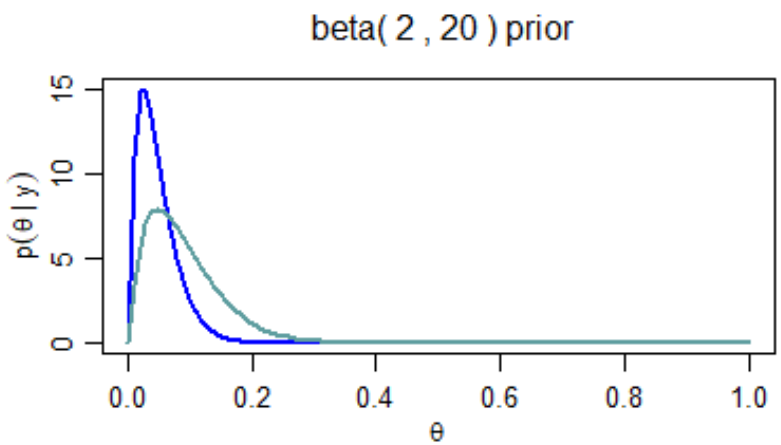
Effect of the prior distribution: for $\theta \sim \mathcal{B}(a, b)$ the posterior is

$$\theta|y \sim \mathcal{B}(a + y, b + n - y)$$









Probability of a Rare Event: Sensitivity Analysis

With the posterior

$$\theta|y \sim \mathcal{B}(a+y, b+n-y)$$

the posterior expectation is

$$\begin{aligned}
\mathrm{E}(\theta|y) &= \frac{a+y}{a+b+n} = \frac{a}{a+b+n} + \frac{y}{a+b+n} = \\
&= \frac{a+b}{a+b+n} \cdot \frac{a}{a+b} + \frac{n}{a+b+n} \cdot \frac{y}{n} = \\
&= \frac{a+b}{a+b+n} \cdot E(\theta) + \frac{n}{a+b+n} \cdot \overline{y} = \\
&= w \cdot \overline{y} + (1-w) \cdot E(\theta)
\end{aligned}$$

Impact of the Prior Distribution

the posterior expectation is a weighted average of prior expectation and data mean

$$\mathrm{E}(\theta|y) = \frac{a+b}{a+b+n} \cdot \mathrm{E}(\theta) + \frac{n}{a+b+n} \cdot \overline{y}$$

- the weight
  - of the data mean w is proportional to n
  - of the prior expectation $1-w$ is proportional to $a+b$
- for fixed
  - n: the weight of the prior mean increases with $a+b$
  - $a+b$: the weight of the sample mean converges to 1 for $n \to \infty$

Parameter Estimation

Point estimation of θ by location parameters of the posterior distribution, e.g. posterior mode, posterior median or posterior mean.

Each of these estimators is the optimal estimator with respect to certain loss function:

- Let $\mathcal{R}(\hat{\theta}(\mathbf{y}), \theta)$ be a function quantifying the loss made when estimating a parameter θ by the point estimate $\hat{\theta}(\mathbf{y})$.
- Select the estimator $\hat{\theta}(\mathbf{y})$ that minimizes the expected loss with respect to the posterior distribution:

$$E\{\mathcal{R}(\hat{\theta}(\mathbf{y}), \theta)|\mathbf{y}\} = \int \mathcal{R}(\hat{\theta}(\mathbf{y}), \theta) p(\theta|\mathbf{y})\, d\theta$$

Parameter Estimation

- The posterior expectation $E\{\boldsymbol{\theta}|\mathbf{y}\}$ is optimal with respect to the quadratic loss function $\mathcal{R}(\hat{\boldsymbol{\theta}}(\mathbf{y}), \boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta})^T(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta})$.
- The posterior mode is the optimal with respect to the 0/1 loss function:

$$\mathcal{R}(\hat{\boldsymbol{\theta}}(\mathbf{y}), \boldsymbol{\theta}) = \begin{cases} 0, & \hat{\boldsymbol{\theta}}(\mathbf{y}) = \boldsymbol{\theta}, \\ 1, & \hat{\boldsymbol{\theta}}(\mathbf{y}) \neq \boldsymbol{\theta}. \end{cases}$$

- In a single parameter problem, the posterior median is optimal under the absolute deviation $\mathcal{R}(\hat{\boldsymbol{\theta}}(\mathbf{y}, \boldsymbol{\theta}) = |\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}|$

Bayesian Interval Estimation

uncertainty on the parameter can be expressed by an interval/region or a dispersion parameter of the posterior

Definition (Credibility interval)

Let $\Theta \subset \mathbb{R}$ (i.e. θ is univariate).

- An interval

$$[l(\mathbf{y}), u(\mathbf{y})]$$

has $100(1-\alpha)\%$ Bayesian coverage if

$$P(l(\mathbf{y}) < \theta < u(\mathbf{y}) | Y = y) = 1 - \alpha$$

- An interval with $100(1-\alpha)\%$ Bayesian coverage is a $100 \cdot (1-\alpha)\%$ credibility interval.

Interpretation: a credibility interval has post-experimental coverage of $100(1-\alpha)\%$

Credibility Regions

- quantile based (equal tailed) interval

$$\left[\theta_{\alpha/2}, \theta_{1-\alpha/2}\right]$$

  where θ_α is the α -quantile of the posterior distribution $p(\theta|y)$
- goal: shortest region with $100(1-\alpha)\%$ coverage

Definition (HPD region = highest posterior density region)

A $100(1-\alpha)\%$ HPD region consists of a subset of the parameter space $s(\mathbf{y}) \subset \Theta$ such that

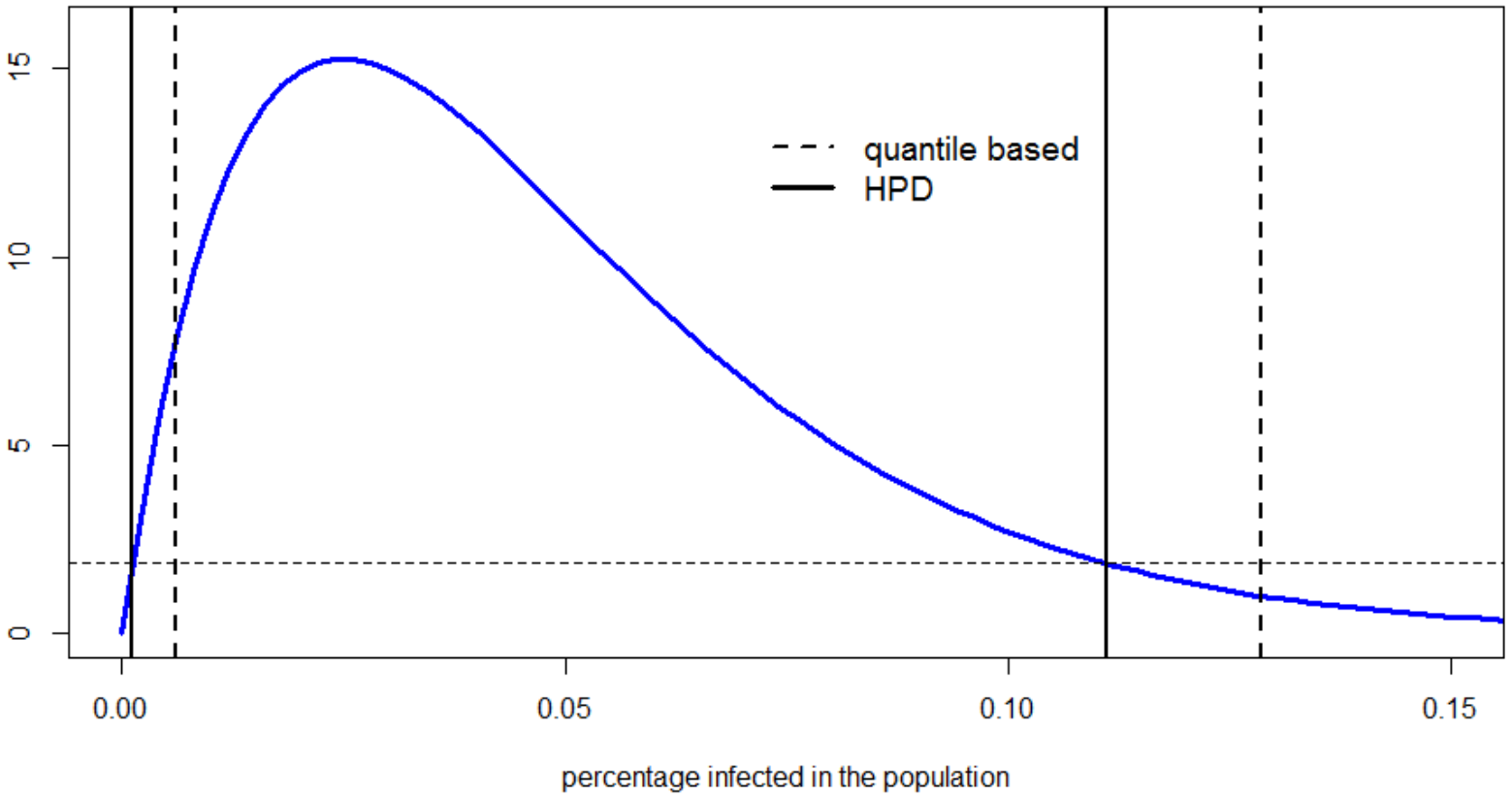
- $P(\theta \in s(y)|Y) = 1-\alpha$
- if $\theta_1 \in s(y)$ and $\theta_2 \notin s(y)$, then $p(\theta_1|Y=y) > p(\theta_2|Y=y)$

Note: a HPD region is not necessarily an interval

Probability of a Rare Event: Credibility Regions

Example: 95% credibility regions for θ

- $(\theta_{0.025}, \theta_{0.975}) = (0.00596, 0.12855)$; length $= 0.1225$
- HPD region $(0.0011, 0.1111)$; length $= 0.1100$



Bayesian and Frequentist Coverage

Definition (Frequentist coverage)

A random interval $[l(Y), u(Y)]$ has $100(1-\alpha)\%$ frequentist coverage if

$$P(l(Y) < \theta < u(Y)) = 1 - \alpha$$

- confidence limits $l(Y)$ and $u(Y)$ are random variables
- frequentist coverage
  - refers to repeated experiments: 95 % of the confidence intervals will contain the true parameter
  - is pre-experimental: after the data are collected the confidence interval either contains the parameter or not
- intervals with $100(1-\alpha)\%$ Bayesian coverage usually have approximate $100(1-\alpha)\%$ frequentist coverage

Probability of a Rare Event: Prediction

goal: prediction of the disease status for a person not yet observed

- let $\tilde{Y} \in [0, 1]$ denote the disease status of this person
- predictive distribution of $\tilde{Y}$ given y

$$P(\tilde{Y} = 1|y) = \int_0^1 P(\tilde{Y} = 1, \theta|y)d\theta =$$
$$= \int_0^1 P(\tilde{Y} = 1|\theta, y)P(\theta|y)d\theta =$$
$$= \int_0^1 \theta P(\theta|y)d\theta = E(\theta|y) = \frac{a+y}{a+b+n}$$
$$P(\tilde{Y} = 0) = 1 - E(\theta|y) = \frac{b+n-y}{a+b+n}$$

Further examples of conjugacy

Conjugate priors exist for the natural parameter of distributions from the exponential family.

likelihood	conjugate prior family
Bino(n, θ)	$\theta \sim$ Beta(a_0, b_0)
Poisson(λ)	$\lambda \sim$ Gamma(a_0, b_0)
$N(\mu, \sigma^2)$	$\mu \sim N(m_0, M_0)$
$N(\mu, \sigma^2)$	$\sigma^2 \sim$ InvGamma(s_0, S_0)

Conjugate Bayesian Analysis

The Normal Model

The Normal Model

Example: Midge wing data (Grogan and Wirth, 1981)

Data: $n = 9$ measurements of wing length (in mm) of a species of midge

$$\mathbf{y} = (1.64, 1.70, 1.72, 1.74, 1.82, 1.82, 1.82, 1.90, 2.08)$$

Model: $y_1, \ldots y_n$ independent

$$y_i = \mu + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}\left(\mu, \sigma^2\right), i = 1, \ldots, n$$

with $\sigma^2 = 0.13^2$.

Likelihood:

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

Bayesian analysis

Prior: Conjugate prior

$$\mu|\sigma^2 \sim \mathcal{N}(m_0, M_0)$$

Posterior: Normal distribution

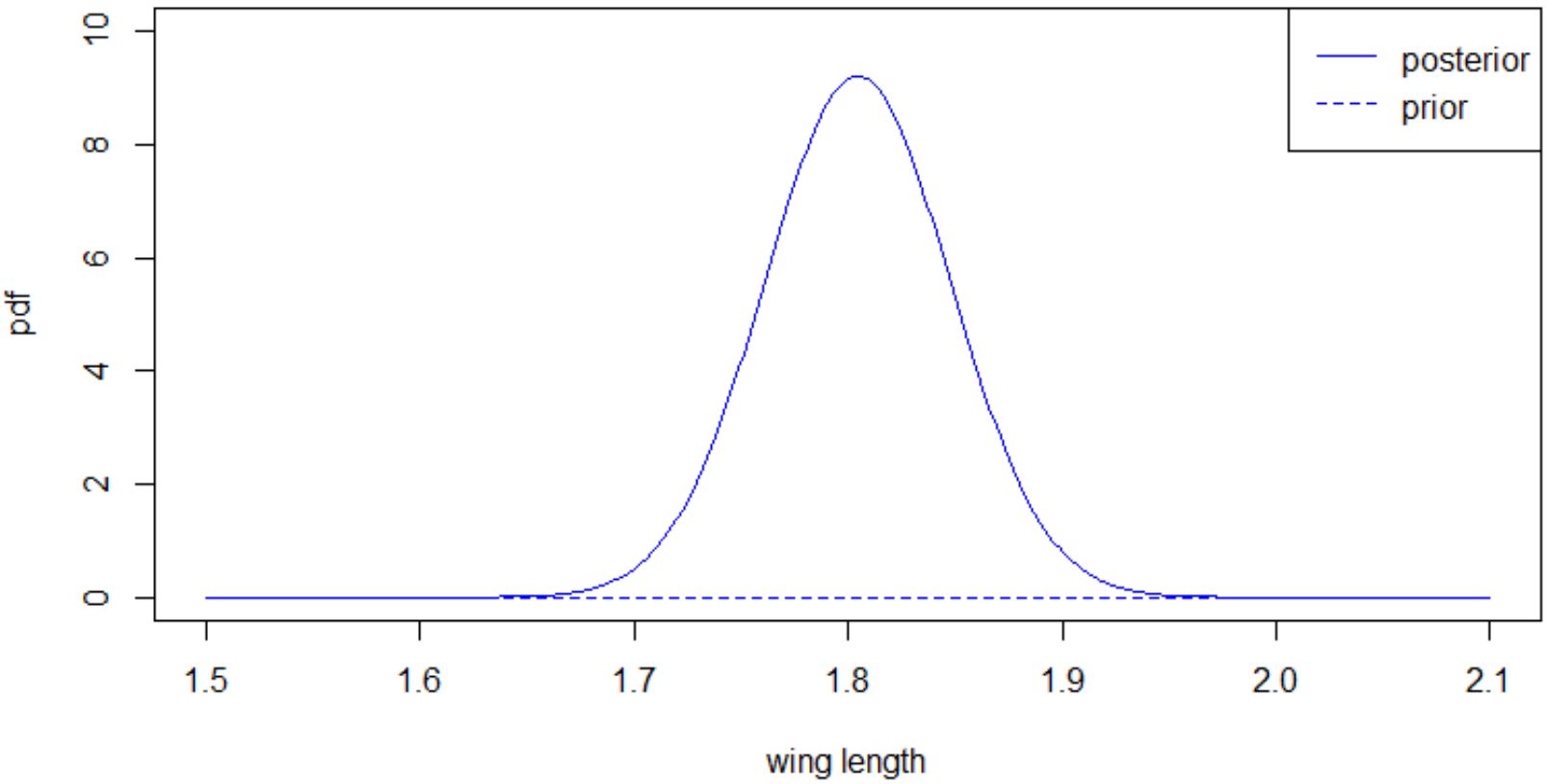
$$\mu|\sigma^2, \mathbf{y} \sim \mathcal{N}(m_n, M_n)$$

with

$$M_n^{-1} = \frac{1}{M_0} + \frac{n}{\sigma^2} \qquad m_n = M_n\left(\frac{m_0}{M_0} + \frac{n}{\sigma^2}\bar{y}\right)$$

Midge data: posterior distribution

$$\mu|\sigma^2, \mathbf{y} \sim \mathcal{N}(m_n, M_n)$$



Joint inference on mean and variance

What if also σ^2 is unknown?
Model: $y_1, \ldots y_n$ independent

$$y_i = \mu + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}\left(\mu, \sigma^2\right),$$

$\Longrightarrow$ joint prior for μ and σ^2 required

Likelihood of the normal model

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

Prior: Conjugate prior

$$\mu|\sigma^2 \sim \mathcal{N}\left(m_0, M_0\sigma^2\right) \qquad \sigma^2 \sim \mathcal{G}^{-1}(s_0, S_0)$$

The inverse Gamma distribution

- parameters: $a, b > 0$
- probability density function (pdf)

$$p(x|\mathcal{G}^{-1}(a,b)) = \frac{b^a}{\Gamma(a)} \frac{1}{x^{a+1}} \exp(-b/x) \quad \text{for } x > 0$$

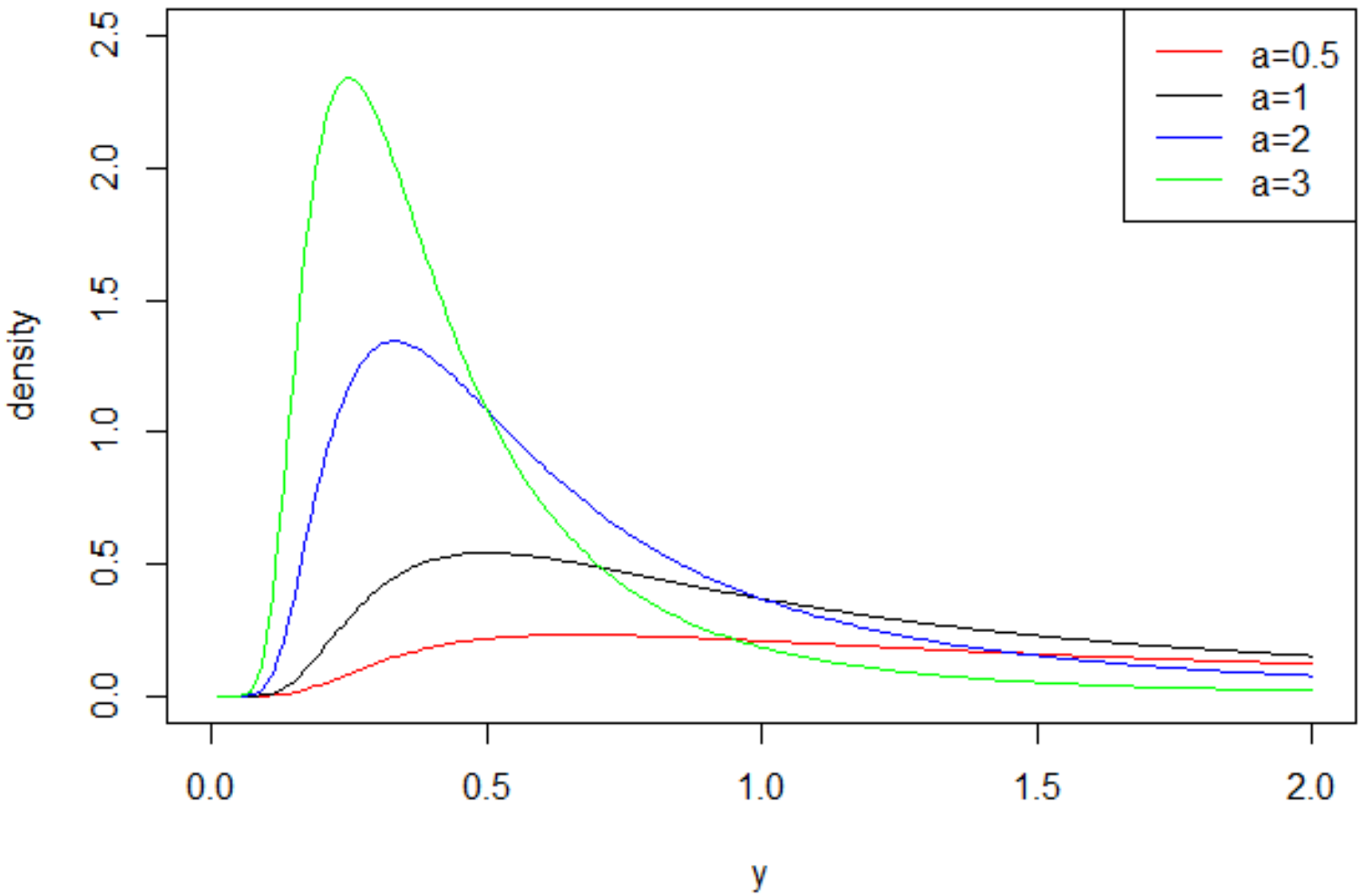
- If $X \sim \mathcal{G}(a,b)$ then $1/X \sim \mathcal{G}^{-1}(a,b)$
- moments

$$\mathrm{E}(X) = \frac{b}{a-1} \quad \text{for } a > 1$$

$$\mathrm{mode}(X) = \frac{b}{a+1}$$

$$\mathrm{Var}(X) = \frac{b^2}{(a-1)^2(a-2)} \quad \text{for } a > 2$$

Pdf of the inverse Gamma distribution



Pdf of the inverse Gamma distribution (different shape a and b=1)

Specification of the prior distribution

- prior knowledge suggests that μ and σ are not too far from 1.9 mm and $(0.1\ \text{mm})^2$
- Reparameterization:

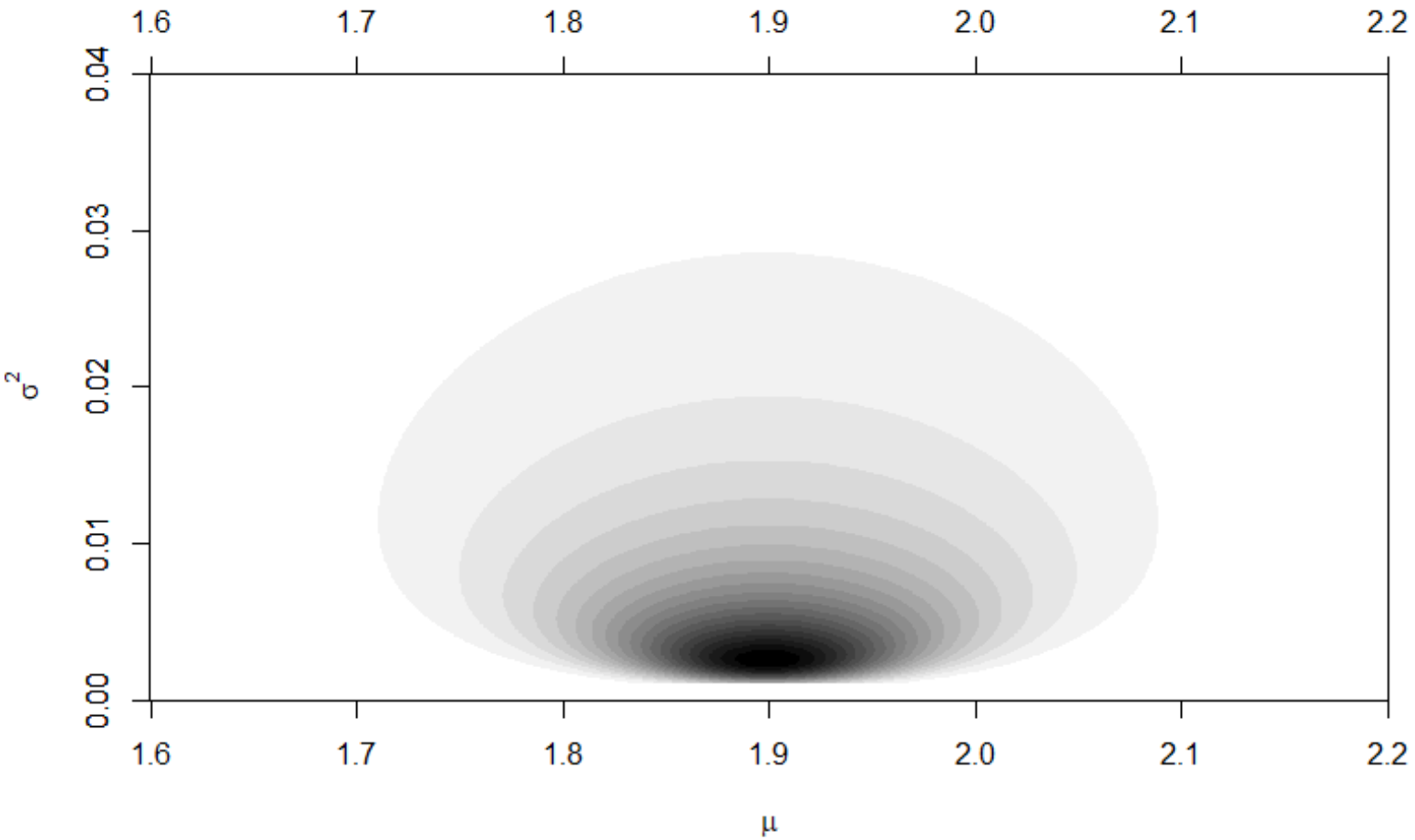
$$\mu|\sigma^2 \sim \mathcal{N}\left(m_0, M_0\,\sigma^2\right)$$
$$\sigma^2 \sim \mathcal{G}^{-1}\left(\nu_0/2, \nu_0/2\,\sigma_0^2\right)$$

- Prior moments of σ^2

$$\mathrm{E}(\sigma^2) = \frac{b}{a-1} = \sigma_0^2 \frac{\nu_0/2}{\nu_0/2-1} \quad \text{for } \nu_0 > 2$$
$$\mathrm{Var}(\sigma^2) = \left(\mathrm{E}(\sigma^2)\right)^2 \frac{1}{\nu_0/2-2} \quad \text{for } \nu_0 > 4$$

Joint prior for mean and variance

prior parameters: $m_0 = 1.9$, $M_0 = 1$; $\nu_0 = 1$, $\sigma_0^2 = 0.01$



Joint posterior distribution

$$\sigma^2|\mathbf{y} \sim \mathcal{G}^{-1}\left(\nu_n/2, \nu_n/2\,\sigma_n^2\right)$$
$$\mu|\sigma^2, \mathbf{y} \sim \mathcal{N}\left(m_n, M_n\,\sigma^2\right)$$

with

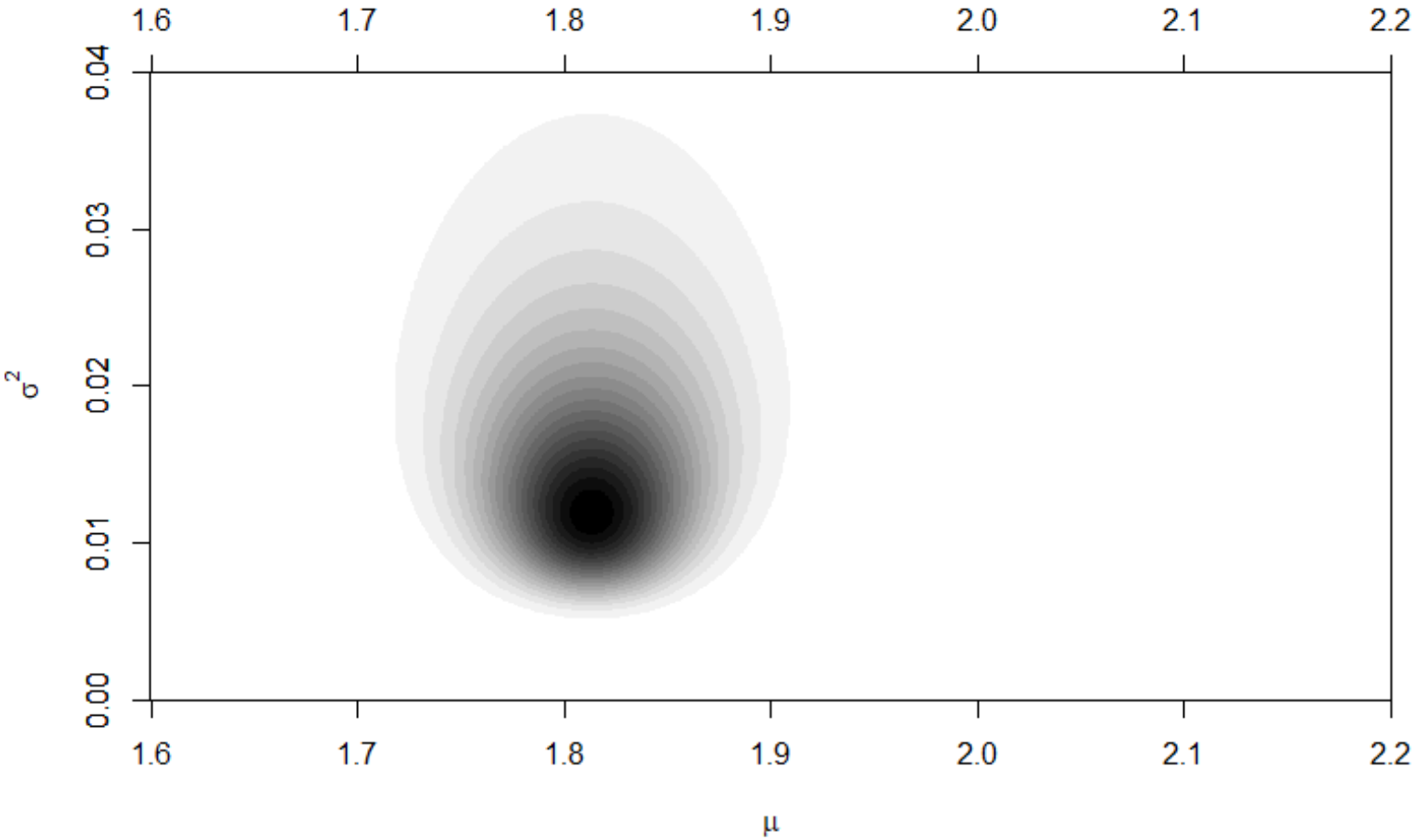
$$M_n^{-1} = 1/M_0 + n$$
$$m_n = M_n(\frac{\mu_0}{M_0} + n\bar{y})$$

and

$$\nu_n = \nu_0 + n$$
$$\sigma_n^2 = \frac{1}{\nu_n}\left(\nu_0\sigma_0^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2 + \frac{1}{M_0 + 1/n}(\bar{y} - m_0)^2\right)$$

Joint Posterior for mean and variance

prior parameters: $m_0 = 1.9$, $M_0 = 1$; $\nu_0 = 1$, $\sigma_0^2 = 0.01$



Part 2: Methods for posterior inference

- Posterior inference for regression models
- Posterior simulation
- MCMC methods
  - Gibbs sampling: linear regression with semi-conjugate prior
  - Data Augmentation: probit model
  - Metropolis Hastings Algorithm: logistic regression
  - Posterior inference based on MCMC samples

Posterior inference for regression models

Bayesian linear regression model

Data: $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$
Model: $y_1, \ldots y_n$ independent

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right),$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$ is the vector of regression coefficients.
Deep learning linear regression: $y_i = \mathbf{W}'\mathbf{x}_i + \varepsilon_i$, i.e. $\boldsymbol{\beta} = \mathbf{W}'$

Conjugate priors:

$$\boldsymbol{\beta}|\sigma^2 \sim N_d(\mathbf{b}_0, \sigma^2 \mathbf{B}_0), \qquad \sigma^2 \sim \mathcal{G}^{-1}(s_0, S_0)$$

Bayesian inference for the linear regression model

Likelihood in matrix form: Define $\mathbf{X} \in \mathbb{R}^{N \times d}$ as the design matrix with rows $\mathbf{x}_i'$:

$$p(\mathbf{y}|\beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)\right)$$

Posterior distribution

$$\begin{aligned} p(\beta, \sigma^2|\mathbf{y}) &\propto p(\mathbf{y}|\beta, \sigma^2)p(\beta, \sigma^2) \\ &\propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)\right) \cdot \\ &\quad (\sigma^2)^{-d/2} \exp\left(-\frac{1}{2\sigma^2}(\beta - \mathbf{b}_0)^T\mathbf{B}_0^{-1}(\beta - \mathbf{b}_0)\right) \cdot \\ &\quad (\sigma^2)^{-(s_0+1)} \exp(-S_0/\sigma^2) \end{aligned}$$

$$\Longrightarrow p(\beta, \sigma^2|\mathbf{y}) = p(\beta|\sigma^2, \mathbf{y})p(\sigma^2|\mathbf{y})$$

Bayesian inference for the linear regression model

- Conditional posterior of $\beta|\sigma^2, \mathbf{y}$:

$$p(\beta|\sigma^2, \mathbf{y}) \propto \exp\left(-\frac{1}{2\sigma^2}(\beta^T(\mathbf{X}'\mathbf{X} + \mathbf{B}_0^{-1})\beta - 2\beta^T(\mathbf{X}^T\mathbf{y} + \mathbf{B}_0^{-1}\mathbf{b}_0))\right)$$

and therefore

$$\beta|\sigma^2, \mathbf{y} \sim N_d(\mathbf{b}_n, \sigma^2\mathbf{B}_n)$$

with parameters

$$\mathbf{B}_n^{-1} = \mathbf{B}_0^{-1} + \mathbf{X}^T\mathbf{X}, \qquad \mathbf{b}_n = \mathbf{B}_n(\mathbf{B}_0^{-1}\mathbf{b}_0 + \mathbf{X}^T\mathbf{y})$$

Bayesian inference for the linear regression model

- Marginal posterior of $\sigma^2|\mathbf{y}$:

$$p(\sigma^2|\mathbf{y}) \propto (\sigma^2)^{-(s_0+n/2+1)} \exp\left(-\frac{1}{\sigma^2}(S_0 + S_{\mathbf{y}}/2)\right)$$

where

$$S_{\mathbf{y}} = \mathbf{y}'\mathbf{y} + \mathbf{b}_0'\mathbf{B}_0^{-1}\mathbf{b}_0 - \mathbf{b}_n'\mathbf{B}_n^{-1}\mathbf{b}_n$$

and therefore

$$\sigma^2|\mathbf{y} \sim \mathcal{G}^{-1}(s_n, S_n)$$

with parameters

$$s_n = s_0 + n/2, \quad S_n = S_0 + S_{\mathbf{y}}/2$$

Bayesian Logit Model

- **Data**: $(y_i, \mathbf{x}_i), i = 1, \ldots, n$
- **Model**
  - $y_1, \ldots y_n$ independent with

$$p(y_i = 1|\beta) = \frac{\exp(\mathbf{x}_i'\beta)}{1 + \exp(\mathbf{x}_i'\beta)} \qquad p(y_i = 0|\beta) = \frac{1}{1 + \exp(\mathbf{x}_i'\beta)}$$

Deep learning logit model:

$$p(y_i = 1|\beta) = \frac{1}{1 + \exp(-\mathbf{x}_i'\beta)} = \frac{\exp(\mathbf{x}_i'\beta)}{\exp(\mathbf{x}_i'\beta)(1 + \exp(-\mathbf{x}_i'\beta))} =$$
$$= \frac{\exp(\mathbf{x}_i'\beta)}{1 + \exp(\mathbf{x}_i'\beta)}$$

Bayesian Logit Model

- Model
  - Likelihood:

$$p(\mathbf{y}|\beta) = \prod_{i=1}^{n} p(y_i|\beta)$$

  - Prior distribution

$$\beta \sim \mathcal{N}(\mathbf{b}_0, \mathbf{B}_0)$$

- **Posterior distribution**

$$p(\beta|\mathbf{y}) \propto p(\mathbf{y}|\beta) \exp\left(-\frac{1}{2}(\beta - \mathbf{b}_0)' \mathbf{B}_0^{-1} (\beta - \mathbf{b}_0)\right)$$

has no closed form

Posterior simulation

Bayesian Inference for Complex Models

- the posterior is available in closed form only in special cases (conjugate analysis)
- usually approximation of the posterior is required
- the posterior is a distribution $\Longrightarrow$ simulation based approximation feasible
  - simulate $\theta^{(1)}, \ldots \theta^{(M)}$ from the posterior $p(\theta|\mathbf{y})$
  - summarize samples from the posterior by descriptive methods
  - approximate characteristics from the posterior by sample statistics e.g. posterior expectation can be approximated by the sample mean

Note: The posterior is completely determined by likelihood and prior!

Posterior Simulation

algorithms for sampling from complex distributions

- importance sampling
- MCMC methods
  - Gibbs sampling
  - data augmentation
  - Metropolis Hastings algorithm
- ABC (approximate Bayesian computation)
- VB (Variational Bayes)

Marginal posterior of the mean in the Normal model

Model:

$$y_i = \mu + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}\left(\mu, \sigma^2\right),$$

two options if interest is only in μ

- integrate over σ^2

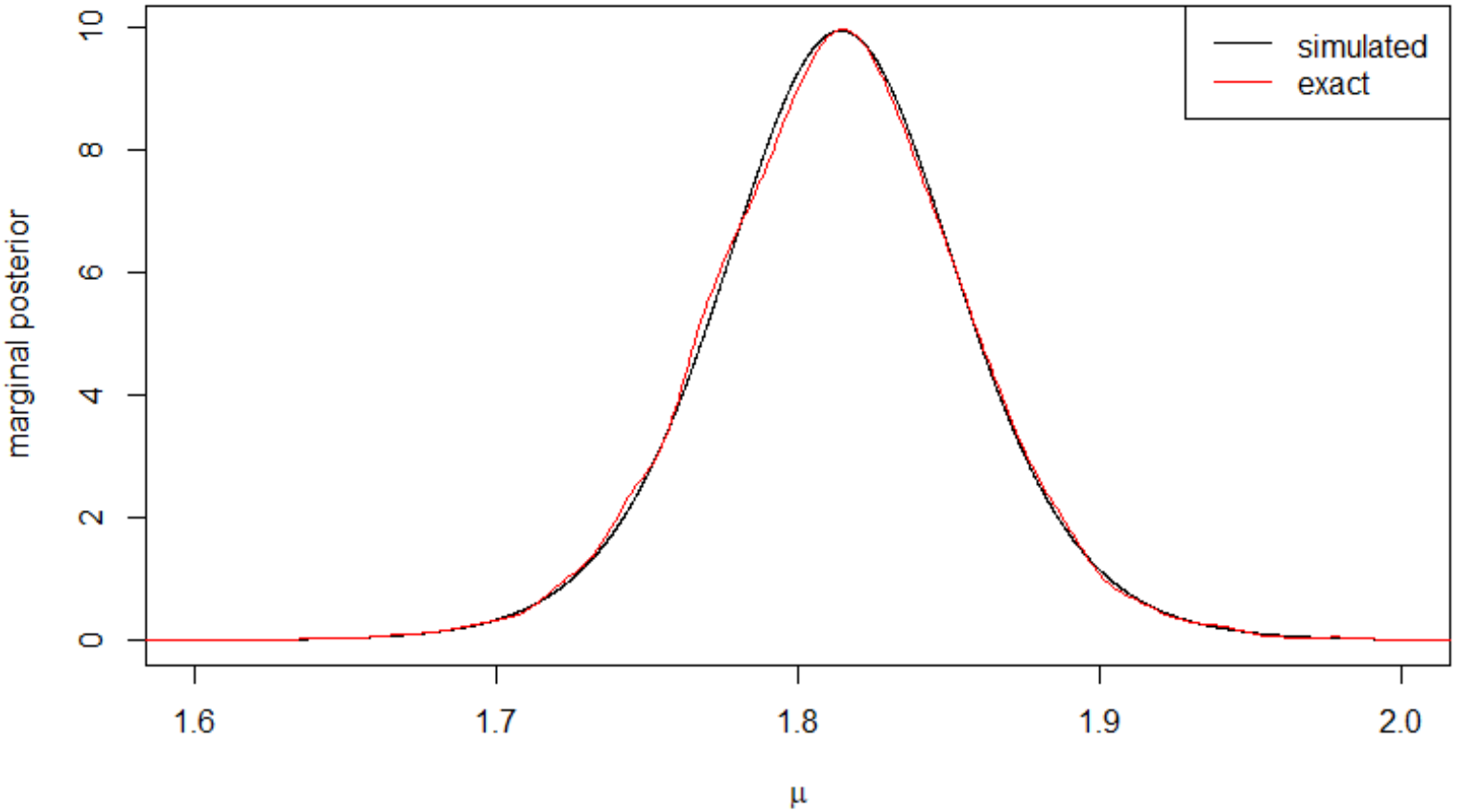
$$p(\mu|\mathbf{y}) = \int p(\mu, \sigma^2|\mathbf{y}) d\sigma^2$$

$\Longrightarrow$ scaled t-distribution

- sample from the posterior
  for m=1, . . . , M
  - sample $(\sigma^2)^{(m)}$ from $p(\sigma^2|\mathbf{y})$
  - sample $(\mu)^{(m)}$ from $p(\mu|(\sigma^2)^{(m)}, \mathbf{y})$

  use $(\mu)^{(1)}, \ldots, (\mu)^M$ to approximate the posterior/interesting quantity from the posterior

Midge data: Marginal posterior of the mean



posterior expectation: exact: 1.814, approximated: 1.8138

MCMC methods

Concept of MCMC methods

- a Markov chain converges to its stationary distribution if it is irreducible, positive recurrent and aperiodic
- this result can be used to to sample from a distribution p
  - generate a Markov chain $Y = \{Y_n, n \geq 0\}$ with p as its stationary distribution
  - after a burn-in period the realizations Y_n of this Markov chain are dependent draws from p
  - use these draws to approximate the distribution p or interesting quantities of p (e.g. the mean)
- MCMC methods
  - Gibbs sampling
  - Data augmentation
  - Metropolis Hastings algorithm

Semiconjugate prior distribution

In the Normal model

$$y_i = \mu + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}\left(\mu, \sigma^2\right),$$

the parameters can be assumed to be a priori independent

$$p(\mu, \sigma^2) = p(\mu)p(\sigma^2)$$

with

$$\mu \sim \mathcal{N}\left(m_0, M_0\right) \qquad \text{and} \qquad \sigma^2 \sim \mathcal{G}^{-1}\left(s_0, S_0\right)$$

But: the posterior is not of closed form
$\Longrightarrow$ use Gibbs sampling

MCMC methods: Gibbs Sampling

Gibbs Sampling

- Gibbs sampling
  - is a method to sample from a p-variate distribution

$$p(\mathbf{y}) = p(y_1, \ldots, y_p)$$

  - in one sweep of the sampler components of $\mathbf{y}$ are updated drawing from the full conditionals

$$p(y_j | y_{\setminus j})$$

    where each component of $y_{\setminus j} = (y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_p)$ is fixed at the most recent draw

- for Bayesian inference sampling from the joint posterior of a multivariate parameter $p(\boldsymbol{\theta}|\mathbf{y})$ is required

Gibbs Sampling from the posterior

Interest in the p-variate posterior $p(\boldsymbol{\theta}|\mathbf{y})$

Algorithm

Choose starting values for $\theta_2^{(0)}, \ldots, \theta_p^{(0)}$ and repeat for $m = 1, \ldots, M$:

- Draw $\theta_1^{(m)}$ from $p(\theta_1|\theta_2^{(m-1)}, \ldots, \theta_p^{(m-1)}, \mathbf{y})$,
- Draw $\theta_2^{(m)}$ from $p(\theta_2|\theta_1^{(m)}, \theta_3^{(m-1)}, \ldots, \theta_p^{(m-1)}, \mathbf{y})$

$$\vdots$$

- Draw $\theta_p^{(m)}$ from $p(\theta_p|\theta_1^{(m)}, \ldots, \theta_{p-1}^{(m)}, \mathbf{y})$.

Full conditionals in the Bayesian Normal model

- full conditional for μ

$$p(\mu|\sigma^2, \mathbf{y}) \propto \exp\left(-\frac{1}{2\sigma^2}\sum(y_i-\mu)^2 - \frac{1}{2M_0}(\mu-m_0)^2\right)$$

i.e.

$$\mu|\sigma^2, \mathbf{y} \sim \mathcal{N}\left((\frac{n}{\sigma^2}+\frac{1}{M_0})^{-1}(\sum y_i/\sigma^2 + m_0/M_0),\ (\frac{n}{\sigma^2}+\frac{1}{M_0})^{-1}\right)$$

- full conditional for σ^2

$$p(\sigma^2|\mu, \mathbf{y}) \propto \left(\frac{1}{\sigma^2}\right)^{s_0+n/2+1} \exp\left(-\frac{S_0+\sum(y_i-\mu)^2/2}{\sigma^2}\right),$$

i.e.

$$\sigma^2|\mu, \mathbf{y} \sim \mathcal{G}^{-1}\left(s_0+n/2, S_0+(\sum(y_i-\mu)^2)/2\right)$$

Gibbs Sampling Steps

Gibbs-Sampling Scheme:

Initialization: Choose a starting value for $\sigma^{2,(0)}$

Repeat for $m = 1, \ldots, M$

1. Draw $\mu^{(m)}$ from $\mathcal{N}(m_n, M_n)$ where

$$M_n = (n/(\sigma^2)^{(m-1)} + 1/M_0)^{-1}$$
$$m_n = M_n(\sum y_i/(\sigma^2)^{(m-1)} + m_0/M_0)$$

2. Draw $(\sigma^2)^{(m)}$ from $\mathcal{G}^{-1}(s_n, S_n)$ where

$$s_n = s_0 + n/2$$
$$S_n = S_0 + (\sum (y_i - \mu^{(m)})^2)/2$$

Blocked Gibbs-Sampling

- components can be blocked, i.e. sampled jointly
- Example: Normal regression model
  Semi-conjugate prior

$$\beta \sim N_d(\mathbf{b}_0, \mathbf{B}_0), \qquad \sigma^2 \sim \mathcal{G}^{-1}(s_0, S_0)$$

Initialization: Chose a starting value for $(\sigma^2)^{(0)}$
Repeat for $m = 1, \ldots, M$:

1. Sample $\beta^{(m)}$ from $\mathcal{N}(\mathbf{b}_n, \mathbf{B}_n)$ where

$$\mathbf{B}_n = (\mathbf{X}'\mathbf{X}/(\sigma^2)^{(m-1)} + \mathbf{B}_0^{-1})^{-1}$$
$$\mathbf{b}_n = \mathbf{B}_n(\mathbf{X}'\mathbf{y}/(\sigma^2)^{(m-1)} + \mathbf{B}_0^{-1}\mathbf{b}_0)$$

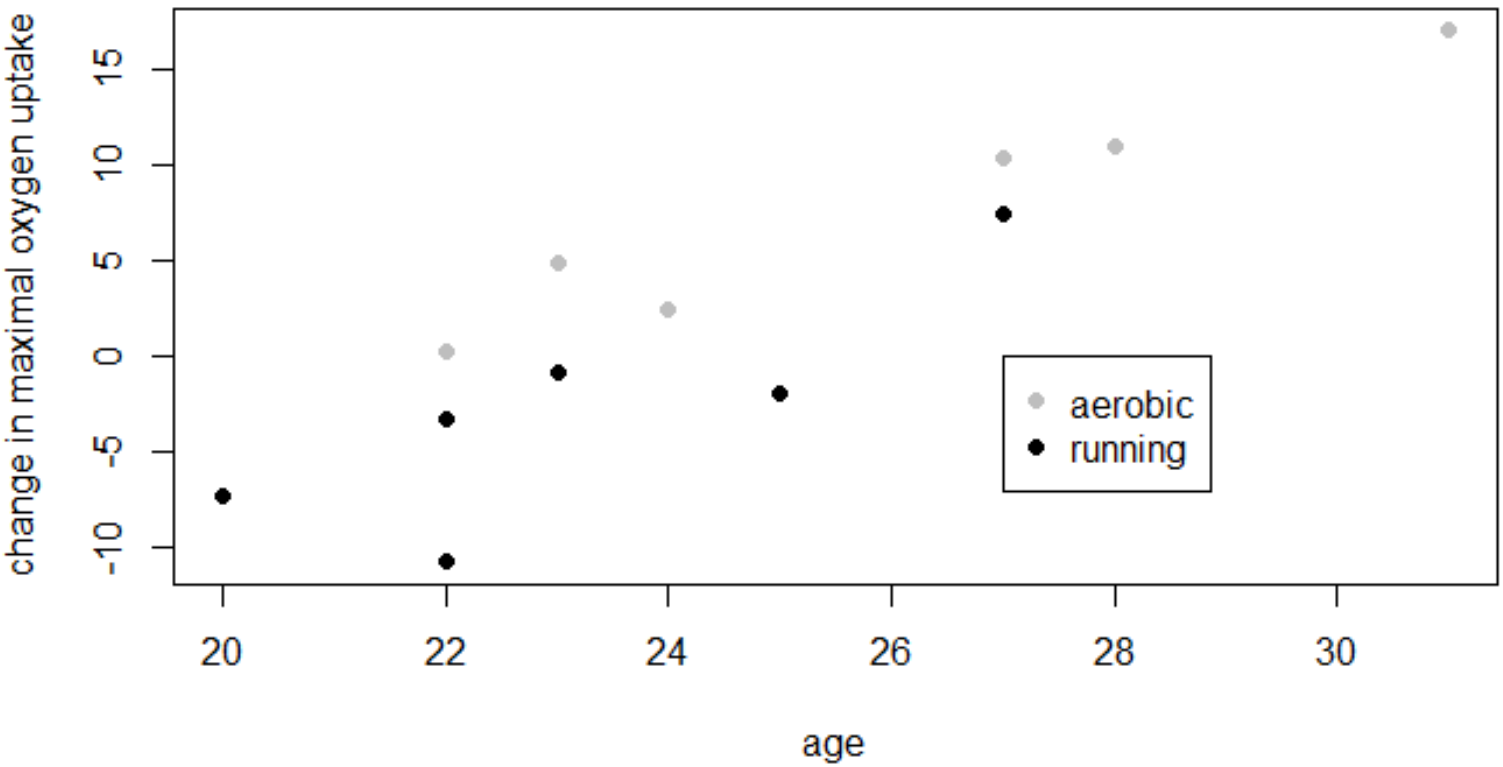
2. Sample $(\sigma^2)^{(m)}$ from $\mathcal{G}^{-1}(s_n, S_n)$ where

$$s_n = s_0 + n/2 \quad \text{and} \quad S_n = S_0 + (\mathbf{y} - \mathbf{X}\beta^{(m)})'(\mathbf{y} - \mathbf{X}\beta^{(m)})/2$$

Example: Oxygene uptake

- randomized study to compare effects of two different exercise regimens
  - six men assigned to a 12-week flat-terrain running program
  - six men assigned to a 12-week step aerobics
- outcome Y maximum oxygen uptake of each subject (measured in liters per minute) in training during and after the program
- interest in effect of `training program`
- `age` might be a confounding factor

Oxygene uptake: Data



Oxygene uptake: Regression modelling

- regression modelling

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i \qquad \varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

  with
  - $x_{i1} = 1 \Longrightarrow \beta_0$ is the intercept
  - $x_{i2} = 1$ if i is assigned to the aerobic program (0 if running program)
  - x_{i3}: age of the subject
  - $x_{i4} = x_{i2} x_{i3}$: interaction of age and program type
- implications of the model

$$\begin{aligned} \mathrm{E}(y_i|\mathbf{x}) &= \beta_1 + \beta_3 \cdot \texttt{age} && \text{if } x_{i2} = 0 \text{ (running)} \\ \mathrm{E}(y_i|\mathbf{x}) &= (\beta_1 + \beta_2) + (\beta_3 + \beta_4) \cdot \texttt{age} && \text{if } x_{i2} = 1 \text{ (aerobic)} \end{aligned}$$

Oxygene uptake: Priors

- uninformative prior for β

$$\beta \sim \mathcal{N}\left(\mathbf{0}, 1000^2\mathbf{I}\right)$$

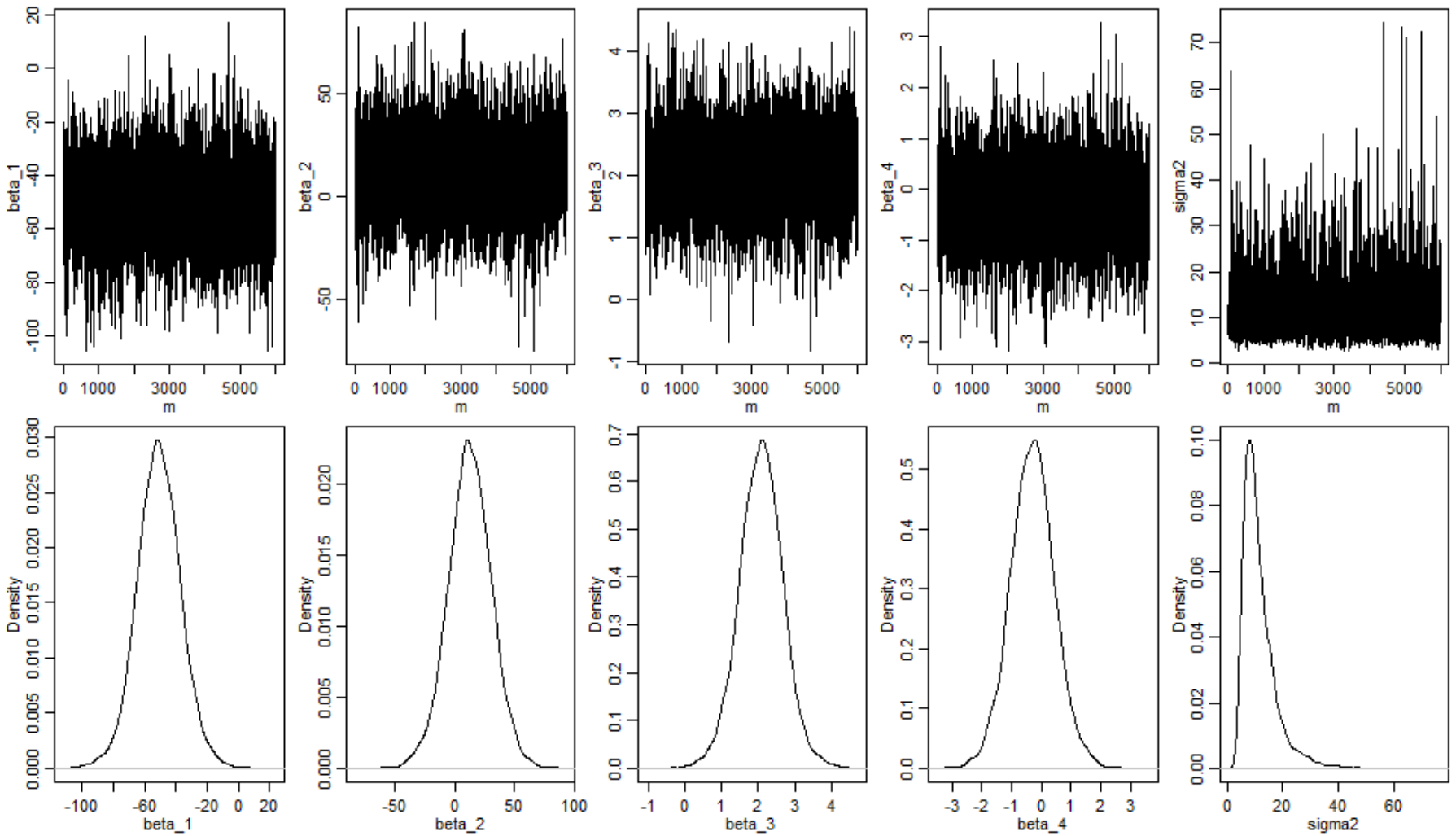
- prior for σ^2 with mode at 6

$$\sigma^2 \sim \mathcal{G}^{-1}(1, 12)$$

Starting value for σ^2: variance of the residuals in linear regression

$$(\sigma^2)^{(0)} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \mathbf{X}_i\hat{\beta})^2$$

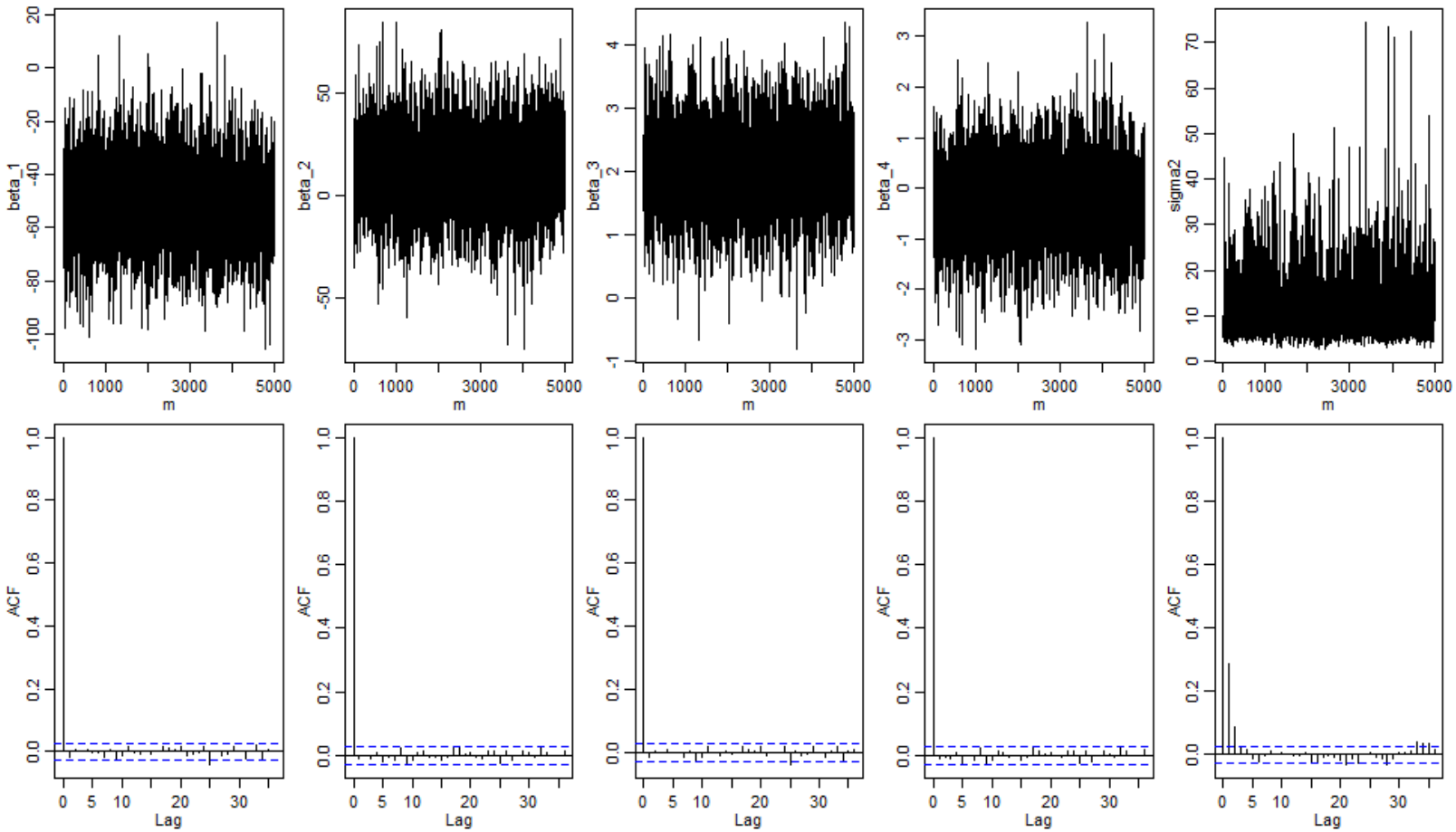
Oxygene uptake: Samples from the posterior



MCMC diagnostics

- optimal method for sampling from the posterior should
  - explore the whole posterior distribution and move around quickly (i.e. mix fast)
  - yield independent draws from the posterior
- MCMC methods produce
  - samples from the posterior after convergence
    $\Longrightarrow$ inspect trace plots and discard samples from burnin
  - correlated draws from the posterior
    $\Longrightarrow$ inspect ACF (autocorrelation function)
  - compute effective sample size (ESS): corresponding number of independent draws
- (almost) uncorrelated draws can be obtained by thinning of the MCMC output - i.e. keep only every k-th draw

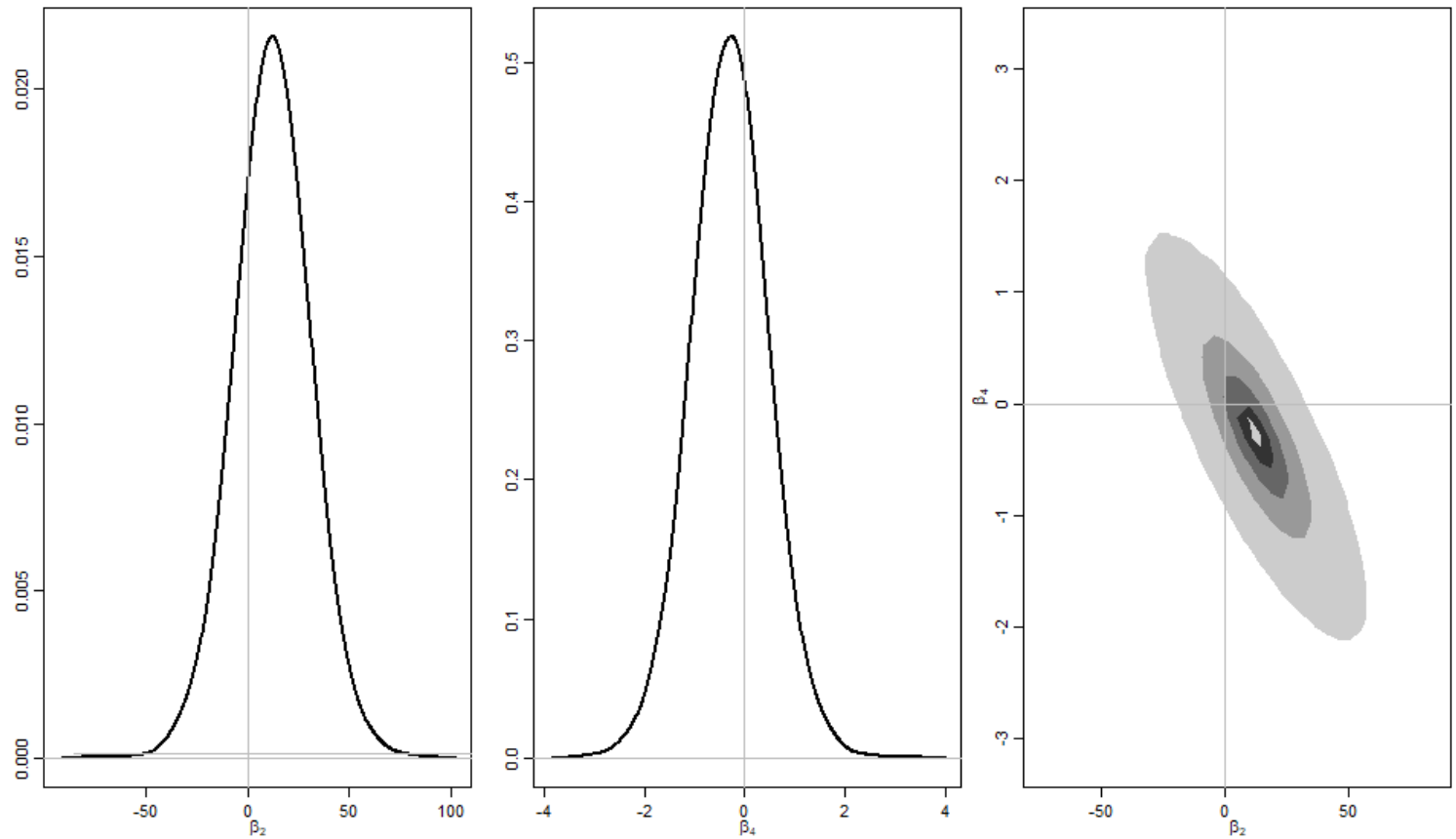
Oxygene uptake: MCMC diagnostics



burnin of 1000 draws discarded $\Longrightarrow$ 5000 MCMC draws

effective Sample Size: almost 5000 for $\beta_1, \ldots \beta_4$; around 2500 for σ^2

Oxygene uptake: Effect of aerobic vs. running



β_2: additional effect of aerobic (vs. running) $\Longrightarrow$ positive effect of aerobic

β_4: modification of age effect for aerobic $\Longrightarrow$ positive effect of aerobic decreases slightly with age

MCMC methods: Data Augmentation

Pima Indian data

Pima Indian Data on Diabetes: `Pima.tr` (R-package `MASS`)

- **Data**: for $n = 200$ women
  `npreg`: number of pregnancies
  `glu`: plasma glucose concentration in an oral glucose tolerance test
  `bp`: diastolic blood pressure (mm Hg)
  `skin`: triceps skin fold thickness (mm)
  `bmi`: body mass index (weight in kg/(height in m^2))
  `ped`: diabetes pedigree function
  `age`: age in years
  `type`: Yes or No, for diabetic according to WHO criteria
- response variable: $Y = 1$ (`type= Yes`) or $Y = 0$ (`type= No`)
- Goal: model $P(Y = 1)$ conditional on covariates x_1 (`npreg`), ..., x_7 (`age`)

Probit Model

- **data:** $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$ with $y_i \in \{0, 1\}$
- **model**
  linear regression model not appropriate for a binary response $\Longrightarrow$ probit model
  - $y_1, \ldots, y_n$ independent with $P(y_i = 1) = \Phi(\mathbf{x}_i'\beta)$
    $\Longrightarrow$ likelihood:

$$p(y_1, \ldots, y_n|\beta) = \prod \left(\Phi(\mathbf{x}_i'\beta)\right)^{y_i} \left(1 - \Phi(\mathbf{x}_i'\beta)\right)^{1-y_i}$$

  - prior: $\beta \sim \mathcal{N}(\mathbf{b}_0, \mathbf{B}_0)$
- **posterior** is not of closed form, as

$$p(\beta|\mathbf{y}) \propto \prod_{i=1}^{n} \left(\Phi(\mathbf{x}_i'\beta)\right)^{y_i} \left(1 - \Phi(\mathbf{x}_i'\beta)\right)^{1-y_i} \exp\left(-\frac{1}{2}(\beta - \mathbf{b}_0)'\mathbf{B}_0^{-1}(\beta - \mathbf{b}_0)\right)$$

Data Augmentation

Problem: posterior of θ not of closed form

- Idea: Introduce auxiliary variables $\mathbf{z}$ such that the posterior $p(\theta|\mathbf{y})$ of interest is the marginal density of the joint posterior $p(\theta, \mathbf{z}|\mathbf{y})$
- Implementation
  - sample $\theta|\mathbf{z}, \mathbf{y}$
  - add a further step to sample $\mathbf{z}|\theta, \mathbf{y}$
- data augmentation is useful if
  - the full conditional distribution $p(\mathbf{z}|\theta, \mathbf{y})$
  - and the full conditional(s) $p(\theta|\mathbf{z}, \mathbf{y})$

  are of closed form and/or easy to sample

Data Augmentation: Algorithm

Interest in the p-variate posterior $p(\theta|\mathbf{y})$

Algorithm

Initialization: Choose a starting values $\theta^{(0)}$ and repeat for m=1,. . . , M

- sample $\mathbf{z}^{(m)}$ from $p(\mathbf{z}|\theta_1^{(m-1)}, \ldots, \theta_p^{(m-1)}, \mathbf{y})$,
- sample $\theta_1^{(m)}$ from $p(\theta_1|\theta_2^{(m-1)}, \ldots, \theta_p^{(m-1)}, \mathbf{z}^{(m)}, \mathbf{y})$,
- . . .
- sample $\theta_p^{(m)}$ from $p(\theta_p|\theta_1^{(m)}, \ldots, \theta_{p-1}^{(m)}, \mathbf{z}^{(m)}, \mathbf{y})$

Results are draws from $p(\theta, \mathbf{z}|\mathbf{y})$
$\Longrightarrow$ use $\theta^{(1)} \ldots, \theta^{(M)}$ to approximate $p(\theta|\mathbf{y})$

NOTE: The sampler can also start with initialization of $\mathbf{z}$

Data augmentation for the probit model

The model

$$z_i \sim \mathcal{N}\left(\mathbf{x}_i'\beta, 1\right)$$
$$y_i = 1_{(0,\infty)}(z_i),$$

where $(z_1, y_1) \ldots, (z_n, y_n)$ are independent, has the likelihood

$$p((z_1, y_1), \ldots, (z_n, y_n)|\beta) = \prod_{i:y_i=0} f_N(z_i; \mathbf{x}_i'\beta, 1)1_{(-\infty,0)}(z_i) \cdot \prod_{i:y_i=1} f_N(z_i; \mathbf{x}_i'\beta, 1)1_{(0,\infty)}(z_i)$$

Data augmentation for the probit model

Marginalizing over $z_1, \ldots, z_n$ yields

$$p(y_1, \ldots, y_n|\beta) = \prod_{i:y_i=0} \int_{-\infty}^{0} f_N(z_i; \mathbf{x}_i'\beta, 1) dz_i \prod_{i:y_i=1} \int_{0}^{\infty} f_N(z_i; \mathbf{x}_i'\beta, 1) dz_i =$$

$$= \prod_{i:y_i=0} \Phi(-\mathbf{x}_i'\beta) \prod_{i:y_i=1} (1 - \Phi(-\mathbf{x}_i'\beta)) =$$

$$= \prod_{i=1}^{n} \left(\Phi(\mathbf{x}_i'\beta)\right)^{y_i} \left(1 - \Phi(\mathbf{x}_i'\beta)\right)^{1-y_i}$$

Data augmentation with latent variables, the latent utilities $z_i, i = 1, \ldots, n$ yields a marginal probit model

Probit Model: full conditional distribution of the latent utilities

$$p(z_1, \ldots, z_n|\beta, \mathbf{y}) \propto p(\mathbf{y}|\beta, z_1, \ldots, z_n)p(z_1, \ldots, z_n|\beta)$$
$$= \prod_{i=1}^{n} p(y_i|z_i)p(z_i|\beta)$$

$\Longrightarrow$ $z_1, \ldots, z_n$ are independent conditional on β and

$$p(z_i|\beta, y_i) \propto \begin{cases} \frac{1}{\sqrt{2\pi}} \exp(-(z_i - \mathbf{x}_i'\beta)^2/2) \cdot 1_{(0,\infty)}(z_i) & \text{for } y_i = 1 \\ \frac{1}{\sqrt{2\pi}} \exp(-(z_i - \mathbf{x}_i'\beta)^2/2) \cdot 1_{(-\infty,0)}(z_i) & \text{for } y_i = 0 \end{cases}$$

Probit Model: full conditionals

1. full conditional distribution of $\mathbf{z} = z_1, \ldots, z_n$: independent with

$$z_i|\beta, y_i \sim \mathcal{N}\left(\mathbf{x}_i'\beta, 1\right) \quad \text{censored to} \quad \begin{cases} (-\infty, 0) & \text{if } y_i = 0 \\ (0, \infty) & \text{if } y_i = 1 \end{cases}$$

2. full conditional of β:

$$p(\beta|\mathbf{z}, \mathbf{y}) \propto p(\mathbf{z}|\beta)p(\beta)$$

posterior of β in linear regression model with response vector $\mathbf{z}$

MCMC sampling for the probit model

Choose a starting value for β and iterate for $m = 1, \ldots, M$:

1. Sample z_i, $i = 1, \ldots, n$ as

$$z_i|\beta, y_i \sim \mathcal{N}\left(\mathbf{x}_i'\beta, 1\right) \quad \text{censored to} \quad \begin{cases} (-\infty, 0) & \text{if } y_i = 0 \\ (0, \infty) & \text{if } y_i = 1 \end{cases}$$

2. Sample β from the posterior of the regression model for $\mathbf{z}$

Pima indian data: Modelling the data

- probit model
- reference values

`npreg` (pregnancies)	0
`glu` (glucose concentration)	100
`bp` (diastolic blood pressure)	80
`skin` (skin fold thickness)	23
`bmi`	25
`ped` (diabetes pedigree)	0.25
`age` (in years)	20

Pima indian data: Bayesian Inference

- uninformative prior for β

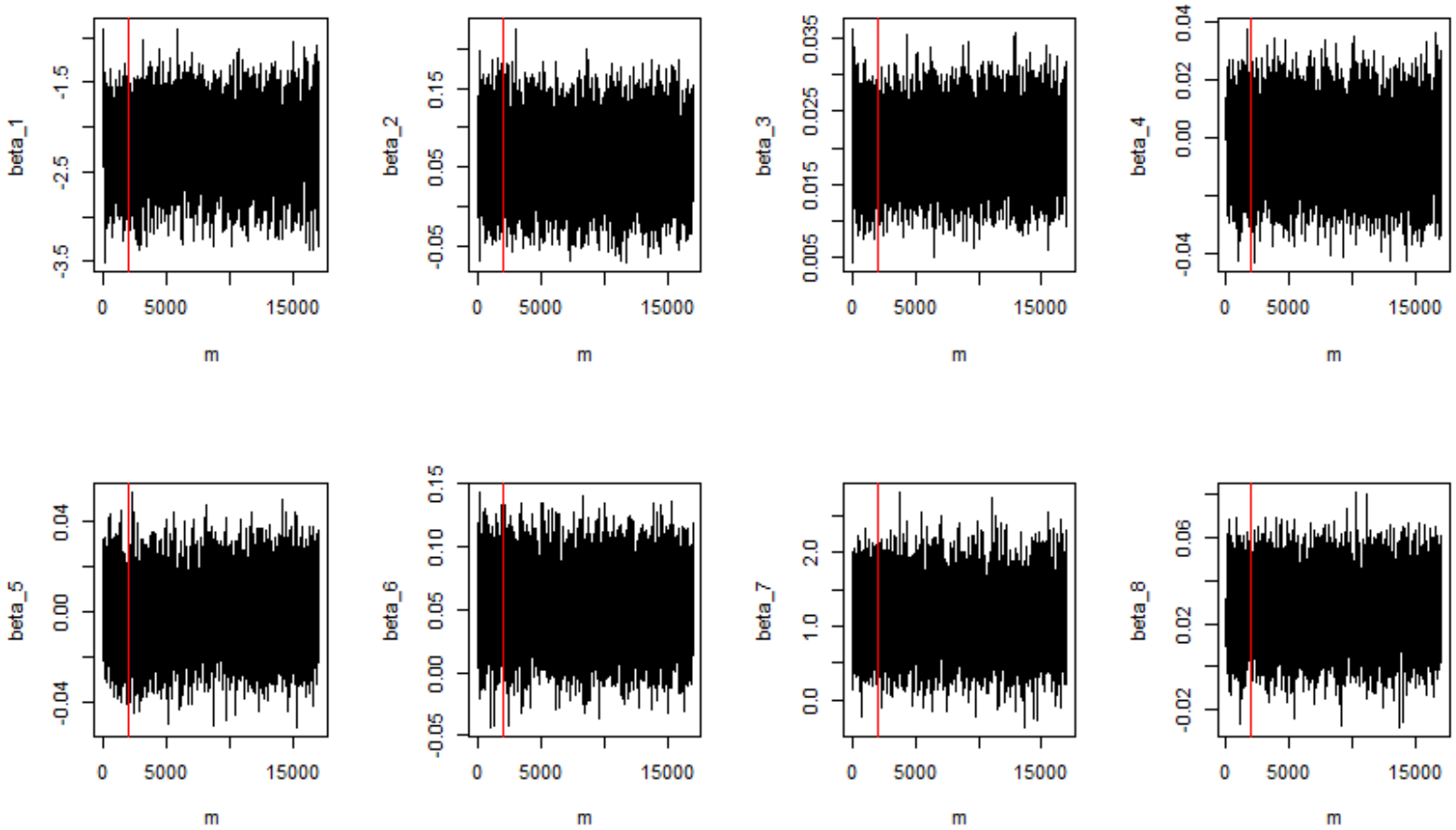
$$\beta \sim \mathcal{N}\left(\mathbf{0}, 100^2\mathbf{I}\right)$$

- Starting values for β

$$\beta = \mathbf{0}$$

- MCMC
  17000 iterations (tentative burnin: 2000)

Pima indian data: Traceplots, Probit model

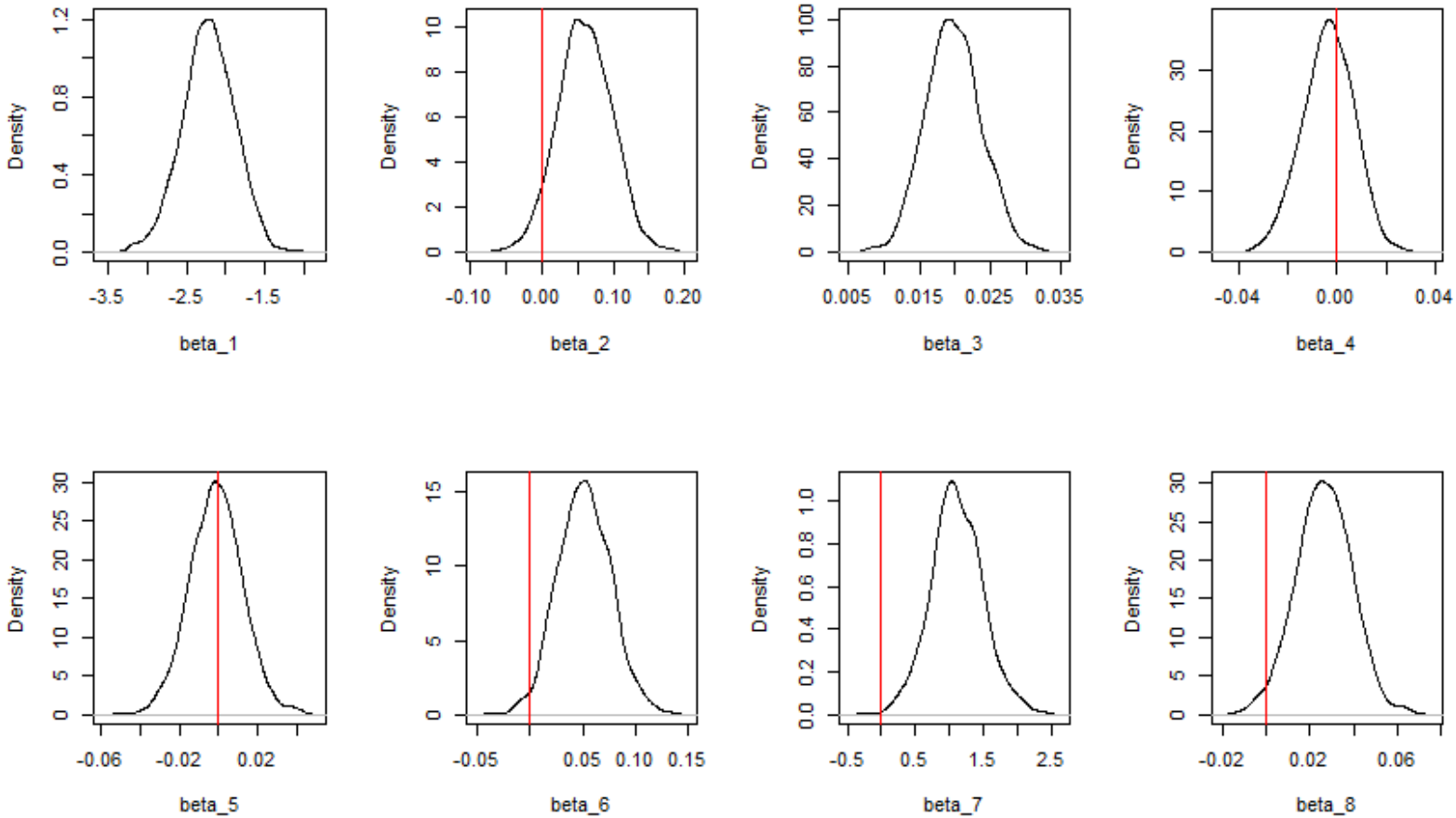


burnin of 2000 draws ok. $\Longrightarrow$ 15000 MCMC draws

effective sample sizes: minimum 2630 for β_1, maximum 4814 for β_2

$\Longrightarrow$ thinning: use every 10th draw

Pima indian data: Kernel density estimates, Probit model



based on 1500 samples (from the thinned chain)

Pima indian data: Results

thinned sample: 1500 draws

	Q2.5%	Q50%	Q97.5%	mean	sd
int	-2.90	-2.22	-1.58	-2.22	0.34
npreg	-0.01	0.06	0.13	0.06	0.04
glu100	0.01	0.02	0.03	0.02	0.00
bp80	-0.02	-0.00	0.02	-0.00	0.01
skin23	-0.03	-0.00	0.03	-0.00	0.01
bmi25	0.00	0.05	0.10	0.05	0.02
ped025	0.36	1.09	1.91	1.11	0.38
age20	0.00	0.03	0.05	0.03	0.01

MCMC methods: Metropolis Hastings Algorithm

Pima Indian data: Logit model

- in medical applications for binary responses usually the logit model is preferred
- in the logit model the probability for $y = 1$ (disease) given covariates $\mathbf{x}$ is specified as

$$P(Y = 1|\mathbf{x}) = \frac{\exp(\mathbf{x}'\beta)}{1 + \exp(\mathbf{x}'\beta)}$$

- interpretation of regression effects
  - β_j is the effect of increasing X_j by 1 on the log odds ratio
  - $\exp(\beta_j)$ is the effect of increasing X_j by 1 on the odds ratio
- risk for $y = 1$ is increases with X_j if $\beta_j > 0$ and decreases with $\beta_j > 0$

The odds of an event A is given as $\frac{P(A)}{1-P(A)}$ the odds ratio of two events A and B is

$$\frac{P(A)/(1 - P(A))}{P(B)/(1 - P(B))}$$

Bayesian inference for the logit model

- with a Normal prior on the regression effects β the posterior is not of closed form
- the logit model also has a representation as a latent utility model: it results when the latent utility has a logistic distribution with mean $\mathbf{x}'\beta$
- But: also the logistic likelihood combined with a Nomal prior does not yield a closed form posterior

$\Longrightarrow$ Posterior inference with the Metropolis-Hastings - Algorithm

The Metropolis-Hastings algorithm

Goal: Generate M draws from a distribution $p(y)$

- the Metropolis-Hastings algorithm generates a homogenous Markov chain with stationary distribution $p(y)$
- to move from $y^{(m-1)}$ to $y^{(m)}$ a candidate is generated from a proposal distribution $q(y^*|y^{(m-1)})$
- this value is accepted, i.e. $y^{(m)} = y^*$ with probability

$$\alpha(y^*|y^{(m-1)}) = \min\left(1, \frac{p(y^*)\, q(y^{(m-1)}|y^*)}{p(y^{(m-1)})\, q(y^*|y^{(m-1)})}\right)$$

otherwise $y^{(m)} = y^{(m-1)}$.

The Metropolis-Hastings algorithm

Algorithm

Choose a starting value $y^{(0)}$ and iterate the following steps for $m = 1, \ldots, M$

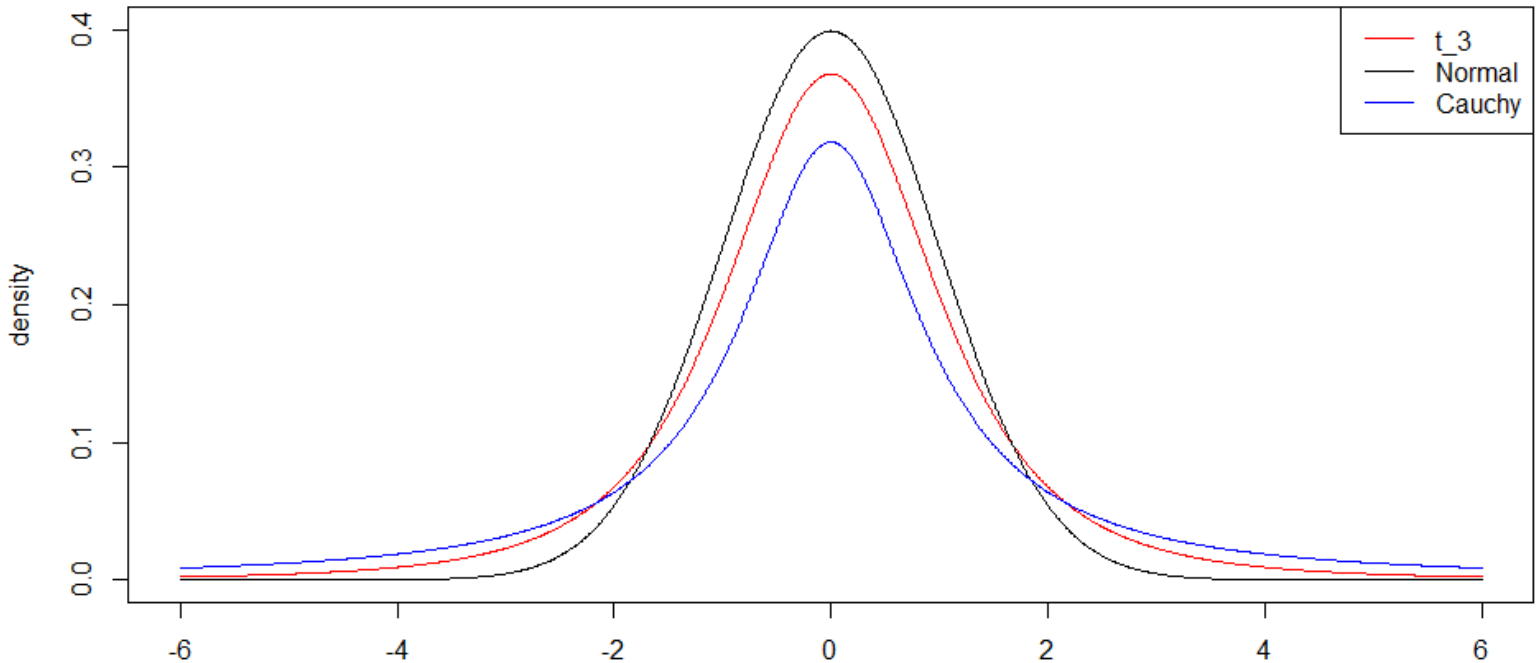
- draw a candidate y^* from the proposal distribution $q(y^*|y^{(m-1)})$
- compute the acceptance probability $\alpha(y^*|x^{(m-1)})$
- generate $u \sim \mathcal{U}([0,1])$
- set $y^{(m)} = y^*$ if $u \leq \alpha(y^*|y^{(m-1)})$, otherwise set $y^{(m)} = y^{(m-1)}$

Note: it can be numerically advantageous to compute the acceptance probability on the log-scale

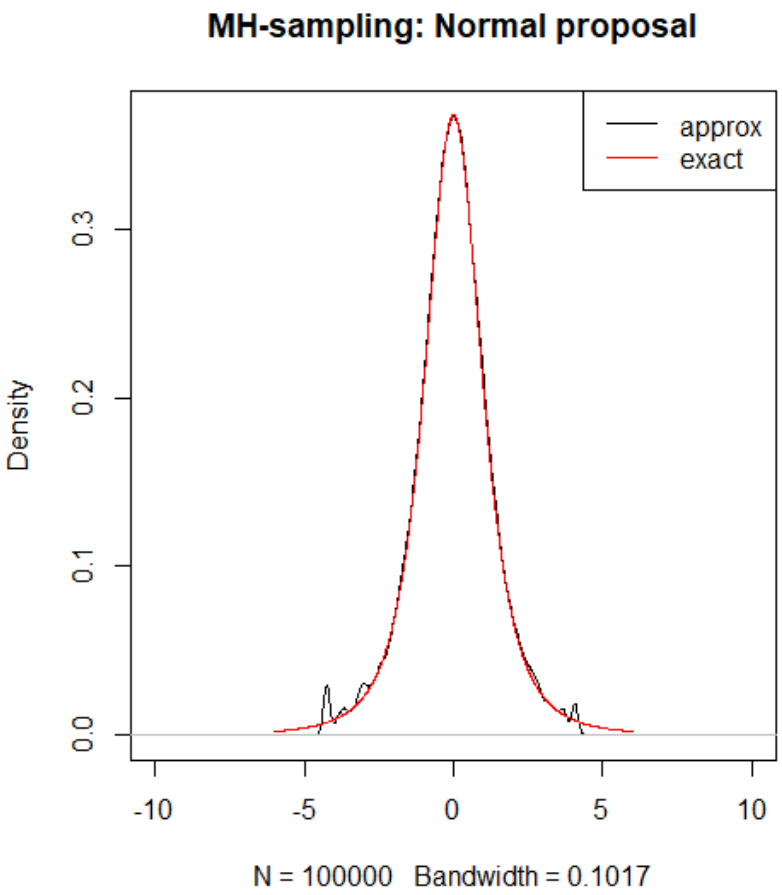
$$\log \alpha(y^*|y^{(m-1)}) = \min\Big(0, \log p(y^*) - \log p(y^{(m-1)}) + \log q(y^{(m-1)}|y^*) - \log q(y^*|y^{(m-1)})\Big)$$

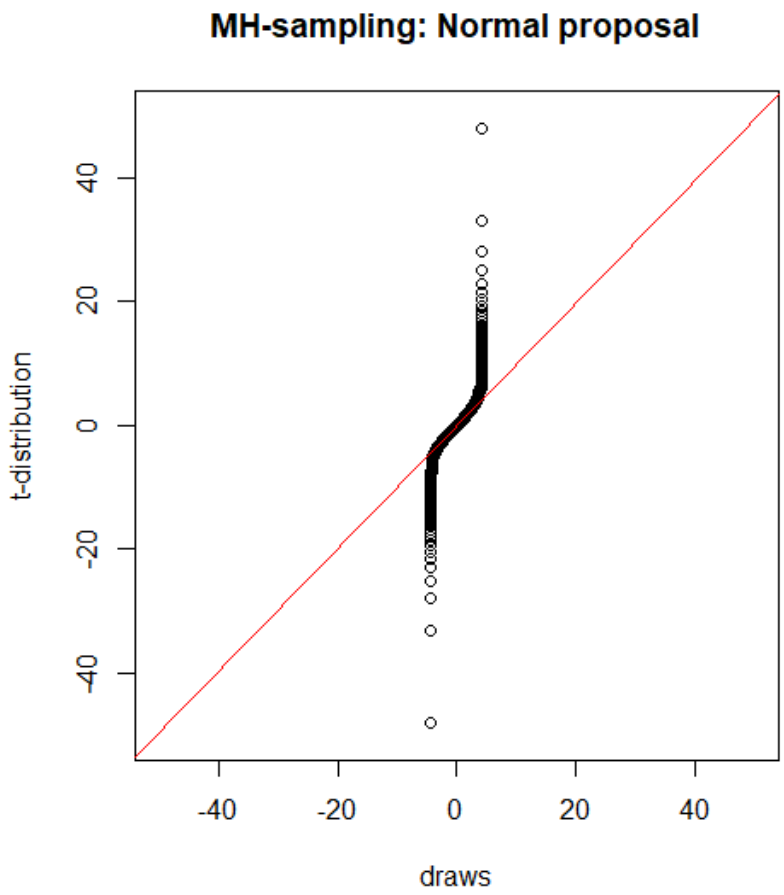
Toy Example: Sample from a t_3-distribution

- standard t_ν - distribution (Student distribution) is symmetric around zero
- ν (the degrees of freedom) determine heavyness of the tails
- special cases
  - Cauchy distribution: $\nu = 1$
  - standard Normal distribution $\nu \to \infty$



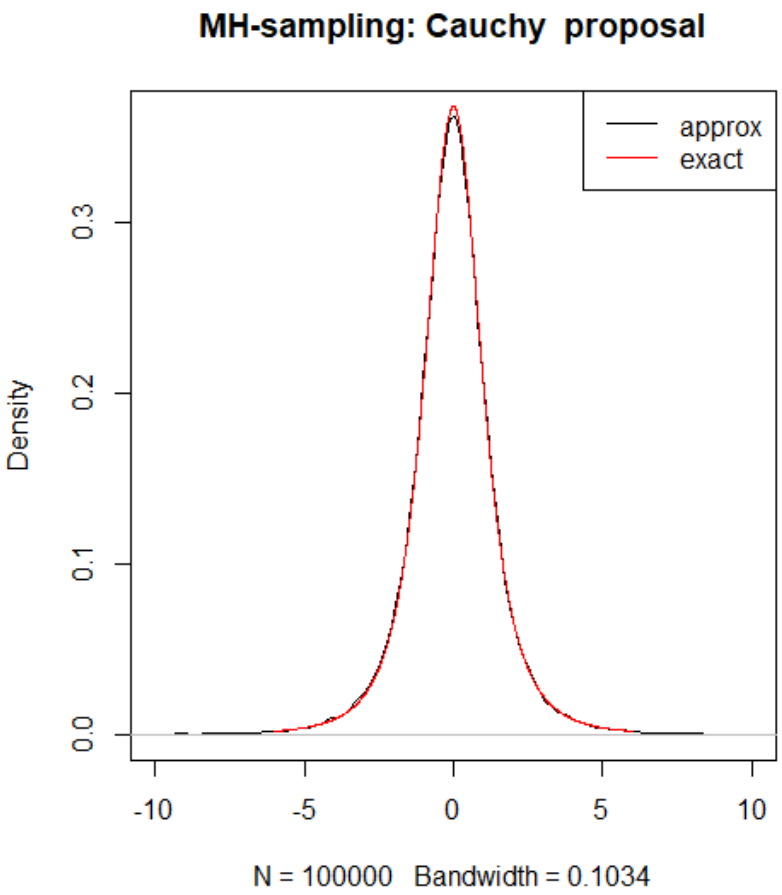
Toy Example: Normal proposal

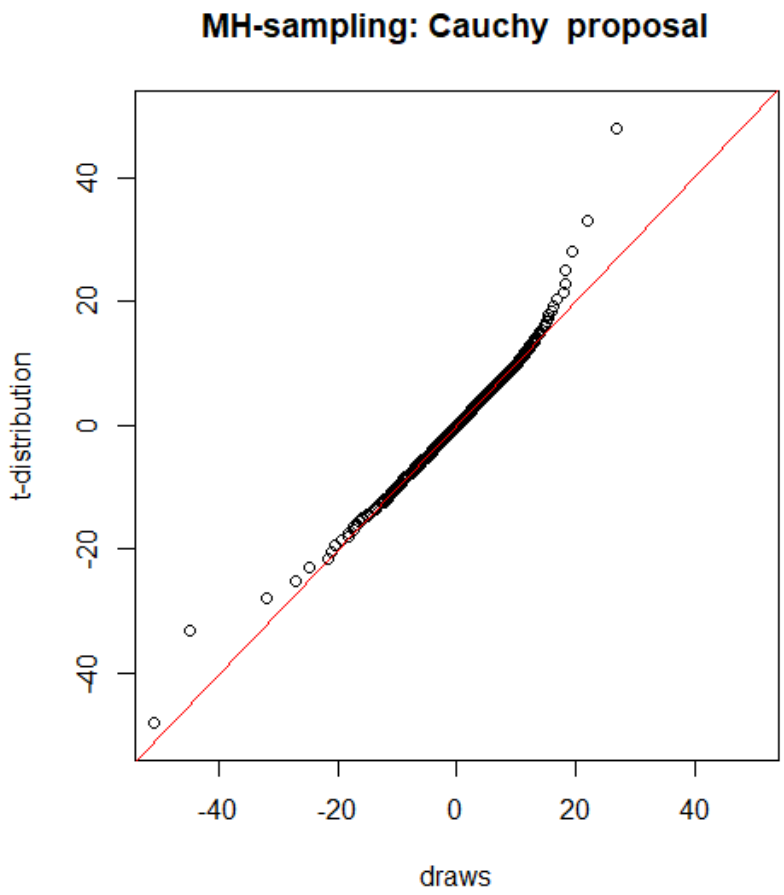




M=100.000 draws: ESS=3121, acceptance rate=89.2%

Toy Example: Cauchy proposal





M=100.000 draws: ESS=80 459, acceptance rate=81.1%

Metropolis-Hastings for posterior inference

Goal: sample from the posterior $p(\boldsymbol{\theta}^*|\mathbf{y})$

- generate a candidate $\boldsymbol{\theta}^*$ from the proposal $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{old})$
- the acceptance probability is then computed as

$$\begin{aligned}\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{old}) &= \min\left(1, \frac{p(\boldsymbol{\theta}^*|\mathbf{y})\, q(\boldsymbol{\theta}^{old}|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{old}|\mathbf{y})\, q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{old})}\right) = \\ &= \min\left(1, \frac{\frac{p(\mathbf{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{p(\mathbf{y})}\, q(\boldsymbol{\theta}^{old}|\boldsymbol{\theta}^*)}{\frac{p(\mathbf{y}|\boldsymbol{\theta}^{old})p(\boldsymbol{\theta}^{old})}{p(\mathbf{y})}\, q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{old})}\right) \\ &= \min\left(1, \frac{p(\mathbf{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)\, q(\boldsymbol{\theta}^{old}|\boldsymbol{\theta}^*)}{p(\mathbf{y}|\boldsymbol{\theta}^{old})p(\boldsymbol{\theta}^{old})\, q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{old})}\right)\end{aligned}$$

- the normalizing constant/marginal likelihood $p(\mathbf{y})$
  - is usually difficult to compute (requires integration over $\boldsymbol{\theta}$)
  - but cancels out and is not required to determine the acceptance probability

Choice of the proposal density

- goals
  - easy to sample
  - high effective sample size of the draws
    $\Longrightarrow$ low correlations of the draws
- different algorithms, e.g.
  - independence sampler
  - random walk samper

Independence Sampler

Idea: generate a proposal which does not depend on θ^{old}
(as in the MH toy-example to sample from the t_3 distribution)

- the proposal $\boldsymbol{\theta}^*$ is generated independent from $\boldsymbol{\theta}^{old}$

$$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{old}) = q(\boldsymbol{\theta}^*).$$

- the acceptance probability is then

$$\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{old}) = \min\left(1, \frac{p(\mathbf{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)\ q(\boldsymbol{\theta}^{old})}{p(\mathbf{y}|\boldsymbol{\theta}^{old})p(\boldsymbol{\theta}^{old})\ q(\boldsymbol{\theta}^*)}\right)$$

- the acceptance propability
  - is 1 if the proposal distribution $q(\boldsymbol{\theta})$ is equal to the posterior $p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$
  - is close to 1 if the proposal is close to the posterior

$\Longrightarrow$ goal: high acceptance rates

Random Walk Sampler

- the proposal is a random walk

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^{old} + \varepsilon, \quad \varepsilon \sim f(\varepsilon),$$

  and hence $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{old}) = f(\boldsymbol{\theta}^* - \boldsymbol{\theta}^{old})$.

- the acceptance probability is

$$\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{old}) = \min\left(1, \frac{p(\mathbf{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)\, f(\boldsymbol{\theta}^{old} - \boldsymbol{\theta}^*)}{p(\mathbf{y}|\boldsymbol{\theta}^{old})p(\boldsymbol{\theta}^{old})\, f(\boldsymbol{\theta}^* - \boldsymbol{\theta}^{old})}\right)$$

The random walk sampler is a Metropolis sampler if $f(\varepsilon)$ is symmetric, e.g. Normal.

Normal Random Walk Sampler

- normal random walk

$$\boldsymbol{\theta}^* \sim \mathcal{N}\left(\boldsymbol{\theta}^{old}, C\right)$$

- properties depend on the scale of $f(\varepsilon)$:
  - small variance $\Longrightarrow$ small steps $\boldsymbol{\theta}^* - \boldsymbol{\theta}^{old}$ with usually high acceptance rates, but high autocorrelations
    extreme case $C \to 0 \Longrightarrow \alpha = 1$
  - large scale $\Longrightarrow$ large steps $\boldsymbol{\theta}^* - \boldsymbol{\theta}^{old} \Longrightarrow$ proposals in the tails, low acceptance rates
- tuning of the variance/covariance necessary
- low / high acceptance rates yield highly correlated draws
  optimal acceptance rate $\approx$ 23% for a multi-dimensional parameter

Logistic regression

- **data**: $\mathbf{y} = (y_1, \ldots y_n)$
- **model**: $y_1, \ldots y_n$ iid with

$$p(y_i = 1) = \frac{\exp(\mathbf{x}_i'\beta)}{1 + \exp(\mathbf{x}_i'\beta)}$$

- Implementation, e.g.
  - tailored proposal
  - random walk proposal
- Alternative: Data augmentation based on the representation of the logit likelihood as a mixture of Normals (Polson et al., 2013)

Pima indian data: MH with tailored proposal

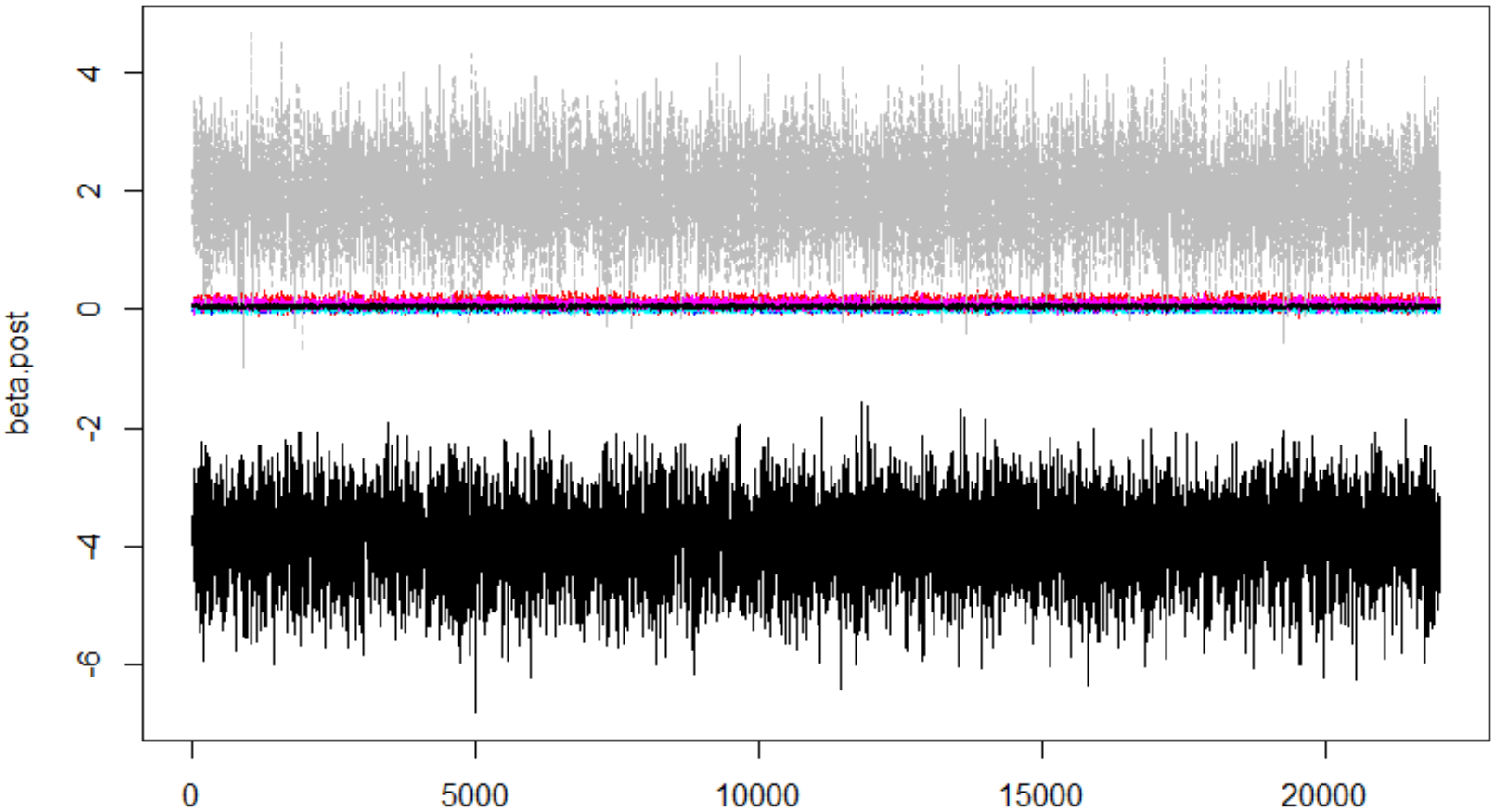
Idea: generate proposals using results from the probit model and the connection between probit and logit model

- logit and probit model result from a latent utility model: the latent utility has a standard Normal distribution in the probit and a standard logistic distribution in the logit model
- the variance is $\pi^2/3$ for the standard logistic distribution and 1 for the standard Normal distribution
- ML estimates for the regression effects in the logit model $\tilde{\beta}$ and the probit model $\hat{\beta}$ are related by

$$\tilde{\beta} \approx \hat{\beta} \frac{\pi}{\sqrt{3}}$$

Tailored proposal: t-distribution centered at the rescaled MLE of the probit model with appropriate covariance matrix

Pima indian data: MH with tailored proposal



M=22000; burnin=2000; ESS $\in (5657, 6563)$

Pima indian data: MH with tailored proposal- results

results (thinning factor 10)

	Q2.5%	Q50%	Q97.5%	mean	sd
int	-5.13	-3.78	-2.64	-3.81	0.63
npreg	-0.02	0.11	0.24	0.11	0.07
glu100	0.02	0.03	0.05	0.03	0.01
bp80	-0.04	-0.01	0.03	-0.01	0.02
skin23	-0.04	-0.00	0.04	-0.00	0.02
bmi25	-0.00	0.09	0.17	0.09	0.04
ped025	0.62	1.90	3.29	1.92	0.70
age20	0.00	0.04	0.09	0.04	0.02

Metropolis-Hastings within Gibbs

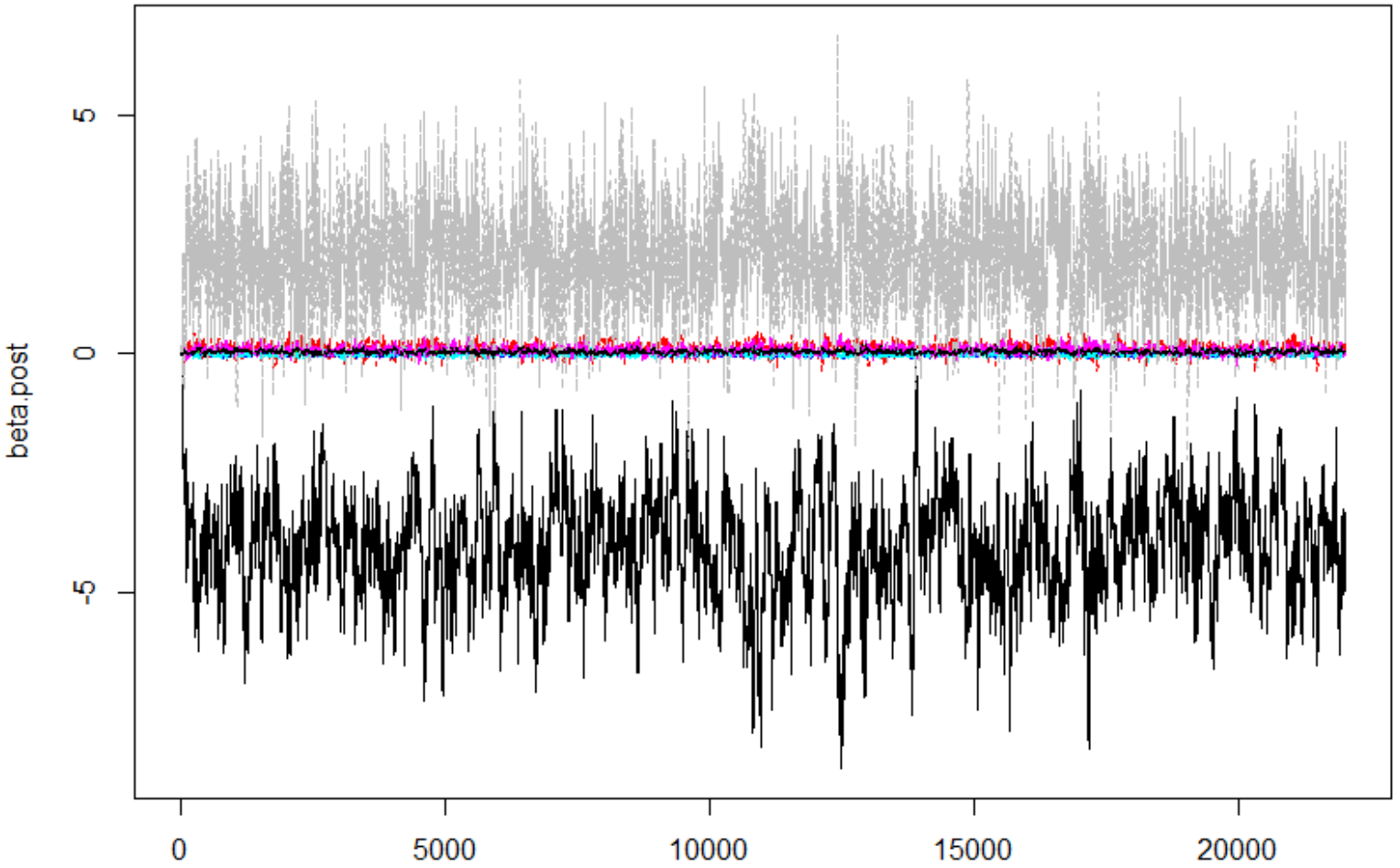
- in a Gibbs sampling scheme each component of the parameter vector is sampled from the full conditionals

  Algorithm:
  - Draw $\theta_1^{(m)}$ from $p(\theta_1|\theta_2^{(m-1)}, \ldots, \theta_p^{(m-1)}, \mathbf{y})$

    $\vdots$
  - Draw $\theta_p^{(m)}$ from $p(\theta_p|\theta_1^{(m)}, \ldots, \theta_{p-1}^{(m)}, \mathbf{y})$.

- any method that produces a draw from the full conditional can be used in each Gibbs sampling step, e.g. direct sampling, accept-reject sampling, MH-algorithm, ...

Pima indian data: componentwise random walk MH



M=22000; burnin=2000; ESS $\in (234, 1085)$

Posterior inference based on MCMC samples

Diagnosing convergence

Convergence cannot be proved from a finite sample but slow convergence is indicated by results of the MCMC sampler

- empirical autocorrelations of the MCMC draws
  - low autocorrelations indicate that the sampler explores the posterior well (mixes well)
  - high autocorrelations indicate stickiness of the algorithm
- start several chains from different starting points and check their convergence
- convergence diagnostic statistics use one or more chains (e.g. Heidelberger-Welch, Gelman-Rubin, Geweke statistic)

Approximation the posterior mean

- burnin =MCMC draws before convergence to the posterior (stationary distribution)
  the burnin sample is not used for inference but discarded
- the posterior mean

$$\mathrm{E}\left(g(\boldsymbol{\theta})|\mathbf{y}\right) = \int g(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$$

  can be approximated by the mean of M MCMC draws (after burnin) as

$$\hat{g}(\boldsymbol{\theta}) = \frac{1}{M}\sum_{m=1}^{M} g(\boldsymbol{\theta}^{(m)})$$

Approximation the posterior mean

- For independent draws from the posterior the sampling variance of $\hat{g}(\boldsymbol{\theta})$ is given as

$$\text{Var}\left(\hat{g}(\boldsymbol{\theta})\right)_{iid} = \text{Var}\left(g(\boldsymbol{\theta})\right)/M$$

  For the mean of M MCMC draws the sampling variance will be larger due to their (usually) positive autocorrelation.

- For dependent draws

$$\text{Var}\left(\hat{g}(\boldsymbol{\theta})\right)_{MCMC} = \text{Var}\left(g(\boldsymbol{\theta})\right)\frac{\tau}{M}$$

  with the inefficiency factor (integrated autocorrelation time) τ given as

$$\tau = 1 + 2\sum_{s=1}^{\infty}\rho_s.$$

  ρ_s is the autocorrelation of the draws $g(\boldsymbol{\theta}^{(m)})$ at lag s.

Effective sample size

- τ is the number of MCMC draws which are equivalent to one iid draw. Usually $\tau > 1$.
- The effective sample size M/τ is the number of independent draws which is equivalent to M MCMC draws.
- Thinning
  - keep only each k-th draw for inference
  - these draws are (nearly) independent

Software for Bayesian Inference

- R functions, e.g.
  - `bayesm`
  - `MCMCPack`
  - `BAMLSS`
  - `bayesSurv`

  see CRAN Task View: Bayesian Inference
- BayesX
  - `http://www.stat.uni-muenchen.de/~bayesx`
  - R-interface `R2BayesX`
- JAGS
  - BUGS and JAGS
  - R-interface `rjags`
- STAN
  - `http://www.mc-stan.org/`
  - R-interface `rstan`

JAGS: Just another Gibbs Sampler

- JAGS
  - source code and binaries for Windows and Mac
    `https://sourceforge.net/projects/mcmc-jags/files/`
  - Current version: 4.3.0
- `R` package `rjags`
  - Bayesian graphical models using MCMC with the `JAGS` library
  - compatible version to `JAGS 4.0.0`
  - `install.packages("rjags")`
- `R` package `coda` (Convergence Diagnosis and Output Analysis)
  - output analysis and diagnostics for MCMC
  - `install.packages("coda")`

Bayesian Analysis with JAGS

- model definition in BUGS language in a seperate file
- read the model with `jags.model`
- update the model with `update` for `jags` objects
- extract samples from the posterior with `coda.samples`
- summarize posterior distribution

for more details see Exercise 8

Pima Indian data: rjags-results

$M = 100,000$ iteration, thinning factor 100

	Q2.5%	Q50%	Q97.5%	mean	sd
int	-5.03	-3.77	-2.69	-3.80	0.61
npreg	-0.03	0.11	0.23	0.11	0.07
glu100	0.02	0.03	0.05	0.03	0.01
bp80	-0.04	-0.01	0.03	-0.01	0.02
skin23	-0.04	-0.00	0.05	-0.00	0.02
bmi25	0.01	0.09	0.17	0.09	0.04
ped025	0.51	1.88	3.40	1.90	0.71
age20	-0.00	0.04	0.09	0.04	0.02

essentially the same results as for MH with tailored proposal

Bayesian Hierarchical Modelling

- the prior allows for structural modelling of parameters
- hierarchical modelling
  - data model: $p(\mathbf{y}|\theta)$
  - prior for θ: $p(\theta|\psi)$ depends on hyperparameters ψ
  - hyper-prior for ψ: $p(\psi|\xi)$
- MCMC estimation via Gibbs sampling requires one further step to sample ψ
  - sample θ from $p(\theta|\psi, \mathbf{y})$
  - sample ψ from $p(\psi|\theta, \mathbf{y})$

Bayesian Hierarchical Modelling: Example

- **data:** $y_1, \ldots, y_n$
- **simple Bayesian model**
  - likelihood: y_i iid. with

$$y_i = \mu + \epsilon_i \qquad \epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

  - prior:

$$\mu \sim \mathcal{N}\left(m_0, M_0\right)$$

  restrictive: same mean for all observations

- **flexible model**

$$y_i = \mu_i + \epsilon_i \qquad \epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

  prior for μ_i?

Modelling the specific means

- Random intercept model

  $i = 1, \ldots, n$ subjects : μ_i is the subject specific level (intercept)

  $$\mu_i \sim \mathcal{N}\left(\mu, \tau^2\right)$$

  with hyperprior for τ^2
- Two component mixture model

  $i = 1, \ldots, n$ subjects : μ_i is the subject specific level (intercept)

  $$\mu_i = \begin{cases} \mu_1 & \text{with probability } \pi \\ \mu_2 & \text{with probability } 1 - \pi \end{cases}$$
- Local level model

  $i = 1, \ldots, n$ time-points: μ_i is the level at time point i

  $$\mu_i = \mu_{i-1} + \eta_i \qquad \eta_i \sim \mathcal{N}\left(0, \tau^2\right)$$

  and hyperprior on τ^2

Further Topics

- Bayesian modelling
  - mixture models for model based clustering
  - dynamic models
- model selection and model averaging, variable selection
- model checking
- priors (objective priors, Jeffreys prior)

Conclusion

Bayesian approach to statistics

- relies on specification of data model and prior distribution
  - allows to incorporate prior information/regularization
  - requires specification of prior distributions for all parameters
- inference is based on the posterior distribution
  - conjugate analysis only in special cases
  - more general: via sampling from the posterior distribution .e.g. by MCMC methods
- is useful for complex, particularly hierarchical models

Literature

- Albert J. (2009). Bayesian computation with R.
- Gelman A., Carlin J.B., Stern H.S., Dunson D. B. , Vehtari A. and Rubin, D.R. (2004). Bayesian Data Analysis. Chapman and Hall
- Held, Leonhard (2008). Methoden der statistischen Inferenz. Likelihood und Bayes. Spektrum Verlag
- Hoff, Peter D. (2009). A First Course in Bayesian Statistical Methods
- Marin, Jean-Michel and Robert Christian P. (2013). Bayesian Essentials with R, Springer
- Bertsch McGrayne (2012). The theory that would not die. Yale University Press.

International Society for Bayesian Analysis (ISBA):
`http:www.bayesian.org`