Bayesian Inference

Chapter 1

Introduction

Bayes theorem for two events

Suppose there are two events A and B, then Bayes theorem says

$$
\begin{aligned}
P(A|B) &= \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\bar{A}) \cdot P(\bar{A})} \\
&= \frac{P(B|A) \cdot P(A)}{P(B)}
\end{aligned}
$$

Bayes theorem for two events

Taxicab example

- 85% of the taxicabs in a city are blue, 15% are green.
- A hit and run accident occured involving a taxi.
- A witness claims a green taxi was responsible.
- A test with the witness revealed that there is a 80% chance that the true color is recognized.
- What is the probability that the taxi was indeed green?

Bayes theorem for two events

Taxicab example

- Define the events

$$
\begin{aligned}
A &= \text{taxi blue} \\
B &= \text{witness recognizes blue taxi.}
\end{aligned}
$$

- We have the following probabilitities

$$
\begin{aligned}
P(\text{taxi blue}) &= P(A) &= 0.85 \\
P(\text{witness blue taxi} \,|\, \text{taxi blue}) &= P(B \,|\, A) &= 0.8 \\
P(\text{witness green taxi} \,|\, \text{taxi green}) &= P(\bar{B} \,|\, \bar{A}) &= 0.8.
\end{aligned}
$$

Bayes theorem for two events

Taxicab example

- Moreover we have

$$
\begin{aligned}
P(\text{taxi green}) &= P(\bar{A}) &&= 0.15 \\
P(\text{witness green taxi} \,|\, \text{taxi blue}) &= P(\bar{B} \,|\, A) &&= 0.2 \\
P(\text{witness blue taxi} \,|\, \text{taxi green}) &= P(B \,|\, \bar{A}) &&= 0.2.
\end{aligned}
$$

- We are interested in the probability

$$P(\text{taxi green} \,|\, \text{witness green taxi}) = P(\bar{A} \,|\, \bar{B}).$$

Bayes theorem for two events

Taxicab example

- Using Bayes theorem we obtain

$$
\begin{aligned}
P(\bar{A}|\bar{B}) &= \frac{P(\bar{B}\,|\,\bar{A}) \cdot P(\bar{A})}{P(\bar{B}|\bar{A}) \cdot P(\bar{A}) + P(\bar{B}|A) \cdot P(A)} \\
&= \frac{0.8 \cdot 0.15}{0.8 \cdot 0.15 + 0.2 \cdot 0.85} \\
&= 0.41.
\end{aligned}
$$

Bayes theorem for more than two events

Assume $A_1, \ldots, A_k$ is a disjunct decomposition of Ω and $P(A_i) > 0$, $P(B) > 0$. Then we have:

$$P(A_j|B) = \frac{P(B|A_j) \cdot P(A_j)}{\sum_{i=1}^{k} P(B|A_i) \cdot P(A_i)} = \frac{P(B|A_j) \cdot P(A_j)}{P(B)}, \quad j = 1, \ldots, k\,.$$

Bayes theorem for more than two events

Monty hall problem

- Suppose you are attending a game show and you can choose one of three doors.
- Behind one of the three doors there is a brandnew car, behind the other doors you find a goat.
- After deciding for a door, the show master opens one of the two remaining doors with a goat behind it.
- The show master offers you now to rethink your choice and change the door. Will you accept his offer?
- In the following we assume you chose door one.
- Assumptions:
  - Show master opens only doors with goats behind them.
  - If there are only goats behind the two remaining goats, the show master chooses with equal probability one of the doors.

Bayes theorem for more than two events

Monty hall problem

- Define the following events:
  $T_1 \hat{=}$ car behind door 1
  $T_2 \hat{=}$ car behind door 2
  $T_3 \hat{=}$ car behind door 3
  $M_2 \hat{=}$ show master opens door 2
  $M_3 \hat{=}$ show master opens door 3
- We have:

$$
\begin{aligned}
P(T_1) &= P(T_2) = P(T_3) = \frac{1}{3} \\
P(M_2 \mid T_1) &= \frac{1}{2} \quad , \quad P(M_3 \mid T_1) = \frac{1}{2} \\
P(M_2 \mid T_2) &= 0 \quad , \quad P(M_3 \mid T_2) = 1 \\
P(M_2 \mid T_3) &= 1 \quad , \quad P(M_3 \mid T_3) = 0
\end{aligned}
$$

Bayes theorem for more than two events

Monty hall problem

- We would like to compute $P(T_1|M_2)$ and $P(T_3|M_2)$, respectively $P(T_1|M_3)$ and $P(T_2|M_3)$
- According to Bayes Theorem we have

$$
\begin{aligned}
P(T_1|M_2) &= \frac{P(M_2|T_1)\cdot P(T_1)}{P(M_2|T_1)\cdot P(T_1) + P(M_2|T_2)\cdot P(T_2) + P(M_2|T_3)\cdot P(T_3)} \\
&= \frac{\frac{1}{2}\cdot\frac{1}{3}}{\frac{1}{2}\cdot\frac{1}{3} + 0\cdot\frac{1}{3} + 1\cdot\frac{1}{3}} \\
&= \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3} \\
P(T_3|M_2) &= \frac{P(M_2|T_3)\cdot P(T_3)}{P(M_2|T_1)\cdot P(T_1) + P(M_2|T_2)\cdot P(T_2) + P(M_2|T_3)\cdot P(T_3)} \\
&= \frac{1\cdot\frac{1}{3}}{\frac{3}{6}} = \frac{2}{3}
\end{aligned}
$$

Bayes theorem in everyday life

- Bayes theorem relevant e.g. in courtroom (famous O.J. simpson process), cancer screening, medical testing for disease (AIDS test, corona rapid tests), etc.
- Use natural frequencies to figure out relevant probabilities rather than Bayes theorem directly
- Reference:
  - Gerd Gigerenzer, 2015: Das Einmaleins der Skepsis. Uber den richtigen Umgang mit Zahlen
  - Gerd Gigerenzer, 2015: Reckoning with Risk. Learning to live with uncertainty

Possible topics for presentations

- Review of the Guardian article *The obscure maths theorem that governs the reliability of Covid testing*
- Bayes theorem in everyday life
- Analyzing (possibly yor own) data using Bayesian linear models (including variable selection).
- Bayesian LASSO versus the original LASSO.
- Deriving some properties of the lecture.
- Bayesian P-splines (and their applications in recent Covid-19 research)

Bayesian Inference

Chapter 2

Review - Likelihood based inference

Situation

- Let $X_1,...,X_n$ be a random sample with probability function or density $f_i(x_i,\boldsymbol{\theta})$.
- The X_i's are independent but not identically distributed.
- Our goal is to estimate $\boldsymbol{\theta}$.

Situation

Normal distribution

The random sample $X_1, ..., X_n$ is assumed to be i.i.d. with $X_i \sim N(\mu, \sigma^2)$, i.e.

$$\boldsymbol{\theta} = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}.$$

Situation

Linear Model

- Y_1,...,Y_n independent with $Y_i \sim N(\mu_i, \sigma^2)$ and

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}.$$

- Define $\theta = \left(\beta_0, \beta_1, \ldots, \beta_k, \sigma^2\right)'$.
- We assume X_{i1},...,X_{ik} non stochastic.

Situation

Binary regression models

- We assume an independent random sample $Y_1, ..., Y_n$ with $Y_i \sim B(1, \pi_i)$ being Bernoulli distributed, i.e. $P(Y_i = 1) = \pi_i$ and $P(Y_i = 0) = 1 - \pi_i$.
- Logit model:

$$\begin{aligned}\pi_i &= \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})} \\ &= \frac{\exp(\boldsymbol{x}_i' \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i' \boldsymbol{\beta})} \\ &= \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}\end{aligned}$$

Situation

Binary regression models

- Probit model

$$\pi_i = \Phi(\beta_0 + \cdots + \beta_k x_{ik}) = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}) = \Phi(\eta_i),$$

  where $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ik})'$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ and $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$.
- Here $\boldsymbol{\theta} = (\beta_0, \ldots, \beta_k)'$.
- For the logit model we further obtain

$$g(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik},$$

  where g is called link function (here the logit-link).

Situation

Binary regression models

- The ratio $\log \frac{\pi_i}{1-\pi_i}$ is called log-odds, which can be regarded as a linear combination of the covariates.
- For the odds ratio $\frac{\pi_i}{1-\pi_i}$ we have a multiplicative model, i.e.

$$\frac{\pi_i}{1-\pi_i} = \exp(\beta_0)\exp(\beta_1 x_{i1}) \cdots \exp(\beta_k x_{ik}).$$

Likelihood

- The likelihood of the sample $X_1,...,X_n$ is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(X_i, \boldsymbol{\theta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\theta})$$

with

$$L_i(\boldsymbol{\theta}) = f_i(X_i, \boldsymbol{\theta}).$$

- The log-likelihood is given by

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\big(f_i(X_i, \boldsymbol{\theta})\big),$$

where

$$\ell_i(\boldsymbol{\theta}) = \log\big(f_i(X_i, \boldsymbol{\theta})\big).$$

Likelihood

Normal distribution

$$
\begin{aligned}
L_i(\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x_i - \mu\right)^2\right) \\
&\propto \frac{1}{\sqrt{\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x_i - \mu\right)^2\right) \\
\ell_i(\mu, \sigma^2) &= \log(1) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left(x_i - \mu\right)^2 \\
&= -\frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left(x_i - \mu\right)^2 \\
\ell(\mu, \sigma^2) &= \sum_{i=1}^{n} \ell_i(\mu, \sigma^2)
\end{aligned}
$$

Likelihood

Linear regression

$$
\begin{aligned}
L_i(\boldsymbol{\beta}, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2\right) \\
&\propto \frac{1}{\sqrt{\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2\right) \\
\ell_i(\boldsymbol{\beta}, \sigma^2) &= -\frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 \\
\ell(\boldsymbol{\beta}, \sigma^2) &= -\frac{1}{2}n\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2
\end{aligned}
$$

Likelihood

Binary regression

We obtain

$$\ell_i(\boldsymbol{\beta}) = \pi_i^{y_i}\big(1-\pi_i\big)^{1-y_i}$$
$$\ell_i(\boldsymbol{\beta}) = y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)$$

for the likelihood and log-likelihood.

Likelihood

Binary regression

Because of

$$\pi_i = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}, \quad 1-\pi_i = \frac{1}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}$$

and

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

we further obtain

$$\begin{aligned}\ell_i(\boldsymbol{\beta}) &= y_i\boldsymbol{x}_i'\boldsymbol{\beta} + \log\left(\frac{1}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}\right)\\ &= y_i\boldsymbol{x}_i'\boldsymbol{\beta} - \log\big(1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\big)\end{aligned}$$

in case of the logit model.

Score function and ML-estimator

The score function is the vector of first derivatives of the log likelihood, i.e.

$$
\begin{aligned}
S_i(\boldsymbol{\theta}) &= \left(\frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \theta_p}\right)' \\
S(\boldsymbol{\theta}) &= \sum_{i=1}^{n} S_i(\boldsymbol{\theta}).
\end{aligned}
$$

The ML-estimator is the solution to the following system of equations

$$S(\boldsymbol{\theta}) = \mathbf{0}.$$

Score function and ML-estimator

Normal distribution

$$\ell_i(\mu, \sigma^2) = -\frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\big(x_i - \mu\big)^2$$

$$\frac{\partial \ell_i(\mu, \sigma^2)}{\partial \mu} = -\frac{1}{2\sigma^2}2(-1)(x_i - \mu) = \frac{1}{\sigma^2}(x_i - \mu)$$

$$\frac{\partial \ell_i(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(x_i - \mu)^2$$

$$\mathsf{S}\begin{pmatrix}\mu \\ \sigma^2\end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2}\sum_{i=1}^{n}\big(x_i - \mu\big) \\ -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}\big(x_i - \mu\big)^2 \end{pmatrix}$$

Score function and ML-estimator

Normal distribution

The ML-estimator is given as the solution to

$$\text{I}: \quad \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right) = 0$$

$$\text{II}: \quad -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2 = 0$$

Score function and ML-estimator

Normal distribution

From I we have $\hat{\mu} = \bar{x}$.

Inserting $\hat{\mu}$=$\bar{x}$ in II yields

$$-\frac{n}{\sigma^2} + \frac{1}{(\sigma^2)^2} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 = 0$$

We obtain

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2.$$

REML estimator

- The ML-estimator for σ^2 in the previous example is biased.
- Alternatively to the ML-estimator for variance parameters, the so called Restricted ML-estimator (REML) is often used.
- It maximizes the marginal likelihood

$$\begin{aligned}\mathrm{RL}(\sigma^2) &= \int_{-\infty}^{\infty} L(\mu, \sigma^2)\,\mathrm{d}\mu \\ &= \int_{-\infty}^{\infty} \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x_i - \mu\right)^2\right)\right) d\mu.\end{aligned}$$

- REML estimators are used regularly for variance parameters, e.g. in linear models or linear mixed models, in the context of likelihood based inference.

Score function and ML-estimator

Linear regression

$$
\begin{aligned}
\ell_i(\boldsymbol{\beta}, \sigma^2) &= -\frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 \\
\ell(\boldsymbol{\beta}, \sigma^2) &= -\frac{1}{2}n\log(\sigma^2) - \frac{1}{2\sigma^2}\left(Y - X\beta\right)'\left(Y - X\beta\right) \\
&= -\frac{1}{2}n\log(\sigma^2) - \frac{1}{2\sigma^2}\left(Y'Y - 2Y'X\beta + \beta'(X'X)\beta\right) \\
\frac{\partial \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \beta} &= -\frac{1}{2\sigma^2}\left(-2X'Y + 2X'X\beta\right) \\
&= -\frac{1}{\sigma^2}\left(X'X\beta - X'Y\right) \\
\frac{\partial \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\left(Y - X\beta\right)'\left(Y - X\beta\right)
\end{aligned}
$$

Score function and ML-estimator

Linear regression

Hence the score function is given by

$$S\begin{pmatrix}\beta \\ \sigma^2\end{pmatrix} = \begin{pmatrix} -\frac{1}{\sigma^2}\left(X'X\beta - X'Y\right) \\ -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(X - X\beta)'(Y - X\beta) \end{pmatrix}$$

To obtain the ML-estimator we solve

$$S\begin{pmatrix}\beta \\ \sigma^2\end{pmatrix} = \begin{pmatrix}\mathbf{0} \\ 0\end{pmatrix}$$

Score function and ML-estimator

Linear regression

We immediately obtain

$$\hat{\beta}_{\text{ML}} = \left(X'X\right)^{-1} X'Y$$
$$\hat{\sigma}^2_{\text{ML}} = \frac{1}{n}(Y - X\hat{\beta})'(Y - X\hat{\beta}).$$

Again $\hat{\sigma}^2_{\text{ML}}$ is biased. The REML estimator

$$\hat{\sigma}^2_{\text{REML}} = \frac{1}{n-k-1}(Y - X\hat{\beta})'(Y - X\hat{\beta})$$

is unbiased and maximizes the marginal likelihood

$$\text{RL}(\sigma^2) = \int \text{L}\begin{pmatrix} \boldsymbol{\beta} \\ \sigma^2 \end{pmatrix} \text{d}\boldsymbol{\beta}.$$

Bayesian Inference

Chapter 3

Basic concepts

Prior and posterior distribution

- The fundamental difference to likelihood-based inference is that the unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ are not considered as fixed, deterministic quantities but as random variables with a *prior distribution*.
- *Prior distribution:* Any (subjective) information about the unknown parameter $\boldsymbol{\theta}$ is expressed by specifying a probability distribution $p(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$. The prior describes the *degree of uncertainty* about the unknown parameters prior to the statistical analysis.
- *Observation model:* The observation model specifies the conditional distribution of observable quantities, that is the random sample variables $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$, given the parameters. The p.d.f. or probability function of this conditional distribution is proportional to the likelihood $L(\boldsymbol{\theta})$ and will be denoted by $p(\boldsymbol{y} \mid \boldsymbol{\theta})$.

Prior and posterior distribution

- Based on the prior and the observation model, Bayes' theorem determines the distribution of $\boldsymbol{\theta}$ after the data are known through the statistical experiment, that is the conditional distribution of $\boldsymbol{\theta}$ given the observations $\boldsymbol{y} = (y_1, \ldots, y_n)'$.
- We obtain

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}} = c \cdot p(\boldsymbol{y} \mid \boldsymbol{\theta})\; p(\boldsymbol{\theta}),$$

with the normalizing constant $c = [\int p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}]^{-1}$. This conditional distribution is called *posterior (distribution)*.

Examples

Poisson Distribution

- Consider an i.i.d. sample $Y_1, \ldots, Y_n$ from a Poisson distribution, i.e. $Y_i \sim Po(\lambda)$.
- The joint probability for the observed sample $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is

$$p(\boldsymbol{y} \mid \lambda) = \frac{1}{y_1! \cdots y_n!} \lambda^{\sum_{i=1}^n y_i} \exp(-n\lambda).$$

- We specify a Gamma distribution with parameters a and b for λ, i.e. $\lambda \sim Ga(a, b)$. It follows that λ has p.d.f.

$$p(\lambda) = k\, \lambda^{a-1} \exp(-b\lambda)$$

with $k = \frac{b^a}{\Gamma(a)}$.

Examples

Poisson Distribution

- The posterior is obtained as

$$
\begin{aligned}
p(\lambda \mid \boldsymbol{y}) &= \frac{p(\boldsymbol{y} \mid \lambda)\, p(\lambda)}{\int p(\boldsymbol{y} \mid \lambda)\, p(\lambda)\, d\lambda} \\
&= c\, \frac{1}{y_1! \cdots y_n!} \lambda^{\sum_{i=1}^n y_i} \exp(-n\lambda)\, k\lambda^{a-1} \exp(-b\lambda).
\end{aligned}
$$

- To determine the type of this distribution, we can ignore all factors that do not depend on λ. This gives

$$
\begin{aligned}
p(\lambda \mid \boldsymbol{y}) &\propto \lambda^{\sum_{i=1}^n y_i} \exp(-n\lambda)\, \lambda^{a-1} \exp(-b\lambda) \\
&= \lambda^{a+\sum_{i=1}^n y_i - 1} \exp(-(b+n)\lambda).
\end{aligned}
$$

Examples

Poisson Distribution

- This has the form of a gamma distribution with parameters $a' = a + \sum_{i=1}^{n} y_i$ and $b' = b + n$, i.e.

$$\lambda \mid \boldsymbol{y} \sim Ga\left(a + \sum_{i=1}^{n} y_i, b + n\right),$$

  and the posterior has the same type of distribution as the prior.
- We call the prior as conjugate to the Poisson model because the posterior is of the same type as the prior.

Examples

Bayesian Logic Model - Diffuse Prior

- We consider a logit model with a single covariate x:

$$Y_i = B(1, \pi_i), \quad \pi_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}, \quad \eta_i = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, n.$$

- Assuming, as usual, (conditionally) independent response variables, the observation model is given by

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}) \propto L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ is the vector of regression coefficients.

Examples

Bayesian Logic Model - Diffuse Prior

- Since estimated regression coefficients are often approximately normally distributed, it is reasonable to assume a two-dimensional normal prior, i.e.

$$p(\boldsymbol{\beta}) \sim N_2(\boldsymbol{m}, \boldsymbol{M})$$

  with prior mean $\boldsymbol{m}$ and prior covariance matrix $\boldsymbol{M}$.
- If results from a previous statistical analysis are available, we could choose the previous point estimate as $\boldsymbol{m}$ and its estimated covariance matrix as $\boldsymbol{M}$.
- If the previous analysis has been carried out some time ago, we may also multiply $\boldsymbol{M}$ with a factor $a > 1$ to express increased uncertainty.

Examples

Bayesian Logic Model - Diffuse Prior

- Increasing the variances in $\boldsymbol{M}$, the normal prior becomes very flat and approximates a uniform distribution.
- In the limiting case the prior becomes proportional to a constant, i.e.

$$p(\boldsymbol{\beta}) \propto \text{const.}$$

  We also write $p(\boldsymbol{\beta}) \propto 1$.
- The integral of this flat prior over $\mathbb{R}^2$ is not finite, so that $p(\boldsymbol{\beta})$ is not a density in the usual sense. Such a prior is called improper or diffuse.

Examples

Bayesian Logic Model - Diffuse Prior

- Such diffuse priors are admissible as long as the posterior, resulting from Bayes' theorem, is a proper distribution. i.e. its integral over $\mathbb{R}^2$ is finite. In a Bayesian logit model this is the case if a finite MLE exists.
- With a flat, diffuse prior the posterior density is

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{\beta})p(\boldsymbol{y} \mid \boldsymbol{\beta})}{\int p(\boldsymbol{\beta})p(\boldsymbol{y} \mid \boldsymbol{\beta})d\,\boldsymbol{\beta}} \propto p(\boldsymbol{y} \mid \boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}.$$

- Although the posterior is proper, it has no known distributional type.

Examples

Normal Distribution with known expectation

- Consider an i.i.d. sample $Y_1, \ldots, Y_n$ from a normal distribution with known expectation μ, i.e. $Y_i \sim \mathrm{N}(\mu, \sigma^2)$.
- In the absence of any prior knowledge regarding the unknown parameter σ^2, one is tempted to assume a diffuse uniform prior over $\mathbb{R}^+$.
- It follows that $p(\sigma^2)$ is improper and proportional to a constant, i.e., $p(\sigma^2) \propto 1$.
- Suppose now, that we had parameterized the Normal distribution in terms of the standard deviation σ rather than the variance.
- According to our recipe for a noninformative prior, we would then assign a uniform distribution for σ, i.e., $p(\sigma) \propto 1$.

Examples

Normal Distribution with known expectation

- We now obtain $p(\sigma^2) \propto (\sigma^2)^{-1/2}$, which is no longer a uniform distribution over $\mathbb{R}^+$.
- Thus, if we parameterize the normal distribution with the variance, our noninformative prior for σ^2 is a uniform distribution. If we parameterize in terms of the standard deviation, we result with a non-uniform distribution for σ^2.
- This means that the prior is not invariant with respect to the specific parametrization.

Jeffreys' Prior

- Jeffreys' prior for a scalar parameter θ is defined to be proportional to the square root of the expected Fisher information, i.e.

$$p(\theta) \propto \sqrt{F(\theta)} = \sqrt{\mathsf{E}(-l''(\theta))}.$$

- Jeffreys' prior solves the invariance problem of the previous example as it is indeed invariant with respect to one to one transformations of the parameter.
- Jeffreys' prior for a scalar parameter can be characterized as a *reference prior*.
- Informally, the reference prior can be characterized as the distribution that maximizes the influence of the data on the posterior.

Jeffreys' Prior

- More specifically, the reference prior maximizes the expected Kullback-Leibler distance of the posterior relative to the prior. In this sense the reference prior is noninformative as the data get maximal weight and the influence of the prior is minimized.
- In situations with vector-valued parameter $\boldsymbol{\theta}$, Jeffreys' prior generalizes to
$$p(\boldsymbol{\theta}) \propto \sqrt{|F(\boldsymbol{\theta})|}.$$
- This prior is still translation invariant, but is in general not the reference prior.
- Instead, we can still base the choice of noninformative priors on the reference prior concept, see e.g. Held and Sabanes Bove (2021) for details.

Normal Distribution with known expectation

- In the case of a normal distribution with known expectation μ, Jeffreys' or the reference prior for σ^2 is obtained as

$$p(\sigma^2) \propto \frac{1}{\sigma^2}.$$

- Now the distribution of the standard deviation can be derived to be $p(\sigma) \propto 1/\sigma$ which is the same distribution as for σ^2.
- If both parameters μ and σ^2 are unknown, we obtain a noninformative prior in the form of the reference prior. This reference prior is given by

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$$

implying a priori independence of μ and σ^2 with distributions $p(\mu) \propto 1$ and $p(\sigma^2) \propto 1/\sigma^2$.

Bayesian Point Estimates

- The posterior mean is given by

$$\hat{\boldsymbol{\theta}} = EW(\boldsymbol{\theta} \mid \boldsymbol{y}) = \int \boldsymbol{\theta}\, p(\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta} = c \cdot \int \boldsymbol{\theta}\, p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

- The posterior mode is the value $\hat{\boldsymbol{\theta}}$ that (globally) maximizes the posterior density, i.e.

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\; p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}).$$

  The second expression shows that no integration is necessary to compute the posterior mode, because the normalizing constant is not needed.

- The posterior median, that is the median of the posterior distribution, is sometimes preferred to the posterior mean because it is more robust against outliers.

Poisson Distribution

- The posterior mean is

$$EW(\lambda \mid \boldsymbol{y}) = \frac{a + \sum_{i=1}^{n} y_i}{b + n}.$$

- The smaller a (in relation to $\sum y_i$) and b (in relation to n), the closer the posterior mean is to the usual MLE $\hat{\lambda} = \bar{y}$.
- The larger the prior information, i.e. the larger a and b are, the more the posterior mean and the MLE differ from each other.

Bayesian Interval Estimates

- For the posterior mean, a natural measure is the posterior variance.
- For the posterior median, the interquartile distance seems to be appropriate to measure its variability.
- In case of the posterior mode, the curvature of the posterior at the mode, i.e. the observed Fisher information, is a natural choice.
- Another way of assessing uncertainty are Bayesian confidence intervals or *credible intervals* or, more generally, *credible regions*: A region $C \subset \boldsymbol{\Theta}$ of the parameter space is said to be a $(1-\alpha)$-credible region for $\boldsymbol{\theta}$ if

$$\mathrm{P}(\boldsymbol{\theta} \in C \mid \boldsymbol{y}) = 1 - \alpha.$$

  If $C \subseteq \mathbb{R}$ is an interval it is called credible interval.
- A credible region contains (at least) a probability mass $1-\alpha$ of the posterior.

Bayesian Inference

Chapter 4

MCMC Methods

Basic Idea

- MCMC methods allow to draw samples from posterior distributions (and, in principle, from any distribution) that are usually not available analytically and to estimate characteristics of the posterior such as the mean, the variance or quantiles, or the posterior density itself.
- The most important advantage compared to more traditional methods of drawing a sample from a distribution, for example importance or rejection sampling, is that samples can be drawn from high-dimensional densities, even for dimensions in the thousands.
- Another advantage is that the normalizing constant, often a high-dimensional integral that cannot be computed with traditional numerical methods, does not have to be known.

Basis Idea

- Let $\boldsymbol{\theta}$ be the unknown vector of parameters in a Bayesian model and $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ the posterior density (we assume here that $\boldsymbol{\theta}$ is continuous).
- Instead of directly drawing an i.i.d. sample from $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, a Markov chain is generated such that the iterations of the transition kernel converge to the posterior of interest.
- In this way random numbers are generated that can be considered as a (correlated) sample from the posterior after some time of convergence, the *burn-in phase*.

Metropolis-Hastings Algorithm

To draw random numbers from the density $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, the Metropolis-Hastings algorithm proceeds as follows:

1. Initialize $\boldsymbol{\theta}^{(0)}$ and the number T of iterations. Set $t = 1$.
2. Draw a random number $\boldsymbol{\theta}^*$ from the proposal density $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)})$ and accept it as the new state $\boldsymbol{\theta}^{(t)}$ with probability

$$\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)}) = \min\left\{\frac{p(\boldsymbol{\theta}^* \mid \boldsymbol{y})\, q(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{y})\, q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)})}, 1\right\},$$

   otherwise set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.
3. Stop if $t = T$, otherwise set $t = t + 1$ and go to 2.

After a *burn-in phase* t_0, the random numbers $\boldsymbol{\theta}^{(t_0+1)}, \ldots, \boldsymbol{\theta}^{(T)}$ can be considered as a (correlated) sample from $p(\boldsymbol{\theta} \mid \boldsymbol{y})$.

Bayesian Logit Model

- We consider the following simulated logit model with two covariates x_1 and x_2:

$$
\begin{aligned}
Y_i &= B(1, \pi_i) \qquad i = 1, \ldots, 500, \\
\pi_i &= \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}, \\
\eta_i &= -0.5 + 0.6\, x_{i1} - 0.3\, x_{i2}.
\end{aligned}
$$

- The covariates x_1 and x_2 are drawn independently from a standard normal distribution.
- We want to construct a Metropolis-Hastings algorithm to estimate the parameter $\boldsymbol{\beta} = (-0.5, 0.6, -0.3)'$ given this simulated data.
- We specify independent diffuse priors $p(\beta_j) \propto$ const.

Bayesian Logit Model

- The posterior is then proportional to the likelihood:

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) \propto \prod_{i=1}^{500} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}.$$

- As a proposal density for the Metropolis-Hastings algorithm we choose a 3-dimensional normal distribution, with the current state $\boldsymbol{\beta}^{(t-1)}$ as its mean.
- For its covariance matrix, we start with the diagonal matrix $\Sigma = diag(0.4^2, 0.4^2, 0.4^2)$.
- Figure 1 (first row) shows the first 2000 random numbers for β_0 and β_1 drawn with this proposal density.
- Since we have specified diffuse priors, Bayes estimates for the regression coefficients should not differ too much from the MLEs. Therefore, the MLEs are displayed as horizontal lines in the plots.

Bayesian Logit Model

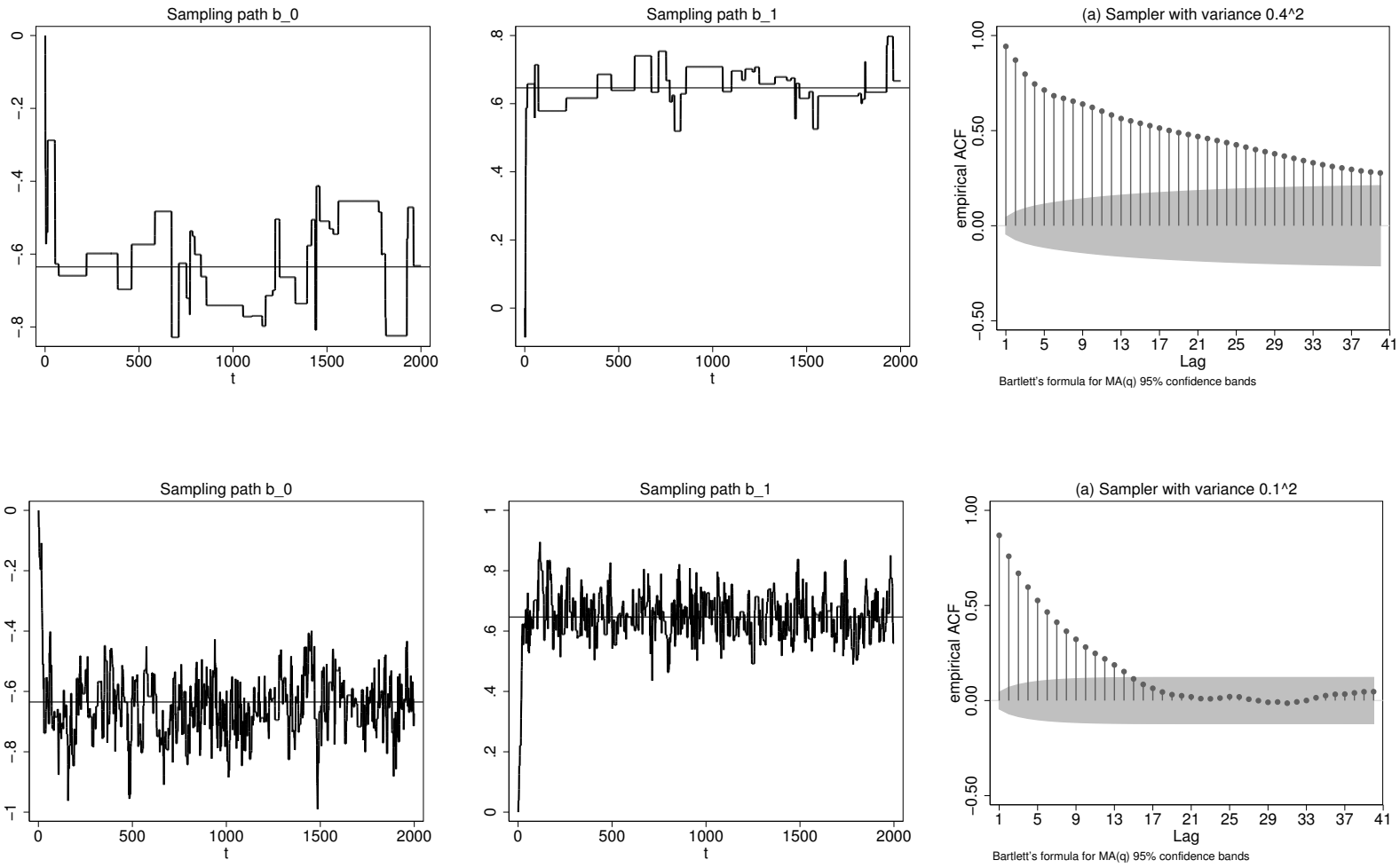

Figure: *Sampling paths for β_0 and β_1 for different MH algorithms. The third column shows the respective autocorrelation functions for β_1.*

Bayesian Logit Model

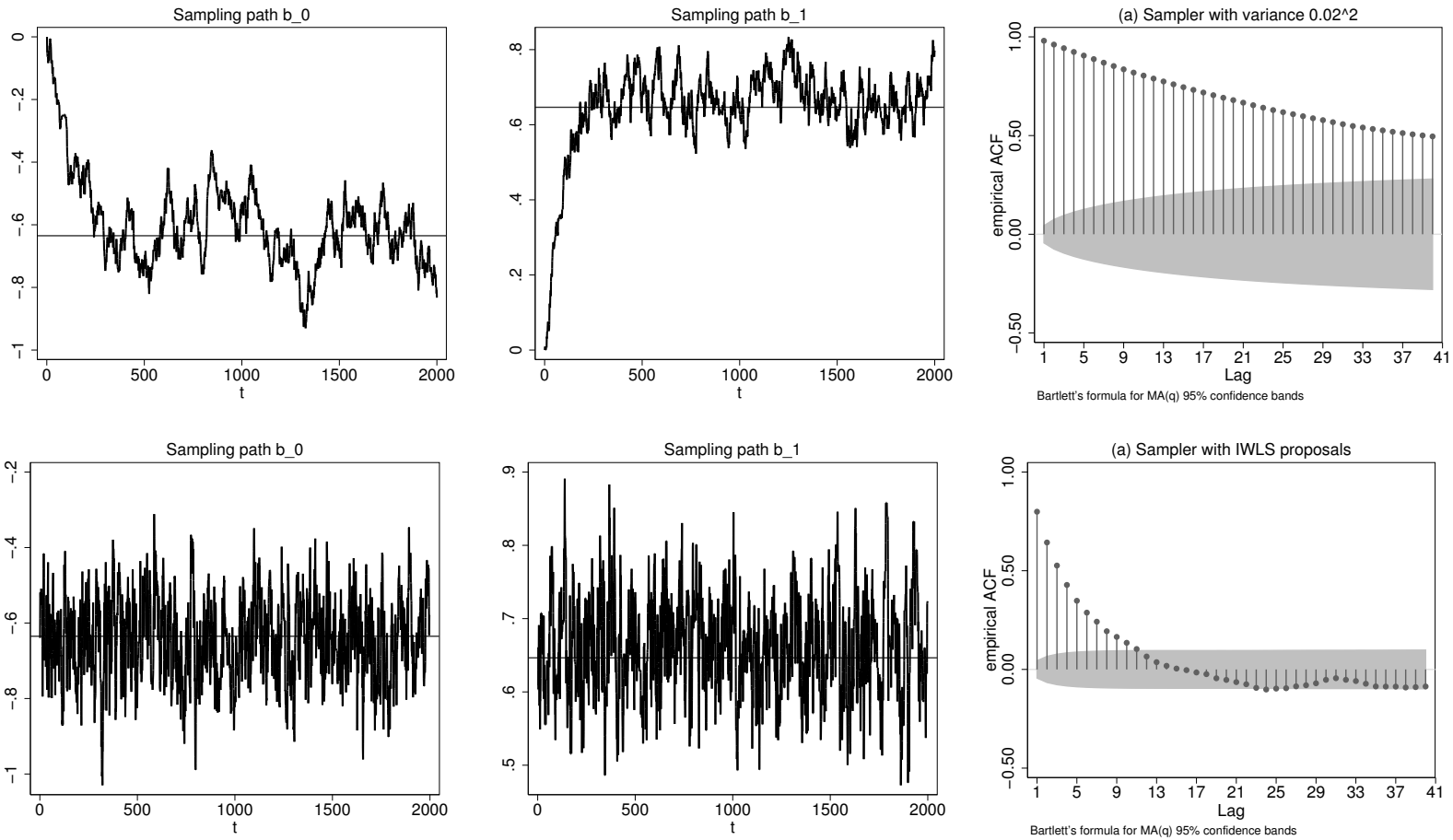

Figure: *Sampling paths for β_0 and β_1 for different MH algorithms. The third column shows the respective autocorrelation functions for β_1.*

Bayesian Logit Model

- Clearly, only a few of the proposed random numbers are accepted with this first algorithm, sometimes the state remains unchanged for more than 100 iterations.
- Thus, the acceptance probabilities are far too small.
- We obtain larger acceptance probabilities if the variances of the proposal density are decreased to $\Sigma = \mathit{diag}(0.1^2, 0.1^2, 0.1^2)$.
- The second row in Figure 1 shows the first 2000 random numbers for β_0 and β_1 resulting from this second MH algorithm.
- We recognize a short burn-in phase of about 50 iterations, followed by reasonable iterations with relatively large acceptance rates.

Bayesian Logit Model

- If we further decrease the variance to $\Sigma = diag(0.02^2, 0.02^2, 0.02^2)$, acceptance rates are further increased, but successive draws remain almost in the same state, see the first row in Figure 2.
- A useful and important tool for assessing the quality of MCMC algorithms is the autocorrelation function of the sample.
- Ideally, autocorrelations should rapidly converge to zero with increasing lags. The smaller the autocorrelation of successive parameters, the better the characteristics of the posterior can be estimated, based on the same length T of the sample.

Bayesian Logit Model

- For practical work, 'thinning' is carried out for the original sample, i.e. only every kth random number is kept in the sample, so that the remaining random numbers are almost uncorrelated. In this way, memory space can be saved without worsening estimation results.
- To generate an approximately uncorrelated sample of size 1000 with our second MH algorithm, we would have to generate about 20000 random numbers after a short burn-in phase and then keep only every 20th random number in the thinned sample.

Bayesian Logit Model

We can conclude the following:

- Small variances of the proposal density lead to high acceptance rates.
- In contrast, acceptance rates become small for large variances.
- For very large or very small variances autocorrelations of successive random numbers are high.
- The art of designing good MH algorithms is therefore the choice of appropriate proposal densities that combine high acceptance rates with low autocorrelations.
- Furthermore, automated methods are desirable that do not require subjective tuning of parameters of the proposal density.

Bayesian Logit Model

- An algorithm with these desirable properties is the MH algorithm based on IWLS proposals, see the last column of Fig. 2.
- Using this algorithm a Markov chain was generated and, after the burn-in phase, 20000 random numbers were drawn. Saving every 20th random number led to a thinned sample of size 1000. Based on this thinned sample all characteristics of interest of the posterior can be approximated.
- To approximate the posterior mean we compute the arithmetic means for the sample, resulting in $\hat{\boldsymbol{\beta}} = (-0.64, 0.65, -0.38)'$.
- Estimation of credible intervals can be based on the quantiles of the sampled random numbers. For example, we obtain 95% credible intervals by choosing the 2.5% quantiles as lower and 97.5% quantiles as upper bounds. This results in the intervals $[-0.87, -0.42]$, $[0.52 - 0.78]$ and $[-0.52, -0.26]$ for the sample generated in our example.

Gibbs Sampler and Hybrid Algorithms

- In many practical applications the parameter vector is high-dimensional.
- The acceptance rates then become rather small, even for well-designed MH algorithms, because a high-dimensional random number has to be accepted or not.
- So-called hybrid algorithms provide a solution to this problem, using a “divide and conquer” strategy.
- The high-dimensional parameter vector $\boldsymbol{\theta}$ is partitioned into smaller blocks $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_S$.
- Separate MH steps are then constructed for these subvectors.

Gibbs Sampler and Hybrid Algorithms

Let $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ be the posterior and assume that $\boldsymbol{\theta}$ is partitioned into S blocks $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_S$. The Gibbs sampler generates random numbers as follows:

1. Specify initial values $\boldsymbol{\theta}_1^{(0)}, \ldots, \boldsymbol{\theta}_S^{(0)}$ and the number of iterations T. Set $t = 1$.
2. For $s = 1, \ldots, S$: Draw random numbers from the full conditionals

$$p(\boldsymbol{\theta}_s \mid \boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_{s-1}^{(t)}, \boldsymbol{\theta}_{s+1}^{(t-1)}, \ldots, \boldsymbol{\theta}_S^{(t-1)}, \boldsymbol{y}).$$

   Note that the most actual states are used in the conditioning set of parameter blocks.
3. Stop if $t = T$, otherwise set $t = t + 1$ and go to 2.

Gibbs Sampler and Hybrid Algorithms

- If it is not possible to directly draw random numbers from some of the full conditionals, then an MH step is included instead.
- For the corresponding block $\boldsymbol{\theta}_s$, a proposal density

$$q_s(\boldsymbol{\theta}_s^* \mid \boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_{s-1}^{(t)}, \boldsymbol{\theta}_s^{(t-1)}, \ldots, \boldsymbol{\theta}_S^{(t-1)})$$

  is chosen and random numbers $\boldsymbol{\theta}_s^*$ are drawn from it.
- They are accepted as new states of the Markov chain with probability

$$\alpha(\boldsymbol{\theta}_s^* \mid \boldsymbol{\theta}_s^{(t-1)}) = \min\left\{\frac{p(\boldsymbol{\theta}_s^* \mid \boldsymbol{\theta}_{-s}^{(t-1)}) q_s(\boldsymbol{\theta}_s^{(t-1)} \mid \boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_{s-1}^{(t)}, \boldsymbol{\theta}_s^*, \ldots, \boldsymbol{\theta}_S^{(t-1)})}{p(\boldsymbol{\theta}_s^{(t-1)} \mid \boldsymbol{\theta}_{-s}^{(t-1)}) q_s(\boldsymbol{\theta}_s^* \mid \boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_{s-1}^{(t)}, \boldsymbol{\theta}_s^{(t-1)}, \ldots, \boldsymbol{\theta}_S^{(t-1)})}, 1\right\},$$

  where $p(\boldsymbol{\theta}_s \mid \boldsymbol{\theta}_{-s}^{(t-1)}) = p(\boldsymbol{\theta}_s \mid \boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_{s-1}^{(t)}, \boldsymbol{\theta}_{s+1}^{(t-1)}, \ldots, \boldsymbol{\theta}_S^{(t-1)}, \boldsymbol{y})$ denotes the full conditional of $\boldsymbol{\theta}_s$.
- Otherwise, $\boldsymbol{\theta}_s^{(t)} = \boldsymbol{\theta}_s^{(t-1)}$ as in the original MH algorithm.

Bayesian Inference

Chapter 5

Model Selection

Classical Approach

- The classical approach for Bayesian model choice is to compare competing models through the *posterior probabilities* of the models.
- Suppose we are given K competing models $M_1, \ldots, M_K$ with associated parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$. By a "model" we mean a set of probability distributions.
- For instance M_j, $j = 1, \ldots, K$, could denote the regression models $\boldsymbol{y} \mid \boldsymbol{\theta}_j, M_j \sim \mathrm{N}(\boldsymbol{X}_j\boldsymbol{\theta}_j, \sigma^2\boldsymbol{I})$ (with known variance σ^2 for simplicity).
- For each Model M_j let $p(\boldsymbol{y} \mid \boldsymbol{\theta}_j, M_j)$ denote the observation model and $p(\boldsymbol{\theta}_j \mid M_j)$ be the prior for the model parameters $\boldsymbol{\theta}_j$.

Classical Approach

- The posterior for $\boldsymbol{\theta}_j$ under model M_j is then given by

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

  where

$$p(\boldsymbol{y} \mid M_j) = \int p(\boldsymbol{y} \mid \boldsymbol{\theta}_j, M_j)\, p(\boldsymbol{\theta}_j \mid M_j)\, d\boldsymbol{\theta}_j \tag{1}$$

  is the *marginal likelihood*.

- For model selection, we additionally have to assign prior probabilities $p(M_j)$ associated with each model M_j.

Classical Approach

- Now the competing models can be compared through the posterior model probabilities given by

$$p(M_j \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid M_j)\, p(M_j)}{p(\boldsymbol{y})} \propto p(\boldsymbol{y} \mid M_j)\, p(M_j)$$

  with

$$p(\boldsymbol{y}) = \sum_{k=1}^{K} p(\boldsymbol{y} \mid M_k)\, p(M_k).$$

- We prefer model M_j against model M_s if $p(M_j \mid \boldsymbol{y}) > p(M_s \mid \boldsymbol{y})$, i.e. if the posterior ratio

$$\frac{p(M_j \mid \boldsymbol{y})}{p(M_s \mid \boldsymbol{y})} = \frac{p(M_j)}{p(M_s)} \frac{p(\boldsymbol{y} \mid M_j)}{p(\boldsymbol{y} \mid M_s)}$$

  is larger than 1.

Bayes Factor

- In cases of equal priors $p(M_1) = p(M_2) = \cdots = p(M_K) = 1/K$ the posterior ratio simplifies to the *Bayes factor*

$$BF_{js}(\boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid M_j)}{p(\boldsymbol{y}|M_s)}.$$

- If none of the competing models is favored prior to analysis of the data, model choice is based on Bayes factors.
- Care has to be taken when using the Bayes factor in combination with improper priors.

Bayes Factor

- Suppose, we assume improper priors

$$p(\boldsymbol{\theta}_j \mid M_j) \propto c_j, \qquad p(\boldsymbol{\theta}_s \mid M_s) \propto c_s$$

  for the model parameters $\boldsymbol{\theta}_j$ and $\boldsymbol{\theta}_s$ of models M_j and M_s. Here c_j and c_s are arbitrary constants.
- The Bayes factor then becomes

$$BF_{js}(\boldsymbol{y}) = \frac{c_j}{c_s} \frac{\int p(\boldsymbol{y} \mid \boldsymbol{\theta}_j, M_j)\, d\boldsymbol{\theta}_j}{\int p(\boldsymbol{y} \mid \boldsymbol{\theta}_s, M_s)\, d\boldsymbol{\theta}_s},$$

  where c_j/c_s is an arbitrary constant. This in turn implies that the Bayes factor is not uniquely defined.
- Thus, improper priors can not be used in the context of Bayesian model choice, at least if Bayes factors or in other words marginal likelihoods are involved.
- Improper priors may be appropriate only if the parameter $\boldsymbol{\theta}$ is the "same" under all models under consideration.

Bayesian Information Criterion

- In many applications, the exact computation of Bayes factors is difficult because computation of the marginal model likelihoods $p(\boldsymbol{y} \mid M_j)$ causes problems.
- An approximation (after multiplying with -2) is

$$-2p(\boldsymbol{y} \mid M_j) \approx -2 \cdot \log(p(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}}_j, M_j)) + \log(n)\, p_j,$$

  where p_j is the dimension of the parameter vector $\boldsymbol{\theta}_j$ and $\hat{\boldsymbol{\theta}}_j$ is the posterior mode.
- The approximation can be derived through a Laplace approximation of the integral in (1) and leads to the Bayesian Information Criterion (BIC).

Bayesian Information Criterion

- For a model with parameter $\boldsymbol{\theta}$, log-likelihood $l(\boldsymbol{\theta})$ and MLE $\hat{\boldsymbol{\theta}}$, the BIC is defined as

$$BIC = -2l(\hat{\boldsymbol{\theta}}) + log(n)\, p.$$

- Among a set of competing models, the model with the smallest BIC will be selected.

Bayesian Information Criterion

- Although derived from a Bayesian perspective, the BIC is not very popular in Bayesian inference. The main reasons are: First, the assumptions underlying the derivation of the BIC as an approximation of marginal log-likelihoods are not sufficiently well fulfilled in complex high-dimensional models.
- Related to this is the problem of determining n in the factor $\log(n)$. It is not always the data sample size: For example, in mixed models, n is the number of individuals.
- Second, when more complex Bayesian models are fitted with MCMC methods, the BIC is not directly available anyway.

Deviance Information Criterion

- A more recent criterion for model choice, that has become quite popular in connection with MCMC inference, is the Deviance Information Criterion (DIC).
- Its popularity is due to the fact that it can be easily computed from MCMC output.
- Let $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}$ denote an MCMC sample from the posterior of the model.

Deviance Information Criterion

- Computation of the DIC is based on two quantities.
- The first is the (unstandardized) deviance

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

  of the model.
- The second is the *effective number* p_D of parameters in the model. It can be estimated through

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}),$$

  where

$$\overline{D(\boldsymbol{\theta})} = \frac{1}{T}\sum_{t=1}^{T} D(\boldsymbol{\theta}^{(t)})$$

  is the average posterior deviance and $D(\bar{\boldsymbol{\theta}})$ is the deviance evaluated at $\bar{\boldsymbol{\theta}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\theta}^{(t)}$.

Deviance Information Criterion

- The DIC is then defined as

$$\text{DIC} = \overline{D(\boldsymbol{\theta})} + p_D = 2\overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}).$$

- As a disadvantage, the DIC value changes for different MCMC random samples.
- Therefore it may happen that model choice by DIC can lead to selecting different models with different MCMC samples. This will be only the case, however, if the DIC values of the models are quite close.

Deviance Information Criterion

Bayesian Logit Model—DIC

- We illustrate the use of DIC with the simulated data from previous Example.
- If we mistakenly omit the covariate x_2 and fit a logit model with x_1 only, then we obtain the (estimated) values $p_D = 1.99$ and $\text{DIC} = 571.6$.
- The effective number of parameters of about 2 is plausible, because we have estimated exactly two parameters β_0 and β_1.
- Fitting the correctly specified model, we obtain $p_D = 2.93$ and $\text{DIC} = 540.3$.
- The effective number of parameters is now about 3, as had to be expected.
- The DIC is now considerably smaller than for the wrong model so that the more complex, true model is selected.

Deviance Information Criterion

Bayesian Logit Model—DIC

- For illustration, we fit the correct model with five further MCMC runs.
- For p_D we obtain the values 3.05, 2.99, 3.15, 2.87 and 3.23 and for the DIC 540.56, 540.42, 540.73, 540.19 and 540.91, respectively.
- We see that the DIC varies between the different MCMC runs, but variability is usually quite low.

Bayesian Inference

Chapter 6

Bayesian Linear Model

Standard Conjugate Analysis

- Our starting point is the classical linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.
- In contrast to classical inference, the Bayesian approach treats the unknown parameters $\boldsymbol{\beta}$ and σ^2 as random variables.
- Thus, the distribution of the response $\mathbf{y}$ can be understood as conditional on the parameters $\boldsymbol{\beta}$ and σ^2, and we obtain the observation model

$$\mathbf{y} \mid \boldsymbol{\beta}, \sigma^{\mathbf{2}} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^{\mathbf{2}}\boldsymbol{I}).$$

Standard Conjugate Analysis

- The standard conjugate prior for linear models is obtained by assuming a multivariate normal prior for the regression coefficients

$$\boldsymbol{\beta} \mid \sigma^2 \sim \mathrm{N}(\boldsymbol{m}, \sigma^2 \boldsymbol{M}),$$

  with known expectation $\boldsymbol{m}$ and covariance matrix $\boldsymbol{M}$, e.g., $\boldsymbol{m} = \mathbf{0}$ and $\boldsymbol{M} = \boldsymbol{I}$.
- A normal distribution seems a natural choice since the distribution of the estimated regression coefficients in the classical linear model is (approximately) multivariate normal.
- For σ^2, we specify an inverse gamma distribution with hyperparameters a and b, i.e.,

$$\sigma^2 \sim IG(a, b). \tag{2}$$

Standard Conjugate Analysis

- To shed light on the specific form of the inverse gamma prior for σ^2, Fig. 3 shows the prior density for various choices of a and b.
- Of particular interest is the case $a = b$ and both values approaching zero. Such a distribution converges to an improper distribution that also results from a general prior construction principle (Jeffreys' prior).
- Another interesting case is when $a = 1$ and b is chosen to be small. In this case, the distribution of $\log(\sigma^2)$ tends to a uniform distribution.

Standard Conjugate Analysis

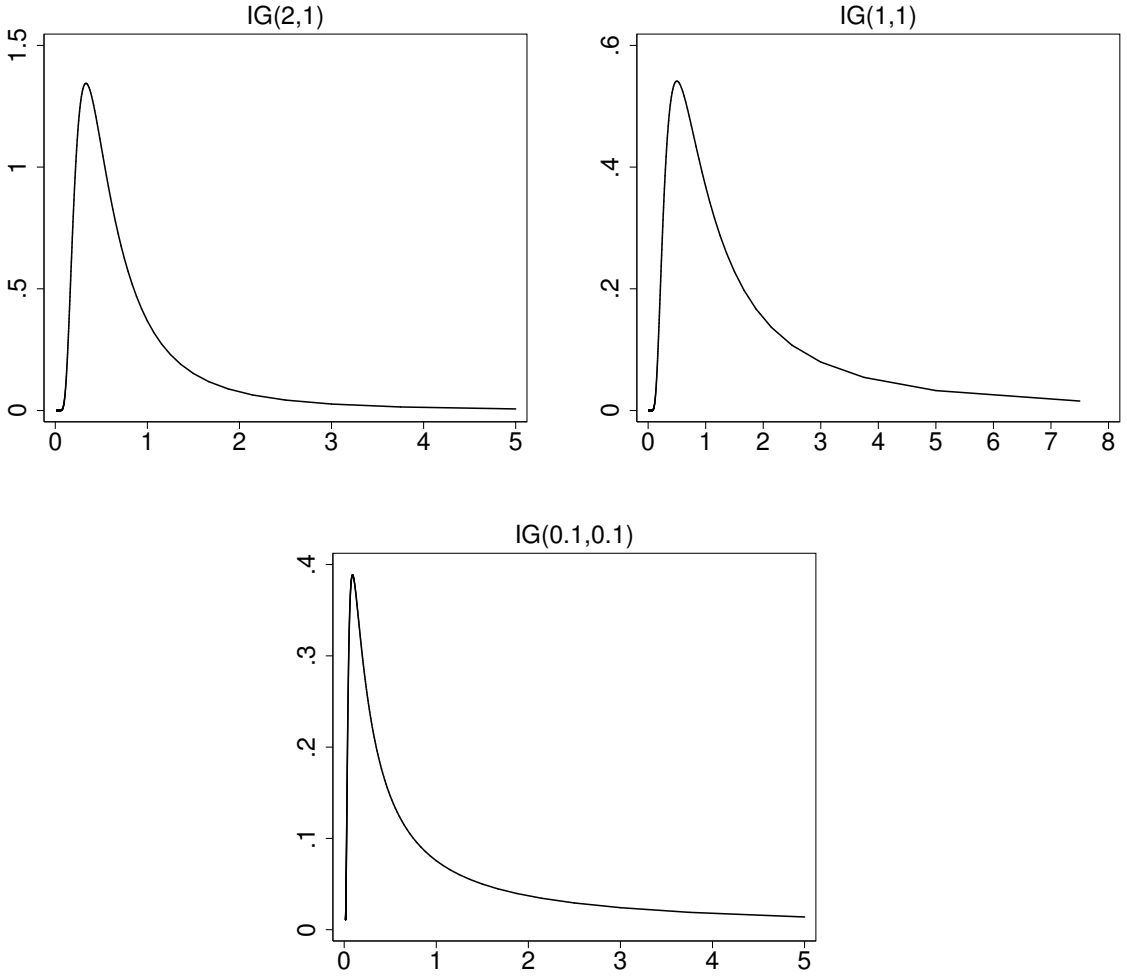

Figure: Inverse gamma prior density for σ^2 for various values of the hyperparameters *a* and *b*

Standard Conjugate Analysis

- The joint prior for $\boldsymbol{\beta}$ and σ^2 is a normal-inverse gamma distribution with density

$$
\begin{aligned}
p(\boldsymbol{\beta}, \sigma^2) &= p(\boldsymbol{\beta} \mid \sigma^2)\, p(\sigma^2) \qquad (3)\\
&= \frac{1}{(2\pi)^{\frac{p}{2}} (\sigma^2)^{\frac{p}{2}} |\boldsymbol{M}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{m})' \boldsymbol{M}^{-1} (\boldsymbol{\beta} - \boldsymbol{m})\right)\\
&\quad \frac{b^a}{\Gamma(a)} \frac{1}{(\sigma^2)^{a+1}} \exp\left(-\frac{b}{\sigma^2}\right)
\end{aligned}
$$

  and parameters $\boldsymbol{m}$, $\boldsymbol{M}$, a, and b. We write $\boldsymbol{\beta}, \sigma^2 \sim NIG(\boldsymbol{m}, \boldsymbol{M}, a, b)$.
- Ignoring all factors in Eq. (3) that are independent of $\boldsymbol{\beta}$ and σ^2, we obtain

$$
p(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{(\sigma^2)^{\frac{p}{2}+a+1}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{m})' \boldsymbol{M}^{-1} (\boldsymbol{\beta} - \boldsymbol{m}) - \frac{b}{\sigma^2}\right) \qquad (4)
$$

  for the density of the normal-inverse gamma prior.

Noninformative prior

- The widely accepted noninformative prior in the linear model is given by
$$p(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}, \tag{5}$$
which is the reference prior that maximizes the expected Kullback-Leibler distance of the posterior distribution relative to the prior.
- Informally, the reference prior can be characterized as the distribution that maximizes the influence of the data on the posterior.
- Since the density (5) cannot be normalized such that it integrates to one, it is an improper prior.
- The prior can be expressed as the product between a uniform prior $p(\boldsymbol{\beta}) \propto 1$ for $\boldsymbol{\beta}$ and the prior $p(\sigma^2) \propto 1/\sigma^2$ for σ^2 so that $\boldsymbol{\beta}$ and σ^2 are a priori stochastically independent.

Noninformative prior

- Note that the prior for σ^2 is equivalent to a uniform prior for $\log(\sigma^2)$.
- Technically, we can identify the noninformative prior (5) with the conjugate $NIG(\boldsymbol{m}, \boldsymbol{M}, a, b)$ prior by setting $\boldsymbol{m} = \mathbf{0}$, $\boldsymbol{M}^{-1} = \mathbf{0}$, $a = -p$, and $b = 0$.
- This is useful for posterior analysis because we can treat the noninformative case within the standard prior.
- Note, however, that we have to be very careful when proceeding this way. When dealing with improper priors, it is important to check whether the resulting posterior is truly proper. For the noninformative prior (5), this is indeed the case.

Noninformative prior

- Another approach to define a noninformative prior is described in O'Hagan (1994). Here we start with the marginal $IG(a, b)$ distribution for σ^2.
- If $a \to 0$ and $b \to 0$ tend to zero, we obtain

$$p(\sigma^2) \propto \frac{1}{\sigma^2}.$$

- For the joint $NIG(\boldsymbol{m}, \boldsymbol{M}, a, b)$ prior, we then have

$$p(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{(\sigma^2)^{\frac{p}{2}+1}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta} - \boldsymbol{m})\right).$$

Noninformative prior

- If rather $\boldsymbol{M}^{-1} = \mathbf{0}$, we arrive at the alternative noninformative prior

$$p(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-(p+2)}.$$

- This can be shown to be Jeffreys' prior.
- Although Jeffreys' prior is usually not used in multiparameter settings, our derivation justifies the widely used choice of a and b as equal to each other and near zero as a weakly informative choice for the prior of σ^2 (and more generally variance parameters).

Posterior Analysis

- Bayesian inference is based on properties of the posterior distribution, i.e., on the conditional distribution of the unknown parameters $\boldsymbol{\beta}$ and σ^2 given the data $\mathbf{y}$.
- The density of the posterior distribution is proportional to the product of the likelihood and the prior distribution, i.e.

$$
\begin{aligned}
p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}) &\propto L(\boldsymbol{\beta}, \sigma^2)\, p(\boldsymbol{\beta} \mid \sigma^2)\, p(\sigma^2) \qquad (6)\\
&\propto \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right)\\
&\quad \frac{1}{(\sigma^2)^{\frac{p}{2}}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta} - \boldsymbol{m})\right)\\
&\quad \frac{1}{(\sigma^2)^{a+1}} \exp(-\frac{b}{\sigma^2}).
\end{aligned}
$$

Posterior Analysis

- The linear model is one of a few examples in which the posterior distribution is analytically tractable, at least for the standard $NIG(\boldsymbol{m}, \boldsymbol{M}, a, b)$ prior.
- We can show that the posterior distribution, like the prior distribution, is a normal-inverse gamma distribution.
- The parameters $\tilde{\boldsymbol{m}}$, $\tilde{\boldsymbol{M}}$, $\tilde{a}$, and $\tilde{b}$ of the distribution are given by

$$\tilde{\boldsymbol{M}} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{M}^{-1})^{-1} \qquad \tilde{\boldsymbol{m}} = \tilde{\boldsymbol{M}}(\boldsymbol{M}^{-1}\boldsymbol{m} + \boldsymbol{X}'\boldsymbol{y}),$$

and

$$\tilde{a} = a + \frac{n}{2} \qquad \tilde{b} = b + \frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{y} + \boldsymbol{m}'\boldsymbol{M}^{-1}\boldsymbol{m} - \tilde{\boldsymbol{m}}'\tilde{\boldsymbol{M}}^{-1}\tilde{\boldsymbol{m}}\right).$$

Posterior Analysis

- Of particular interest is the posterior mean

$$\hat{\boldsymbol{\beta}}_B = \mathsf{E}(\boldsymbol{\beta} \mid \boldsymbol{y}) = \tilde{\boldsymbol{m}} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{M}^{-1})^{-1}(\boldsymbol{M}^{-1}\boldsymbol{m} + \boldsymbol{X}'\boldsymbol{y})$$

  as a point estimate of $\boldsymbol{\beta}$.
- Using the matrix $\boldsymbol{A} = (\mathbf{X}'\mathbf{X} + \boldsymbol{M}^{-1})^{-1}\mathbf{X}'\mathbf{X}$, we can write the Bayes estimator as a weighted average of the prior expectation $\boldsymbol{m}$ and the least squares estimator $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}}_B = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{m} + \boldsymbol{A}\hat{\boldsymbol{\beta}}.$$

- To interpret the Bayes estimator, note that the diagonal elements of $\boldsymbol{M}$ contain (up to the factor σ^2) the prior variances of $\boldsymbol{\beta}$.
- The greater the diagonal elements of $\boldsymbol{M}$ (i.e., the variances of $\boldsymbol{\beta}$), the smaller are the elements of $\boldsymbol{M}^{-1}$.

Posterior Analysis

- In the limit $\boldsymbol{M}^{-1} \to \mathbf{0}$, the matrix $\boldsymbol{A}$ approaches the identity matrix, and $\hat{\boldsymbol{\beta}}_B$ the ordinary least squares estimator.
- On the contrary, small elements in $\boldsymbol{M}$ (corresponding to small variances of $\boldsymbol{\beta}$) imply that the matrix $\boldsymbol{A}$ approaches the zero matrix and $\boldsymbol{I} - \boldsymbol{A}$ the identity matrix.
- The Bayes estimator is then identical with the prior mean $\boldsymbol{m}$.
- This gives us the following interpretation of $\hat{\boldsymbol{\beta}}_B$: The smaller the prior information about $\boldsymbol{\beta}$, i.e., the greater the diagonal elements of $\boldsymbol{M}$, the closer is $\hat{\boldsymbol{\beta}}_B$ to the least squares estimator. The larger the prior information, i.e., the smaller the diagonal elements of $\boldsymbol{M}$, the more the prior mean $\boldsymbol{m}$ dominates $\hat{\boldsymbol{\beta}}_B$.
- For the noninformative prior with $\boldsymbol{m} = \mathbf{0}$ and $\boldsymbol{M}^{-1} = \mathbf{0}$, the Bayes estimator coincides with the least squares estimator, i.e.,

$$\hat{\boldsymbol{\beta}}_B = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \hat{\boldsymbol{\beta}}_{LS}.$$

Full Conditional Densities and MCMC Inference

- We develop a Gibbs sampler, that consecutively draws random numbers from the full conditional distributions of $\boldsymbol{\beta}$ and σ^2.
- We obtain $\boldsymbol{\beta} \mid \cdot \sim \mathrm{N}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right)$ where

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \left(\frac{1}{\sigma^2}\mathbf{X}'\mathbf{X} + \frac{\mathbf{1}}{\sigma^{\mathbf{2}}}\boldsymbol{M}^{-\mathbf{1}}\right)^{-1}. \tag{7}$$

and

$$\boldsymbol{\mu}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\left(\frac{1}{\sigma^2}\mathbf{X}'\boldsymbol{y} + \frac{\mathbf{1}}{\sigma^{\mathbf{2}}}\boldsymbol{M}^{-\mathbf{1}}\boldsymbol{m}\right). \tag{8}$$

- We further obtain $p(\sigma^2 \mid \cdot) \sim IG(a', b')$ with parameters

$$a' = a + \frac{n}{2} + \frac{p}{2} \tag{9}$$

and

$$b' = b + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{m})'\boldsymbol{M}^{-1}(\boldsymbol{\beta} - \boldsymbol{m}). \tag{10}$$

Full Conditional Densities and MCMC Inference

Summarizing, we obtain the following Gibbs sampler:

1. Define initial values $\boldsymbol{\beta}^{(0)}$ and $(\sigma^2)^{(0)}$. Set $t = 1$.
2. Sample $\boldsymbol{\beta}^{(t)}$ by drawing from the Gaussian full conditional with covariance matrix (7) and mean (8), where σ^2 is replaced by the current state of the chain $(\sigma^2)^{(t-1)}$.
3. Sample $(\sigma^2)^{(t)}$ by drawing from the inverse gamma full conditional with parameters a' and b' given by Eqs. (9) and (10), where $\boldsymbol{\beta}$ is replaced by the current state of the chain $\boldsymbol{\beta}^{(t)}$.
4. Stop if $t = T$, otherwise set $t = t + 1$ and go to 2.

A derivation of the Gibbs sampler can be found in Fahrmeir et al. (2022).

Software

- Functions `bayesLMRef` and `bayesLMConjugate` of the `R` package `spBayes`.
- Software package `BayesX` for noninformative priors only (see also the `R` interface `R2BayesX`).
- Function `zlm` of the `R` package `BMS` (Zellner's g-prior only).

Munich Rent Index—Quality of Kitchen – frequentist approach

- The rent index is updated every two years, with the collection of new data.
- For financial reasons, the update during 2001 only consisted of data for 1,500 apartments.
- Due to the smaller sample size, a complete redesign of the rent index was not possible. Instead the following procedure was chosen:
  - The same explanatory variables from the 1999 rent index were used, implying that the structure of the rent index did not change.
  - Potential changes in the effects of regressors across data sets need to be examined.
  - To do so, both data sets of 1999 and 2001 were analyzed simultaneously, and changes in covariate effects have been investigated with the help of interactions.

Munich Rent Index—Quality of Kitchen – frequentist approach

- We will illustrate the approach using the quality of the kitchen (*kitchen*), with categories "kitchen below average" (reference category), "normal kitchen" (dummy variable *nkitchen*), and premium kitchen (dummy variable *pkitchen*).
- We apply the model

$$\begin{aligned} rentsqm_i \quad = \quad & \beta_0 + \beta_1 \cdot area_i^{-1} + \beta_2 \cdot yearc_i + \beta_3 \cdot yearc_i^2 + \beta_4 \cdot yearc_i^3 + \\ & \beta_5\, year01 + \beta_6\, nkitchen + \beta_7\, pkitchen + \\ & \beta_8\, nkitchen \cdot year01 + \beta_9 pkitchen \cdot year01 + \varepsilon_i . \end{aligned} \tag{11}$$

- Hence, the model consists of the transformed living area $1/area$, a cubic polynomial for year of construction, and the two kitchen dummies *nkitchen* ("normal kitchen") and *pkitchen* ("premium kitchen").
- The dummy variable *year01* specifies whether an observation has been taken from the year 2001 ($year01 = 1$) or from the year 1999 ($year01 = 0$).

Munich Rent Index—Quality of Kitchen – frequentist approach

- We obtain the estimate

$$\widehat{rentsqm} = \cdots - 0.26\, year01 + 0.91\, nkitchen + 1.09\, pkitchen + 0.41\, nkitchen \cdot year01 + 0.74\, pkitchen \cdot year01.$$

- The results can be summarized as follows, on a per square meter basis:
  - In 2001, apartments with a below average kitchen are approximately 0.26 Euro per square meter cheaper than apartments in 1999.
  - Apartments with a normal kitchen are approximately $-0.26 + 0.41 = 0.15$ Euro more expensive in 2001 than in 1999.
  - Apartments with a premium kitchen are approximately $-0.26 + 0.74 = 0.48$ Euro more expensive in 2001 than in 1999.

Munich Rent Index—Quality of Kitchen – Bayesian approach

- For the Bayesian approach, we again use model (11) without the interactions.
- We first develop a Bayesian version of the model based solely on the data for 1999.
- At that time we have no prior information regarding the unknown parameters. We therefore use the noninformative prior (5).
- Although the model could in principle be estimated analytically, we used the Gibbs sampler for inference since most available software packages do not support the analytical solution.

Munich Rent Index—Quality of Kitchen

- Using the function `bayesLMRef` of the `R` package `spBayes`, we obtained the results of Table 1.
- Up to sampling imprecision, the posterior mean is identical to the least squares estimator (as suggested by the analytically derived posterior).
- The posterior standard deviations and the quantiles are also very close to the respective least squares standard errors and the 95% confidence intervals.
- Note that the analytical posterior mean (and mode) coincides exactly with the least squares estimator, while the posterior standard deviation and quantiles are slightly different from their least squares counterparts.

Munich Rent Index—Quality of Kitchen

Variable	Coefficient	Standard deviation	2.5% Quantile	97.5% Quantile
invarea	122.5417	5.5877	-111.5955	-133.7277
yearc	-0.0861	0.0351	-0.1549	-0.0174
yearc2	0.0015	0.0007	0.0002	0.0028
yearc3	0.0000	0.0000	0.0000	0.0000
nkitchen	0.9274	0.1258	0.6770	1.1840
pkitchen	1.1022	0.1873	-0.7410	1.4718

Table: Munich rent index: estimation results based on the noninformative prior (5) for the parameters. Results are based on the data collected in 1999

Munich Rent Index—Quality of Kitchen

- Turning our attention to the new data collected in 2001 for the rent index update, it seems natural to use the posterior values obtained with the data from 1999 as prior information for the new analysis.
- That is we estimate a Bayesian linear model using the data collected in 2001 together with a $NIG(\boldsymbol{m}, \boldsymbol{M}, a, b)$ prior with parameters derived from the posterior obtained with the 1999 data.
- Clearly, $\boldsymbol{m}$ should be the empirical mean vector of the sampled regression coefficients in the MCMC sampler for the 1999 data.
- Since the prior covariance matrix is given by $\sigma^2\boldsymbol{M}$, we set $\boldsymbol{M} = 1/\hat{\sigma}^2\boldsymbol{S}$, where $\boldsymbol{S}$ is the empirical covariance matrix of the MCMC samples for the 1999 data and $\hat{\sigma}^2$ is the empirical mean of the samples for σ^2.

Munich Rent Index—Quality of Kitchen

- To obtain prior values for a and b, we note that the mean and variance of the $IG(a, b)$ prior for σ^2 are given by

$$\begin{aligned} \mathrm{E}(\sigma^2) &= \frac{b}{a-1}, \\ \mathrm{Var}(\sigma^2) &= \frac{b^2}{(a-1)^2(a-2)} = \mathrm{E}(\sigma^2)^2 \frac{1}{a-2}. \end{aligned}$$

- Solving for a and b yields

$$\begin{aligned} a &= \frac{\mathrm{E}(\sigma^2) + 2\mathrm{Var}(\sigma^2)}{\mathrm{Var}(\sigma^2)}, \\ b &= (a-1)\mathrm{E}(\sigma^2). \end{aligned}$$

- Based on the analysis for the 1999 data, we can now replace $\mathrm{E}(\sigma^2)$ and $\mathrm{Var}(\sigma^2)$ by their posterior estimates $\hat{\sigma}^2$ and $s^2_{\hat{\sigma}^2}$ to obtain prior values for a and b.

Munich Rent Index—Quality of Kitchen

- Using these values for the $NIG(\boldsymbol{m}, \boldsymbol{M}, a, b)$ prior, we arrive at the results given in Table 2.
- The table is obtained from the function `bayesLMConjugate` in the `R` package `spBayes`. For comparison, we additionally included the least squares estimates for the 2001 data.
- In particular for the kitchen dummies, the least squares estimates for the 2001 data differ considerably from those for the 1999 data (see Table 1).
- The Bayes estimator is a compromise between the least squares results for the 1999 and the 2001 data.

Munich Rent Index—Quality of Kitchen

- Although the new data based on 2001 have an impact on the estimates for *nkitchen* and *pkitchen*, the Bayes estimator is closer to the least squares estimate for the 1999 data.
- This is a clear result of the prior which pulls the posterior mean to a certain extent towards the prior mean, which is identical to the 1999 least squares estimate.
- We also observe that the posterior standard deviations of the regression coefficients are considerably lower than the least squares standard errors. This is again a result of the use of additional prior information.

Munich Rent Index—Quality of Kitchen

Variable	LS 2001		Bayes 2001	
	Coeff.	**Std.**	**Coeff.**	**Std.**
invarea	125.8373	7.2360	124.3540	3.3618
yearc	-0.0335	0.0480	-0.0631	0.0208
yearc2	0.0004	0.0009	0.0010	0.0004
yearc3	0.0000	0.0000	0.0000	0.0000
nkitchen	1.2944	0.1701	1.0327	0.0795
pkitchen	1.7935	0.2714	1.2910	0.1193

Table: Munich rent index: comparison of the Bayes estimate with informative prior and the least squares estimate for the data collected in 2001

Regularization Priors

- A number of “regularization priors” have been proposed in the literature.
- Here, we will develop some of the most widely used Bayesian regularization priors.
- Throughout, the observation model is a classical linear model given by

$$\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2 \sim NV(\beta_0 \mathbf{1} + \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}, \sigma^2 \boldsymbol{I}),$$

  where $\tilde{\boldsymbol{X}}$ is the $n \times k$-design matrix excluding the column of ones for the intercept and $\tilde{\boldsymbol{\beta}}$ is the corresponding vector of regression coefficients excluding β_0.

Regularization Priors

- Since the intercept is not subject to regularization, we assume a noninformative (diffuse) prior, i.e.,

$$p(\beta_0) \propto const.$$

- We also assume that the intercept β_0 is independent of the other regression coefficients $\tilde{\boldsymbol{\beta}}$.
- For the variance parameter σ^2, we specify the usual inverse gamma prior with hyperparameters a and b, i.e., $\sigma^2 \sim IG(a, b)$.

Bayesian Ridge Regression

- Ridge regression minimizes the penalized least squares criterion

$$PLS(\boldsymbol{\beta}) = (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}}. \tag{12}$$

  Usually the intercept is not penalized so that here the penalty is restricted to $\tilde{\boldsymbol{\beta}} = (\beta_1, \ldots, \beta_k)'$.
- The penalty shrinks the parameters towards zero in order to reduce the variance of the least squares estimator at the cost of a (hopefully small) bias.
- The amount of penalization is governed by the parameter λ. Small values for λ correspond to negligible penalization, whereas large values lead to strong penalization.

Bayesian Ridge Regression

- Using a particular prior for the regression coefficients, we obtain a Bayesian version of ridge regression.
- We assume a priori independent regression coefficients β_j, $j = 1, \ldots, k$, and set

$$\tilde{\boldsymbol{\beta}} \mid \tau^2 \sim \mathrm{N}(\mathbf{0}, \tau^2 \boldsymbol{I}). \tag{13}$$

- Since the prior for the intercept is improper, the joint prior for $\boldsymbol{\beta} = (\beta_0, \tilde{\boldsymbol{\beta}})'$ is also improper.
- Maximizing the corresponding posterior with respect to $\boldsymbol{\beta}$ is equivalent to minimizing the penalized least squares criterion (12) for fixed $\lambda = \sigma^2/\tau^2$.

Bayesian Ridge Regression

- While in the frequentist approach to ridge regression, the penalty parameter λ is estimated outside of the optimization criterion, e.g., via cross validation, the Bayesian approach allows for simultaneous inference for the regression coefficients and the amount of penalization measured through τ^2.
- This is facilitated by defining an additional prior for τ^2. A convenient and flexible choice is another inverse gamma distribution $\tau^2 \sim IG(a_{\tau^2}, b_{\tau^2})$, similar to the conjugate prior for σ^2 outlined above.
- The advantage of this specification is that the full conditional for τ^2 is again an inverse gamma distribution allowing for straightforward simulation-based MCMC inference.

Bayesian Ridge Regression

- Introducing an additional prior for τ^2, however, changes the interpretation of the prior.
- This is illustrated in Fig. 4 (left panel) which displays the log-prior density, for a single parameter β_j, conditional on τ^2 and the marginal log-prior with the variance parameter τ^2 integrated out.
- The marginal log-prior is quite different to the Gaussian conditional log-prior. It shows a distinct peak at zero with sharp declines.

Bayesian Ridge Regression

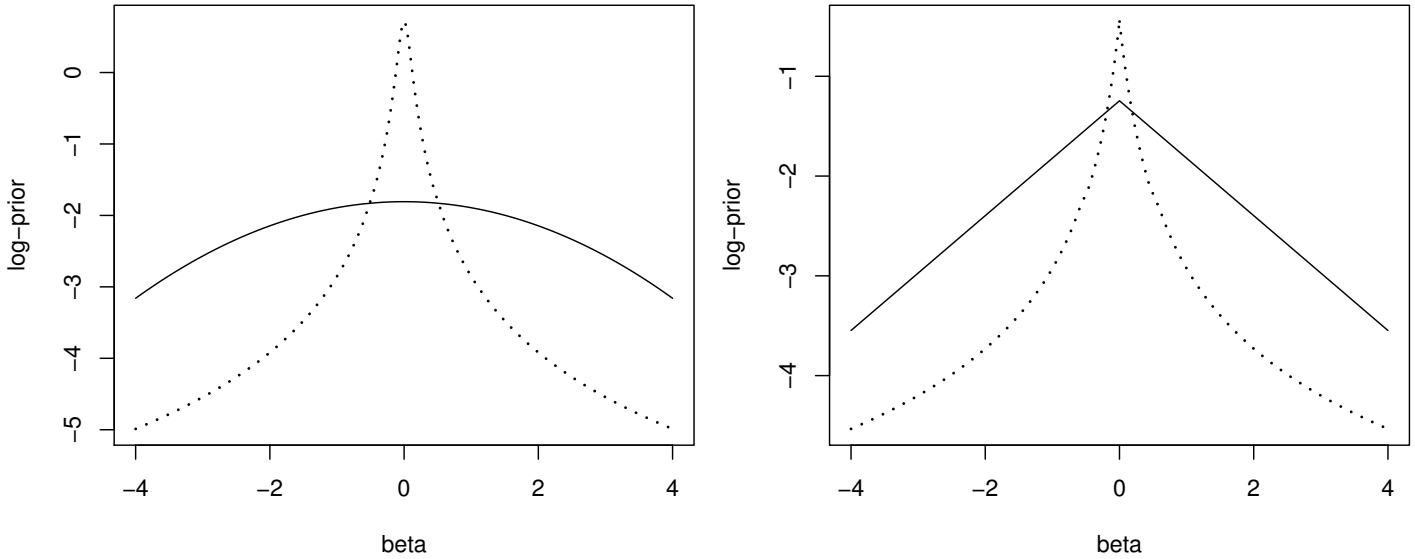

Figure: Conditional (*solid lines*) and marginal (*dashed lines*) log-priors for the ridge (*left panel*) and the LASSO prior (*right panel*). In case of the ridge prior, the plots are based on $a = 0.28$ and $b = 0.005$ for the inverse gamma prior. In case of the LASSO, the plots correspond to $a = 0.08$ and $b = 0.001$. With these choices, roughly 90% of the probability mass are contained in the interval $[-4, 4]$. The hyperparameters are chosen such that the differences between conditional and marginal distributions and between ridge and LASSO are best visible

Bayesian Lasso

- The LASSO replaces the quadratic penalty of ridge regression by the sum of absolute values leading to the penalized least squares criterion

$$PLS(\beta) = (\boldsymbol{y} - \mathbf{X}\beta)'(\boldsymbol{y} - \mathbf{X}\beta) + \lambda \sum_{\mathbf{j=1}}^{\mathbf{k}} |\beta_{\mathbf{j}}|. \tag{14}$$

- Similar to ridge regression, we define a prior for the regression coefficients such that the corresponding posterior mode is obtained by minimizing (14).
- We again assume (conditional) independence among the regression coefficients and arrive at the prior

$$\tilde{\boldsymbol{\beta}} \mid \tau_1^2, \ldots, \tau_k^2 \sim NV(\mathbf{0}, diag(\tau_1^2, \ldots, \tau_k^2)), \tag{15}$$

where now each regression coefficient β_j has its own variance τ_j^2.

Bayesian Lasso

- The variance parameters are assumed mutually independent with priors
$$\tau_j^2 \mid \sim \mathit{Expo}(0.5\lambda^2);$$
- The marginal distribution for β_j, obtained by integrating over τ_j^2, is a Laplace distribution with location parameter 0 and scale parameter $1/\lambda$, i.e., $p(\beta_j) \propto \exp(-\lambda|\beta_j|)$.
- Based on this prior specification, it can be shown that the posterior mode for $\boldsymbol{\beta}$ with fixed penalty parameter λ corresponds to minimizing the penalized least squares criterion (14).

Bayesian Lasso

- Similar to Bayesian ridge regression, we can assign a hyperprior for λ that allows for simultaneous estimation of the regression coefficients and the amount of penalization.
- Since the precision of the regression coefficients is given by $\text{Var}(\beta_j)^{-1} = 2\lambda^2$, we assign a gamma distribution to λ^2, i.e., $\lambda^2 \sim GV(a_\lambda, b_\lambda)$.
- Summarizing, the joint prior for $\boldsymbol{\beta}$, τ_j^2, $j = 0, \ldots, k$, and λ factors as

$$p(\boldsymbol{\beta}, \tau_1^2, \ldots, \tau_k^2, \lambda) = p(\beta_0)\, p(\tilde{\boldsymbol{\beta}} \mid \tau_1^2, \ldots, \tau_k^2)\, p(\tau_1^2 \mid \lambda) \cdot \ldots \cdot p(\tau_k^2 \mid \lambda)\, p(\lambda^2).$$

Bayesian Lasso

- We finally compare the LASSO prior with the ridge prior.
- The right panel of Fig. 4 shows the log-prior $\log p(\beta_j \mid \lambda)$, for a single parameter β_j, conditional on the parameter λ and with the marginal log-prior with λ integrated out.
- As stated, the conditional prior is a Laplace distribution and therefore quite different from the Gaussian conditional log-prior in the case of ridge regression (see the left panel of the figure).
- Somewhat surprisingly, the marginal log-priors appear to be similar, although the LASSO prior still has heavier tails than the ridge prior.
- This is the reason why the Bayesian variants of ridge regression and the LASSO behave often very similar in empirical studies;

Posterior inference regularization priors

1 *Initialization:*

- Define initial values $\boldsymbol{\beta}^{(0)}$, $(\sigma^2)^{(0)}$ and $(\tau^2)^{(0)}$ (ridge), $(\tau_1^2)^{(0)}$, $\ldots, (\tau_k^2)^{(0)}, \lambda^{(0)}$ (LASSO).
- Set $t = 1$ and specify the number of iterations T.

2 *Sample $\boldsymbol{\beta}$:* Draw $\boldsymbol{\beta}^{(t)} \mid \cdot \sim \mathrm{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$ with $\boldsymbol{\mu}_\beta$ and $\mathring{}_\beta$ given by

$$\boldsymbol{\Sigma}_\beta = \left(\frac{1}{(\sigma^2)^{(t-1)}} \mathbf{X}'\mathbf{X} + \boldsymbol{K} \right)^{-1} \qquad \boldsymbol{\mu}_\beta = \frac{1}{(\sigma^2)^{(t-1)}} \boldsymbol{\Sigma}_\beta \mathbf{X}'\boldsymbol{y}.$$

Here $\boldsymbol{K} = 1/\tau^2 diag(0, 1, \ldots, 1)$ in case of ridge regression and $\boldsymbol{K} = (0, 1/\tau_1^2, \ldots, 1/\tau_k^2)$ in case of LASSO.

Posterior inference regularization priors

3 *Sample σ^2:* Sample $(\sigma^2)^{(t)}$ from the full conditional of σ^2 which is inverse gamma with parameters

$$a^{new} = a + \frac{n}{2}, \qquad b^{new} = b + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\beta^{(t)})'(\boldsymbol{y} - \boldsymbol{X}\beta^{(t)}).$$

4 *Sample variance parameters:*

- For the ridge prior draw $(\tau^2)^{(t)} \mid \cdot$ from an inverse gamma distribution with parameters

$$a^{new} = a + \frac{k}{2}, \qquad b^{new} = b + \frac{1}{2}\tilde{(}\beta^{(t)})'\tilde{\beta}^{(t)}.$$

- For the LASSO prior, sample

$$\begin{aligned}
(1/\tau_j^2)^{(t)} \mid \cdot \quad &\sim \quad \text{InvGauss}(\tfrac{|\lambda^{(t-1)}|}{|\beta_j^{(t)}|}, (\lambda^{(t-1)})^2), \\
(\lambda^2)^{(t)} \mid \cdot \quad &\sim \quad \text{G}\left(a + k, b + \tfrac{1}{2}\sum_{j=1}^{k}(\tau_j^2)^{(t)}\right).
\end{aligned}$$

1 Stop if $t = T$, otherwise set $t = t + 1$ and proceed with step 2.

Prices of Used Cars—Bayesian Ridge and LASSO

- This example compares the Bayesian variants of ridge regression and the LASSO with their classical counterparts.
- Using the software package `BayesX`, we obtained the estimates displayed in Table 3 together with their 95% credible intervals.
- The Bayesian ridge estimates behave quite similar to classical ridge regression. Both variants provide comparable results which are also very close to the unpenalized least squares estimate for these data.
- On the other hand, both LASSO variants show pronounced differences.
- Most striking is that the Bayesian LASSO does not allow removal of a covariate from the model, as is possible with the classical LASSO.

Prices of Used Cars—Bayesian Ridge and LASSO

- The reason is that the Bayesian LASSO point estimator is the posterior mean or median (rather than the posterior mode) estimated via MCMC.
- Since the posterior for the regression coefficients is typically skewed, the posterior mean and median will not coincide with the posterior mode, and, as a consequence, will always be different from zero.
- This is a distinct disadvantage of the Bayesian LASSO, as the main attraction of the classical LASSO is lost: the ability to perform variable selection.
- On the other hand, the sampling-based approach provides richer information regarding the posterior, such as posterior standard deviations and quantiles.

Prices of Used Cars—Bayesian Ridge and LASSO

- Finally we note that the Bayesian LASSO appears to induce less shrinkage than the classical LASSO.
- In fact, Bayesian ridge regression and LASSO show quite similar shrinkage behavior.
- Indeed, Table 3 shows that both regularization variants, albeit conceptually different, produce almost identical posterior estimates.
- This is in agreement with our theoretical findings, see Fig. 4.

Prices of Used Cars—Bayesian Ridge and LASSO

Variable	Bayesian ridge		Bayesian LASSO		Ridge	LASSO	LS
	Coeff.	95% CI	Coeff.	95% CI	Coeff.	Coeff.	Coeff.
ageop1	-0.682	(-0.802,-0.559)	-0.694	(-0.813, -0.563)	-0.672	-0.682	-0.709
ageop2	0.165	(0.052,0.280)	0.163	(0.052,0.270)	0.164	0.150	0.172
ageop3	0.014	(-0.089,0.128)	0.011	(-0.099,0.113)	0.015	-	0.016
kilometerop1	-0.428	(-0.541, -0.308)	-0.424	(-0.545,-0.303)	-0.425	-0.412	-0.437
kilometerop2	0.140	(0.028,0.252)	0.126	(0.012,0.244)	0.138	0.110	0.142
kilometerop3	0.013	(-0.103,0.125)	0.009	(-0.085,0.108)	0.010	-	0.009
TIA	-0.005	(-0.021,0.011)	-0.004	(-0.020,0.012)	-0.005	-	-0.005
extras1	-0.093	(-0.332,0.152)	-0.075	(-0.306,0.124)	-0.104	-0.036	-0.114
extras2	-0.030	(-0.257,0.211)	-0.022	(-0.240, 0.200)	-0.042	-	-0.031

Table: Prices of used cars: posterior mean and 95% credible intervals for Bayesian ridge regression and LASSO. For comparison the last three columns contain results for classical ridge and LASSO regression, as well as the least squares estimator

Classical Bayesian Model Choice

- Suppose we are given a number of potential covariates, and there is uncertainty as to which of the covariates should enter the model.
- For k possible regressors, there are 2^k different models when the intercept β_0 is always included. Denote the different models by M_r, $r = 1, \ldots, 2^k$.
- More specifically, M_r is given by

$$\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2, M_r = \boldsymbol{y} \mid \beta_0, \tilde{\boldsymbol{\beta}}_r, \sigma^2, M_r \sim \mathrm{N}(\beta_0 \mathbf{1} + \tilde{\boldsymbol{X}}_r \tilde{\boldsymbol{\beta}}_r, \sigma^2 \boldsymbol{I}),$$

  where the $n \times k_r$-design matrix $\tilde{\boldsymbol{X}}_r$ consists of all k_r covariates included in M_r, and $\tilde{\boldsymbol{\beta}}_r$ is the corresponding vector of regression coefficients.
- As usual, the vector $\boldsymbol{\beta}$ is the full vector of regression coefficients (including the intercept). Those components of $\boldsymbol{\beta}$ not contained in $\tilde{\boldsymbol{\beta}}_r$ are zero in model M_r.

Classical Bayesian Model Choice

- We assign the normal-inverse gamma prior to β_0, $\tilde{\boldsymbol{\beta}}_r$, and the error variance σ^2, such that

$$p(\beta_0, \tilde{\boldsymbol{\beta}}_r, \sigma^2 \mid M_r) = p(\beta_0, \tilde{\boldsymbol{\beta}}_r \mid \sigma^2, M_r)\, p(\sigma^2)$$

  with $\beta_0, \tilde{\boldsymbol{\beta}}_r \mid \sigma^2, M_r \sim \mathrm{N}(\boldsymbol{m}_r, \sigma^2 \boldsymbol{M}_r)$, and $\sigma^2 \mid M_r = \sigma^2 \sim IG(a, b)$.
- For completeness, we combine the zero components of $\boldsymbol{\beta}$ in the $(k - k_r)$-dimensional vector $\tilde{\boldsymbol{\beta}}_{-r}$ with a Dirac prior at $(0, \ldots, 0)'$, i.e.

$$\tilde{\boldsymbol{\beta}}_{-r} \mid M_r \sim Dirac(0, \ldots, 0).$$

  The Dirac prior concentrates all probability mass onto the point $(0, \ldots, 0)'$, i.e., $\mathrm{P}(\tilde{\boldsymbol{\beta}}_{-r} = (0, \ldots, 0)' \mid M_r) = 1$ and zero otherwise.
- Thus the prior for $\boldsymbol{\beta}$ and σ^2 under model M_r factors into

$$p(\boldsymbol{\beta}, \sigma^2 \mid M_r) = p(\tilde{\boldsymbol{\beta}}_{-r} \mid \beta_0, \tilde{\boldsymbol{\beta}}_r, \sigma^2, M_r)\, p(\beta_0, \tilde{\boldsymbol{\beta}}_r, \sigma^2 \mid M_r).$$

Classical Bayesian Model Choice

- To compare models using their posterior probabilities, we additionally have to assign prior probabilities to each model M_r.
- A popular prior is
$$p(M_r) = \theta^{k_r}(1-\theta)^{k-k_r}, \tag{16}$$
i.e., every possible covariate enters the model independently and with inclusion probability $\theta \in (0, 1)$.
- The "natural" choice $\theta = 1/2$ results in a uniform prior $p(M_r) = 1/2^k$, i.e., each model M_r has the same prior probability.

Classical Bayesian Model Choice

- The prior (16) with inclusion probability θ implies a certain prior distribution on the size of the models, denoted by S.
- Let δ_j, $j = 1, \ldots, k$, be inclusion indicators with $\delta_j = 1$ if covariate x_j is included in the model and 0 otherwise.
- Then δ_j has a Bernoulli distribution, i.e., $\delta_j \sim B(1, \theta)$, and the model size S has a binomial distribution with

$$S = \sum_{j=1}^{k} \delta_j \sim B(k, \theta),$$

resulting in a prior mean model size of $\mathrm{E}(S) = \theta \cdot k$ and variance $\mathrm{Var}(S) = \theta \cdot (1 - \theta) \cdot k$.

Classical Bayesian Model Choice

- In light of these results, the choice $\theta = 1/2$ seems less natural than suggested at first sight.
- In particular for a large number k of possible predictors, the prior expected model size appears to be much higher than one would typically expect in applications.
- For instance, for $k = 50$ potential covariates, the prior expected model size of $\mathrm{E}(S) = 25$ seems to be far too high for most applications.
- A convenient way to elicit the model prior (16) is to specify the prior mean model size $\mathrm{E}(S)$ and then to set $\theta = \mathrm{E}(S)/k$.
- For instance, if we have $k = 20$ potential regressors and assume a priori a model size of $\mathrm{E}(S) = 5$, then we must set $\theta = 5/20 = 1/4$.

Classical Bayesian Model Choice

- Based on our prior assumptions, the posterior probability for model M_r is given by

$$p(M_r \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid M_r)\, p(M_r)}{\sum_{h=1}^{2^k} p(\boldsymbol{y} \mid M_h)\, p(M_h)}, \tag{17}$$

  where $p(\boldsymbol{y} \mid M_r)$ is the marginal likelihood obtained as

$$p(\boldsymbol{y} \mid M_r) = \int p(\boldsymbol{y} \mid \beta_0, \tilde{\boldsymbol{\beta}}_r, \sigma^2, M_r)\, p(\beta_0, \tilde{\boldsymbol{\beta}}_r, \sigma^2 \mid M_r)\, d\beta_0\, d\tilde{\boldsymbol{\beta}}_r\, d\sigma^2.$$

- It can be shown that $p(\boldsymbol{y} \mid M_r)$ is multivariate t-distributed with $2a$ degrees of freedom, location parameter $\boldsymbol{X}_r \boldsymbol{m}_r$, and dispersion matrix $\boldsymbol{I} + \boldsymbol{X}_r \boldsymbol{M}_r \boldsymbol{X}_r'$, where $\boldsymbol{X}_r = (\mathbf{1}\, \tilde{\boldsymbol{X}}_r)$.

Classical Bayesian Model Choice

- The Bayes factor for two competing models M_r and M_s is then obtained as
$$BF_{rs} = \frac{p(\boldsymbol{y} \mid M_r)}{p(\boldsymbol{y} \mid M_s)}.$$
- With $\tilde{\boldsymbol{M}}_r = (\boldsymbol{X}_r'\boldsymbol{X}_r + \boldsymbol{M}_r^{-1})^{-1}$, we have

$$BF_{rs} = \left(\frac{|\tilde{\boldsymbol{M}}_r||\boldsymbol{M}_s|}{|\tilde{\boldsymbol{M}}_s||\boldsymbol{M}_r|}\right)^{1/2} \left(\frac{2a + (\boldsymbol{y} - \boldsymbol{X}_s\boldsymbol{m}_s)'(\boldsymbol{I} - \boldsymbol{X}_s\tilde{\boldsymbol{M}}_s\boldsymbol{X}_s')(\boldsymbol{y} - \boldsymbol{X}_s\boldsymbol{m}_s)}{2a + (\boldsymbol{y} - \boldsymbol{X}_r\boldsymbol{m}_r)'(\boldsymbol{I} - \boldsymbol{X}_r\tilde{\boldsymbol{M}}_r\boldsymbol{X}_r')(\boldsymbol{y} - \boldsymbol{X}_r\boldsymbol{m}_r)}\right)^{a+n/2}. \tag{18}$$

Classical Bayesian Model Choice

- Once the posterior probabilities $p(M_r \mid \boldsymbol{y})$ are computed for every model M_r under consideration, there are several ways to summarize the results.
- If the primary focus is on selecting one single (preferably sparse) model, often the model M_* with *highest posterior probability* is taken and inference for the regression coefficients is based on the posterior $p(\boldsymbol{\beta}_*, \sigma^2 \mid \boldsymbol{y}, M_*)$ conditional on model M_*.
- If model selection is done by minimizing the BIC, we exactly follow this strategy.
- However, Barbieri and Berger (2004) point out that often the model M_* with highest posterior probability is not optimal in terms of prediction.
- They show that the optimal predictive model is often the *median probability model*. This model consists of those covariates with posterior probability of 1/2 and higher for being in the model.

Classical Bayesian Model Choice

- Both approaches, however, ignore model uncertainty. In many applications, there are a number of models which are close in terms of posterior probabilities.
- If this is the case, inference for $\boldsymbol{\beta}$ (or any other quantity of interest) is better conducted by *model averaging* where the models are weighted by their posterior probability.
- More specifically, the posterior is given by

$$p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}) = \sum_{r=1}^{2^k} p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}, M_r)\, p(M_r \mid \boldsymbol{y}), \tag{19}$$

where $p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}, M_r)$ is the conditional posterior under model M_r and $p(M_r \mid \boldsymbol{y})$ is the corresponding posterior model probability given in Eq. (17).

Classical Bayesian Model Choice

- While models with high posterior probability make important contributions to the posterior, those with negligible posterior probability will contribute only very little information.
- If most models coincide in their posterior assessment of specific subvectors of $\boldsymbol{\beta}$, this assessment will also carry over to the model-averaged estimate.
- While the computation of the posteriors $p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}, M_r)$ and $p(M_r \mid \boldsymbol{y})$ required to obtain Eq. (19) is straightforward for a particular model M_r, it may be prohibitive for *all models*.

Classical Bayesian Model Choice

- The problem is that the number of possible models grows exponentially with k.
- For $k \leq 25$ regressors, enumeration of all models under consideration is usually possible in the available software packages. If k exceeds 25, more sophisticated algorithms are necessary.
- The model space then can be explored via MCMC simulation techniques. In doing so, we usually do not visit all models, but rather those models with relatively high posterior probabilities. One such Monte Carlo approach is the MC^3 algorithm of Madigan and York (1995), see the `R` package `BMS`.
- Other `R` packages for Bayesian variable selection (partly based on other methodology than described here) are `BAS` and `BMA`.

Prices of Used Cars—Bayesian Model Averaging (1)

- We illustrate the Bayesian approach for model choice using the data on the price of used cars.
- We assume possibly nonlinear effects for the variables *age* and *kilometer* modeled through the orthogonal cubic polynomials *ageop1*, *ageop2*, *ageop3* and *kilop1*, *kilop2*, *kilop3*.
- Together with the regressors *TIA*, *extras1*, and *extras2*, we have $k = 9$ potential covariates.
- We used the package `BMS` in `R` for the analysis; see Zeugner (2010) for a tutorial.

Prices of Used Cars—Bayesian Model Averaging (1)

- We started with a uniform prior for the models, i.e. $\theta = 1/2$ in Eq. (16).
- Table 4 provides a summary of the preliminary results. From left to right, the columns correspond to the variable names, the posterior inclusion probabilities (PIP), the posterior estimates for the regression coefficients together with their standard deviations, the probability of a positive sign for the respective coefficient, and (for comparison) the least squares estimates.
- The PIP is the ratio between the number of iterations with models that include a particular covariate and the total number of MCMC iterations.
- Similarly, the probability of a positive sign reflects the ratio between iterations with positive sign for a covariate and the total number of iterations.

Prices of Used Cars—Bayesian Model Averaging (1)

Variable	PIP	Mean	Std.dev	Cond. pos sign	LS
ageop1	1.00	-0.7065	0.0621	0	-0.7085
ageop2	0.94	0.1823	0.0721	1	0.1716
ageop3	0.07	0.0009	0.0156	1	.0162
kilop1	1.00	-0.4345	0.0616	0	-.4366
kilop2	0.61	0.0872	0.0827	1	0.1417
kilop3	0.07	0.0011	0.0160	1	0.0090
TIA	0.08	-0.0003	0.0025	0	-0.0051
extras1	0.11	-0.0127	0.0565	0	-0.1135
extras2	0.07	-0.0029	0.0374	0	-0.0315

Table: Prices of used cars: posterior inclusion probabilities (PIP), model averaged estimated coefficients and standard deviations, probabilities of positive sign, and (for comparison) the ordinary least squares estimates. The results are based on a uniform prior for the models ($\theta = 0.5$).

Prices of Used Cars—Bayesian Model Averaging (1)

- For the covariates *ageop1*, *ageop2*, *kilop1*, and *kilop2* with high inclusion probabilities (> 0.5), the Bayesian estimator averaged over the models is quite close to the ordinary least squares estimator.
- For the remaining covariates, the inclusion probabilities are very low, and the model averaged estimates are shrunken to zero compared to least squares.
- As can be seen from Table 5, the covariates with high inclusion probability also define the two models with by far the highest posterior probabilities.
- The top model, with posterior model probability of 0.37, consists exactly of the four covariates with inclusion probabilities greater than 0.5.

Prices of Used Cars—Bayesian Model Averaging (1)

- In this case, the median probability model, that consists of those covariates with posterior probability of 1/2 and higher for being in the model, coincides with the model with largest posterior probability.
- The second model, with posterior probability 0.25, additionally excludes the variable *kilop2*.
- All other models have comparably low posterior probabilities.

Prices of Used Cars—Bayesian Model Averaging (1)

- We conclude the example by an investigation of the sensitivity of results on prior assumptions, particularly on the prior expected model size $\mathrm{E}(S)$.
- The upper panel of Fig. 5 compares the PIP for the three models with $\theta = 1/2$, $\theta = 2/9$, and $\theta = 8/9$ corresponding to expected model sizes $\mathrm{E}(S) = 4.5$, $\mathrm{E}(S) = 2$, and $\mathrm{E}(S) = 8$, respectively.
- The results are quite sensitive to the choice of θ and $\mathrm{E}(S)$. Except for *ageop1*, *ageop2*, and *kilop1*, the PIP differ considerably, e.g., for *kilop2* from 0.33 ($\mathrm{E}(S) = 2$), then 0.61 ($\mathrm{E}(S) = 4.5$), to 0.92 ($\mathrm{E}(S) = 8$).
- We will explain the reasons of this undesirable behavior below.

Prices of Used Cars—Bayesian Model Averaging (1)

Variable	T1	T2	T3	T4	T5
ageop1	+	+	+	+	+
ageop2	+	+	+	+	+
ageop3	-	-	-	-	-
kilop1	+	+	+	+	+
kilop2	+	-	+	+	+
kilop3	-	-	-	-	-
TIA	-	-	-	+	-
extras1	-	-	+	-	-
extras2	-	-	-	-	+
PMP	0.375	0.250	0.040	0.032	0.030

Table: Prices of used cars: top five models T1-T5 with highest posterior probabilities. The results are based on a uniform prior for the models ($\theta = 0.5$). A plus (minus) sign indicates that the variable is included in (excluded from) the model. The last row displays the posterior model probabilities (PMP)

Prices of Used Cars—Bayesian Model Averaging (1)

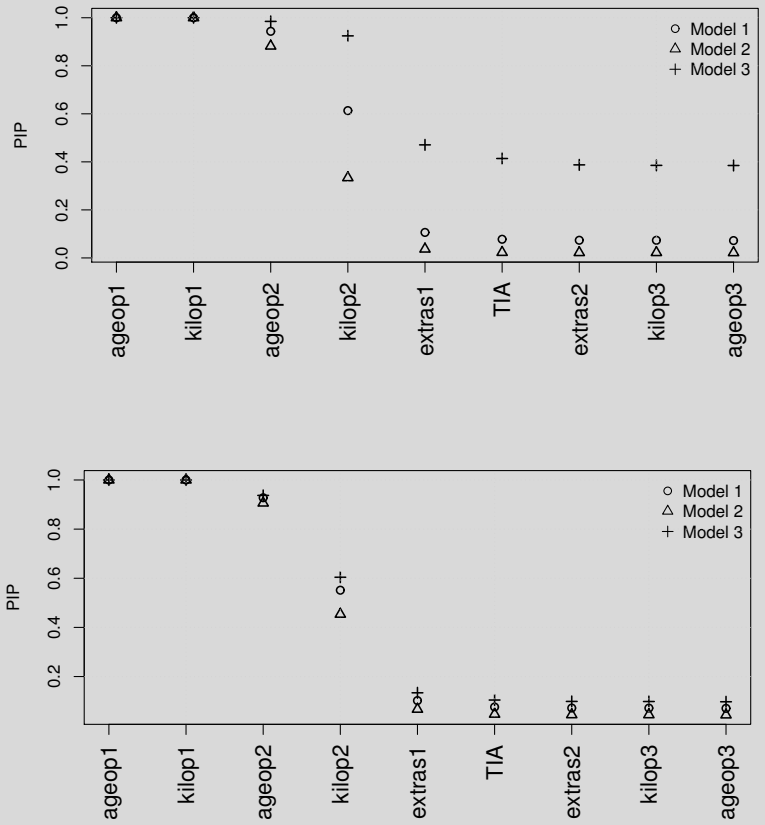

Figure: Prices of used cars: comparison of posterior inclusion probabilities with $\theta = 1/2$ (Model 1, corresponding to expected model size $\mathrm{E}(S) = 4.5$), $\theta = 2/9$ (Model 2, $\mathrm{E}(S) = 2$) and $\theta = 8/9$ (Model 3, $\mathrm{E}(S) = 8$). The *upper panel* corresponds to fixed θ. The *lower panel* corresponds to a beta hyperprior for θ

Classical Bayesian Model Choice

- The example shows that the results can be strongly affected by the prior choice for θ and the corresponding prior on the model size.
- Indeed, if θ is chosen as $\theta = \mathrm{E}(S)/k$, then the induced prior for the model size S places relatively small probability mass on model sizes that are moderately far away from the mean $\mathrm{E}(S)$.
- This is illustrated with the following figures which show, both with $k = 10$ regressors (left column) and with $k = 40$ regressors (right column), some priors for the model size based on different choices for θ (dashed lines).
- In particular, with an asymmetric prior for expected model size $\mathrm{E}(S)$ and for $k = 40$, a broad range of possible model sizes has virtually no prior probability mass.
- As a result, estimates are often quite sensitive to the choice of the prior for expected model size.

Classical Bayesian Model Choice

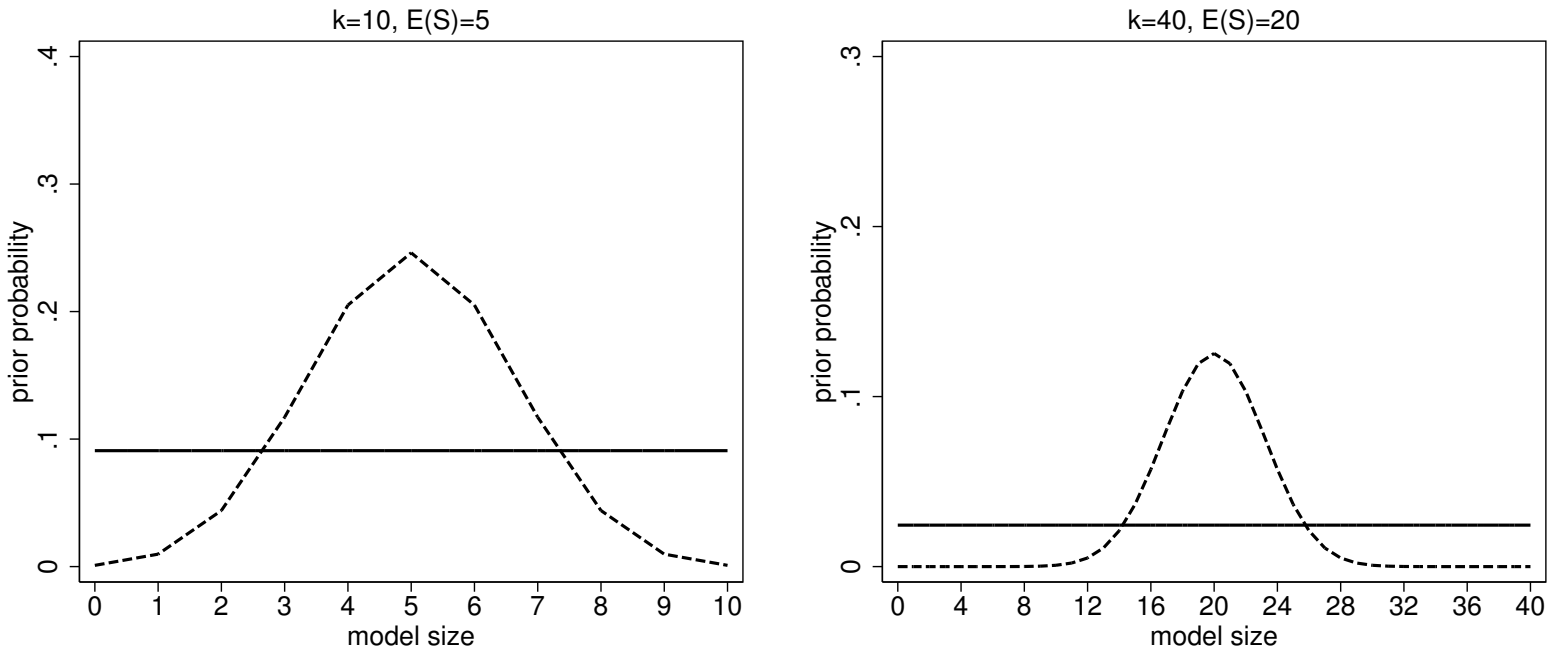

Figure: Priors for model size. The *left column* shows various priors for model size with $k = 10$ potential regressors; the right column corresponds to $k = 40$. The *solid lines* correspond to a $Beta(a, b)$ hyperprior for θ. The *dashed lines* correspond to fixed θ.

Classical Bayesian Model Choice

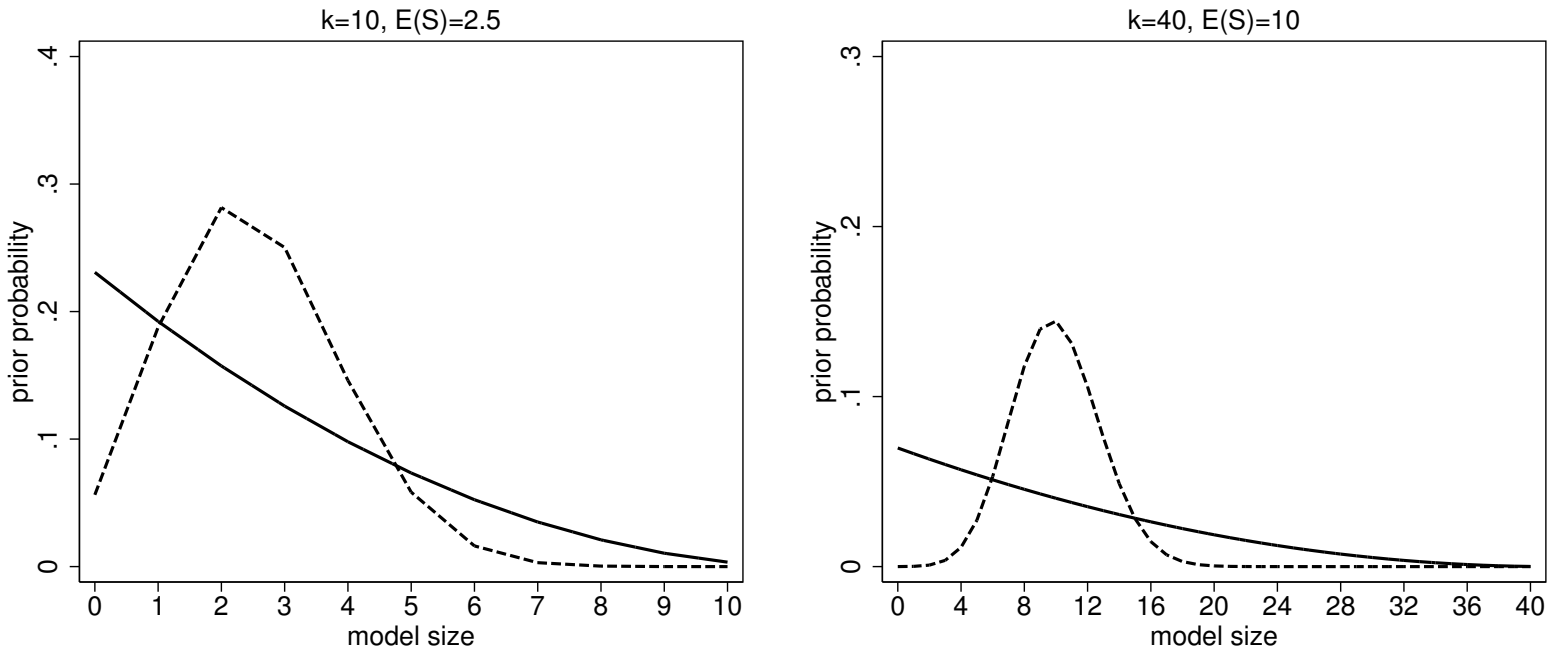

Figure: Priors for model size. The *left column* shows various priors for model size with $k = 10$ potential regressors; the right column corresponds to $k = 40$. The *solid lines* correspond to a $Beta(a, b)$ hyperprior for θ. The *dashed lines* correspond to fixed θ.

Classical Bayesian Model Choice

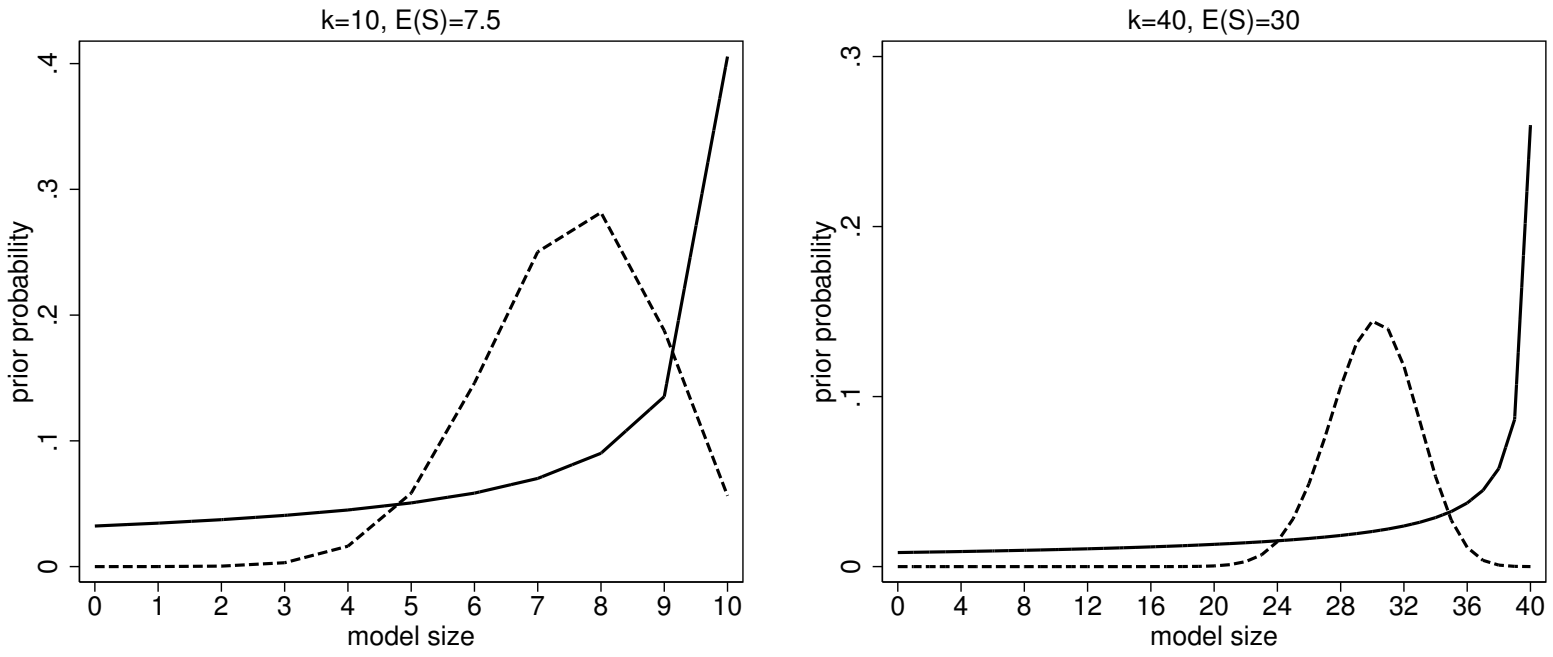

Figure: Priors for model size. The *left column* shows various priors for model size with $k = 10$ potential regressors; the right column corresponds to $k = 40$. The *solid lines* correspond to a $Beta(a, b)$ hyperprior for θ. The *dashed lines* correspond to fixed θ.

Classical Bayesian Model Choice

- To reduce the dependence of results on the prior choice, a remedy is to introduce a hyperprior in a further stage of the hierarchy.
- In our case, a flexible and convenient choice for θ is to assume a beta distribution with hyperparameters a and b, i.e. $\theta \sim Beta(a, b)$.
- Since $S \,|\, \theta \sim B(k, \theta)$, the marginal distribution for S is beta-binomial, i.e. $S \sim BetaB(k, a, b)$ with prior model size distribution

$$\mathrm{P}(S = s) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+k)} \binom{k}{s} \Gamma(a+s)\Gamma(b+k-s),$$

and mean and variance given by

$$\begin{aligned} \mathrm{E}(S) &= \tfrac{a}{a+b} k, \\ \mathrm{Var}(S) &= \tfrac{ab(a+b+k)}{(a+b)^2(a+b+1)} k. \end{aligned} \tag{20}$$

Classical Bayesian Model Choice

- To ease initiation of the prior, we fix $a = 1$, which still allows for very flexible priors. To choose b, we rearrange $\mathrm{E}(S)$ in Eq. (20) to obtain $b = (k - \mathrm{E}(S))/\mathrm{E}(S)$. Hence, similar to fixed θ, we can choose the prior expected model size to fully specify the prior.
- As can be seen from the figures above, the (marginal) prior for model size S shows much more variability compared to fixed θ so that all possible model sizes have positive probability mass.

Prices of Used Cars—Bayesian Model Averaging (2)

- We rerun the three regressions of our example, again with prior expected model size $\mathrm{E}(S) = 2, 4.5, 8$, but now with a beta hyperprior for θ.
- The bottom panel of Fig. 5 compares the PIP for the nine possible covariates.
- The probabilities are now quite close to each other. An exception are the PIP for *kilop2*, where there is some inclusion uncertainty.
- For two of the three model priors (with $\mathrm{E}(S) = 4.5$ and $\mathrm{E}(S) = 8$) the best model with highest posterior probability consists of the covariates *ageop1*, *ageop2*, *kilop1*, and *kilop2*.
- The model that excludes *kilop2* is the second best model. It is also the best model for prior expected model size $\mathrm{E}(S) = 2$.
- All other models have very low posterior probabilities.

Prices of Used Cars—Bayesian Model Averaging (2)

- In summary, we are quite certain that only two covariates, the age of the car and the kilometer reading, are relevant predictors for the price.
- A second-order polynomial is sufficient to model the nonlinear effects of the two covariates.
- There is some uncertainty whether a linear effect for kilometer reading is sufficient. Note also that the quadratic effect of kilometer reading is already close to linearity.

Bayesian Inference

Chapter 7

Appendix: Some distributions

Gamma distribution

A continuous nonnegative random variable X is said to have a gamma distribution with parameters $a > 0$ and $b > 0$ if it has pdf

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp\left(-bx\right), \qquad x > 0.$$

The mean and variance are given by

$$\begin{aligned} \mathrm{E}(X) &= a/b, \\ \mathrm{Var}(X) &= a/b^2. \end{aligned}$$

The mode is $(a-1)/b$ (for $a > 1$). We write $X \sim G(a, b)$.

Gamma distribution

An alternative parametrization of the pdf, depending on $\mu = \mathsf{E}(X)$ and scale parameter $\nu > 0$ is

$$f(x) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^{\nu} x^{\nu-1} \exp\left(-\frac{\nu}{\mu} x\right), \qquad x > 0.$$

This alternative pdf is used, for example, for gamma regression models.

Inverse Gamma distribution

If $Y \sim G(a, b)$, then $X = 1/Y$ has an inverse gamma distribution with pdf

$$f(x) = \frac{b^a}{\Gamma(a)} x^{-(a+1)} \exp\left(-b/x\right), \qquad x > 0.$$

The mean and variance are given by

$$\begin{aligned} \mathrm{E}(X) &= b/(a-1), \quad a > 1, \\ \mathrm{Var}(X) &= b^2/((a-1)^2(a-2)), \quad a > 2. \end{aligned}$$

We write $X \sim IG(a, b)$.

Normal-Inverse Gamma Distribution

Let $\boldsymbol{Y}$ be a $p \times 1$ dimensional random vector and S be a random variable. The random vector $\boldsymbol{X} = (\boldsymbol{Y}, S)'$ is said to have a normal-inverse gamma distribution with parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, a and b if

$$\begin{aligned} \boldsymbol{Y} \mid S &\sim N(\boldsymbol{\mu}, S\boldsymbol{\Sigma}), \\ S &\sim IG(a, b). \end{aligned}$$

We write $\boldsymbol{X} = (\boldsymbol{Y}, S)' \sim NIG(\boldsymbol{\mu}, \boldsymbol{\Sigma}, a, b)$. The density of the distribution is given by

$$\begin{aligned} f(\boldsymbol{y}, s) &= \frac{1}{(2\pi)^{\frac{p}{2}} |s\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2s}(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right) \\ &\quad \times \frac{b^a}{\Gamma(a)} \frac{1}{s^{a+1}} \exp\left(-\frac{b}{s}\right). \end{aligned}$$

Normal-Inverse Gamma Distribution

The $NIG(\boldsymbol{\mu}, \boldsymbol{\Sigma}, a, b)$-distribution has the following properties:

1. $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{\mu}$.
2. $\mathrm{Cov}(\boldsymbol{Y}) = b/(a-1)\boldsymbol{\Sigma}$.
3. $\mathrm{E}(S) = b/(a-1)$ provided that $a > 1$.
4. $\mathrm{Var}(S) = b^2/[(a-1)^2(a-2)]$ provided that $a > 2$.
5. $\boldsymbol{Y} \sim \mathrm{t}(2a, \boldsymbol{\mu}, b/a\boldsymbol{\Sigma})$.

Multivariate t-Distribution

- A continuous p-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_p)'$ is said to have a multivariate t-distribution with ν degrees of freedom, location parameter $\boldsymbol{\mu}$ and (positive definite) dispersion matrix $\boldsymbol{\Sigma}$, if it has pdf

$$f(\boldsymbol{x}) = |\boldsymbol{\Sigma}|^{-\frac{1}{2}} (\nu\pi)^{-\frac{p}{2}} \frac{\Gamma((\nu+p)/2)}{\Gamma(\nu/2)} \left(1 + \frac{(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{\nu}\right)^{-(\nu+p)}$$

- We write $\boldsymbol{X} \sim \mathrm{t}(\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- The expectation is $\boldsymbol{\mu}$ (provided that $\nu > 1$) and the covariance matrix is $\nu/(\nu-2)\boldsymbol{\Sigma}$ (provided that $\nu > 2$).

Multivariate t-Distribution

- Note that a diagonal dispersion matrix $\boldsymbol{\Sigma}$ corresponds to uncorrelated components of the random vector $\boldsymbol{X}$. In contrast to the multivariate normal distribution the components are, however, not stochastically independent.
- Any subvector of $\mathbf{X}$ has a (multivariate) t-distribution with ν degrees of freedom and the corresponding subvector of $\boldsymbol{\mu}$ and the submatrix of $\mathring{}$ as location parameter and dispersion matrix, respectively.

Beta Distribution

- A continuous random variable X is said to have a beta distribution with parameters $a > 0$ and $b > 0$ if it has probability function

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \qquad x \in (0,1),$$

  where $\Gamma(\cdot)$ is the gamma function.
- The mean and the variance are given by

$$\begin{aligned} \mathrm{E}(X) &= \frac{a}{a+b}, \\ \mathrm{Var}(X) &= \frac{ab}{(a+b)^2(a+b+1)}. \end{aligned}$$

- We write $X \sim \mathrm{Beta}(a, b)$.
- For $a = b = 1$, we obtain a uniform distribution on the interval $(0, 1)$.

Beta-Binomial Distribution

- A discrete random variable X is said to have a beta-binomial distribution with parameters $n \in \{1, 2, \dots\}$, $a > 0$, $b > 0$ if it has probability function

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+n)} \binom{n}{x} \Gamma(a+x)\Gamma(a+n-x) \qquad x = 0, 1, 2, \dots$$

- The mean and the variance are given by

$$\begin{aligned} \mathrm{E}(X) &= n\frac{a}{a+b}, \\ \mathrm{Var}(X) &= n\frac{ab}{(a+b)^2}\frac{a+b+n}{a+b+1}. \end{aligned}$$

- We write $X \sim \mathrm{BetaB}(n, a, b)$.

Beta-Binomial Distribution

- For $a = b = 1$, the beta-binomial distribution corresponds to a discrete uniform distribution on $0, 1, \ldots, n$, i.e. $f(x) = 1/(n+1)$.
- The beta-binomial distribution arises as a mixture distribution. Suppose $X \mid \pi \sim B(n, \pi)$ and $\pi \sim Beta(a, b)$, then $X \sim BetaB(n, a, b)$.

Bayesian Inference

Chapter 8

References

- Barbieri, M.M. and Berger, J.O (2004): Optimal predictive model selection. *The Annals of Statistics*, 32, 870–897.
- **Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2022): Regression. Models, Methods and Applications (second edition). Springer Verlag**
- Fernandez, C., Ley, E. and Steel, M. F. J. (2001): Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100, 381–427.
- Foster, D.P. and George, E.I. (1994): The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22, 1947–1975.
- George, E.I. and Foster, D.P. (2000): Calibration and empirical Bayes variable selection. *Biometrica*, 87, 731–747.
- **Held, L. and Sabanés Bové, D., 2021: Likelihood and Bayesian Inference, Springer Verlag.**

- Madigan, D. and York, J. (1995): Bayesian graphical models for discrete data. *International Statistical Review*, 63, 215–232.
- O'Hagan, A. (1994): Kendall's Advanced Theory of Statistics Vol. 2B: Bayesian Inference. Arnold.
- Zeugner, S. (2010): Bayesian Model Averaging with BMS. Available at http://bms.zeugner.eu/tutorials/bms.pdf.