Introduction (review?) to Bayesian Statistics

Vanda Inácio

University of Edinburgh

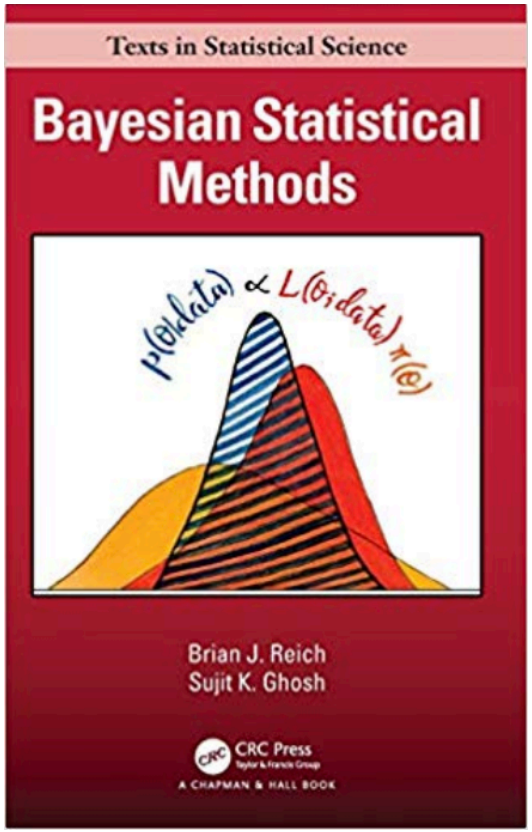
December 2023

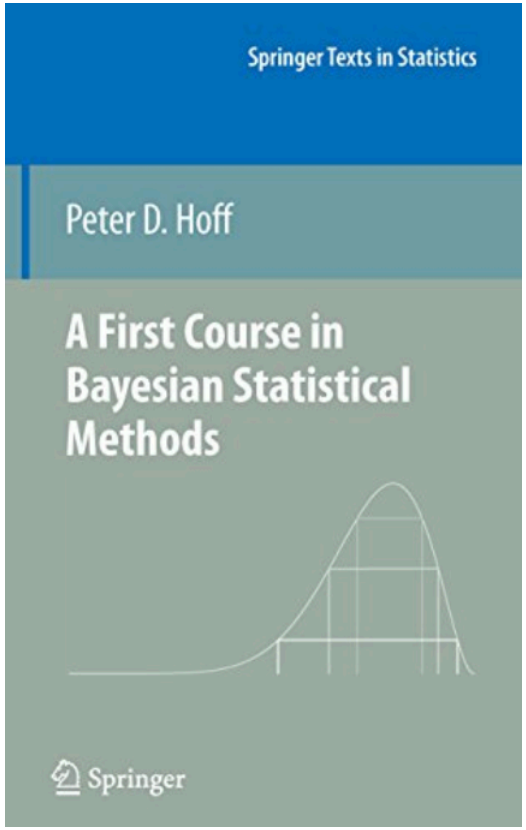
General aspects

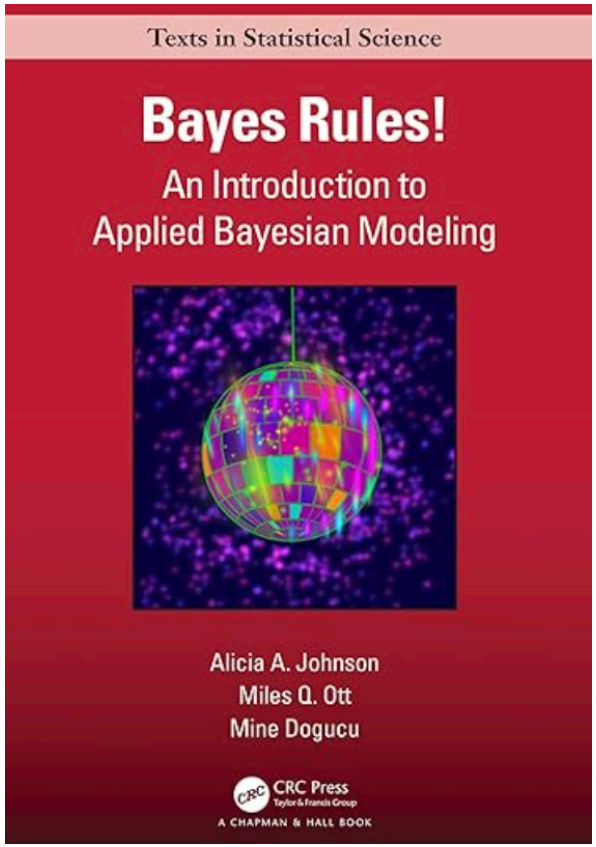
- ↪ General aspects
- ↪ Point and interval estimation
- ↪ Choice of the prior distribution
- ↪ Bayesian computational tools
- ↪ Bayesian model comparison criteria
- ↪ Bayesian model checking

Useful textbooks

Covers from amazon.co.uk

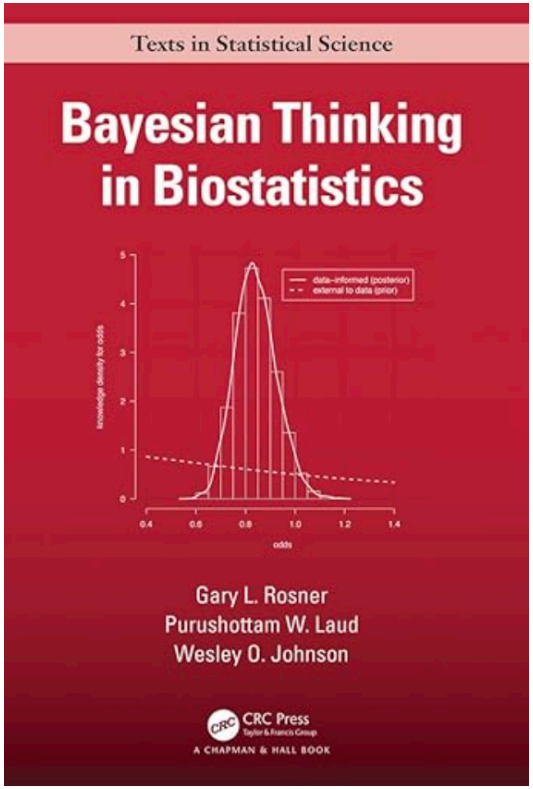


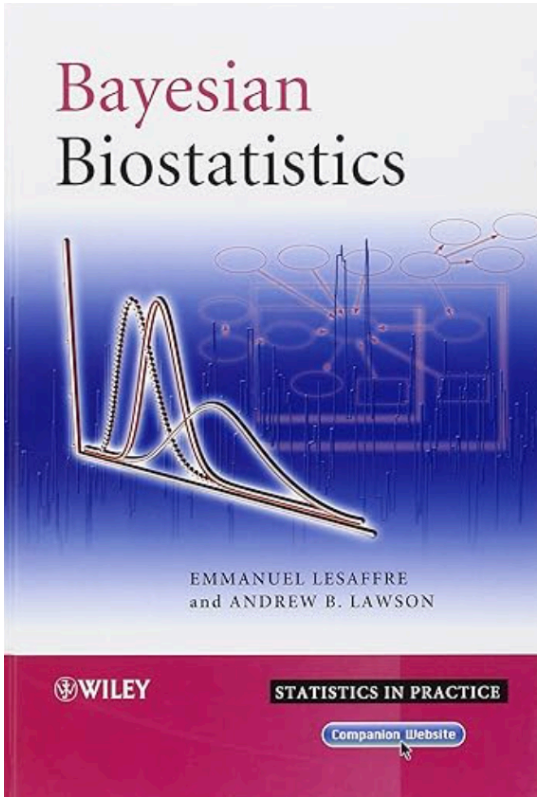




Useful textbooks

Covers from amazon.co.uk





General aspects

- ↪ In classical/frequentist statistics, data are regarded as random and a sampling model is chosen.
- ↪ Given the parametric (sampling) model, the parameters of such model are considered fixed, typically unknown, constants.
- ↪ Inference about the parameters is often done through the likelihood function. Maximum likelihood estimates are then obtained.
- ↪ The frequentist approach can be regarded as a procedure that quantifies uncertainty in terms of repeating the process that generated the data many times.

General aspects

- ↪ In Bayesian statistics, both the data and the parameters of the model are treated as random.
- ↪ As such, a probability distribution expressing the prior belief about which values of the parameter are plausible must be specified.
- ↪ The prior belief can range from basically ‘ignorance’ to a strong subjective belief.
- ↪ The prior distribution is combined with the likelihood, through the Bayes theorem, to form the posterior distribution (of the parameters given the data).
- ↪ The posterior distribution contains all relevant information about a model’s parameters, and thus all statistical inference should be based on it.

Posterior distribution

- Let $y_1, \ldots, y_n$ be a random sample and let $\boldsymbol{\theta}$ be a continuous parameter vector (or a scalar).

- The main 'ingredients' to conduct Bayesian inference about $\boldsymbol{\theta}$ are the likelihood function, which we will be denoting by $L(\boldsymbol{\theta}; \mathbf{y})$, and the prior distribution, $p(\boldsymbol{\theta})$.

- Bayes theorem gives a 'recipe' to conduct posterior inference about $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{L(\boldsymbol{\theta}; \mathbf{y}) p(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} L(\boldsymbol{\theta}; \mathbf{y}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}. \tag{1}$$

- The interpretation of (1) is as follows: when the prior information about θ, expressed by $p(\boldsymbol{\theta})$, is combined with the observed data, expressed through the likelihood, the prior information is updated and expressed by $p(\boldsymbol{\theta} \mid \mathbf{y})$, the posterior distribution.

Posterior distribution

- The denominator in equation (1) (also known as normalising constant) ensures that $p(\boldsymbol{\theta} \mid \mathbf{y})$ is a proper density, i.e., that it integrates to one.

- Note that this normalising constant $\int_{\boldsymbol{\Theta}} L(\boldsymbol{\theta}; \mathbf{y}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$ is nothing more than the marginal likelihood of the data or the marginal density.

- Given the denominator in (1) does not depend on the parameter (which is integrated out), we can rewrite

$$p(\theta \mid \mathbf{y}) \propto L(\theta; \mathbf{y}) p(\theta).$$

- In words,

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

Summarising the posterior distribution – univariate case

Point estimation

- A univariate posterior is best summarised with a plot because this retains all information about the parameter.
- However, it is also common to report both a point and an interval estimate.
- Common one number summaries of the posterior used as point estimates are the posterior mean, median, or mode.
- For symmetric posterior distributions, the mean and the median will, of course, be identical.
- For symmetric unimodal posterior distributions, the three measures will coincide.
- For asymmetric posterior distributions, the choice is less clear, although the median is often preferred since it is intermediate to the mode (which considers only the value corresponding to the maximum value of the density) and the mean (which often gives too much weight to extreme outliers).

Summarising the posterior distribution – univariate case

Point estimation

- More formally, Bayesian point estimation requires specifying a loss function.
- If θ is the true value of the parameter and $\widehat{\theta}(y)$ is the estimator, then the loss function is defined as $\ell(\theta, \widehat{\theta}(y))$.
- For instance, we might choose the squared error loss $\ell(\theta, \widehat{\theta}(y)) = \{\theta - \widehat{\theta}(y)\}^2$, or the absolute loss $\ell(\theta, \widehat{\theta}(y)) = |\theta - \widehat{\theta}(y)|$.
- The Bayes risk is the expected loss (with respect to the posterior distribution of θ)

$$R(\widehat{\theta}(y)) = \mathbf{E}[\ell(\theta, \widehat{\theta}(y)) \mid \mathbf{y}] = \int_{\Theta} \ell(\theta, \widehat{\theta}(y)) p(\theta \mid \mathbf{y}) \mathrm{d}\theta.$$

- The Bayes estimator estimator is the estimator $\widehat{\theta}(y)$ that minimises the Bayes risk

$$\widehat{\theta}_{\text{Bayes}} = \arg\min_{\widehat{\theta}(y)} \mathbf{E}[\ell(\theta, \widehat{\theta}(y)) \mid \mathbf{y}].$$

↪ Common estimators and their corresponding loss functions:

↪ Mean: $\widehat{\theta}_{\text{Bayes}} = \mathbb{E}[\theta \mid \mathbf{y}] = \int_{\Theta} \theta p(\theta \mid \mathbf{y}) \mathrm{d}\theta$ minimises the **squared** loss.

↪ Median: $\int_{\widehat{\theta}_{\text{Bayes}}}^{\infty} p(\theta \mid \mathbf{y}) \mathrm{d}\theta = \frac{1}{2}$ minimises the **absolute** loss.

↪ Mode: $\widehat{\theta}_{\text{Bayes}} = \arg\max p(\theta \mid \mathbf{y})$ (also called maximum a posteriori–MAP) minimises the **'zero–one'** loss, $\ell(\theta, \widehat{\theta}(y)) = 1 - \delta(\theta - \widehat{\theta}(y))$.

- The posterior mode is analogous to the MLE, since it is the parameter value that maximizes the posterior pdf, while the MLE maximizes the likelihood. In large samples, they will often be very close to each other.

- Historically, squared error loss functions have been popular due to their analytical tractability for obtaining MSE in frequentist calculations. But that is no longer a real issue due to modern high powered computing.

- The absolute error loss is preferable since it is on the scale of the data.

- The zero one error loss is interesting, but perhaps not in line with intuition that involves assigning greater penalties in estimation for being very far away from the truth.

- Let us check that, in fact, the posterior mean minimises the expected squared-error loss.
- To alleviate notation I will be using $\widehat{\theta}$ instead of $\widehat{\theta}(\mathbf{y})$. We have

$$\ell(\theta, \widehat{\theta}) = \left(\theta - \widehat{\theta}\right)^2 = \theta^2 - 2\theta\widehat{\theta} + \widehat{\theta}^2.$$

- Therefore,

$$\mathbb{E}[\ell(\theta, \widehat{\theta}) \mid \mathbf{y}] = \mathbb{E}[\theta^2 \mid \mathbf{y}] - 2\widehat{\theta}\mathbb{E}[\theta \mid \mathbf{y}] + \widehat{\theta}^2.$$

- We want to find the $\widehat{\theta}$ that minimises the above Bayesian risk. Let us take the derivative

$$\frac{\mathrm{d}}{\mathrm{d}\widehat{\theta}}\mathbf{E}[\ell(\theta, \widehat{\theta}) \mid \mathbf{y}] = -2\mathbb{E}[\theta \mid \mathbf{y}] + 2\widehat{\theta},$$

which implies that

$$\frac{\mathrm{d}}{\mathrm{d}\widehat{\theta}}\mathbb{E}[\ell(\theta, \widehat{\theta}) \mid \mathrm{y}] = 0 \Rightarrow \widehat{\theta} = \mathbb{E}[\theta \mid \mathbf{y}].$$

- Checking the second derivative

$$\frac{\mathrm{d}^2}{\mathrm{d}\widehat{\theta}^2} = 2.$$

- So $\widehat{\theta} = \mathbb{E}[\theta \mid \mathbf{y}]$ minimises the expected squared error loss.

↪ A $100(1-\alpha)\%$, for $\alpha \in (0,1)$, **credible interval** is any interval (L, U) such that

$$\int_L^U p(\theta \mid \mathbf{y}) \mathrm{d}\theta = 1 - \alpha.$$

↪ This definition allows direct probability statements about the likelihood of θ falling in (L, U). We can say that the probability of θ belonging to the interval (L, U) given the data $\mathbf{y}$ is $1 - \alpha$.

↪ This is in contrast to the usual frequentist confidence interval, for which the corresponding interpretation would be that, if for a larger number of datasets collected in the same way as ours, we calculate the interval (L, U), then about $(1 - \alpha) \times 100\%$ of them will contain the true value of θ.

↪ Note that there are infinitely many intervals (L, U) with coverage $1 - \alpha$.

↪ The easiest to compute is the equal tailed interval for which

$$\frac{\alpha}{2} = \int_{-\infty}^{L} p(\theta \mid \mathbf{y}) \mathrm{d}\theta = \int_{U}^{\infty} p(\theta \mid \mathbf{y}) \mathrm{d}\theta.$$

↪ An alternative is the highest posterior density (HPD) interval which searches over L and U to minimise the interval width $U - L$, while maintaining the posterior coverage to be $100(1 - \alpha)\%$.

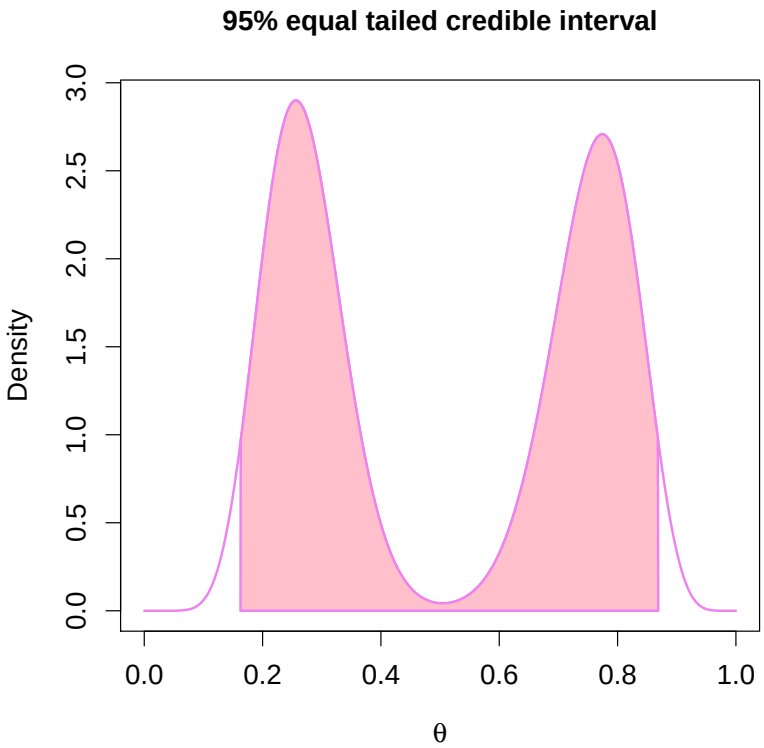
↪ The HPD interval is a credible interval that, in addition, satisfies

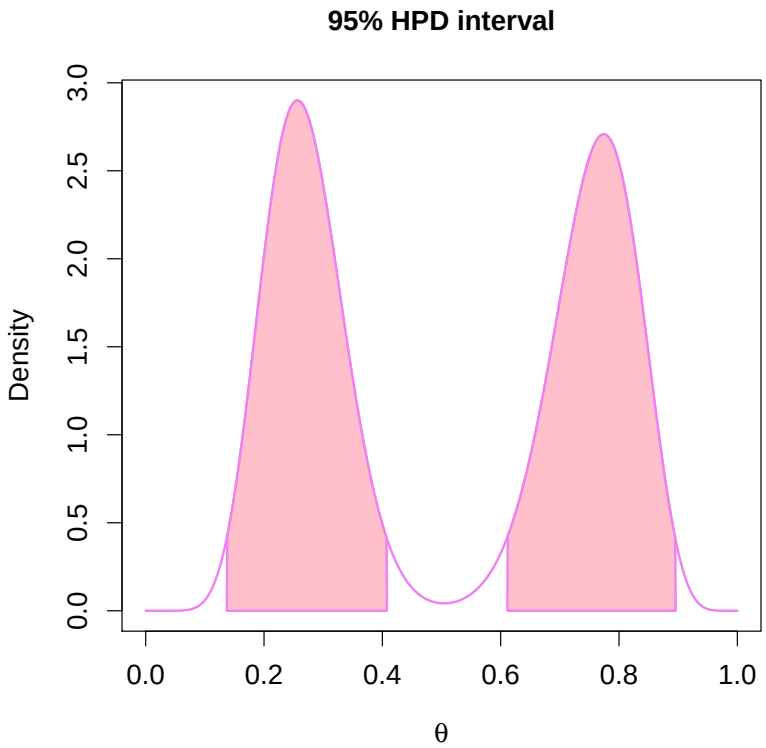
$$p(\theta_1 \mid \mathbf{y}) \geq p(\theta_2 \mid \mathbf{y}), \quad \forall \theta_1 \in (L, U) \text{ and } \theta_2 \notin (L, U).$$

↪ If the posterior is unimodal and symmetric, the equal tailed and HPD intervals are the same.

Summarising the posterior distribution – univariate case

Interval estimation





Summarising the posterior distribution – multivariate case

- In the case of a high dimensional parameter, plotting the posterior distribution will be challenging, if not impossible.
- A typical remedy is to marginalise out other parameters and summarise univariate marginal distributions with plots, point estimates, credible intervals, and maybe plots of a few bivariate marginal distributions.
- For instance, suppose that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ and that we are interested to make inferences about θ_1. The marginal posterior for θ_1 is

$$p(\theta_1 \mid \mathbf{y}) = \int \cdots \int p(\theta_1, \ldots, \theta_p \mid \mathbf{y}) \mathrm{d}\theta_2 \ldots \mathrm{d}\theta_p.$$

- In practice, this integral might not be feasible to calculate, but we can 'overcome the problem' by simulation.

Summarising the posterior distribution – multivariate case

↪ The main idea is to randomly draw from the joint posterior distribution

$$\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(S)} \sim p(\boldsymbol{\theta} \mid \mathbf{y}),$$

where each $\boldsymbol{\theta}^{(s)} = (\theta_1^{(s)}, \ldots, \theta_p^{(s)})'$ and collect all together the first component of each draw, i.e., $\theta_1^{(1)}, \ldots, \theta_1^{(S)}$ then these are a sample of $p(\theta_1 \mid \mathbf{y})$.

↪ We will learn in a few slides how to simulate from the joint posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{y})$.

Choice of the prior distribution

- Choosing the prior distribution is obviously important and uniquely Bayesian.
- Some type of prior distributions include:
  - Informative/expert priors.
  - Non-informative/vague priors.
  - Conjugate priors.

Choice of the prior distribution

Informative/expert priors

- A major advantage of the Bayesian approach is the ability to include expert prior information.
- This can come from either an expert in the field or from past data (literature, pilot study, etc).
- We call elicitation to the process of extracting prior knowledge in a suitable manner to permit the formulation of a suitable prior distribution that represents the expert/historical information as accurately as possible.
- Spiegelhalter et al. (2004), sections 5.2–5.4, contain a good discussion on how priors might be elicited from experts or historical data.

Choice of the prior distribution

Informative/expert priors

- ↪ A basic principle in specifying informative distributions is that we will specify information about parameters that are in some sense, on the scale of the data, which are observable.
- ↪ When random sampling data, Y, from a population, the population mean, $\mathbb{E}(Y \mid \mu) = \mu$, are on the scale of the data. Since we can observe values of $Y = y$, it is possible to think about and to know something about μ, especially if we have sampled the same or similar populations in the past.
- ↪ If we have observed data in the past on the proportion of individuals with a particular characteristic, then we are able to specify information about that parameter before we observe new data from the same population.
- ↪ On the other hand, if the data, Y, are modelled as $Y \sim \text{LN}(\mu, \sigma^2)$, the parameter μ is not on the scale of the data. Since the median of the distribution of Y is e^{μ}, it is e^{μ} that is on the scale of the data.
- ↪ In this situation, we would elicit prior information about the median of the observed data, and then induce a prior distribution on μ from that.

Choice of the prior distribution

Informative/expert priors

- What about if we ask more than one expert and they don't agree?
- Suppose we are interested in a quantity θ and that expert j recommends prior $\theta \sim \mathrm{N}(\mu_j, \sigma_j)$.
- We could weight the experts using a mixture model

$$p(\theta) = \sum_{j=1}^{J} \omega_j \mathrm{N}(\theta \mid \mu_j, \sigma_j^2),$$

where ω_j is the weight given to expert j ($\sum_{j=1}^{J} \omega_j = 1$).

Choice of the prior distribution

Non-informative priors: improper priors

- What should we do if we do not have any prior information concerning the parameter of interest?
- As a concrete example, assume $Y_i \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$, with σ^2 known and μ unknown and to be estimated.
- In the absence of any knowledge it is tempting to use $p(\mu) = c$ ($c > 0$) for all possible $\mu \in (-\infty, \infty)$.
- This can be thought either as an approximation to a uniform prior distribution with bounds tending to infinity or a normal distribution with variance tending to infinity.
- However, for any c, we have that $\int_{-\infty}^{\infty} p(\mu) \text{d}\mu = \infty$, and so the prior is improper.

Choice of the prior distribution

Non-informative priors: improper priors

- Note that
$$L(\mu) \propto e^{-\frac{n}{2\sigma^2}(\mu-\bar{y})^2}.$$
- Then $p(\mu \mid \mathbf{y}) \propto L(\mu)p(\mu)$, which results in $\mu \mid \mathbf{y} \sim \text{N}(\bar{\mathbf{y}}, \sigma^2/n)$.
- As this example shows, an improper prior can lead to a proper posterior distribution, but we should be cautious when using improper priors and posterior propriety must always be checked.
- Note that in this case posterior inferences would be identical to frequentist inferences since
$$0.95 = \Pr(\bar{\mathbf{y}} - 1.96\,\sigma/\sqrt{n} < \mu < \bar{\mathbf{y}} + 1.96\,\sigma/\sqrt{n} \mid \mathbf{y}),$$
which implies that $\bar{\mathbf{y}} \pm 1.96\,\sigma/\sqrt{n}$ is a 95% probability interval for μ.
- This interval formula is identical to the comparable frequentist confidence interval formula, but as previously described, with a very different interpretation.

Choice of the prior distribution

Non-informative priors: improper priors

- ↪ Alternatively, we could have just placed a N(0, b^2) prior on μ with a very large value for b, say $b = 1000$.
- ↪ In most if not all situations, the posterior would be virtually identical to this one.
- ↪ The problem with overly diffuse comes up if we model a probability, say θ, as $\log\{\theta/(1-\theta)\} = \beta$, and then let $\beta \sim$ N(0, b^2) with $b = 1000$.
- ↪ This is related to Bayesian logistic regression modelling, which has been used by many. This particular specification has been used in the literature going back to the 1990's, before it was realized that it is a terribly dis-informative prior.
- ↪ It is a prior on the probability θ that will be U shaped with huge mass very close to 0 and also very close to 1, unless n is quite large. Posterior inferences necessarily shrink towards zero or one, for no good reason.

Choice of the prior distribution

Conjugate priors

- A prior and a likelihood pair are conjugate if the resulting posterior distribution is a member of the same family of distributions as the prior.

- That is, suppose that the prior distribution $p(\boldsymbol{\theta})$ belongs to a class of parametric distributions $\mathcal{F}$. We say that the prior distribution $p(\boldsymbol{\theta})$ is conjugate with respect to the likelihood $L(\boldsymbol{\theta}; \mathbf{y})$ if the posterior distribution also belongs to $\mathcal{F}$, i.e.,

$$p(\boldsymbol{\theta}) \in \mathcal{F} \Rightarrow p(\boldsymbol{\theta} \mid \mathbf{y}) \in \mathcal{F}.$$

- Because both the prior and posterior are members of the same family, the update from the prior to the posterior affects only the parameters that index the family.

Choice of the prior distribution

Conjugate priors

- There are long lists of conjugacies that we should be aware of:

  `https://en.wikipedia.org/wiki/Conjugate_prior`

- Conjugate priors are not unique. For example, the beta prior is conjugate for both the binomial and negative binomial likelihood.

- Conjugate priors are somewhat limited because not all likelihood functions have a known conjugate prior and most conjugate pairs are for small examples with only a few parameters.

- However, in fancier models, conjugate priors facilitate Gibbs sampling, which is the easiest Bayesian computational algorithm.

Choice of the prior distribution

Conjugate priors: beta–binomial model

↪ Let us suppose $Y \sim \text{Bin}(n, \theta)$, whose likelihood is

$$L(\theta; y) = \binom{n}{y} \theta^y (1-\theta)^{n-y}, \quad 0 < \theta < 1, \quad y \in \{0, 1, \ldots, n\}.$$

↪ The kernel of the likelihood resembles a beta distribution, and so we might suspect that a beta distribution is the conjugate prior for the binomial likelihood.

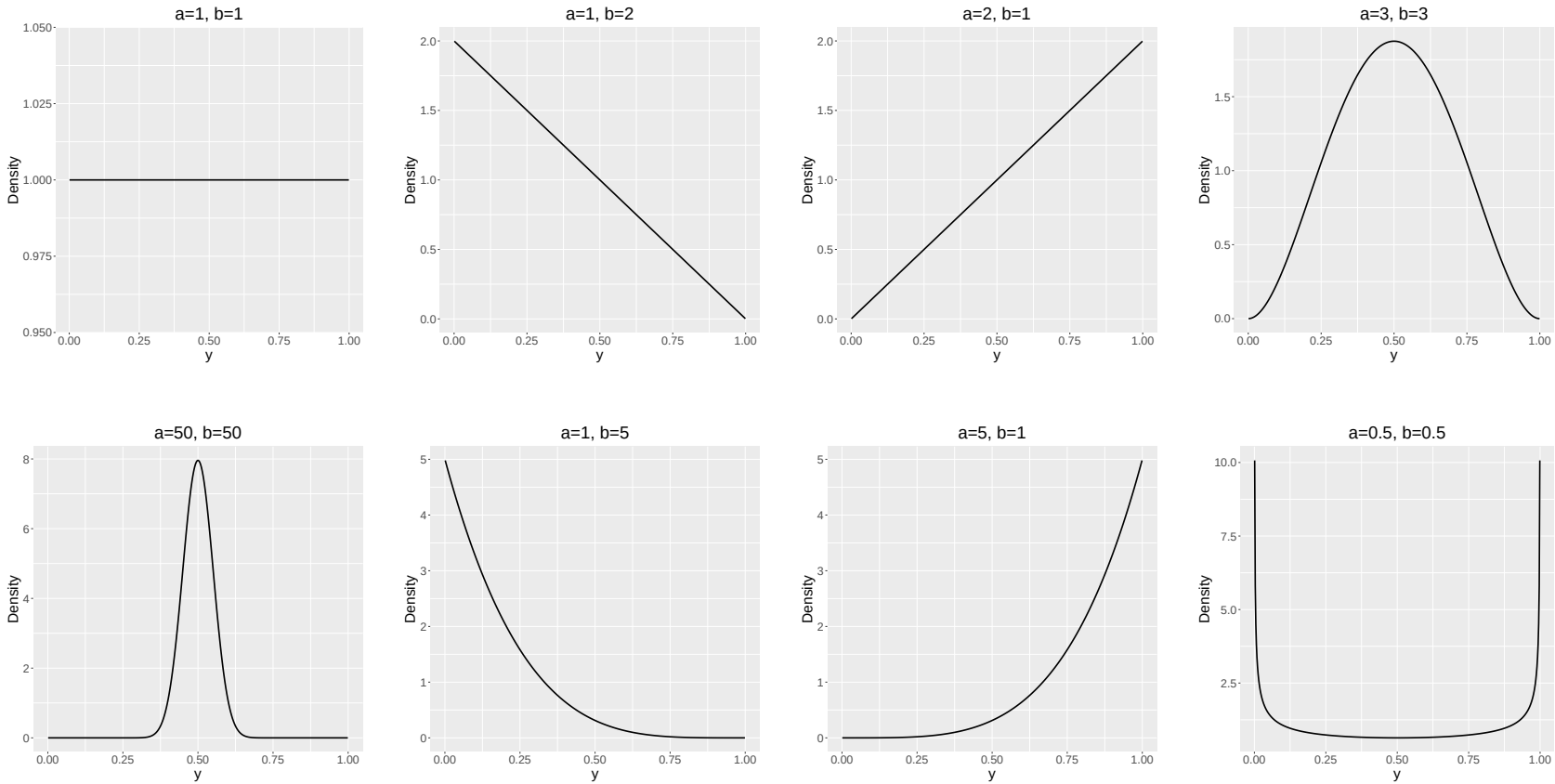
↪ Let $\theta \sim \text{Beta}(a, b)$. The density function is

$$p(\theta) = \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}.$$

Choice of the prior distribution

Conjugate priors: beta–binomial model

↪ The beta distribution can take several shapes.



Choice of the prior distribution

Conjugate priors: beta–binomial model

- We can interpret the prior parameters in the following way:
  - a: prior number of successes.
  - b: prior number of failures.
  - $a + b$: prior number of trials.
  - $\mathbb{E}(\theta) = \frac{a}{a+b}$: prior mean.
- This interpretation may help in the elicitation of the parameters a and b of the beta distribution.

Choice of the prior distribution

Conjugate priors: beta–binomial model

↪ To obtain the posterior distribution we can follow two routes:

1. The hard way
   - ↪ Derive the marginal likelihood of the data $p(y) = \int L(\theta; y)p(\theta)\mathrm{d}\theta$.
   - ↪ Derive $p(\theta \mid y) = L(\theta; y)p(\theta)/p(y)$.
2. The easy way
   - ↪ Derive $L(\theta; y)p(\theta)$ and recognise it as the kernel of some known distribution.

Choice of the prior distribution

Conjugate priors: beta–binomial model

↪ Let us start with the hard way and derive, firstly, the marginal likelihood of the data.

$$
\begin{aligned}
p(y) &= \int L(\theta; y)p(\theta)\mathrm{d}\theta \\
&= \int \binom{n}{y}\theta^{y}(1-\theta)^{n-y}\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta \\
&= \binom{n}{y}\frac{1}{B(a,b)}\int \theta^{a+y-1}(1-\theta)^{b+n-y-1}\mathrm{d}\theta \\
&= \binom{n}{y}\frac{B(a+y,b+n-y)}{B(a,b)}\int \frac{1}{B(a+y,b+n-y)}\theta^{a+y-1}(1-\theta)^{b+n-y-1}\mathrm{d}\theta \\
&= \binom{n}{y}\frac{B(a+y,b+n-y)}{B(a,b)},
\end{aligned}
$$

which is known as the beta–binomial distribution.

Choice of the prior distribution

Conjugate priors: beta–binomial model

↪ Now

$$
\begin{aligned}
p(\theta \mid y) &= \frac{L(\theta; y)p(\theta)}{p(y)} \\
&= \frac{\binom{n}{y} \frac{1}{B(a,b)} \theta^{a+y-1}(1-\theta)^{b+n-y-1}}{\binom{n}{y} \frac{B(a+y,b+n-y)}{B(a,b)}} \\
&= \frac{\theta^{a+y-1}(1-\theta)^{b+n-y-1}}{B(a+y, b+n-y)}.
\end{aligned}
$$

↪ Thus, $\theta \mid y \sim \text{Beta}(a+y, b+n-y)$.

Choice of the prior distribution

Conjugate priors: beta–binomial model

↪ Now, through the easy way

$$\begin{aligned} p(\theta \mid y) &\propto L(\theta; y)p(\theta) \\ &\propto \theta^{y}(1-\theta)^{n-y}\theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{a+y-1}(1-\theta)^{b+n-y-1}, \end{aligned}$$

which we recognise as the kernel of a beta distribution with parameters $a+y$ and $b+n-y$, i.e., $\theta \mid y \sim \text{Beta}(a+y, b+n-y)$.

Choice of the prior distribution

Conjugate priors: beta–binomial model

↪ The posterior mean is

$$\begin{aligned}\mathbb{E}(\theta) &= \frac{a+y}{a+b+n} \\ &= \frac{a}{a+b+n} + \frac{y}{a+b+n} \\ &= \frac{a+b}{a+b+n}\left(\frac{a}{a+b}\right) + \frac{n}{a+b+n}\left(\frac{y}{n}\right) \quad = \omega\mathbb{E}(\theta) + (1-\omega)\widehat{\theta}_{\text{MLE}}, \quad \omega = \frac{a+b}{a+b+n}.\end{aligned}$$

↪ Thus, the posterior mean is a weighted average of the prior mean and the MLE.

↪ This confirms the intuition that for any a and b, if the sample size n is small, the posterior mean is approximately the prior mean and, as the sample size increases, the posterior mean becomes closer to the MLE (the sample proportion).

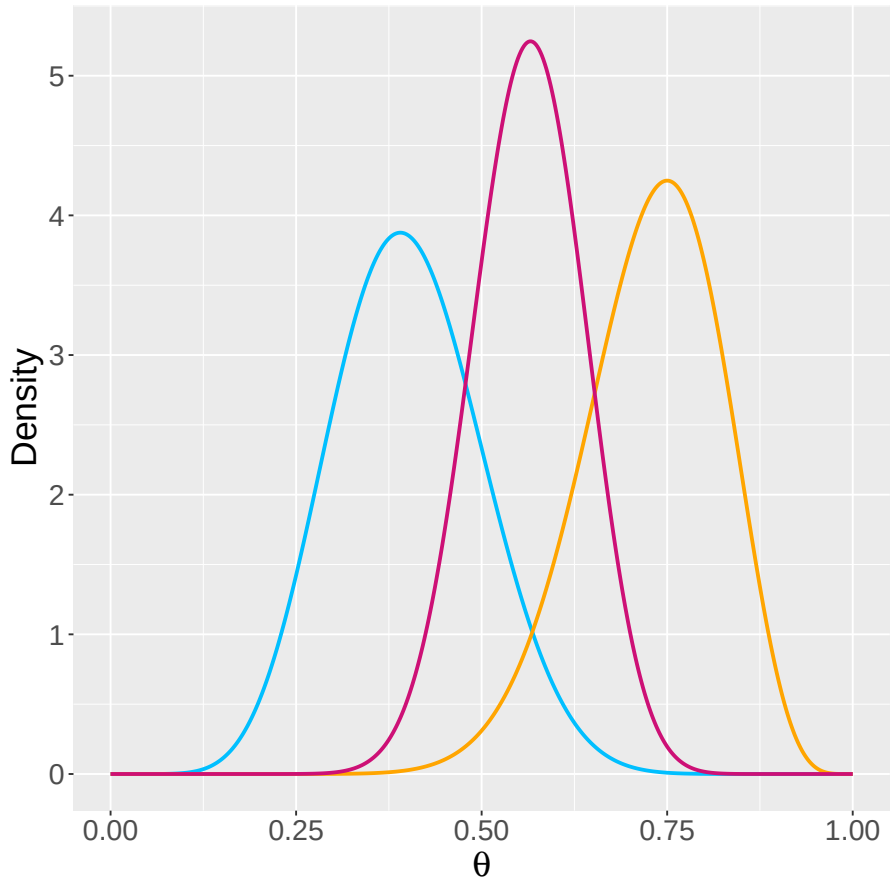
Choice of the prior distribution

Conjugate priors: beta–binomial model

- Consider a drug to be given for relief of chronic pain.
- Experience with similar compounds has suggest that response rates, say θ, between 0.2 and 0.6 could be feasible.
- Interpret this as a distribution with mean 0.4 and standard deviation 0.1.
- A Beta(9.2, 13.8) distribution has these properties.
- Suppose we treat $n = 20$ volunteers with the compound and observe $y = 15$ positive responses.
- The parameters of the Beta distribution are updated to $9.2 + 15 = 24.2$ and $13.8 + 20 - 15 = 18.8$.

Choice of the prior distribution

Conjugate priors: beta–binomial model



Blue line: prior distribution. Pink line: posterior distribution. Orange line: scaled likelihood.

Choice of prior distribution

Conjugate priors: beta-binomial model

- ↪ Note that the likelihood, although a function of the parameter, it is not a density and so, in particular, does not integrate to one.
- ↪ In order to plot the likelihood along with the prior and posterior distributions, it is convenient that the three are in the same scale
- ↪ Therefore, we have rescaled the likelihood function so that it integrates to one.

Choice of the prior distribution

Conjugate priors: normal model, unknown mean, known variance

↪ Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$, with σ^2 fixed.

↪ The likelihood is

$$L(\mu; \sigma^2, \mathbf{y}) = \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \mu)^2 \right\} \right\}$$
$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\}.$$

↪ Let us consider $\mu \sim \text{N}(\mu_0, \sigma_0^2)$, with μ_0 and σ_0^2 fixed.

Choice of the prior distribution

Conjugate priors: normal model, unknown mean, known variance

↪ The posterior distribution is

$$
\begin{aligned}
p(\mu \mid \sigma^2, \mathbf{y}) &\propto L(\mu; \sigma^2, \mathbf{y}) p(\mu) \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{ -\frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right\} \\
&\propto \exp\left\{ -\frac{1}{2\sigma^2} (-2\mu n\bar{y} + n\mu^2) - \frac{1}{2\sigma_0^2} (\mu^2 - 2\mu\mu_0) \right\} \\
&= \exp\left\{ -\frac{1}{2} \left(\frac{-2\mu n\bar{y}\sigma_0^2 + n\mu^2\sigma_0^2 + \mu^2\sigma^2 - 2\mu\mu_0\sigma^2}{\sigma^2\sigma_0^2} \right) \right\} \\
&= \exp\left\{ -\frac{1}{2} \left(\frac{\mu^2(n\sigma_0^2 + \sigma^2) - 2\mu(n\bar{y}\sigma_0^2 + \mu_0\sigma^2)}{\sigma^2\sigma_0^2} \right) \right\} \\
&= \exp\left\{ -\frac{1}{2} \left(\frac{\mu^2 - 2\mu(n\bar{y}\sigma_0^2 + \mu_0\sigma^2)/(n\sigma_0^2 + \sigma^2)}{\sigma^2\sigma_0^2/(n\sigma_0^2 + \sigma^2)} \right) \right\}
\end{aligned}
$$

Choice of the prior distribution

Conjugate priors: normal model, unknown mean, known variance

↪ Let

$$m = \frac{n\bar{y}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2}.$$

↪ The posterior distribution takes then the form

$$\begin{aligned}
p(\mu \mid \sigma^2, \mathbf{y}) &\propto \exp\left\{-\frac{1}{2}\left(\frac{\mu^2 - 2\mu m}{\sigma^2\sigma_0^2/(n\sigma_0^2 + \sigma^2)}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{\mu^2 - 2\mu m + m^2 - m^2}{\sigma^2\sigma_0^2/(n\sigma_0^2 + \sigma^2)}\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{\mu^2 - 2\mu m + m^2}{\sigma^2\sigma_0^2/(n\sigma_0^2 + \sigma^2)}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{(\mu - m)^2}{\sigma^2\sigma_0^2/(n\sigma_0^2 + \sigma^2)}\right)\right\}
\end{aligned}$$

Choice of the prior distribution

Conjugate priors: normal model, unknown mean, known variance

- We recognise the last line in the posterior distribution expression in the previous slides as being the kernel of a normal distribution with mean $m = \frac{n\bar{y}\sigma_0^2+\mu_0\sigma^2}{n\sigma_0^2+\sigma^2}$ and variance $\sigma^2\sigma_0^2/(n\sigma_0^2+\sigma^2)$.

- That is,

$$\mu \mid \sigma^2, \mathbf{y} \sim \mathrm{N}\left(\frac{\frac{\mu_0}{\sigma_0^2}+\frac{n\bar{y}}{\sigma^2}}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}\right)$$

- The posterior precision (inverse of the variance) $\tilde{\tau}_\mu$ is the sum of the prior precision and the data precision, i.e.,

$$\tilde{\tau}_\mu = \sigma_0^{-2} + n\sigma^{-2}.$$

Choice of the prior distribution

Conjugate priors: normal model, unknown mean, known variance

↪ Notice that for the posterior mean we have

$$\tilde{\mu} = \frac{\sigma_0^{-2}}{\sigma_0^{-2} + n\sigma^{-2}}\mu_0 + \frac{n\sigma^{-2}}{\sigma_0^{-2} + n\sigma^{-2}}\bar{y},$$

and so the posterior mean is a weighted average of the prior mean and of the sample mean. The weight of the sample mean is n/σ^2, the sampling variance of the sample mean. The weight of the prior mean is $1/\sigma_0^2$, the prior variance.

↪ **Remark**: The improper prior we have used in slide 22 is the limit when $\sigma_0^2 \to \infty$, as

$$\lim_{\sigma_0^2 \to \infty} \mathrm{N}(\mu_0, \sigma_0^2) \to c,$$

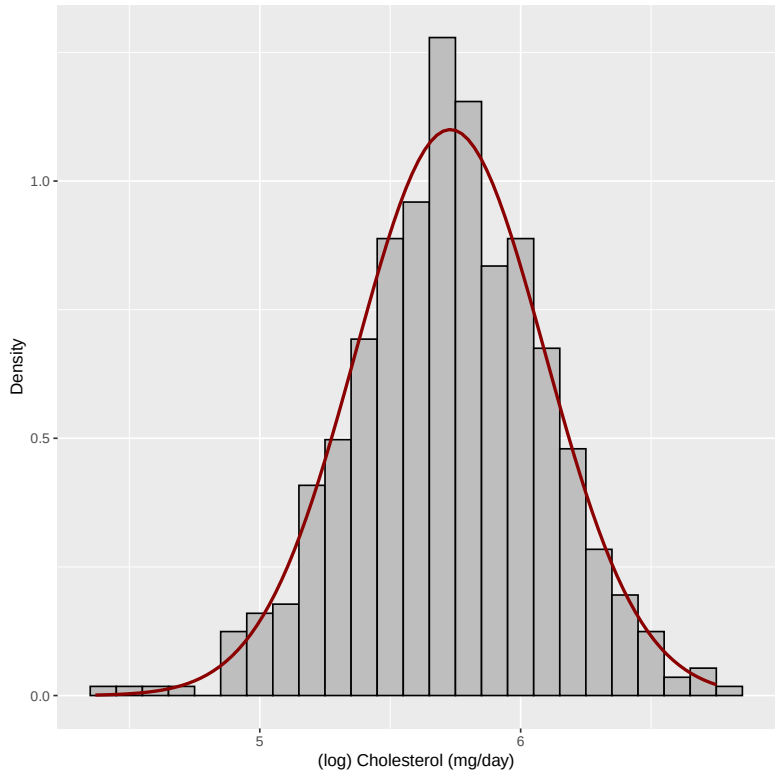
and

$$\lim_{\sigma_0^2 \to \infty} \mathrm{N}\left(\frac{\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \right) \to \mathrm{N}(\bar{y}, \sigma^2/n).$$

Choice of the prior distribution

Conjugate priors: normal model, unknown mean, known variance

- We use as an example the (log) cholesterol levels of 563 bank employees, collected in 1990, in Belgium (Lesaffre and Lawson, 2012, Bayesian Biostatistics).

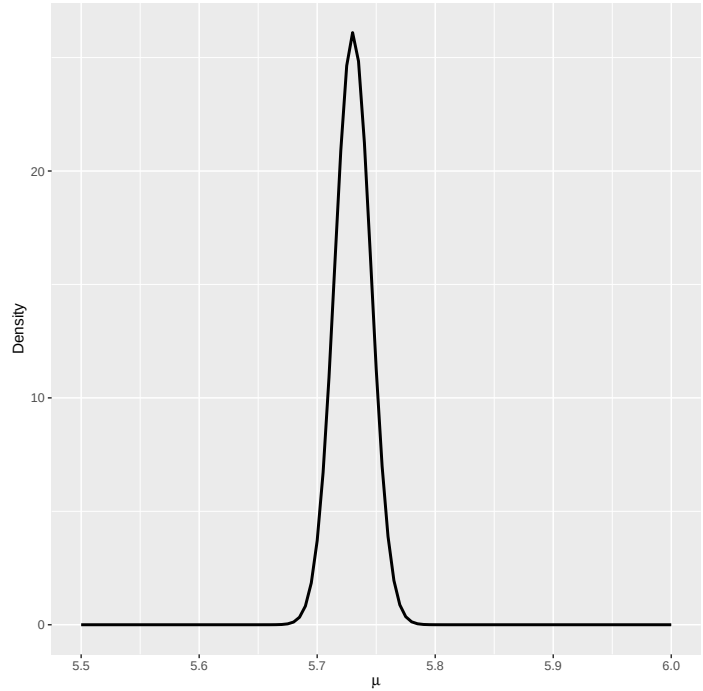


- Histogram of the log cholesterol levels along with the normal fit.

Choice of the prior distribution

Conjugate priors: normal model, unknown mean, known variance

↪ Letting $\sigma^2 = s^2 = 0.132$ and the prior mean and variance to be 0 and 100, i.e., $\mu_0 = 0$ and $\sigma_0^2 = 100$, respectively, the resulting posterior density for μ is shown below.



↪ We further have that $\mathbb{E}(\mu \mid \sigma^2, \mathbf{y}) = 5.730$ and that a 95% credible interval for μ is $(5.700, 5.760)$.

Choice of the prior distribution

Conjugate priors: normal model, known mean, unknown variance

- We now turn attention to estimating a normal variance assuming the mean is fixed.
- As before $Y_i \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$, but now μ id fixed. The likelihood function is the same as before, but now is viewed as a function of σ^2

$$L(\sigma^2; \mu, \mathbf{y}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}.$$

- The prior distribution for σ^2 should be a distribution with support on $(0, \infty)$.
- Let us consider $\sigma^2 \sim \text{IG}(a, b)$, where IG stands for inverse gamma and whose density is

$$p(\sigma^2) = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-(a+1)}e^{-b/\sigma^2}, \quad 0 < \sigma^2 < \infty, \quad a, b > 0.$$

- Note that if $X \sim \text{Gamma}(a, b)$, then $Y = \frac{1}{X} \sim \text{IG}(a, b)$.

Choice of the prior distribution

Conjugate priors: normal model, known mean, unknown variance

↪ The resulting posterior distribution is

$$\begin{aligned} p(\sigma^2 \mu, \mathbf{y}) &\propto L(\sigma^2; \mu, \mathbf{y}) p(\sigma^2) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\} \frac{b^a}{\Gamma(a)} (\sigma^2)^{-(a+1)} e^{-b/\sigma^2} \\ &\propto (\sigma^2)^{(-a+n/2+1)} \exp\left\{ -\frac{1}{\sigma^2} \left(b + \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^2 \right) \right\}. \end{aligned}$$

↪ Sure enough, this is the kernel of an inverse gamma density

$$\sigma^2 \mid \mu, \mathbf{y} \sim \text{IG}\left(a + \frac{n}{2}, b + \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^2 \right).$$

Choice of the prior distribution

Conjugate priors: normal model, known mean, unknown variance

- Since the precision parameter in a normal density is just the inverse of the variance, if we do inference on one, we can easily derive the corresponding inference for the other.

- Specifically, if $\tau = 1/\sigma^2$, and the prior on σ^2 is specified as an IG(a, b) distribution, then the equivalent prior induced on τ is

$$\tau \sim \text{Gamma}(a, b).$$

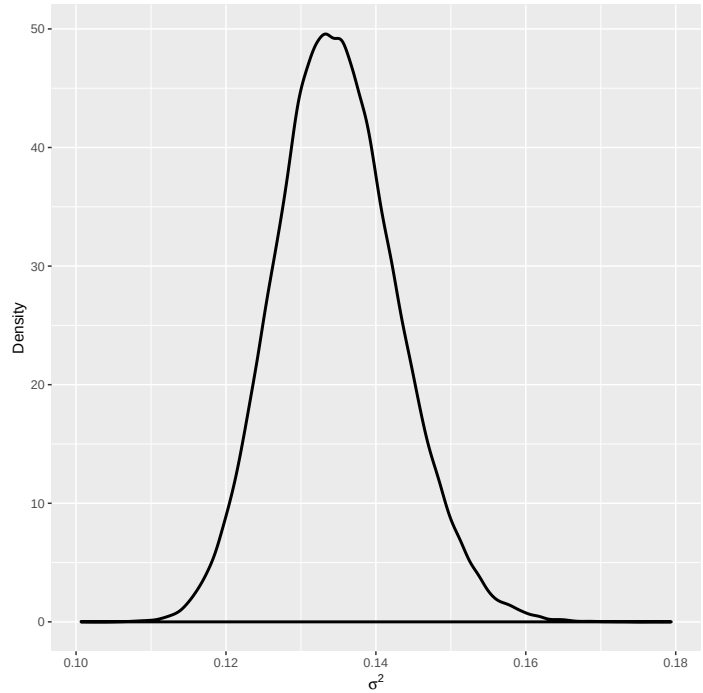
- The resulting posterior distribution is

$$\tau \mid \mu, \mathbf{y} \sim \text{Gamma}\left(a + \frac{n}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2\right).$$

Choice of the prior distribution

Conjugate priors: normal model, known mean, unknown variance

↪ Letting $\mu = \bar{y} = 5.730$ and $\sigma^2 \sim \text{IG}(a, b)$ with $a = 1$ and $b = 1$, leads to the following posterior distribution



↪ We further have that $\mathbb{E}(\sigma^2 \mid \mu, \mathbf{y}) = 0.135$ and a 95% credible interval for σ^2 is $(0.120, 0.152)$.

Normal model: unknown mean, unknown variance

- Let us know consider the more realistic situation which corresponds to the case where both μ and σ^2 are unknown.

- The likelihood is still the same, but now regarded as a function of both μ and σ^2 and not only one of them

$$L(\sigma^2; \mu, \mathbf{y}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}.$$

- We now have to specify the joint prior distribution $p(\mu, \sigma^2)$.

- We will use the independence prior and, as such,

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

- Other options do exist, with some popular ones being

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}, \qquad p(\mu, \sigma^2) = p(\mu \mid \sigma^2)p(\sigma^2).$$

Normal model: unknown mean, unknown variance

Gibbs sampling

↪ Letting $\mu \sim \text{N}(\mu_0, \sigma_0^2)$ and $\sigma^2 \sim \text{IG}(a, b)$, the joint posterior distribution is

$$\begin{aligned} p(\mu, \sigma^2 \mid \mathbf{y}) &\propto L(\mu\sigma^2; \mathbf{y}) p(\mu) p(\sigma^2) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\} \\ &\quad \times \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{ -\frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right\} \\ &\quad \times \frac{b^a}{\Gamma(a)} (\sigma^2)^{-(a+1)} e^{-b/\sigma^2}, \end{aligned}$$

which we do not recognise as being the kernel of a known distribution.

↪ However, if σ^2 is assumed known, we know how to draw from $\mu \mid \sigma^2, \mathbf{y}$ and we also know, if we assume μ to be known, how to draw from $\sigma^2 \mid \mu, \mathbf{y}$.

↪ Gibbs sampling (proposed by Geman & Geman, 1984) allows us to sample from the joint posterior distribution $p(\mu, \sigma^2 \mid \mathbf{y})$.

Normal model: unknown mean, unknown variance

Gibbs sampling

- To simulate from the joint posterior using Gibbs sampling we can start with initial values for both parameters, say $\mu^{(0)}$ and $\sigma^{2(0)}$ and proceed by alternating between sampling $\mu \mid \sigma^2, \mathbf{y}$ and then sampling $\sigma^2 \mid \mu, \mathbf{y}$ (or the other way around) until S samples have been collected.

- The distributions $\mu \mid \sigma^2, \mathbf{y}$ and $\sigma^2 \mid \mu, \mathbf{y}$ are called the full conditional distributions of μ and σ^2, respectively.

- Formally, given $\mu^{(s)}$ and $\sigma^{2(s)}$ at iteration s, the values of μ and σ^2 at iteration $s+1$ are generated as follows:

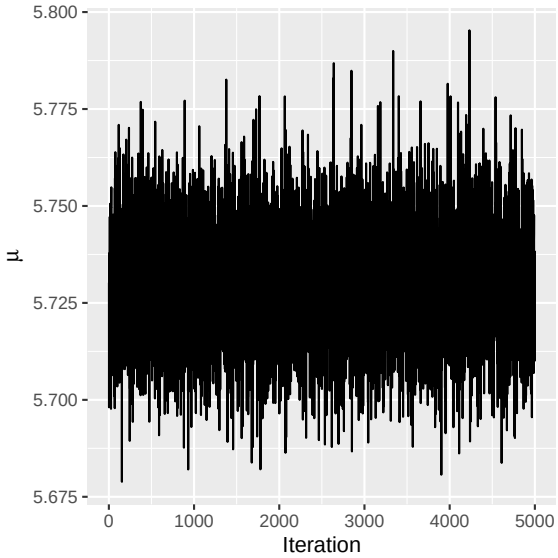
  - Sample $\mu^{(s+1)} \sim p(\mu \mid \sigma^{2(s)}, \mathbf{y})$.

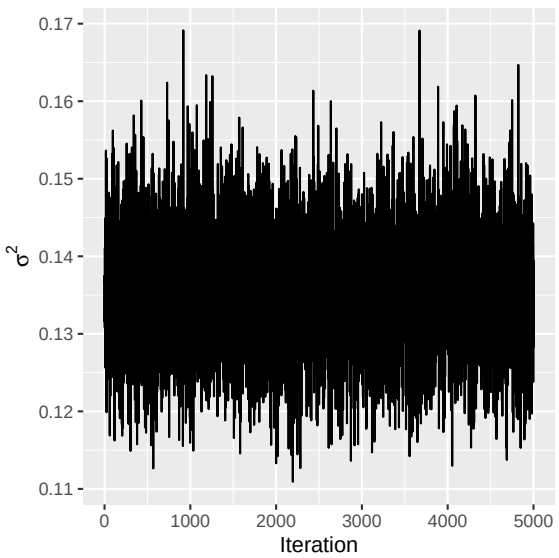
  - Sample $\sigma^{2(s+1)} \sim p(\sigma^2 \mid \mu^{(s+1)}, \mathbf{y})$.

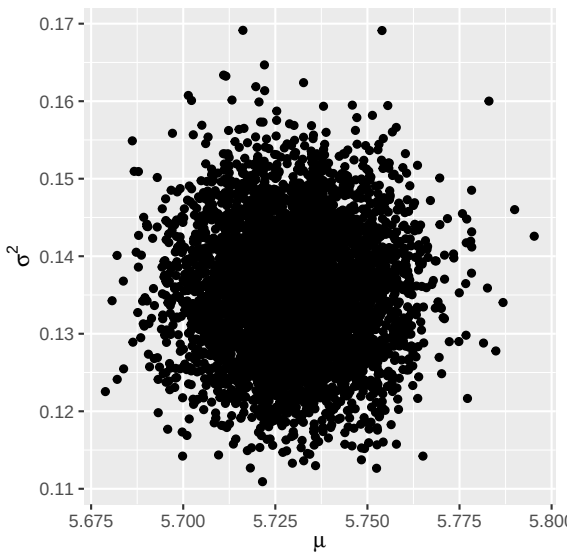
- This results in S posterior samples $\theta^{(1)}, \ldots, \theta^{(S)}$ to summarise the posterior, where $\theta^{(s)} = (\mu^{(s)}, \sigma^{2(s)})$.

Normal model: unknown mean, unknown variance

Gibbs sampling

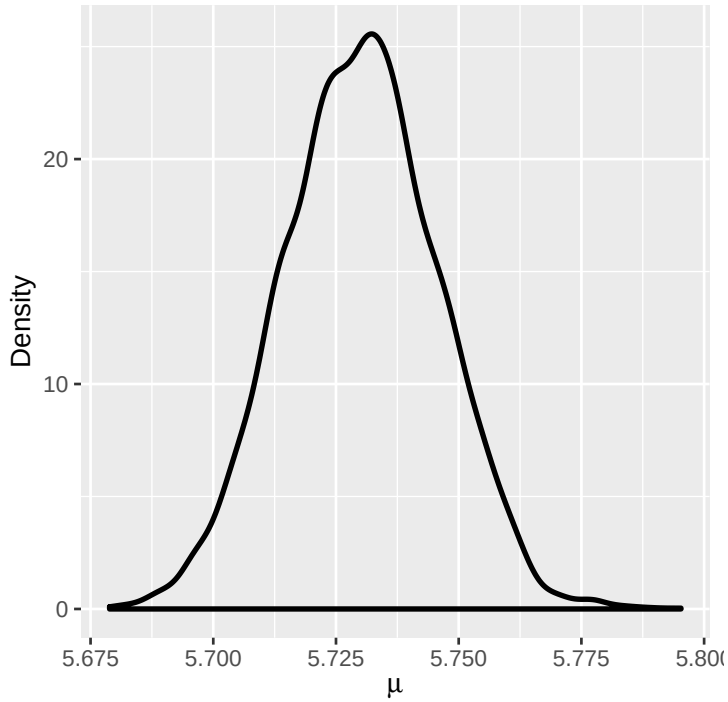


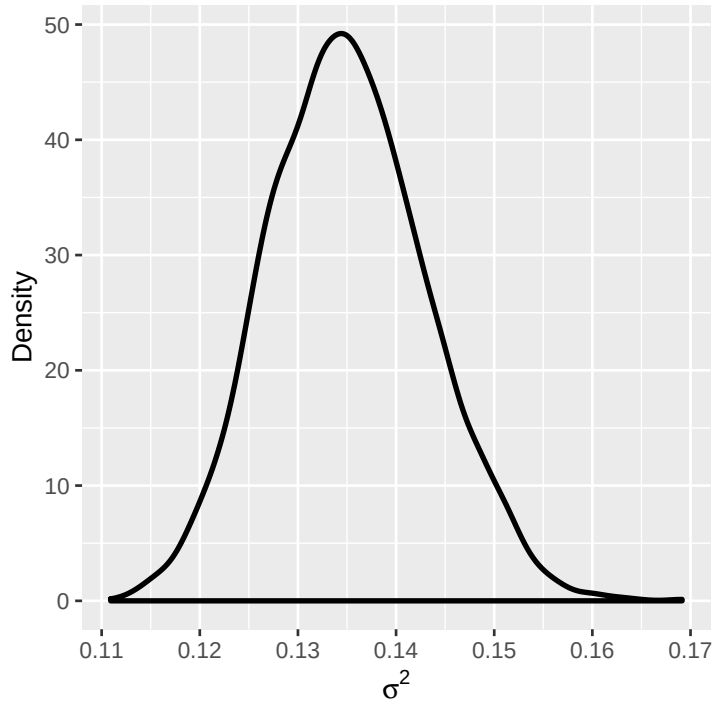




Normal model: unknown mean, unknown variance

Gibbs sampling: revisiting the cholesterol levels example





Gibbs sampling: revisiting the cholesterol levels example

↪ A general recipe for Gibbs sampling is

1. Set initial values $\boldsymbol{\theta} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})$.
2. For $s = 1, \ldots, S$
   - ↪ Draw $\theta_1^{(s)} \sim p(\theta_1 \mid \theta_2^{(s-1)}, \theta_3^{(s-1)}, \ldots, \theta_p^{(s-1)}, \mathbf{y})$.
   - ↪ Draw $\theta_2^{(s)} \sim p(\theta_2 \mid \theta_1^{(s)}, \theta_3^{(s-1)}, \ldots, \theta_p^{(s-1)}, \mathbf{y})$.
   - ↪ $\ldots$
   - ↪ Draw $\theta_p^{(s)} \sim p(\theta_p \mid \theta_1^{(s)}, \theta_2^{(s)}, \ldots, \theta_{p-1}^{(s)}, \mathbf{y})$.

Gibbs sampling

- Because each step cycles through the parameters and updates them given the current values of all other parameters, the samples are not independent.
- All sampling is performed conditionally on the previous iterations' values and so the samples form a Markov chain, and hence the name Markov chain Monte Carlo (MCMC).
- The convergence of the p-tuple obtained at iteration s, $(\theta_1^{(s)}, \ldots, \theta_p^{(s)})$, to a draw from the joint posterior distribution occurs under mild regular conditions that are generally satisfied for most statistical models (see, e.g., Geman and Geman, 1984, or Roberts and Smith, 1993).
- Gibbs sampling reduces the hard problem of sampling from a multivariate distribution to the one of sampling from several univariate distributions.
- Of course, this assumes that the univariate full conditional distributions are available and that they are easy to sample from.

Metropolis–Hastings sampling

- What should we do if some/all full conditional distributions are not available in closed form?
- For such cases, other methods have been proposed, with Metropolis–Hastings being the most widely used.
- The Metropolis–Hastings (MH) algorithm replaces draw from the exact full conditional distribution with a draw from a proposal distribution followed by an acceptance/rejection step.
- For instance, in the previous example, the Gibbs update of μ is a sample $\mu^{(s)} \mid \sigma^{2(s-1)}, \mathbf{y}$, whereas MH takes a candidate draw $\mu^* \sim q(\mu^* \mid \mu^{(s-1)})$ that is conditioned on the current value of μ (and potentially σ^2 and/or $\mathbf{y}$). Note that the proposal distribution should have the same support as the parameter.
- Obviously, the candidate value cannot be blindly accepted because the proposal distribution may be only little related to the posterior distribution.

Metropolis–Hastings sampling

↪ To correct for this, the candidate value is accepted with probability $\min\{1, r\}$, where r is the acceptance ratio

$$
\begin{aligned}
r &= \frac{p(\mu^*, \sigma^{2(s-1)} \mid \mathbf{y})}{p(\mu^{(s-1)}, \sigma^{2(s-1)} \mid \mathbf{y})} \frac{q(\mu^{(s-1)} \mid \mu^*)}{q(\mu^* \mid \mu^{(s-1)})} \\
&= \frac{L(\mu^*, \sigma^{2(s-1)}; \mathbf{y}) p(\mu^*) p(\sigma^{2(s-1)})}{L(\mu^{(s-1)}, \sigma^{2(s-1)}; \mathbf{y}) p(\mu^{(s-1)}) p(\sigma^{2(s-1)})} \frac{q(\mu^{(s-1)} \mid \mu^*)}{q(\mu^* \mid \mu^{(s-1)})}.
\end{aligned}
$$

↪ Equivalently, the acceptance/rejection step generates $u \sim \text{Unif}(0, 1)$ and accepts the candidate if $u < r$.

↪ If $r \geq 1$, then the candidate value μ^* will be accepted with probability one. If $r < 1$, then μ^* is accepted with probability r (and so we set $\mu^{(s)} = \mu^*$) and rejected with probability $1 - r$ (and so we set $\mu^{(s)} = \mu^{(s-1)}$).

↪ In practice, we usually compute the ratio of acceptance in the log scale to avoid numerical instabilities.

Metropolis–Hastings sampling

↪ The Metropolis–Hastings algorithm may be, more generally, described as follows.

1. Set initial values $\boldsymbol{\theta} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})$.
2. For $s = 1, \ldots, S$
   - ↪ Simulate $\theta_1^* \sim q_1(\theta_1 \mid \theta_1^{(s-1)})$.
   - ↪ Compute the acceptance ratio

$$r_1 = \frac{L(\theta_1^*, \theta_2^{(s-1)}, \ldots, \theta_p^{(s-1)}; \mathbf{y}) p(\theta_1^*, \theta_2^{(s-1)}, \ldots, \theta_p^{(s-1)})}{L(\theta_1^{(s-1)}, \theta_2^{(s-1)}, \ldots, \theta_p^{(s-1)}; \mathbf{y}) p(\theta_1^{(s-1)}, \theta_2^{(s-1)}, \ldots, \theta_p^{(s-1)})} \frac{q_1(\theta_1^{(s-1)} \mid \theta_1^*)}{q_1(\theta_1^* \mid \theta_1^{(s-1)})}.$$

   - ↪ Generate $u_1 \sim \text{Unif}(0, 1)$. If $u_1 < \min 1, r_1$, set $\theta_1^{(s)} = \theta_1^*$, otherwise set $\theta_1^{(s)} = \theta_1^{(s-1)}$.
   - ↪ $\ldots$
   - ↪ Simulate $\theta_p^* \sim q_p(\theta_p \mid \theta_p^{(s-1)})$.
   - ↪ Compute the acceptance ratio

$$r_1 = \frac{L(\theta_1^{(s)}, \theta_2^{(s)}, \ldots, \theta_p^*; \mathbf{y}) p(\theta_1^{(s)}, \theta_2^{(s)}, \ldots, \theta_p^*)}{L(\theta_1^{(s)}, \theta_2^{(s)}, \ldots, \theta_p^{(s-1)}; \mathbf{y}) p(\theta_1^{(s)}, \theta_2^{(s)}, \ldots, \theta_p^{(s-1)})} \frac{q_p(\theta_p^{(s-1)} \mid \theta_p^*)}{q_p(\theta_p^* \mid \theta_p^{(s-1)})}.$$

   - ↪ Generate $u_p \sim \text{Unif}(0, 1)$. If $u_p < \min 1, r_p$, set $\theta_p^{(s)} = \theta_p^*$, otherwise set $\theta_p^{(s)} = \theta_p^{(s-1)}$.

Metropolis–Hastings sampling

- The MH sampler can be applied more generally than Gibbs sampling, but at the cost of having to select and tune the proposal distribution for each parameter (or block of parameters).
- A popular choice for the proposal distribution is the random walk normal distribution

$$\theta_j^* \mid \theta_j^{(s-1)} \sim \mathrm{N}(\theta_1^{(s-1)}, c_j^2).$$

- The name random walk proposal distribution comes from the fact that it simply adds a normal noise to the current value of the chain.
- Ideally, we would like to tune the algorithm so that the acceptance rate is not too high or too low. A rule of thumb is to accept around 20–50% of the candidate values.
- We have a high acceptance rate when we make small movements not too far away from the current value of the chain, but if the chain moves slowly, it will take a long time to sample the whole distribution.
- On the contrary, if we make larger movements chances are high that we propose candidate values 'living' on regions of low probability and then we will have a high rate of rejected candidate values.

Metropolis–Hastings sampling

↪ Note that the random walk normal proposal distribution is symmetric

$$
\begin{aligned}
q(\theta_j^{(s-1)} \mid \theta_j^*) &= \frac{1}{\sqrt{2\pi c_j^2}} \exp\left\{ -\frac{1}{2c_j^2} \left(\theta_j^{(s-1)} - \theta_j^*\right)^2 \right\} \\
&= \frac{1}{\sqrt{2\pi c_j^2}} \exp\left\{ -\frac{1}{2c_j^2} \left(\theta_j^* - \theta_j^{(s-1)}\right)^2 \right\} \\
&= q(\theta_j^* \mid \theta_j^{(s-1)}),
\end{aligned}
$$

↪ Because of such symmetry, the MH acceptance ratio simplifies as the term $\frac{q_j(\theta_j^{(s-1)} \mid \theta_j^*)}{q_j(\theta_j^* \mid \theta_j^{(s-1)})}$ is equal to one.

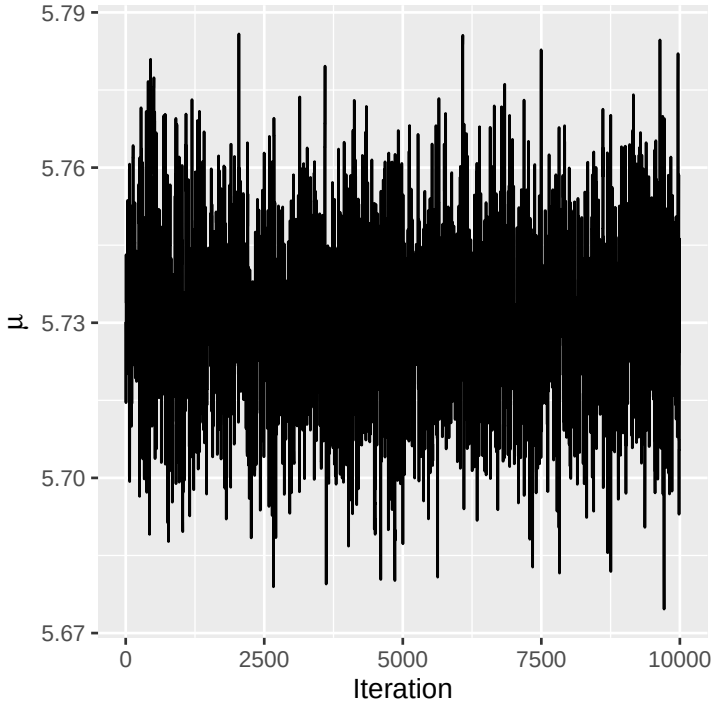
Metropolis–Hastings sampling

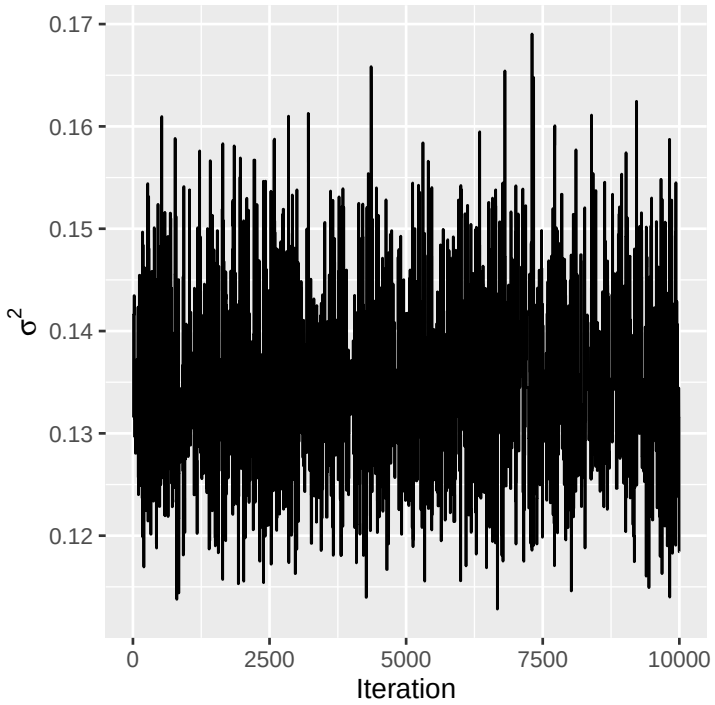
Revisiting the cholesterol levels example

- ↪ As a toy illustration, and although not needed for this case, we will use MH to sample from the joint posterior $p(\mu, \sigma^2 \mid \mathbf{y})$.
- ↪ We need to select the proposal distributions and their variance. For μ we will use a normal distribution centred on the previous values of the chain and with variance 0.05.
- ↪ For σ^2 we will also use a normal distribution centred on the previous value and with the same variance (0.05). Note that we can never accept a negative value and so in the `R` code we ensure that every time a negative value is sampled, the corresponding acceptance ratio is zero.
- ↪ For the specific seed that we have used, this has led to acceptance probabilities of 35% for μ and 20% for σ^2.

Metropolis–Hastings sampling

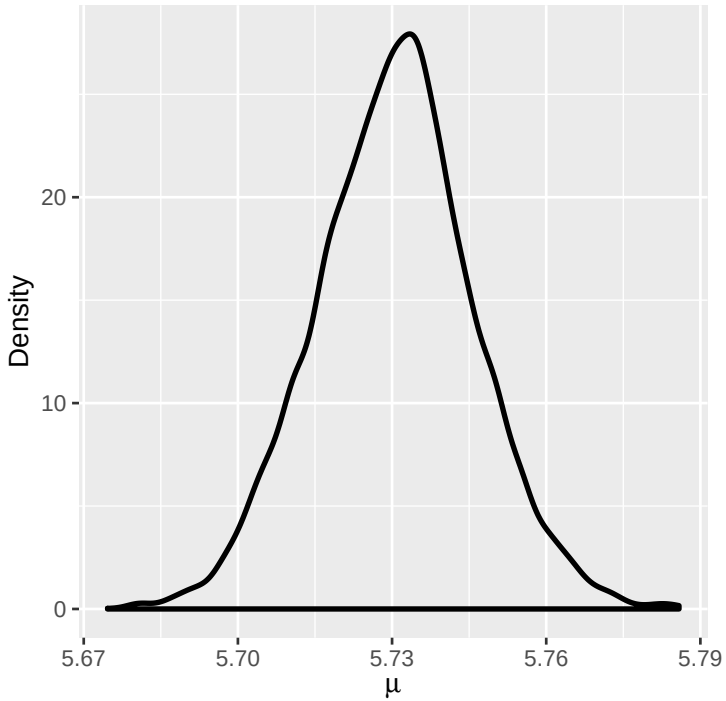
Revisiting the cholesterol levels example

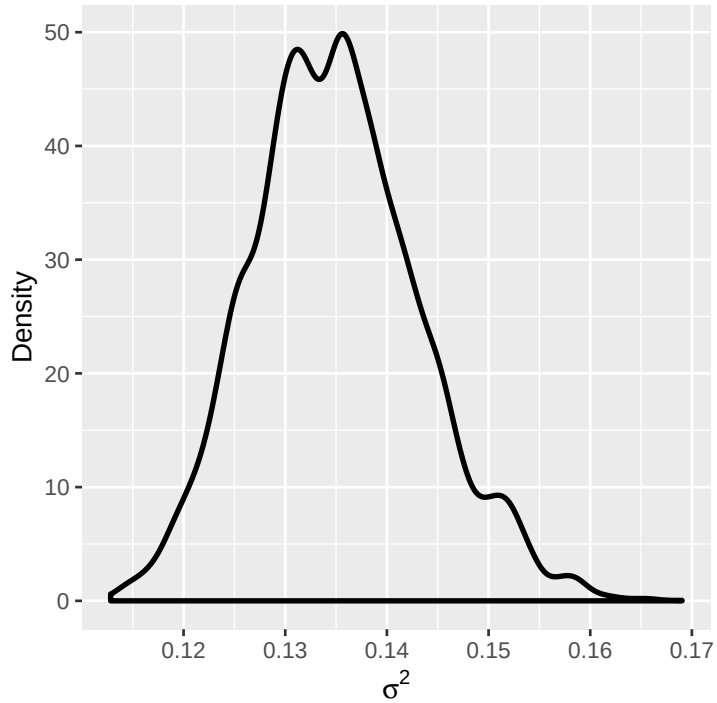




Metropolis–Hastings sampling

Revisiting the cholesterol levels example





Metropolis–Hastings sampling

↪ Sometimes it might be more efficient to update the parameters in block(s).

↪ In such case, and assuming that we would update in block all p parameters (which is only realistic for p small), we would simulate the candidate vector from a $p-$dimensional proposal distribution, say $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(s-1)})$.

↪ We would simultaneously accept or reject the whole p dimensional candidate vector and the acceptance ratio is therefore computed as

$$r = \frac{L(\boldsymbol{\theta}^*; \mathbf{y}) p(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}^{(s-1)}; \mathbf{y}) p(\boldsymbol{\theta}^{(s-1)})} \frac{q(\boldsymbol{\theta}^{(s-1)}) \mid \boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(s-1)})}.$$

Metropolis–Hastings sampling

- ↪ It is also possible to do a kind of hybrid algorithm combining both Gibbs and Metropolis–Hastings steps.
- ↪ For instance, if for some parameters their corresponding full conditional distributions are available in closed form we can use Gibbs sampling, while for the other parameters for which the full conditional distributions are not recognisable, we would use MH. A toy example is given in the supplementary file.
- ↪ It is worth mentioning that the Gibbs sampling is a particular case of the Metropolis–Hastings where the proposal distribution is the full conditional distribution and all candidate values are accepted with probability one.

Convergence diagnostics

Trace plot

- One basic diagnostic is to monitor the *traceplots*: the plot of the iteration versus the sampled values of the parameter.
- If all values are within a zone without strong periodicity and especially tendencies, then there is no evidence of lack of convergence.
- All sampled values before the chain gets stable (formally, before the chain reaches the stationary distribution) should be discarded. This is known as the *burn-in* period.

Convergence diagnostics

Autocorrelation function

- ↪ The autocorrelation of the chains provide a measure of how fast the chains are mixing.
- ↪ Ideally, the draws would be independent of each other, but the Markovian nature of the sampler induces dependence between successive draws.
- ↪ A somewhat crude, yet reasonably effective, way of dealing with autocorrelation is to only keep every k draws from the posterior and discard the rest; this is known as thinning the chain.
- ↪ The advantages of thinning are both simplicity and a reduction in memory usage as working with large chains can be, depending on the problem, burdensome.
- ↪ The disadvantage is that we are clearly throwing away information; thinning can never be as efficient as using all the iterations.

Convergence diagnostics

Effective sample size

- The effective sample size (ESS) of a parameter is the number of independent draws from the posterior distribution that the Markov chain is equivalent to.
- Formally, for parameter θ_j the ESS is computed as

$$\text{ESS}_j = \frac{S}{1 + 2\sum_{l=1}^{\infty} \rho_j(l)}, \qquad \rho_j(l) = \text{Cor}(\theta_j^{(s)}, \theta_j^{(s-l)}).$$

- The `R` package `coda` computes the ESS for a given MCMC chain (through the function `effectiveSize`).

Convergence diagnostics

Geweke's criterion

- Geweke's diagnostic criterion is used to detect non-convergence of the chain.
- The criterion has its routes in time series and it uses a two sample t-test to compare the mean of the chain between batches at the beginning (say the first 10% of the sampled values) and at the end of the samples (say the last 50% of the samples).
- Geweke shown that, under the null hypothesis that the means are the same for the two batches, the test statistic follows a standard normal distribution, and so values, in magnitude, larger than 2/3 are of concern.

Software for Bayesian computation

- There are several general purpose MCMC packages that can be called from R and that avoid to code everything from scratch (e.g., deriving all full conditionals for Gibbs sampling or tuning the variance of the proposal distribution in Metropolis–Hastings), including:
  - `JAGS` (Just Another Gibbs sampling)
  - `OpenBugs`
  - `Stan`
  - `NIMBLE`
- These packages are not specific to a single model, rather they take a script specifying the likelihood and prior as input and use this information to construct an MCMC sampler for the specified model.
- `JAGS` will be used sometimes throughout this course to illustrate some of the methods.

Software for Bayesian computation

↪ In general, the steps to using `JAGS` to produce samples are:

1. Download `JAGS` at
   `http://sourceforge.net/projects/mcmc-jags/files/`
2. Install the `rjags` package in `R`, which should automatically find your JAGS installation.
3. Specify the statistical model (likelihood and prior) using the model command.
4. Compile the model using `jags.model`.
5. Generate samples using `update` and `jags.samples`.

↪ There are, at least, 3 `R` interfaces for JAGS: `R2jags`, `rjags`, and `runjags`. The model syntax is the same for the three interfaces, what changes is how relevant information is extracted. We will stick to `rjags`.

↪ There is an example for the normal model with both mean and variance unknown in the supplementary materials file.

Bayesian model comparison criteria

- In most data analyses there will be a competition among models for the same data, as some models will be more appropriate for a given data set than others.
- A classic model selection/fit problem might involve a specific sample of log biomarker outcomes. We could compare the fit of a log-normal distribution to that of a gamma distribution and decide to select the model that fit better.
- When working with the nonparametric models we will use some model comparison criteria to assist in a number of tasks (e.g., selecting the number of components in a finite mixture model).
- If covariate information is available and believed to be possibly associated with biomarker outcomes, then choosing between models with and without an interaction term would be warranted.
- There are several available criteria, such as the DIC (deviance information criterion), WAIC (widely applicable information criterion), and LPML (log pseudo marginal likelihood).
- Because the WAIC can be regarded as an improvement of the DIC, throughout we will only be using the LPML and the WAIC.

Bayesian model comparison criteria

LPML

- The main ingredient to calculate the LPML is the conditional predictive ordinate (CPO). For the ith observation, the CPO_i is defined as

$$\text{CPO}_i = f\left(y_i \mid \mathbf{y}_{(-i)}\right) = \int_{\Theta} f(y_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{y}_{(-i)}) \mathrm{d}\boldsymbol{\theta},$$

  where $\mathbf{y}_{(-i)}$ is $\mathbf{y} = (y_1, \ldots, y_n)$ with the ith observation omitted, and $p(\boldsymbol{\theta} \mid \mathbf{y}_{(-i)})$ is the posterior density of $\boldsymbol{\theta}$ based on the data $\mathbf{y}_{(-i)}$.

- Thus CPO_i is the marginal posterior predictive density of y_i given $\mathbf{y}_{(-i)}$ and can be interpreted as the height of this marginal density at y_i.

- A higher value of CPO_i under one model for implies a better fit of that model for the ith observation, $i = 1, \ldots, n$.

Bayesian model comparison criteria

LPML

↪ Gelfand and Dey (1994)showed that CPO_i is easily estimated from a posterior sample via

$$\text{CPO}_i \approx \left\{ \frac{1}{S} \sum_{s=1}^{S} \frac{1}{f(y_i \mid \boldsymbol{\theta}^{(s)})} \right\}^{-1}.$$

↪ The result is based on showing that CPO_i as defined in the previous slide can be rewritten as

$$\text{CPO}_i = \left\{ \int_{\boldsymbol{\Theta}} \frac{1}{f(y_i \mid \boldsymbol{\theta})} p(\boldsymbol{\theta} \mid \mathbf{y}) \mathrm{d}\boldsymbol{\theta} \right\}^{-1} = \left\{ E_{\boldsymbol{\theta}|\mathbf{y}} \left[\frac{1}{f(y_i \mid \boldsymbol{\theta})} \right] \right\}^{-1}.$$

↪ The LPML defined as $\text{LPML} = \sum_{i=1}^{n} \log(\text{CPO}_i)$ gives an aggregate summary measure of a model's predictive ability and the larger the LPML is, the better the fit of the model under consideration.

Bayesian model comparison criteria

WAIC

- The Widely Applicable information criterion (WAIC), also known as the Watanabe-Akaike information criterion, was introduced by Watanabe (2010) and further studied by Gelman et al. (2014).

- The WAIC comprises two key components: a raw goodness-of-fit measure and a penalty term that depends on the number of effective parameters. The goodness of fit is measured by the log pointwise predictive density, given by

$$\text{lppd} = \sum_{i=1}^{n} \log p(y_i \mid \mathbf{y}), \quad p(y_i \mid \mathbf{y}) = \int f(y_i \mid \theta) p(\theta \mid \mathbf{y}) \mathrm{d}\theta.$$

- One potential problem of choosing a model based only on the lppd is that observation y_i is being predicted using all of the data y, including y_i itself.

- For example, it is well-known that in regression models, using a fit of the model that includes the observation whose future counterpart is being predicted could lead to overly optimistic predictions.

- Remember that when computing CPO_i, we used $p(y_i \mid \mathbf{y}_{-i})$, which correctly excludes the observation y_{-i} when calculating the predictive density of y_i.

Bayesian model comparison criteria

WAIC

- In order to correct for this over-optimism, Watanabe suggested subtracting the following penalty from lppd

$$p_{\text{WAIC}} = \sum_{i=1}^{n} \text{Var}_{\theta|y} \log f(y_i \mid \theta).$$

- Gelman et al. (2014) discuss two possible penalty adjustments but argue in favour of this one. They expect numerical approximation to it to be stable. These authors also suggest multiplying both the lppd and p_{WAIC} by -2 so that the criterion is on the deviance scale, similar to DIC (and AIC), leading to

$$\begin{aligned} \text{WAIC} &= -2\text{lppd} + 2p_{\text{WAIC}} \\ &= -2\sum_{i=1}^{n} \log p(y_i \mid \mathbf{y}) + 2\sum_{i=1}^{n} \text{Var}_{\theta|y} \log f(y_i \mid \theta). \end{aligned}$$

- Similarly to DIC, from a pool of candidate models, the model with the lowest WAIC is to be preferred.

Bayesian model comparison criteria

WAIC

↪ The main ingredients of the WAIC criterion are easily computed from a sample of draws from the posterior distribution

$$\text{lppd} \doteq \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} f\left(y_i \mid \theta^{(s)}\right) \right),$$

$$p_{\text{WAIC}} \doteq \sum_{i=1}^{n} V_{s=1}^{S} \{\log f(y_i \mid \theta^{(s)})\},$$

where $V_{s=1}^{S}$ represents the sample variance, $V_{s=1}^{S}(a) = \frac{1}{S-1} \sum_{s=1}^{S} (a_s - \bar{a})^2$.

Bayesian model checking

Posterior predictive checks

- ↪ Model comparison criteria, such as the LPML and WAIC are useful to assess the relative merits of competing models. However, it is also mandatory that to assess the quality of the selected model in absolute terms.

- ↪ We discuss here posterior predictive checks, introduced by Rubin (1984) and further developed by Gelman et al. (1996) as they are widely popular.

- ↪ Posterior predictive checks evaluate a model's goodness of fit by comparing datasets simulated from the fitted model to the observed dataset.

- ↪ The logic behind posterior predictive checks is as follows: if a model (likelihood and prior) fits the observed data well, i.e., if a model is a good approximation to the (unknown) data generation process, then replicated data simulated from the model should be similar to the observed data.

Bayesian model checking

Posterior predictive checks

- The replicated data, say y_{rep}, is sampled from the posterior predictive distribution

$$p(y_{\text{rep}} \mid y) = \int p(y_{\text{rep}} \mid \theta) p(\theta \mid y) \mathrm{d}\theta.$$

- To simplify comparisons, the replicated dataset should be of the same dimension as the observed data. Moreover, if the model includes covariates, these will be the same for y and for y_{rep}.

- To evaluate the similarity between the replicated and observed data, we can use a test statistic, $T(y)$, which is a test quantity that depends only on the data (or the replicated data), and which measures misfit between y and the model.

- For example, we may consider $T(y) = \min(y)$ if we wish to investigate the model's fit in the left tail.

Bayesian model checking

Posterior predictive checks

- ↪ Implementation of posterior predictive checks is relatively straightforward.
- ↪ Given an MCMC sample from the posterior distribution of the parameters, $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$, we can sample replicated data $y_{\text{rep}}^{(s)}$ from $p(y_{\text{rep}} \mid \theta^{(s)})$ and compute $T(y_{\text{rep}}^{(s)})$, for $s = 1, \ldots, S$.
- ↪ A visual goodness of fit check plot the histogram of the replicated values $T(y_{\text{rep}}^{(s)})$ with the observed $T(y)$ superimposed. If the value of the test statistic based on the observed data is very distant from the center of the distribution of the test statistics calculated across the replicated datasets, then doubt is cast on some aspect of the model.

Bayesian model checking

Posterior predictive checks

- The choice of the test statistics is crucial.
- The most effective ones are those that verify modeling assumptions. It is important to note that the test statistic(s) should not focus on aspects that are parameterized by the model, as these are automatically fit in the posterior distribution.
- For example, when fitting a normal model with unknown mean and variance to the data, we should not select the test statistic to be the sample mean or the sample variance.
- However, choosing T to be the sample skewness is informative, as it provides a way to check one of the modeling assumptions inherent in the use of the normal distribution: its symmetry.

Bayesian model checking

Posterior predictive checks

- Posterior predictive checks use the data twice: first, for computing the posterior distribution of the parameters, which serves as the basis for computing the posterior predictive distribution and second, for computing the observed test statistic.

- When there is sufficient data, one could split the data into two parts: one for model fitting and the other for computing the observed test statistic; for example, one may leave out a random 10% or 15% of the data for validation purposes. This is also the basic idea behind the holdout predictive checks recently proposed by Moran et al. (2023).

- In applications where sample size is not large, data splitting is potentially problematic. We should be aware that if we do not split the data, diagnostics are likely to be conservative.

Bayesian model checking

Posterior predictive checks

↪ Returning to the cholesterol levels example (both mean and variance unknown).