Bayesian Data Analysis

Daniel Paulin & Nicolò Margaritella

University of Edinburgh

Semester 2, 2020/2021

With thanks to Jonathan Gair, Ruben Amoros-Salvador, Ken Newman, Vanda Inácio and Natalia Bochkina for much of the material.

Scope

- The goal of this course is to provide practical experience of applying Bayesian analyses to a range of statistical models.
- The statistical analyses will be conducted using the widely used computer packages JAGS and R-INLA.
- Topics:
  - Brief overview of main Bayesian ideas.
  - Introduction to JAGS.
  - Introduction to INLA.
  - Linear and generalised linear models (fixed effects).
  - Hierarchical Bayesian models: linear and generalised linear models with random effects.
  - Further topics (which might include spatial and temporal models).

General information

- **Lecturers**: Daniel Paulin & Nicolò Margaritella

- **Lectures**: Recordings posted by Monday mornings on Learn during odd weeks.

- **Workshops**: Fridays even weeks on Zoom:
  22 January, 5 February, 26 February, 12 March, 26 March.
  3 available time slots: 9:00-11:00, 11:00 - 13:00, 14:00 - 16:00.

- **Additional information**:

  Workshops will be conducted using www.kaggle.com (similar environment to Noteable). **Please go to www.kaggle.com now, register, and open a new notebook. Select File/Language as R and File/Editor Type as Notebook. Click on File/Upload to add the Workshop problems from Learn.**

  The workshop problems will be available 2 weeks before the workshops on Learn.

  At the end of each workshop, you must download your work from Kaggle, and upload it in Learn in Workshop Submissions.

General information

- **Office hours**: Tuesdays every week from 12:00-13:00 on Zoom.
  Thursdays weeks 5-6 and 10-11 from 9:00-10:00 on Zoom.

- **Piazza**: The course has a Piazza page, feel free to post short questions of common interest there. For more complex questions that require some context to formulate, communication through Piazza can be very cumbersome. You should pose them during the workshops or office hours.

- **Email**: Other forms of communication such as discussions during office hours, workshops or on Piazza are preferred due to the large number of students enrolled in this course.

Assessment

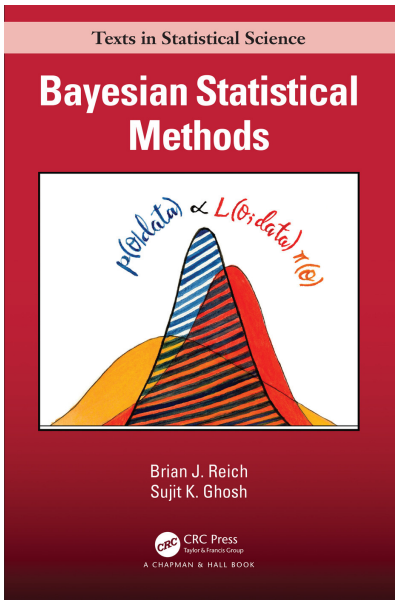
- ↪ 100% coursework: two homework assignments.
- ↪ Homework assignment 1 (50%) will be issued in week 5 (week beginning February 8). Solutions should be uploaded to Learn by 17:00 on Tuesday 2 March, 2021 (week 8).
- ↪ Homework assignment 2 (50%) will be issued in week 10 (week beginning March 22). Solutions should be uploaded to Learn by 17:00 on Monday 12 April, 2021 (week 13).
- ↪ Both homework assignments are to be done **individually** and consist of a sample of applied problems requiring the use of statistical software. A report containing all the analyses and conclusions should be delivered, along with your code that should run without error and fully reproduce all results in the report. Reports containing work that has not been done individually will be sent to the Academic Misconduct Officer.
- ↪ The homework assignments will be impossible to solve without a good understanding of the material in the lectures and workshops. **Regular study is essential.**
- ↪ **Students who discuss their solutions to the homework problems on Piazza will have their posts removed, and their access to the Piazza page revoked.**

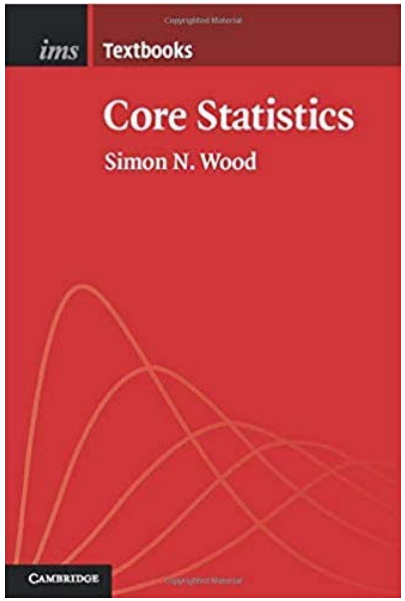
Study plan (100 hours in total)

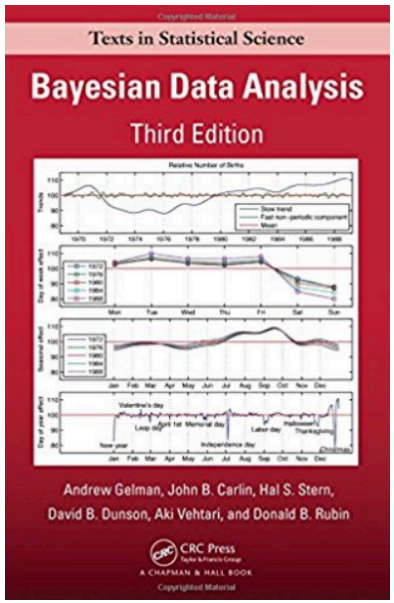
- ↪ Weeks 1 - 2 : 4 h study of lecture materials, 6 h work on Workshop 1 problems, 2 h participation in Workshop 1, 1 h on Piazza/office hours
- ↪ Weeks 3 - 4 : 4 h study of lecture materials, 6 h work on Workshop 2 problems, 2 h participation in Workshop 2, 1 h on Piazza/office hours
- ↪ Weeks 5 - 6 (3 weeks as include Flexible learning week) :
  4 h study of lecture materials, 6 h work on Workshop 3 problems, 2 h participation in Workshop 3, **15 h work on Homework 1**, 1 h on Piazza/office hours
- ↪ Weeks 7 - 8 : 4 h study of lecture materials, 6 h working on Workshop 4 problems, 2 h participating in Workshop 4, 1 h on Piazza/office hours
- ↪ Weeks 9 - 12 : 8 h study of lecture materials (Lectures 5 and 6), 6 h work on Workshop 5 problems, 2 h participation in Workshop 5, **15 h work on Homework 2**, 2 h on Piazza/office hours

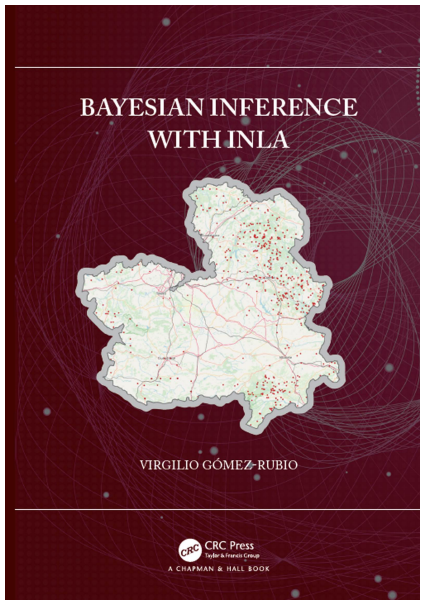
Recommended textbooks

- ↪ The course material (slides, worksheets, and other support material) contain all the information needed for the course.
- ↪ However, there are a wide variety of books, at all levels, that cover the material in this course that may be of interest to supplement the lecture material and/or provide additional examples.
- ↪ Suitable books include

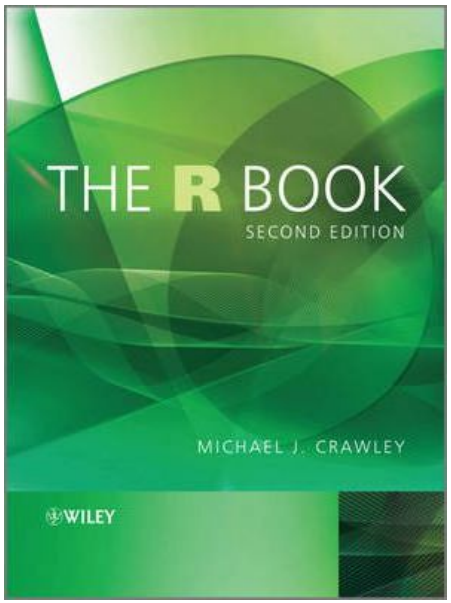


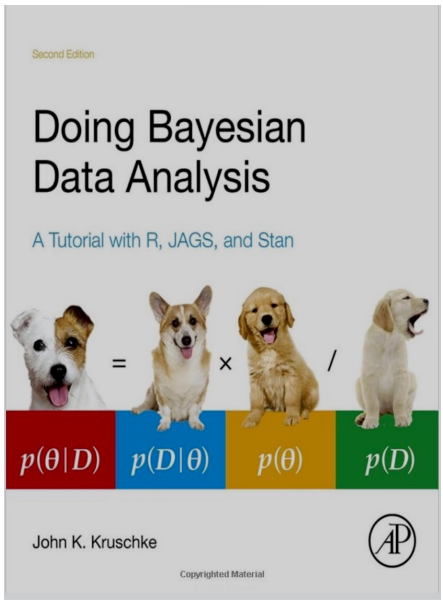


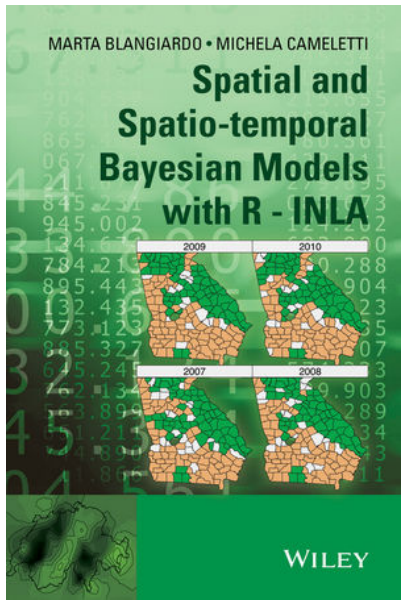


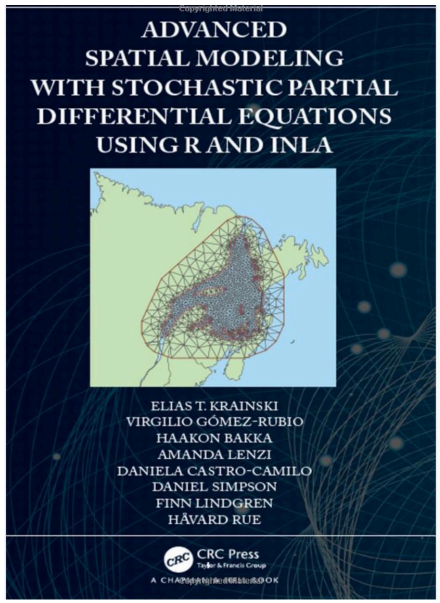


Additional textbooks



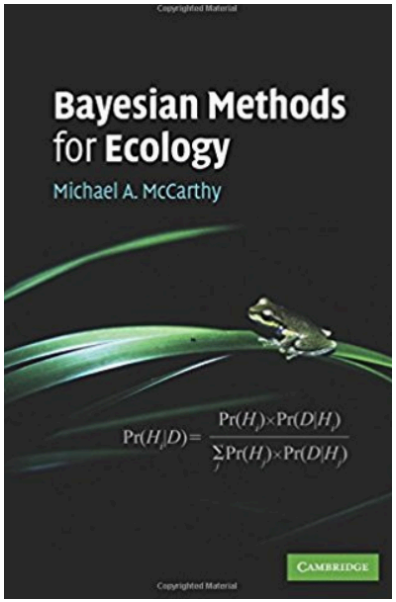


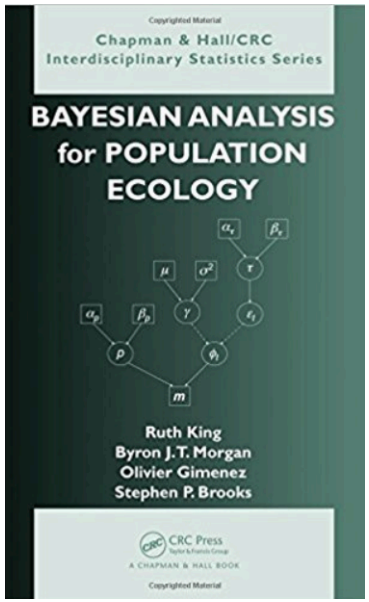


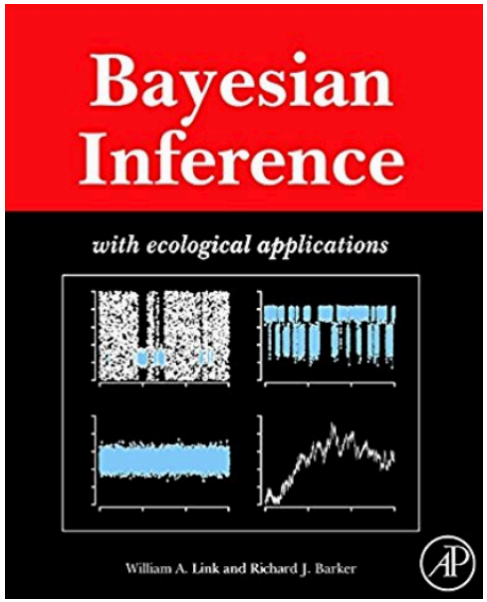


Additional textbooks

Ecological applications (covers from `amazon.co.uk`)

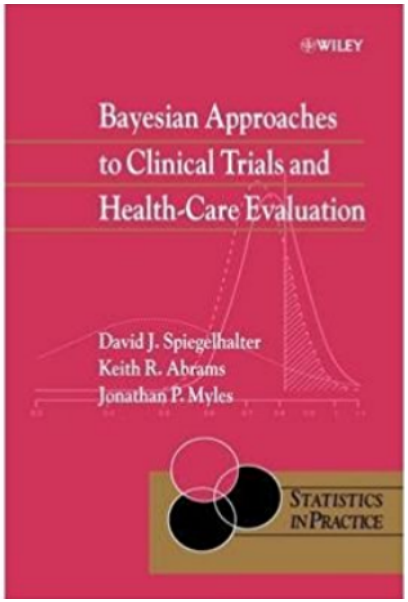


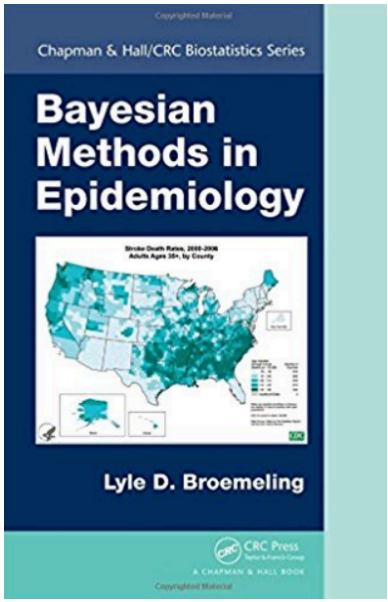


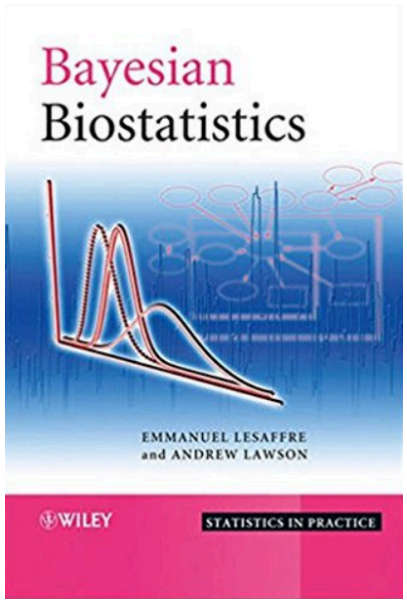


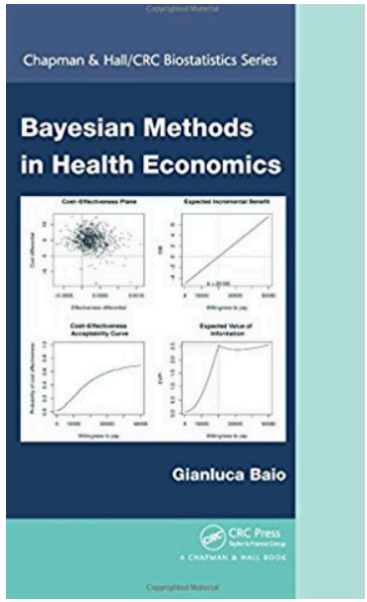
Additional textbooks

Medical/epidemiological applications (covers from `amazon.co.uk`)



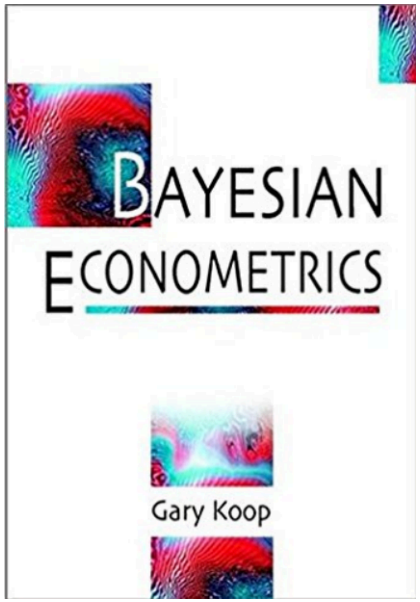


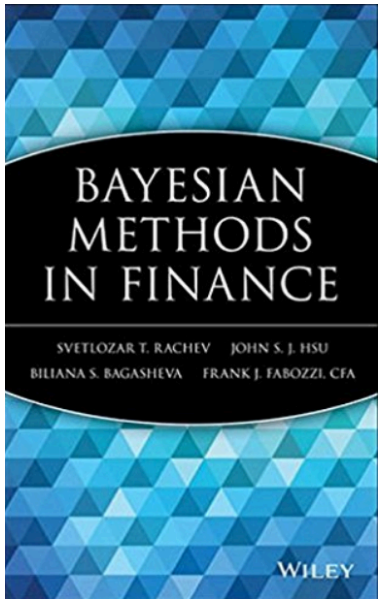


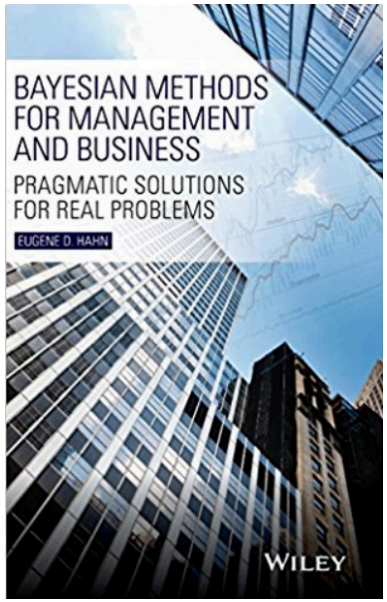


Additional textbooks

Finance/business/econometric applications (covers from `amazon.co.uk`)

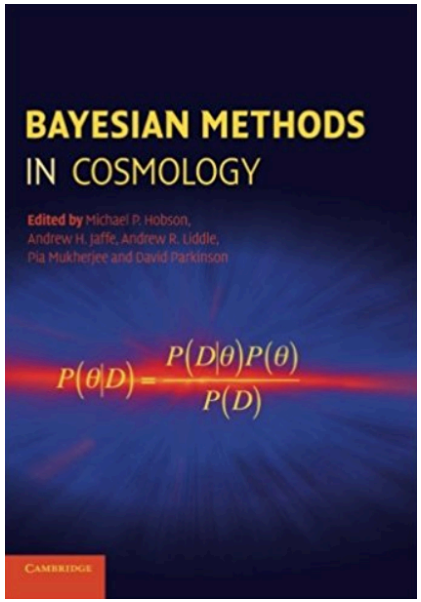


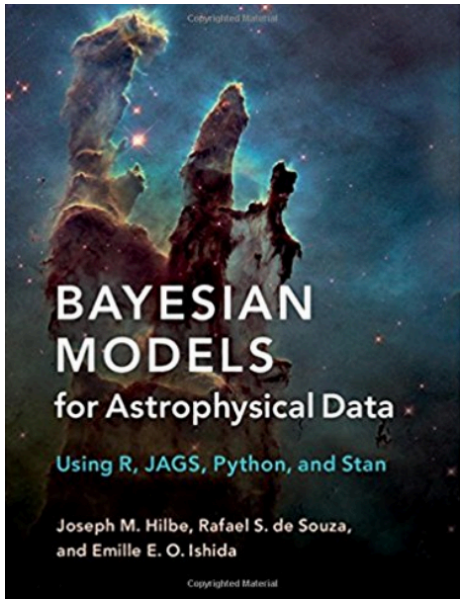




Additional textbooks

Astrophysics/cosmology applications (covers from `amazon.co.uk`)





Outline

1. Review of Bayesian inference
   - Choice of prior distribution
   - Predictive inference

2. Bayesian computing
   - Deterministic approximation methods
   - Monte Carlo methods
   - MCMC convergence diagnostics

Review of Bayesian inference

- Bayesian methods have been widely applied in many areas of science (e.g., medicine, finance, ecology, physics, psychology, etc).
- Motivations for adopting Bayesian approach vary:
  - natural and coherent way of thinking about science and learning,
  - pragmatic choice that is suitable for the problem in hand.
- Spiegelhalter et al. (2004) define a Bayesian approach as

  'the explicit use of external information in the design, monitoring, analysis, interpretation, and reporting of a [scientific investigation].'

- These authors argue that a Bayesian approach is
  - more flexible in adapting to each unique situation,
  - more efficient in using all available evidence,
  - more useful in providing relevant quantitative summaries,

  than traditional (frequentist) methods.

Review of Bayesian inference

Bayesian vs frequentist statistics

- ↪ The **frequentist** approach can be regarded as a procedure that quantifies uncertainties (p-value, confidence interval, etc) if the process that generated the data is repeated many times.
- ↪ Parameters are fixed and unknown, only the data is random.
- ↪ Aims to be objective.
- ↪ **Bayesian** represent their uncertainty about parameters with probability distributions and treat them as random variables.
- ↪ We can thus, under a Bayesian framework, make probability statements about model parameters.
- ↪ This is in contrast with the frequentist framework where probability statements only concern the data.

↪ Both approaches have pros and cons. When both are applicable they are unlikely to give different answers.

Review of Bayesian inference

Main components

- Suppose we have a parameter $\theta \in \Theta$ on which we wish to make inference.
- The main 'ingredients' of Bayesian inference are:
  - The prior distribution, $\pi(\theta)$, which represents the initial beliefs concerning the parameter prior to any data being observed.
  - The likelihood $f(\mathbf{y} \mid \theta)$, which plays a key role in all statistical inference, both Bayesian and frequentist. Represents the information contained in the data $\mathbf{y}$ about the parameter θ.
  - The posterior distribution, $p(\theta \mid \mathbf{y})$, which updates the prior beliefs with respect to θ, following the data $\mathbf{y}$ being observed.
- Bayesian learning combines past experience (prior) with new data (likelihood) in a mathematically coherent way (Bayes' Theorem) to obtain the current state of knowledge (posterior).

Review of Bayesian inference

Bayes' theorem

↪ Bayes' theorem (continuous version) tells us

$$p(\theta \mid \mathbf{y}) = \frac{f(\mathbf{y} \mid \theta)\pi(\theta)}{m(\mathbf{y})},$$

where $m(\mathbf{y}) = \int_\Theta f(\mathbf{y} \mid \theta)\pi(\theta)\mathrm{d}\theta$ is the evidence (the marginal distribution of the data).

↪ Since the marginal distribution does not depend on θ, we can write

$$p(\theta \mid \mathbf{y}) \propto f(\mathbf{y} \mid \theta)\pi(\theta),$$

i.e., the posterior distribution is proportional to the likelihood times the prior distribution.

Review of Bayesian inference

Summarising posterior distributions

- All inference about the parameter(s) of interest, θ, is (are) based on the posterior distribution.
- The information contained in the posterior distribution can be summarised in different ways as appropriate to the inference goal, e.g.
  - Means, standard deviations, medians.
  - Probability of exceeding a certain threshold, say θ_0, $\Pr(\theta > \theta_0 \mid \mathbf{y})$.
  - Credibility intervals.

Review of Bayesian inference

Decision theory

- ↪ Quantities derived from the posterior can be used as estimators of the unknown parameter(s). But, which of the possible summary statistics is 'best'?
- ↪ This is only a meaningful question if we have some notion of the 'cost' of making an error when estimating the value of the parameter. Decision theory provides such a framework by introducing the notion of a 'loss function'.
- ↪ Let $L(\theta, a)$ be the loss associated with using a as the estimate, when the true value is θ. Note that for simplicity we only write a but, in fact, it is $a(\mathbf{y})$.
- ↪ The corresponding Bayes estimator is then chosen to minimise the expected loss with respect to the posterior distribution.
- ↪ Mathematically, the Bayes estimator, $\widehat{\theta}$, is defined such that

$$\begin{aligned}\widehat{\theta}(\mathbf{y}) &= \text{argmin}_{a\in\Theta}\mathbb{E}_{\text{post.}}\left[L(\theta, a)\right] \\ &= \text{argmin}_{a\in\Theta}\left[\int_{\theta\in\Theta} L(\theta, a)p(\theta \mid \mathbf{y})\text{d}\theta\right].\end{aligned}$$

Review of Bayesian inference

Decision theory

↪ Three commonly used loss functions and corresponding Bayes estimators are:

Loss	Bayes estimate
$L(\theta, a) = (\theta - a)^2$ (quadratic loss)	$\widehat{\theta} = \mathbb{E}_{\text{post.}}(\theta) = \int_{\theta \in \Theta} \theta p(\theta \mid \mathbf{y}) \mathrm{d}\theta$
$L(\theta, a) = \lvert\theta - a\rvert$ (absolute error loss)	$\widehat{\theta} = \text{median}_{\text{post}}(\theta)$
$L(\theta, a) = I(\theta \neq a)$ (zero/one loss)	$\widehat{\theta} = \arg\max_\theta p(\theta \mid \mathbf{y})$

↪ Appropriate loss functions for hypothesis testing and interval estimation also exist.

Choice of prior distribution

- Picking the prior is obviously important and uniquely Bayesian. Priors must be specified on all the parameters that the statistician is interested in.
- Common types of priors include (these categories are not mutually exclusive):
  - Informative/expert priors
  - Non-informative/vague priors
  - Conjugate priors
- It is also important to try several priors in a sensitivity analysis.

Choice of prior distribution

Informative/expert priors

- A major advantage of the Bayesian approach is the ability to include expert prior information.
- This can be designed to reflect the subjective opinion of an expert in the field or by using past data (literature, pilot study, etc).
- The process of extracting prior knowledge in a suitable manner to permit the formulation of a prior distribution that represents the expert/historical information as accurately as possible is called **elicitation**.
- Spiegelhalter et al. (2004), Sections 5.2, 5.3, and 5.4, contain a good discussion on how priors might be elicited from experts or historical data.

Choice of 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 $\theta \sim \mathrm{N}(\mu_j, \sigma_j)$.
- One approach is to weight the experts using a *mixture model*

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

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

Choice of prior distribution

Non-informative/vague priors

- What should we do if we do not have any prior information concerning the parameter of interest?
- Bayes himself suggested that when this is the case, the Uniform prior should be used, so that $\pi(\theta) = c$, for all θ.
- In this case we clearly have

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

  i.e., the posterior distribution has the same shape as the likelihood function.
- In this case the mode of the posterior distribution coincides with the MLE.
- If the parameter space Θ is unbounded, this prior will be improper for any choice of c, i.e., $\int_\Theta \pi(\theta) \mathrm{d}\theta = \infty$. Improper priors can be used but we must check if the resulting posterior is proper (i.e., if it can be normalized to integrate to one).

Choice of prior distribution

Non-informative/vague priors

- ↪ As argued by Raiffa and Schlaiffer (1961), if one is ignorant about θ, one should also be ignorant about θ^2, and one cannot find a distribution that is uniform on both θ and θ^2.
- ↪ To see this, suppose that we place a Uniform prior on $\theta \in [0, 1]$, so that $\pi(\theta) = 1$. The corresponding prior on $\psi = \theta^2$ is $p(\psi) = \frac{1}{2\sqrt{\psi}}$, which is obviously non-uniform on ψ.

Choice of prior distribution

Non-informative/vague priors: Jeffreys' prior

- Jeffreys (1961) proposed a class of priors that are invariant to transformations.
- Jeffreys' prior is $\pi(\theta) \propto \sqrt{I(\theta)}$, where $I(\theta)$ is the expected Fisher information.
- Remember that

$$I(\theta) = \mathbb{E}\left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log f(\mathbf{Y} \mid \theta)\right)^2,$$

  which in regular cases equals to $I(\theta) = -\mathbb{E}\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f(\mathbf{Y} \mid \theta)\right)$.

- For the case of a multivariate parameter vector, say $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$, Jeffreys' prior is given by

$$\pi(\boldsymbol{\theta}) \propto \sqrt{\det I(\boldsymbol{\theta})},$$

  with $I(\boldsymbol{\theta}) = -\mathbb{E}\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} \log f(\mathbf{Y} \mid \boldsymbol{\theta})\right)$.

Choice of prior distribution

Non-informative/vague priors: Jeffreys' prior

- We will now calculate the Jeffreys' prior for binomial data.
- Let us suppose that $Y \sim \text{Bin}(n, \theta)$, where Y denotes the number of 'successes' out of n trials, and where the probability of success is θ.
- The likelihood is

$$f(y \mid \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y},$$

and the log likelihood is

$$\log f(y \mid \theta) = C + y \log \theta + (n-y) \log(1-\theta).$$

- The first derivative is

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log f(y \mid \theta) = \frac{y}{\theta} - \frac{n-y}{1-\theta},$$

and the second derivative is

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f(y \mid \theta) = -\frac{y}{\theta^2} - \frac{n-y}{(1-\theta)^2}.$$

Choice of prior distribution

Non-informative/vague priors: Jeffreys' prior

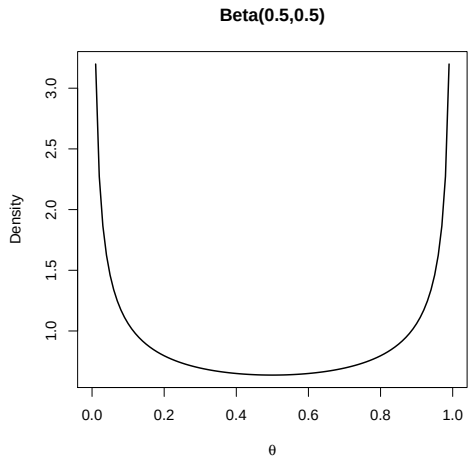
- Thus,

$$I(\theta) = -\mathbb{E}\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(Y \mid \theta)\right) = \frac{1}{\theta^2}\mathbb{E}(Y) + \frac{1}{(1-\theta)^2}(n - \mathbb{E}(Y)).$$

- Remember that $Y \sim \text{Bin}(n, \theta)$, implies $E(Y) = n\theta$.
- Therefore,

$$I(\theta) = \frac{n}{\theta(1-\theta)}.$$

- We can then conclude that Jefrreys's prior is $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, which is a Beta$(1/2, 1/2)$ distribution and gives greater plausibility to values near 0 and 1 than to values in between (see figure below).



Choice of prior distribution

Non-informative/vague priors: Jeffreys' prior

↪ Let us now compute Jeffreys' prior for a normal distribution with mean θ (assuming that the variance σ^2 is known).

↪ With $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{N}(\theta, \sigma^2)$, the likelihood is

$$
\begin{aligned}
f(\mathbf{y} \mid \theta) &= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \theta)^2 \right\} \right] \\
&= (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta)^2 \right\},
\end{aligned}
$$

and the corresponding log likelihood is

$$
\log f(\mathbf{y} \mid \theta) = C - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta)^2.
$$

Choice of prior distribution

Non-informative/vague priors: Jeffreys' prior

↪ The first and second derivatives are, respectively

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log f(\mathbf{y} \mid \theta) = \frac{n}{\sigma^2}(\bar{y} - \theta),$$

and

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f(\mathbf{y} \mid \theta) = -\frac{n}{\sigma^2}.$$

↪ The expected Fisher information is then $I(\theta) = \frac{n}{\sigma^2}$, which does not depend on θ, thus implying that $\pi(\theta) \propto 1$, $\forall\theta$ ($\Rightarrow$ improper prior).

Choice of prior distribution

Non-informative/vague priors: Jeffreys' prior

- We will now consider the Poisson case. Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$, $\theta > 0$.
- The likelihood and log likelihood are, respectively, given by

$$f(\mathbf{y} \mid \theta) = \prod_{i=1}^{n} \left\{ \frac{e^{-\theta}\theta^{y_i}}{y_i!} \right\} = \frac{e^{-n\theta}\theta^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!},$$

and

$$\log f(\mathbf{y} \mid \theta) = -n\theta + \sum_{i=1}^{n} y_i \log \theta + C.$$

- The first derivative is

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log f(\mathbf{y} \mid \theta) = -n + \frac{1}{\theta} \sum_{i=1}^{n} y_i,$$

while the second derivative is

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f(\mathbf{y} \mid \theta) = -\frac{\sum_{i=1}^{n} y_i}{\theta^2}.$$

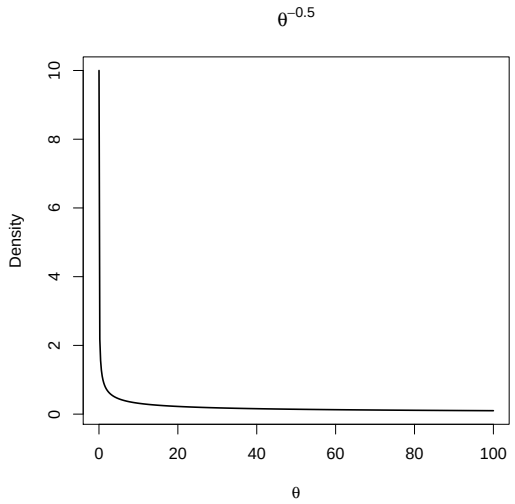
Choice of prior distribution

Non-informative/vague priors: Jeffreys' prior

↪ The expected Fisher information is then

$$I(\theta) = \frac{1}{\theta^2} n\mathbb{E}[Y] = \frac{n}{\theta},$$

implying that $\pi(\theta) \propto \theta^{-1/2}$.



↪ Jeffreys' prior is improper in this case, but it can be approximated by a Gamma distribution with parameters $\alpha = 1/2$ and $\beta \to 0$.

Choice of prior distribution

Non-informative/vague priors: Jeffreys' prior

- ↪ Jeffreys' prior is objective in that there is no prior tuning. It is a means of constructing a prior in the absence of prior information.
- ↪ Although Jeffreys' prior has the desirable property of being invariant to reparameterisations it can lead to improper priors.
- ↪ Alternative vague or non-informative prior distributions often have a reasonable mean for the distribution, but with a large variance parameter.
- ↪ Several different priors may be considered, each of which may be described to be vague or non-informative, and the sensitivity of the posterior on these priors investigated.

Choice of prior distribution

Conjugate priors

- A conjugate prior leads to a posterior from the same parametric family as the prior.
- There are long lists of conjugacies that we should be aware of

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

- Conjugate priors are used often for computational convenience because the posterior has a closed form.
- In fancier models, conjugate priors facilitate Gibbs sampling which is the easiest Bayesian computational algorithm.

Choice of prior distribution

Conjugate priors: beta-binomial model

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

$$f(y \mid \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

↪ It is mathematically convenient to use a Beta(a, b) prior distribution for θ because it has a similar form to the binomial likelihood. Its density function is

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

↪ If $\theta \sim \text{Beta}(a, b)$, then

$$\mathbb{E}(\theta) = \frac{a}{a+b},$$

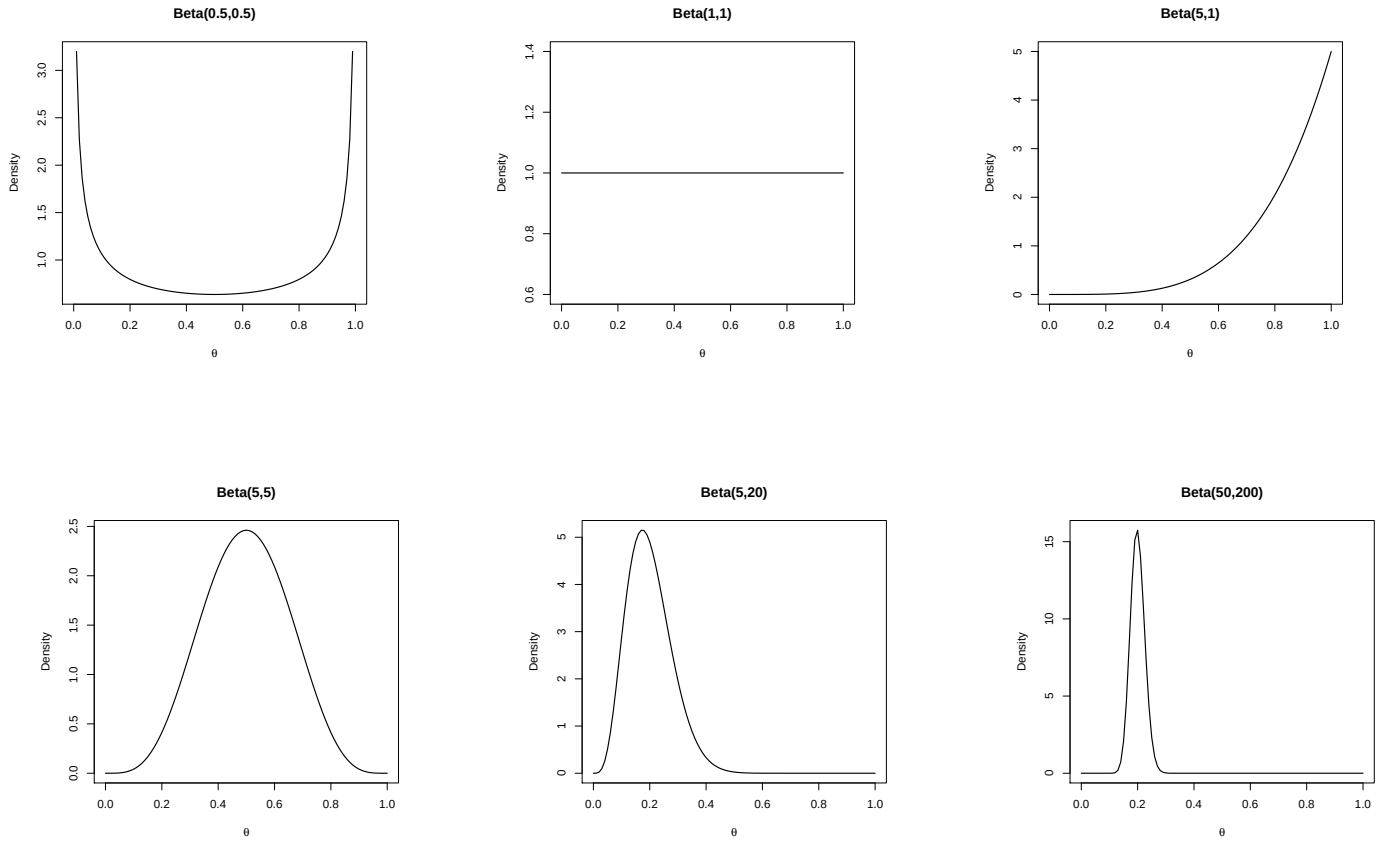
and variance

$$\text{var}(\theta) = \frac{ab}{(a+b)^2(a+b+1)}.$$

Choice of prior distribution

Conjugate priors: beta-binomial model

↪ The beta distribution can take several different shapes.



Choice of prior distribution

Conjugate priors: beta-binomial model

↪ Combining the beta prior distribution for θ with the binomial likelihood results in the following posterior distribution

$$\begin{aligned} p(\theta \mid y) &\propto f(y \mid \theta)\pi(\theta) \\ &= \binom{n}{y}\theta^{y}(1-\theta)^{n-y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\ &\propto \theta^{a+y-1}(1-\theta)^{b+n-y-1}. \end{aligned}$$

↪ That is, the posterior distribution is another Beta distribution

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

Choice of prior distribution

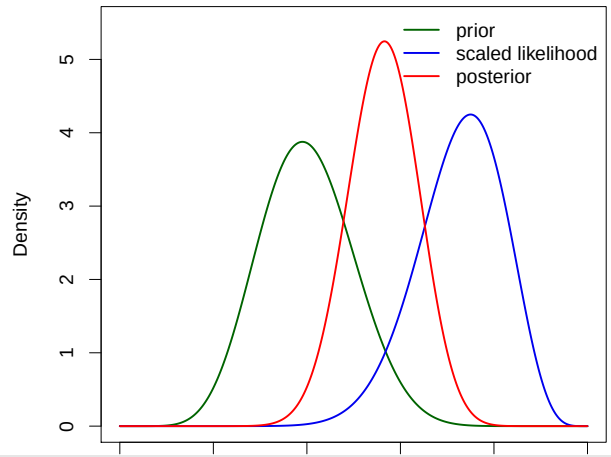
Conjugate priors: beta-binomial model

- The prior mean is $E(\theta) = \frac{a}{a+b}$.
- The data mean (MLE) is y/n.
- The posterior mean $\mathbb{E}[\theta \mid y] = \frac{a+y}{a+b+n}$.
- We can interpret prior information as being equivalent to having observed a successes in $a + b$ prior trials.
- With fixed a and b, as y and n increase, $\mathbb{E}[\theta \mid y] \to \frac{y}{n}$ (the MLE), and the variance tends to zero.
- This is a general phenomenon: as n increases, the posterior distribution gets more concentrated and the likelihood dominates the prior.

Choice of prior distribution

Conjugate priors: beta-binomial model

- Consider a drug to be given for relief of chronic pain.
- Experience with similar compounds has suggested that response rates, say θ, between 0.2 and 0.6 could be feasible.
- One way to turn this information into a prior is to interpret it as a distribution with mean $(0.2 + 0.6)/2 = 0.4$ and standard deviation $(0.6 - 0.2)/4 = 0.1$.
- 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 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.
- See implementation in the `R` script for Lecture 1.
- An alternative way to interpret the information that "response rates between 0.2 and 0.6 could be feasible" as a prior is to consider the effect of the prior, i.e. given the prior $\text{Beta}(a, b)$, the posterior is $\theta \mid y \sim \text{Beta}(a + y, b + n - y)$. This means that the prior effectively corresponds to data confirming a successes out of $a + b$ trials. Since the interval for the feasible response rates ([0.2,0.6]) is rather large, it makes sense to choose the number of trials (a+b) rather low. So for example $\text{Beta}(4, 6)$ would be sensible, as this corresponds to 4 successes out of 10 trials. Such an approach can be applied even when it is difficult to link the information to the parameters directly.

Choice of prior distribution

Conjugate priors: poisson-gamma model

- Suppose we have $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$.
- We have already seen that the likelihood is

$$f(\mathbf{y} \mid \theta) = \prod_{i=1}^{n} \left\{ \frac{e^{-\theta} \theta^{y_i}}{y_i!} \right\}$$

- The kernel of the Poisson likelihood (as a function of θ) has the same form as that of a Gamma(a, b) prior for θ

$$\pi(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta}.$$

- This parameterisation of the Gamma distribution in terms of the shape parameter a and rate parameter b, has mean a/b and variance a/b^2.

Choice of prior distribution

Conjugate priors: poisson-gamma model

↪ With this prior, the posterior is

$$\begin{aligned} p(\theta \mid \mathbf{y}) &\propto f(\theta \mid \mathbf{y})\pi(\theta) \\ &= \left\{ \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{y_i}}{y_i!} \right\} \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} \\ &\propto e^{-n\theta}\theta^{\sum_{i=1}^{n} y_i} \theta^{a-1} e^{-b\theta} \\ &= \theta^{a+\sum_{i=1}^{n} y_i - 1} e^{-\theta(b+n)} \end{aligned}$$

↪ We recognise this as the kernel of a gamma distribution with parameters $a + n\bar{y}$ and $b + n$, that is

$$\theta \mid \mathbf{y} \sim \text{Gamma}(a + n\bar{y}, b + n).$$

Choice of prior distribution

Conjugate priors: poisson-gamma model

↪ Note that

$$\mathbb{E}(\theta \mid \mathbf{y}) = \frac{a + n\bar{y}}{b + n} = \bar{y}\left(\frac{n}{n+b}\right) + \frac{a}{b}\left(1 - \frac{n}{n+b}\right).$$

↪ The posterior mean is then a compromise between prior mean a/b and the MLE $\bar{y}$.

Choice of prior distribution

Conjugate priors: normal-normal model

- Let us now consider $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{N}(\theta, \sigma^2)$, with σ^2 known. Further, let $\theta \sim \text{N}(\mu_0, \sigma_0^2)$.
- We've already seen the likelihood

$$f(\mathbf{y} \mid \theta) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta)^2 \right\}.$$

- The posterior is

$$\begin{aligned} p(\theta \mid \mathbf{y}, \sigma^2) &\propto f(\mathbf{y} \mid \theta)\pi(\theta) \\ &\propto \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i} (y_i - \theta)^2 \right\} \exp\left\{ -\frac{1}{2\sigma_0^2} (\theta - \mu_0)^2 \right\} \\ &\propto \exp\left\{ -\frac{1}{2\sigma^2\sigma_0^2} \left[\theta^2(n\sigma_0^2 + \sigma^2) - 2\theta(n\bar{y}\sigma_0^2 + \mu_0\sigma^2) \right] \right\} \end{aligned}$$

Choice of prior distribution

Conjugate priors: normal-normal model

↪ This can be recognised as the density of a normal distribution with mean

$$\mu_n = \frac{n\bar{y}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}},$$

and variance

$$\sigma_n^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2} = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}.$$

↪ The posterior parameters μ_n and σ_n^2 combine the prior parameters μ_0 and σ_0^2 with terms from the data.

↪ For instance, notice that

$$\mu_n = \frac{\tau_0}{\tau_0 + n\tau}\mu_0 + \frac{n\tau}{\tau_0 + n\tau}\bar{y},$$

where $\tau_0 = 1/\sigma_0^2$ (prior precision) and $\tau = 1/\sigma^2$ (sampling precision). The posterior mean is an average of the prior mean and the sample mean, weighted by their precisions.

Choice of prior distribution

Conjugate priors: normal-gamma model

- Suppose now that θ is known (which is an unrealistic scenario in practice), but the variance σ^2 is unknown.
- It is often convenient in Bayesian statistics to work with the precision, $\tau = 1/\sigma^2$.
- Let us take $\tau \sim \text{Gamma}(a, b)$.
- Then the posterior of τ is

$$\begin{aligned} p(\tau \mid \mathbf{y}, \theta) &\propto f(\mathbf{y} \mid \tau)\pi(\tau) \\ &\propto \tau^{n/2} \exp\left\{ -\frac{\tau}{2} \sum_{i=1}^{n} (y_i - \theta)^2 \right\} \tau^{a-1} e^{-b\tau} \\ &= \tau^{a+n/2-1} \exp\left\{ -\tau \left(b + \frac{1}{2} \sum_{i} (y_i - \theta)^2 \right) \right\}. \end{aligned}$$

- That is,

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

Choice of prior distribution

Conjugate priors: normal-gamma model

↪ A common choice is to take both a and b very small, then the posterior is approximately

$$\tau \mid \mathbf{y}, \theta \sim \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\sum_i (y_i - \theta)^2\right),$$

and so

$$\mathbb{E}[\tau \mid \mathbf{y}, \theta] = \left(\frac{1}{n}\sum_i (y_i - \theta)^2\right)^{-1},$$

so that the posterior expectation of the precision is (approximately) the sample precision (but with a divisor of n and not $n - 1$).

Choice of prior distribution

Conjugate priors: mixture priors

- Conjugate priors are convenient, but sometimes they might be not flexible enough.
- Mixtures of conjugate priors are a good alternative and they are actually quite flexible.
- Fortunately, mixtures of conjugate priors are also conjugate. A mixture prior is

$$\pi(\theta) = \sum_{j=1}^{J} \omega_j p_j(\theta|\psi_j),$$

  where p_j are conjugate priors with different hyperparameters, ψ_j, and the mixture weights $\omega_j \in [0, 1]$ sum to one.

- **Question**: If $Y \sim \text{Bin}(n, \theta)$ and $\pi(\theta) = \pi\text{Beta}(\theta \mid a_1, b_1) + (1 - \pi)\text{Beta}(\theta \mid a_2, b_2)$, then $\theta \mid y$ is ... (see first practical lab next week).
- **Additional reading**: see Section 2 of Bayesian Statistical Methods for more examples on the choice of prior distribution.

Predictive inference

- We now consider prediction.
- The prior predictive distribution of $\mathbf{y}$ is

$$m(\mathbf{y}) = \int_{\Theta} f(\mathbf{y} \mid \theta)\pi(\theta)\mathrm{d}\theta,$$

  also called marginal distribution of the data, usually in the frequentist approach.
- It is useful to check whether the model (likelihood+prior) gives (un)reasonable predictions.
- The posterior predictive distribution of a future observation z, given the data $\mathbf{y}$ is

$$\begin{aligned} f(z \mid \mathbf{y}) &= \int_{\Theta} f(z \mid \theta, \mathbf{y})p(\theta \mid \mathbf{y})\mathrm{d}\theta \\ &= \int_{\Theta} f(z \mid \theta)p(\theta \mid \mathbf{y})\mathrm{d}\theta, \end{aligned}$$

  where the second equality is due to the conditional independence of Y and Z.
- The predictive distribution only depends on z and $\mathbf{y}$.

Predictive inference

- Let us consider a beta binomial experiment, where $Y \sim \text{Bin}(n, \theta)$ and $\theta \sim \text{Beta}(a, b)$.
- We have seen that the posterior distribution $\theta \mid y \sim \text{Beta}(a + y, b + n - y)$.
- For ease of notation, let $c = a + y$ and $d = b + n - y$.
- Now let us consider a further random quantity Z for which we judge $Z \sim \text{Bin}(m, \theta)$ and Z and Y are conditionally independent given θ.
- So, having conducted n trials, we consider conducting a further m trials.
- We seek the predictive distribution of Z (the number of successes in m new trials) given the observed y.

Predictive inference

↪ We have

$$
\begin{aligned}
f(z \mid y) &= \int_{\Theta} f(z \mid \theta) p(\theta \mid y) \mathrm{d}\theta \\
&= \int_0^1 \binom{m}{z} \theta^z (1-\theta)^{m-z} \frac{1}{B(c,d)} \theta^{c-1} (1-\theta)^{d-1} \mathrm{d}\theta \\
&= \binom{m}{z} \frac{1}{B(c,d)} \int_0^1 \theta^{c+z-1} (1-\theta)^{d+m-z-1} \mathrm{d}\theta \\
&= \binom{m}{z} \frac{B(c+z, d+m-z)}{B(c,d)} \\
&= \binom{m}{z} \frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)} \frac{\Gamma(c+z)\Gamma(d+m-z)}{\Gamma(c+d+m)}.
\end{aligned}
$$

↪ We say that $Z \mid y$ is the Binomial-Beta distribution with parameters c, d, and m.

↪ **Additional reading**: see Section 5.2 of Bayesian Theory by Bernardo and Smith for more examples of posterior predictive distributions from the exponential family.

Bayesian computing

- As we've seen, Bayesian inference centres around the posterior distribution

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{f(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \propto f(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

- Bayesian inference thus requires summarising the posterior distribution.
- When there are more than a few parameters, this requires advanced computational tools.
- We will first briefly discuss some methods that are deterministic. These usually rely on some approximation, such as
  - Bayesian Central Limit Theorem,
  - Numerical integration.
- Monte Carlo methods are more general, but they can become slow in high dimensions:
  - Direct sampling.
  - Markov chain Monte Carlo methods.

Bayesian computing

Bayesian central limit theorem

- In a frequentist approach it is common to compute the MLE, and then assume (based on asymptotic theory) that its sampling distribution is Gaussian for inference.
- The Bayesian central limit theorem can be used in the same way to summarise a posterior.
- **Bayesian central limit theorem**: Suppose $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(\mid \boldsymbol{\theta})$ and that the prior $\pi(\boldsymbol{\theta})$ and the likelihood $f(\mathbf{y} \mid \boldsymbol{\theta})$ are positive and twice differentiable near $\widehat{\boldsymbol{\theta}}_{\text{post}}$, the posterior mode of $\boldsymbol{\theta}$. Then for large n

$$p(\boldsymbol{\theta} \mid \mathbf{y}) \sim \text{N}\left(\widehat{\boldsymbol{\theta}}_{\text{post}}, [I^{\text{post}}(\boldsymbol{\theta}, \mathbf{y})]^{-1}\right),$$

where

$$I^{\text{post}}(\boldsymbol{\theta}, \mathbf{y}) = -\left[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log p(\boldsymbol{\theta} \mid \mathbf{y})\right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{post}}}.$$

Bayesian computing

Bayesian central limit theorem

- ↪ Other forms of the normal approximation are also often used.
- ↪ In particular, a powerful new type of models called latent Gaussian models have became widely used in the last two decades.
- ↪ In such models, we have a few dimensional (up to 15) non-Gaussian parameters called hyperparameters, and a large dimensional (up to 10^{12} in recent years) latent random vector that is Gaussian when conditioned on the hyperparameters. The observations are allowed to be nonlinear functions of the latent variables.
- ↪ The fact that the latent random vector is conditionally Gaussian allows for efficient and accurate computations using various clever tricks. These form the basis of the INLA (Integrated Nested Laplace Approximations) approach, which will be discussed in detail starting from Lecture 3.

Bayesian computing

Numerical integration

- ↪ The vast majority of posterior summaries are integrals of the posterior (posterior mean, variance, probability above zero, etc).
- ↪ There is a vast literature on approximating integrals.
- ↪ These work great in medium dimensions, say models with 1-15 parameters.
- ↪ The simplest method is a grid approximation. Suppose we have an unnormalised posterior distribution we want to sample from. The grid approximation works as follows:
  1. Divide the unnormalised posterior area into m grids.
  2. Evaluate each grid point at the unnormalised posterior.
  3. Sample grid points with unnormalised posterior ordinates as the probabilities.

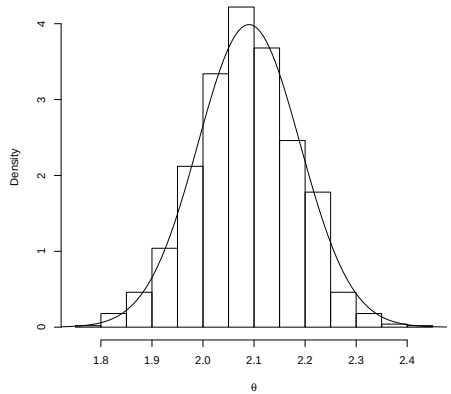
Bayesian computing

Numerical integration

- Suppose $y_1, \ldots, y_n$ is observed data and that $Y_i \overset{\text{iid}}{\sim} \text{N}(\theta, 1)$, and $\theta \sim \text{N}(\mu_0 = 0, \sigma_0^2 = 10^2)$.
- We do not need numerical integration here since we do know that $p(\theta \mid \mathbf{y})$ is actually available in closed form, namely

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

- This is only a toy example to illustrate the use of the method.



- See implementation in the `R` script for Lecture 1.

Bayesian computing

Monte Carlo sampling

- ↪ Monte Carlo sampling is the predominant method of Bayesian inference because it avoids asymptotic approximations and can be used in high-dimensions.
- ↪ The main idea is to approximate posterior summaries by drawing samples from the posterior distribution, and then averaging quantities of interest over these samples to approximate integrals of those quantities over the posterior.
- ↪ We have seen in last lecture that the best estimator of $h(\boldsymbol{\theta})$ in MSE is $\int_{\boldsymbol{\theta}} h(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta}$.
- ↪ For example, if $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(S)}$ are samples from $p(\boldsymbol{\theta} \mid \mathbf{y})$, then the mean $\overline{\boldsymbol{\theta}}$ of S samples can be used to approximate the posterior mean. More generally, we use estimators of the form $(h(\boldsymbol{\theta}^{(1)}) + \ldots + h(\boldsymbol{\theta}^{(S)}))/S$ for $\int_{\boldsymbol{\theta}} h(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta}$.
- ↪ Many argue that this form of approximation is superior to asymptotic approximations because the Bayes CLT requires the sample size of the dataset to go to infinity and the Monte Carlo approximation requires the number of simulated values to go to infinity.
- ↪ In most cases, $S \to \infty$ is cheaper and more realistic than $n \to \infty$.
- ↪ But how to draw samples from some arbitrary distribution $p(\boldsymbol{\theta} \mid \mathbf{y})$?

Bayesian computing

Monte Carlo sampling: Rejection sampling

- ↪ Suppose we wish to generate values $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(S)}$ from the posterior $p(\theta \mid \mathbf{y})$ but this is not a known distribution.
- ↪ Rejection sampling takes samples from a distribution that resembles the posterior and is easy to sample from (say a normal approximation using the CLT), and thins those samples to obtain draws from the posterior distribution.
- ↪ The approximate density, $g(\theta)$, is called the envelope function.
- ↪ Let M be a constant so that $p(\theta \mid \mathbf{y}) \leq Mg(\theta)$ for all θ. Then $p(\theta \mid \mathbf{y})$ resides in the envelope.

Bayesian computing

Monte Carlo sampling: Rejection sampling

↪ The rejection sampling algorithm is:

1. Sample $\theta \sim g(\theta)$ and draw $u \sim \text{Unif}(0, 1)$.
2. Set $\alpha = \frac{p(\theta|\mathbf{y})}{Mg(\theta)}$.
3. If $u \leq \alpha$, accept θ.

Repeat until you have accepted S (pre-determined) values. The accepted values $\theta^{(1)}, \ldots, \theta^{(S)}$ are a random sample from $p(\theta \mid \mathbf{y})$.

Bayesian computing

Monte Carlo sampling: Rejection sampling

- ↪ It is sometimes tricky to choose a good envelope/proposal distribution $g(\theta)$. A basic requirement is that it should have support at least as large as $p(\theta \mid \mathbf{y})$ and preferably heavier tails than $p(\theta \mid \mathbf{y})$.

- ↪ It is desirable to choose M as small as possible for efficiency reasons - in fact, the expected number of samples taken until one is accepted is exactly M. Note that $M \geq \frac{p(\theta|\mathbf{y})}{g(\theta)}$ for all θ, so that the optimal M is simply

$$M^{\text{opt}} = \max_{\theta} \left(\frac{p(\theta \mid \mathbf{y})}{g(\theta)} \right).$$

- ↪ Rejection sampling also works when the normalisation constant is unknown. In this case, we have an unnormalised posterior density $\tilde{p}(\theta \mid \mathbf{y})$. If it satisfies that for a normalised proposal distribution $g(\theta)$,

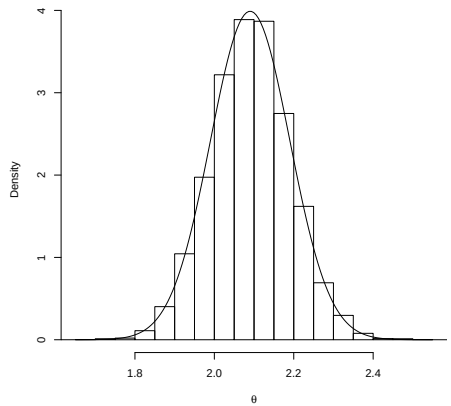
$$\tilde{p}(\theta \mid \mathbf{y}) \leq Mg(\theta) \quad \text{for every } \theta,$$

then samples from $g(\theta)$ accepted with probability $\alpha = \frac{\tilde{p}(\theta|\mathbf{y})}{Mg(\theta)}$ are distributed according to $p(\theta \mid \mathbf{y})$. This works because we can essentially incorporate the normalisation constant into the M term. The expected number of samples taken until one is accepted in this case is M/Z, where $Z = \int_\theta \tilde{p}(\theta \mid \mathbf{y}) d\theta$ is the normalising constant.

Bayesian computing

Monte Carlo sampling: Rejection sampling

- ↪ We return to the toy example used to illustrate the grid approximation to the posterior.
- ↪ In this context, we choose as an envelope distribution, a normal distribution with mean equal to the mean of the data and standard deviation equal to the standard deviation of the data.



- ↪ See implementation in the `R` script for Lecture 1.
- ↪ Although direct sampling methods such as rejection sampling can be very efficient in small dimensions, they become typically exponentially slow in higher dimensions.
- ↪ More general methods are needed.

Bayesian computing

MCMC

- ↪ MCMC techniques are based on the construction of a Markov chain that eventually 'converges' to the target distribution (called the stationary or equilibrium distribution) which, in our case, is the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{y})$.
- ↪ This is the main way to distinguish MCMC algorithms from direct simulation methods, which provide samples directly from the target posterior distribution.
- ↪ Moreover, the MCMC output is a dependent sample since it is generated from a Markov chain, in contrast to the output of direct methods, which are independent samples.
- ↪ Finally, MCMC methods incorporate the notion of an iterative procedure (for this reason they are frequently called iterative methods) since in every step they produce values depending on the previous one.
- ↪ We will briefly cover two different MCMC methods: Gibbs sampler and Metropolis-Hastings.

Bayesian computing

MCMC: Gibbs sampler

- Gibbs sampling was proposed in the early 1990s (Geman and Geman, 1984; Gelfand and Smith, 1990) and fundamentally changed Bayesian computing.
- Gibbs sampling is attractive because it can sample from high-dimensional posteriors.
- The main idea is to break the problem of sampling from the high-dimensional joint distribution into a series of samples from low-dimensional conditional distributions.
- The algorithm begins by setting initial values for all parameters, $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})$.
- Variables are then sampled one at a time from their full conditional distributions

$$p(\theta_j \mid \theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p, \mathbf{y}), \qquad j = 1, \ldots, p.$$

- Rather than 1 sample from the p-dimensional joint distribution, we make p samples from the 1-dimensional conditional distributions (sometimes using rejection sampling).
- The process is repeated until the required number of samples have been generated.

Bayesian computing

MCMC: Gibbs sampler

↪ Formally, the algorithm is:

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

↪ Then for s sufficiently large $(\theta_1^{(s)}, \ldots, \theta_p^{(s)}) \overset{\text{approx.}}{\sim} p(\theta_1, \ldots, \theta_p \mid \mathbf{y})$.

↪ The convergence of the p-tuple obtained at iteration s, $(\theta_1^{(s)}, \ldots, \theta_p^{(s)})$ to a draw from a joint posterior distribution occurs under mild regularity conditions that are generally satisfied for most statistical models (see, e.g., Geman and Geman, 1984, or Roberts and Smith, 1993).

Bayesian computing

MCMC: Gibbs sampler

- Suppose we have data $y_1, \ldots, y_n$ such that $Y_i \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown.

- We need to specify the joint prior distribution $p(\mu, \sigma^2)$. Common choices are:

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

- We will use the independent prior (second option) and assume $\mu \sim \text{N}(\mu_0, \sigma_0^2)$, and $\sigma^2 \sim \text{IG}(a, b)$ (inverse Gamma distribution).

- We have that

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

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

Bayesian computing

MCMC: Gibbs sampler

↪ The algorithm is then

1. Set initial values $\mu^{(0)}$ and $(\sigma^{(0)})^2$.
2. For $s = 1, \ldots, S$
   - Draw $\mu^{(s)} \mid \mathbf{y}, (\sigma^{(s-1)})^2 \sim \text{N}\left(\frac{\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{\mathbf{y}}}{(\sigma^{(s-1)})^2}}{\frac{1}{\sigma_0^2} + \frac{n}{(\sigma^{(s-1)})^2}}, \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{(\sigma^{(s-1)})^2}} \right)$.
   - Draw $(\sigma^{(s)})^2 \mid \mathbf{y}, \mu^{(s)} \sim \text{IG}\left(a + \frac{n}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu^{(s)})^2 \right)$.

↪ See implementation in the R script for Lecture 1.

Bayesian computing

MCMC: Gibbs sampler

- ↪ Later in the lecture we will look more closely at convergence diagnostics.
- ↪ One basic diagnostic is to monitor the *traceplots*: plots of the generated values of the parameters versus the iteration number.
- ↪ If all values are within a zone without strong periodicity and (especially) a trend, then there is no evidence of lack of convergence.

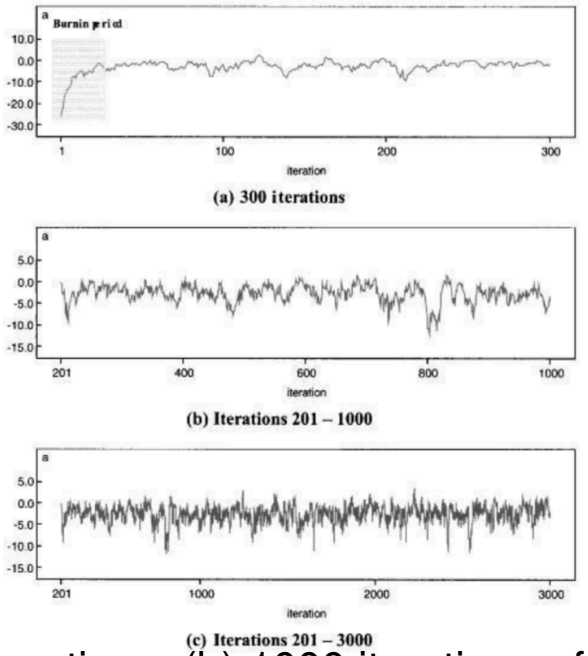


Figure: Traceplots for (a) 300 iterations, (b) 1000 iterations after discarding the first 200, and (c) 3000 iterations after discarding the first 200. Image from Ntzoufras (2009).

Bayesian computing

MCMC: Gibbs sampler

- In the first trace plot, we can clearly see the burnin period (within the gray box), which must be discarded from the final sample. After this period the generated sampled values are stabilised within a zone.

- In the second plot, the initial 200 iterations have been discarded to monitor the sampled values which demonstrate much better behaviour with small periodicities (up and down periods in the graph).

- Finally, generated values of the last trace plot are more convincing in terms of convergence, with all generated values within a parallel zone and no obvious tendencies or periodicities.

Bayesian computing

MCMC: Gibbs sampler

- ↪ Updating parameters one at a time can lead to high autocorrelation.
- ↪ The lag h autocorrelation function (ACF) for parameter θ_j is

$$\rho_j(h) = \lim_{s \to \infty} \text{Cor}(\theta_j^{(s)}, \theta_j^{(s-h)}),$$

i.e. it is the correlation between the given parameter value in the Markov chain separated by h iterations. The term $h > 1$ is usually referred to as lag.

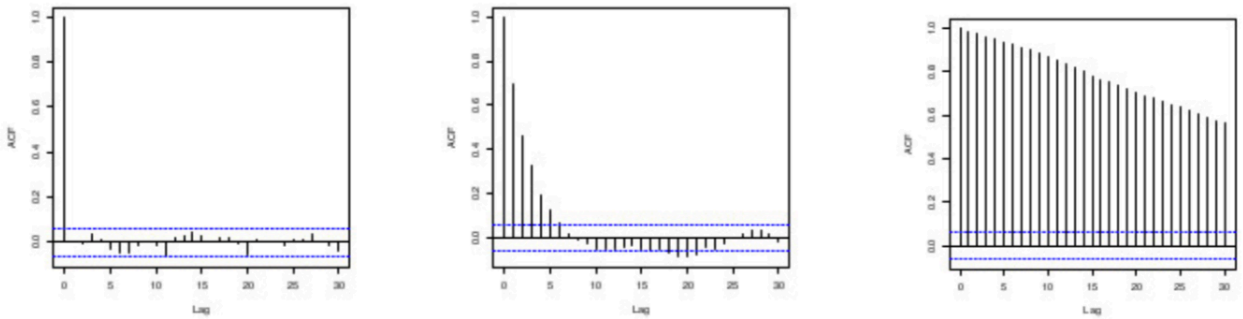


Figure: Sample ACF plots representing (a) ideal mixing, (b) typical good mixing, (c) poor mixing. Image from Prof. R. King.

Bayesian computing

MCMC: Gibbs sampler

- ↪ Note that the autocorrelation function is always equal to 1 for the value $h = 0$, since $\text{Cor}(\theta_j^{(s)}, \theta_j^{(s)}) = 1$.
- ↪ Ideally, for efficient Markov chains, there should be a fast decrease in the value of the autocorrelation function as the lag increases.
- ↪ This would imply that there is little relationship between values of the Markov chain within a small number of iterations.
- ↪ Conversely, poorly mixing chains will typically have a very shallow gradient in the ACF plot, with high autocorrelation values for even relatively large values of h.

Bayesian computing

MCMC: Gibbs sampler

- ↪ A simple, yet reasonably effective, method for 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–saving and working with large chains can be burdensome.
- ↪ The disadvantage is that we are clearly throwing away information; thinning can never be as efficient as using all the iterations.

Bayesian computing

MCMC: Metropolis–Hastings algorithm

- ↪ Gibbs sampling requires drawing a sample from each full conditional distribution.
- ↪ In cases where the conditional distributions are conjugate sampling is straight-forward. But what if they are not conjugate?
- ↪ We could make draws from the conditional distributions using rejection sampling. This works well if there are only a few non-conjugate parameters but can be difficult to tune.
- ↪ Other methods have been proposed, with Metropolis–Hastings being the most widely used.

Bayesian computing

MCMC: Metropolis–Hastings algorithm

- ↪ Metropolis et al. (1953) first formulated the Metropolis algorithm, and the paper was originally published in the physical sciences.
- ↪ Later, Hastings (1970) generalised the original method to what is now known as the Metropolis–Hastings algorithm.
- ↪ The latter is considered to be the general formulation of all MCMC methods.
- ↪ Green (1995) further generalised the Metropolis–Hastings algorithm by introducing reversible jump Metropolis–Hastings algorithms for sampling from parameter spaces with unknown dimension.

Bayesian computing

MCMC: Metropolis–Hastings algorithm

↪ The algorithm is summarised as follows:

1. Set initial values $\theta^{(0)}$.
2. For $s = 1, \ldots, S$, repeat the following steps:
   - Set $\boldsymbol{\theta} = \boldsymbol{\theta}^{(s-1)}$.
   - Draw candidate parameter values $\boldsymbol{\theta}^*$ from a proposal distribution $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta})$.
   - Calculate

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

   - Generate $u \sim \text{Unif}(0, 1)$.
   - Set

$$\boldsymbol{\theta}^{(s)} = \begin{cases} \boldsymbol{\theta}^* & \text{if } u \leq \alpha, \\ \boldsymbol{\theta} & \text{if } u > \alpha. \end{cases}$$

Bayesian computing

MCMC: Metropolis–Hastings algorithm

- In the original Metropolis algorithm, only symmetric proposals of the type $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}) = q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*)$ were considered.
- Random walk Metropolis is a special case of the algorithm with $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}) = g(\boldsymbol{\theta}^* - \boldsymbol{\theta})$ and the function g satisfying $g(\mathbf{x}) = g(-\mathbf{x})$.
- With such a proposal distribution the kernel driving the chain is a random walk since the candidate values are of the form

$$\boldsymbol{\theta}^* = \boldsymbol{\theta} + \mathbf{z}, \quad \mathbf{z} \sim g.$$

- Both cases result in an acceptance probability that depends only on the posterior (target) distribution

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \min\left\{1, \frac{f(\mathbf{y} \mid \boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{f(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}\right\}.$$

- A common proposal of this type is a multivariate normal $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}) = \mathrm{N}_p(\boldsymbol{\theta}, \mathbf{S}_{\boldsymbol{\theta}})$.

Bayesian computing

MCMC: Metropolis–Hastings algorithm

↪ As an implementation example, we consider the case where $y_1, \ldots, y_n$ are drawn from a normal distribution with known variance and unknown mean. The prior for the mean is normally distributed. Of course, as we have already seen, we do not need Metropolis type algorithms to sample from the posterior in this case, but it serves as an illustrative example.

↪ See implementation in the `R` script for Lecture 1.

Bayesian computing

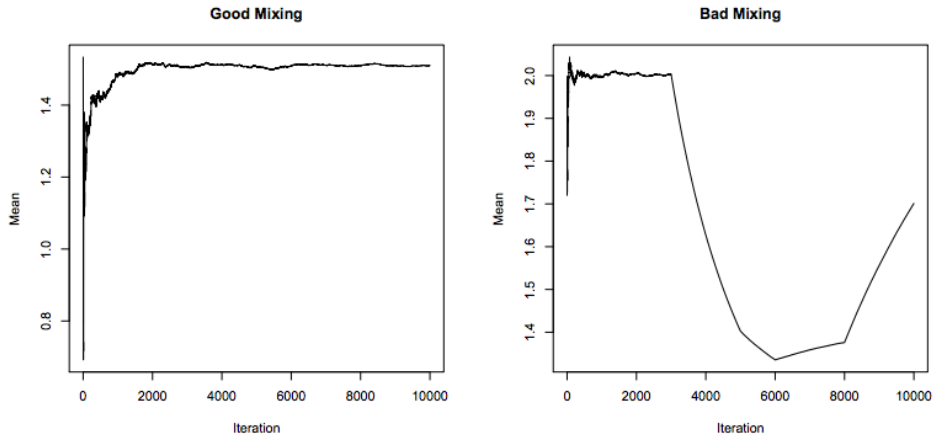
More on MCMC convergence diagnostics

- From the theory of Markov chains (more details were given in the Bayesian theory course), we expect the chains to eventually converge to the stationary distribution, which is also our target distribution. In this case the target is the posterior distribution.
- However, there is no guarantee that the chain has converged after S draws.
- How do we know whether the chain has actually converged?
- We can never be sure, but there are several tests we can do to check if the chain appears to be converged.

Bayesian computing

More on MCMC convergence diagnostics

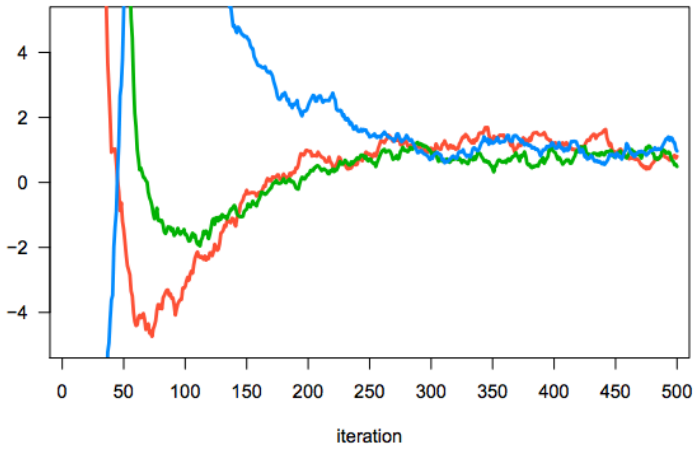
- ↪ We've already seen the use of traceplots and autocorrelation functions.
- ↪ Another quick check is to look at the running mean plots to check how well our chains are mixing.
- ↪ A running mean plot is a plot of the mean of all draws up to the current iteration versus the iteration number.



Bayesian computing

More on MCMC convergence diagnostics

- Many approaches for detecting convergence, both formal and informal, rest on the idea of starting multiple Markov chains and observing whether they come together and start to behave similarly (if they do, we can pool the results from each chain).



- This plot indicates that the three chains converge to the posterior after around $S = 300$ iterations; certainly, however, the number of iterations required to reach convergence depends on the initial values.

- It is typically recommended (e.g., Gelman and Rubin, 1992) to use overdispersed initial values, meaning 'more variable than the target distribution' i.e., the posterior.

Bayesian computing

More on MCMC convergence diagnostics

- ↪ Although looking at trace plots is certainly useful, it is also desirable to obtain an objective, quantifiable measure of convergence.
- ↪ Numerous methods exist, although we will focus on the measure originally proposed in Gelman and Rubin (1992).
- ↪ The basic idea is to quantify the between-chain and the within-chain variability of a quantity of interest. If the chains have converged, these measures will be similar; otherwise, the between-chain variability will be larger.

Bayesian computing

More on MCMC convergence diagnostics

↪ The basic idea of the estimator is as follows (the actual estimator makes a number of modifications to account for degrees of freedom):

- Let B denote the standard deviation of the pooled sample of all MS iterations, where M is the number of chains and S is the number of iterations in each chain.
- Let W denotes the average of the within-chain standard deviations.
- Quantify convergence with

$$\widehat{R} = \frac{B}{W}.$$

↪ If $\widehat{R} \gg 1$, this is clear evidence that the chains have not converged.

↪ As $S \to \infty$, $\widehat{R} \to 1$; $\widehat{R} < 1.05$ is widely accepted as implying sufficient convergence for practical purposes.

Bayesian computing

More on MCMC convergence diagnostics

↪ More details (along with other convergence measures) are given in the `coda` package, whose `gelman.diag` function provides, in addition to $\widehat{R}$ itself:

- An upper confidence interval for $\widehat{R}$.
- A multivariate extension of $\widehat{R}$ for quantifying convergence of the entire posterior.

Bayesian computing

More on MCMC convergence diagnostics

- ↪ The obvious downside to running multiple chains is that it is inefficient: we intentionally force our sampler to spend extra time in a non-converged state, which in turn requires much more burn-in.
- ↪ The obvious upside, however, is that it provides us with some measure of confidence that we are actually drawing samples from the posterior.
- ↪ But we reiterate that (without additional assumptions about the posterior) no method can truly prove convergence; diagnostics can only detect failure to converge.

Bayesian computing

More on MCMC convergence diagnostics

- We may use $\widehat{R}$, then, as a guide to how long we must run our chains until convergence.
- The obvious next question is: how long must we run our chains to obtain reasonably accurate estimates of the posterior?
- If we could obtain iid draws from the posterior, estimating the Monte Carlo standard error (at least, of the posterior mean) is straightforward: letting σ_{post} denote the posterior standard deviation, the MCSE is $\sigma_{\text{post}}/\sqrt{S}$.
- But, this will underestimate the true MC standard error due to autocorrelation in the samples generated using MCMC.
- There are various approaches to obtain better estimate of the MC error, with possibly the most popular being the *batches* mean method.
- **Additional reading**: for additional material on Bayesian computing, see Sections 10-12 of Bayesian Data Analysis by Gelman et al.