

# Introduction to Bayesian Analysis using R (probabilistic programming)

Dr Jackie Wong Siaw Tze\*

\* Department of Mathematical Sciences, University of Essex

IADS Summer School, 29 July 2022, University of Essex.

- 1 Introduction
- 2 Section 1: Basic Bayesian Concept and Computation  
**[9.30-10.30 am]**
- 3 Section 2: Posterior Inferences **[11-12.30 pm]**
- 4 Section 3: MCMC Methods **[1.30-3 pm]**
- 5 Section 4: Practical Examples **[3.30-5 pm]**
- 6 Conclusion

# Learning Objectives

By the end of today, you should be able to:

- State the **conceptual difference** between Frequentist and Bayes
- Understand and apply **Bayes Theorem**
- Define the terms **prior** and **posterior** distributions
- **Derive posterior** distributions given the likelihood and prior
- Perform Bayesian inferences using posterior: **point estimation, credible intervals, Monte Carlo** methods...

# Learning Objectives

By the end of today, you should be able to:

- Understand **the role of prior** distributions
- Assess the behavioral change of posterior w.r.t. prior parameters (**sensitivity analysis**)
- **Objective VS Subjective** Bayes
- Understand the concept of **non-informative** priors
- Define **conjugacy**

# Learning Objectives

By the end of today, you should be able to:

- Summarize posterior distributions using **sample-based** methods
- Understand and apply **MCMC methods** such as **Metropolis-Hastings, Gibbs algorithms** etc.
- Perform convergence checks using graphical diagnostics (e.g. **trace plots, auto-correlations**, etc.)
- **Implement MCMC methods** for some basic statistical models using R
- Apply Bayesian analysis on several **practical examples**

# Background

- Bayesian statistics takes its name from Rev. Thomas Bayes (1702-1761)
- A presbyterian minister, he published a paper posthumously "An essay towards solving a problem in the doctrine of chances" (1763)
- It included a form of Bayes Theorem.
- The approach was also independently developed by Laplace approximately 50 years later.
- For a while in the 20th century Bayesian statistics was overshadowed by Frequentist approach with the likes of Fisher (1890- 1962), Neyman (1894-1981), Pearson (1857-1936) etc.
- A key paper Gelfand and Smith (1990) changed all that and together with advances in high computational powers the Bayesian approach is increasingly appealing in various disciplines.

# Section 1: Basic Bayesian Concept and Computation [9.30-10.30 am]

# A motivational example...

- Suppose we are interested in predicting rainfall tomorrow.
- Observed data in the past:

$$y_i = \{1, 0, 1, 0, 0, 0, 1, 1, 0\}$$



# Conceptual Difference: Frequentist VS Bayesian

	Frequentist	Bayesian
Parameter	$\theta$ is <b>unknown but fixed</b>	$\theta$ is unknown so should be treated as a <b>random variable</b>
Interpretation of probability	<b>Relative frequency</b>	<b>Subjective</b> quantity representing individual's <b>belief</b> of "likelihood"
Intervals	Confidence intervals ( <b>not probabilistic</b> )	Credible intervals ( <b>probabilistic</b> )
Computation	Based primarily on <b>data likelihood</b>	Based on a <b>combination</b> of <b>data likelihood</b> and <b>prior</b>
Inferences	typically focused on point estimates (MLE or LSE)	Posterior distributions, where quantities can be derived from

# Bayes Theorem

- In its simplest form, **Bayes' Theorem** stipulates that

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

- Alternatively, we can write in terms of random variables:

$$P(X \in A|Y \in B) = \frac{P(Y \in B|X \in A)P(X \in A)}{P(Y \in B)}.$$

- Then, in the form of densities:

$$f(x|y) = \frac{f(y|x)f(x)}{f(y)}.$$

- How did this become an inference paradigm?

# Example of Bayes Theorem

- A rare disease affect 1 in 1000 adults.
- A diagnostic test has been developed:
  - For a diseased individual the test will be positive 99% of the time.
  - For a healthy individual the the test will be positive only 2% of the time.
- If a new individual (with unknown disease status) is tested positive, what is the probability s/he has the disease?

# Example of Bayes Theorem

- Let  $D$  be the event that an individual has the disease and  $+$ ,  
– be the events corresponding to a positive or negative test.
- We know

# Example of Bayes Theorem

- We can interpret the above analysis in the following way
- We have some **prior knowledge** about the disease (i.e. its probability distribution is known a priori)
- We collected data through an experiment
- We can then **update our knowledge about the disease given the result of the experiment** (i.e. obtain its probability distribution a posteriori)
- This is the **Bayesian way of thinking**

# Bayesian Inference

Generally speaking, for statistical inference:

- Let  $\mathbf{y}$  denote observed data and replace  $X$  with  $\theta$  denoting the (unobserved) parameters.
- We have

$$\pi(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)\pi(\theta)}{f(\mathbf{y})},$$

where

- $\pi(\theta|\mathbf{y})$ : posterior distribution ;
  - $f(\mathbf{y}|\theta)$ : likelihood;
  - $\pi(\theta)$ : prior distribution;
  - $f(\mathbf{y}) = \int f(\mathbf{y}|\theta)\pi(\theta)d\theta$ : marginal likelihood (key to Bayesian model selection).
- Notice that the denominator does not depend on the parameter of interest  $\theta$ , so we commonly write

$$\pi(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)\pi(\theta).$$

# Bayesian Inference

- Again, we have a natural inference paradigm here. **You infer about what you don't know given what you have seen.**
- By contrast, classical inference uses sampling distributions,  $T(\mathbf{y}; \theta)$ , a statistic given  $\theta$ : **imagine what you “might” see given what you don't know.**
- Posterior inference - benefits of an entire distribution (see later).
  - For parameters:  $\pi(\theta|\mathbf{y})$
  - For prediction:  $\pi(Y_0|\mathbf{y}) = \int f(Y_0|\mathbf{y}, \theta)\pi(\theta|\mathbf{y})d\theta$
  - For model selection:

$$\frac{\pi(M=2|\mathbf{y})}{\pi(M=1|\mathbf{y})} = \frac{f(\mathbf{y}|M=2)}{f(\mathbf{y}|M=1)} \times \frac{\pi(M=2)}{\pi(M=1)},$$

where  $f(\mathbf{y}|\mathbf{M}) = \int f(\mathbf{y}|\theta_M)\pi(\theta_M)d\theta_M$  is the marginal likelihood.

# Bayesian Inference

- More generally, we would have a prior distribution  $\pi(\theta|\boldsymbol{\lambda})$ , where  $\boldsymbol{\lambda}$  is a vector of hyperparameters.
- Can think of  $\theta$  as the process of interest with some parts known and some parts unknown.
- Then, this is a hierarchical specification,

$$f(\mathbf{y}|\text{process}, \theta)f(\text{process}, \theta|\boldsymbol{\lambda})\pi(\theta|\boldsymbol{\lambda})\pi(\boldsymbol{\lambda})$$



# Bayesian Inference

- If  $\lambda$  is known, the posterior of  $\theta$  is

$$\pi(\theta|\mathbf{y}, \lambda) = \frac{f(\mathbf{y}|\theta)\pi(\theta|\lambda)}{\int f(\mathbf{y}|\theta)\pi(\theta|\lambda)d\theta}$$

- If  $\lambda$  is not known, a second stage hyperprior distribution  $\pi(\lambda)$  will be required,

$$\pi(\theta, \lambda|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)\pi(\theta|\lambda)\pi(\lambda)}{\int \int f(\mathbf{y}|\theta)\pi(\theta|\lambda)\pi(\lambda)d\theta d\lambda}.$$

- Alternatively, we might replace  $\lambda$  in  $\pi(\theta|\mathbf{y}, \lambda)$  by an estimate,  $\hat{\lambda}$ , this is called **empirical Bayes analysis**.

# Bayesian Inference: Example 1

Suppose the random variables  $Y_1, \dots, Y_n \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$ .

- We observed data  $\mathbf{y} = (y_1, \dots, y_n)$ .
- Consider the prior distribution  $\theta \sim \text{Beta}(\alpha, \beta)$ , i.e.

$$\pi(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}, \quad 0 < \theta < 1,$$

where

$$E(\theta) = \frac{\alpha}{\alpha + \beta}, \quad \text{Var}(\theta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

- What is the posterior distribution of  $\theta$ ,  $\pi(\theta|\mathbf{y})$ ?

# Bayesian Inference: Example 1

# Bayesian Inference: Example 1

# Bayesian Inference: Example 1

- If we observed  $\mathbf{y} = (1, 1, 1, 1, 1, 0)$  so  $\sum_{i=1}^n x_i = 5$  and  $n = 6$ .
- And assume, say,  $\alpha = \beta = 1$ , i.e.  $\theta \sim \text{Beta}(1, 1)$ .

# Bayesian Inference: Example 1

- We can also show that the posterior variance tends to 0 as  $n \rightarrow \infty$ :

# Bayesian Inference: Example 1

- R illustrations:

# Bayesian Inference: Example 2

Suppose the random variables  $Y_1, \dots, Y_n \stackrel{iid}{\sim} N(\theta, \sigma^2)$ , where  $\theta$  is an **unknown** parameter but  $\sigma^2$  is **known**.

- Assume observed data  $\mathbf{y} = (y_1, \dots, y_n)$  and the prior  $\theta \sim N(\mu, \tau^2)$ .
- It is easy to show (using the method before) that the posterior is:

$$\theta|\mathbf{y} \sim N\left(\frac{\sigma^2/n}{\sigma^2/n + \tau^2}\mu + \frac{\tau^2}{\sigma^2/n + \tau^2}\bar{y}, \frac{1}{n/\sigma^2 + 1/\tau^2}\right).$$

- Note that
  - The posterior mean  $E(\theta|\bar{y})$  is a weighted average of the prior mean  $\mu$  and the data mean  $\bar{y}$ , with weights depending on the **relative uncertainty**.
  - The posterior precision (reciprocal of variance) is  $n/\sigma^2 + 1/\tau^2$ , which is the sum of the precisions of likelihood and prior.



# Bayesian Inference: Example 2

As an illustration, let  $\mu = 2$ ,  $\tau^2 = 1$ ,  $\bar{y} = 6$ ,  $\sigma^2 = 1$ .

- When  $n = 1$ , prior and likelihood receive equal weight.
- When  $n = 10$ , the data begin to dominate the prior.
- When  $n \rightarrow \infty$ , the posterior variance goes to zero.

See R for figures.

# Some remarks on priors

- Setting  $\tau^2 = \infty$  corresponds to an arbitrarily vague (or “non-informative”) prior. The posterior is then

$$\theta|\mathbf{y} \sim N(\bar{y}, \sigma^2/n),$$

the same as the likelihood! The limit of the (normal) prior here is a uniform (or flat) prior, and thus, the posterior is the normalized likelihood.

- The flat prior  $\pi(\theta) \propto 1$  is **improper** since  $\int \pi(\theta)d\theta = +\infty$ .
- However, as long as the posterior is integrable (i.e.  $\int f(\mathbf{y}|\theta)\pi(\theta)d\theta < \infty$ ), as in the case here, then an improper prior can be used!
- **Caution:** Improper priors should be avoided if we are to perform Bayesian model comparison!

# Some remarks on priors

## More about non-informative priors...

- Similar could be observed in Example 1, we may set  $\alpha = \beta = 1$ , i.e.  $\theta \sim \text{Beta}(1, 1)$  as the prior distribution, which is a  $\text{Uniform}(0, 1)$  distribution (proper).
- The result of this is a posterior distribution of  $\theta | \mathbf{y} \sim \text{Beta}(\sum_{i=1}^n y_i, n - \sum_{i=1}^n y_i)$ .
- The use of non-informative priors typically leads to similar inference as the classical approach (**correspondence to classical inference**).

# Some remarks on priors

More about non-informative priors...

- Sometimes, the **Jeffrey's rule** of non-informative priors can be used:  $\pi(\theta) \propto |I(\theta)|^{\frac{1}{2}}$ , where  $I(\theta) = E[-\frac{d^2}{d\theta^2} \log f(\mathbf{y}|\theta)]$  is the Fisher information.
- Jeffrey's rule for:
  - Example 1:  $\theta \sim \text{Beta}(\frac{1}{2}, \frac{1}{2})$  (proper).
  - Example 2:  $\pi(\theta) \propto 1$  (improper).
- Regardless, it is also generally true that as  $n \rightarrow \infty$  (data accrues), the influence of prior diminishes and the posterior is **dominated** by the likelihood.

# Some remarks on priors

- So how do we choose priors? Prior robustness and sensitivity...
- **Informative** VS **non-informative**.
- This is somewhat related to 2 broadly opposing views on Bayesian probability.
  - **Objective** Bayesians: see probability as an objective measure of the plausibility, irrespective of the individual making the assessment.
  - **Subjective** Bayesians: see probability as a “personal belief”.
- Improper priors are appealing but dangerous ... may lead to improper posterior which violates the rule of total probability.
- Always there is *some* prior information.
- Prior elicitation? E.g. priors based on previous experiments (previous posteriors as current priors).

# Some remarks on priors

- The priors we have seen in both examples are **conjugate**: this leads to a posterior distribution for  $\theta$  that is a member of the same distributional family as the prior.
- To reiterate the conjugacy:  $Y|\theta \sim N(\theta, \sigma^2)$ ,  $\theta \sim N(\mu, \tau^2)$ , then marginally,  $Y \sim \text{Normal distribution}$  and the posterior,  $\theta|y \sim \text{Normal}$ .
- For variance parameters, with  $Y|\sigma^2 \sim N(\mu, \sigma^2)$  where  $\mu$  is known and  $\sigma^2$  is unknown. If we set the prior  $\sigma^2 \sim \text{Inverse Gamma}(IG)$ , then the posterior is also  $\sigma^2|y \sim \text{Inverse Gamma}$ .
- Avoid using  $IG(\epsilon, \epsilon)$  with small  $\epsilon$ . Almost improper and can lead to almost improper posteriors and various computational issues (see Gelman, 2006).
- Other conjugacies: Poisson with Gamma; Binomial with Beta.

## Section 2: Posterior Inferences [11-12.30 pm]

# Posterior Inference

- **Point estimation** - Bayes estimates, Monte Carlo estimates
- **Interval estimation** - Credible intervals
- **Density estimation** - Kernel density plots
- **Probability statements**
- **Bayesian hypothesis testing**
- **Bayesian model comparison/selection/averaging** -  
Marginal likelihoods, Bayes factors



# Posterior Inference - Point estimation

- To provide a **point summary** of the posterior distribution.
- Typically an appropriate measure of centrality, e.g. posterior mean, median, mode.
- More formally, the Bayes estimator is defined as

$$E[L(\theta, \tilde{\theta})|\mathbf{y}] = \int L(\theta, \tilde{\theta})\pi(\theta|\mathbf{y})d\theta,$$

where  $L(\theta, \tilde{\theta})$  is the **loss function**.

- Usually,  $L(\theta, \tilde{\theta}) = 0$  if  $\theta = \tilde{\theta}$  (“getting it right”) and  $L(\theta, \tilde{\theta})$  increases as  $|\theta - \tilde{\theta}|$  increases.

# Posterior Inference - Point estimation

- Different loss functions lead to different point estimates.
- Here are some common choices:

$L(\theta, \tilde{\theta})$	Graphical illustration	Point estimate
$L(\theta, \tilde{\theta}) = (\theta - \tilde{\theta})^2$		Posterior mean, $E[\theta \mathbf{y}]$
$L(\theta, \tilde{\theta}) = \begin{cases} 0, & \text{if }  \theta - \tilde{\theta}  \leq \delta \\ 1, & \text{if }  \theta - \tilde{\theta}  > \delta \text{ for some small } \delta \end{cases}$		Posterior mode, $\arg \max_{\theta} \pi(\theta \mathbf{y})$
$L(\theta, \tilde{\theta}) =  \theta - \tilde{\theta}  = \begin{cases} \theta - \tilde{\theta}, & \text{if } \theta > \tilde{\theta} \\ \tilde{\theta} - \theta, & \text{if } \theta \leq \tilde{\theta} \end{cases}$		Posterior median, $\tilde{\theta}$ such that $\int_{-\infty}^{\tilde{\theta}} \pi(\theta \mathbf{y}) d\theta = 0.5$

# Posterior Inference - Point estimation

Recall Example 1, what are the Bayes estimates?



# Posterior Inference - Point estimation

- Suppose instead we are interested in evaluating the expectation of a general function  $h(\theta)$  with respect to posterior,

$$E[h(\theta)|\mathbf{y}] = \int h(\theta)\pi(\theta|\mathbf{y})d\theta.$$

- Can use **numerical** methods.
- This can be very difficult, but  
if we can draw posterior samples,  $\{\theta^{(1)}, \dots, \theta^{(m)}\}$ ,  
we can estimate  $E[h(\theta)|\mathbf{y}]$  by

$$\bar{h} = \frac{1}{m} \sum_{i=1}^m h(\theta^{(i)}).$$

- This is known as the **Monte Carlo (MC) estimate**.

# Posterior Inference - Point estimation

- Basic idea of statistics:  
**estimates unknowns by drawing samples.**
- It can be shown by *Law of Large Numbers* that  $\bar{h} \rightarrow E[h(\theta)|\mathbf{y}]$  as  $m \rightarrow \infty$ .
- The numerical standard error (nse) of  $\bar{h}$  is

$$\text{nse}(\bar{h}) \approx \sqrt{\frac{\sigma_h^2}{m} \underbrace{\left[ 1 + 2 \sum_{i=1}^m \rho_I(h) \right]}_{=1 \text{ if independent samples}}},$$

where  $\sigma_h^2$  and  $\rho_I(h)$  are respectively the variance and lag- $I$  auto-correlation estimated from  $\{h(\theta^{(i)})\}$ .

- $\text{nse}(\bar{h}) \rightarrow 0$  as  $m \rightarrow \infty$ , **consistency** :)

# Posterior Inference - Interval estimation

- In Bayesian statistics, **credible interval** is an interval within which an unknown parameter value falls with a pre-specified probability.
- Analogous with confident interval in classical statistics, but different philosophically.
- Common types of credible intervals:

- **Equal-tailed intervals** - seek  $q_U$  and  $q_L$  such that

$$\int_{-\infty}^{q_L} \pi(\theta|\mathbf{y})d\theta = \frac{\alpha}{2} \text{ and } \int_{q_U}^{\infty} \pi(\theta|\mathbf{y})d\theta = \frac{\alpha}{2}.$$

Then  $P(q_L < \theta < q_U|\mathbf{y}) = 1 - \alpha$ , and we refer this interval  $(q_L, q_U)$  as a  $1 - \alpha$  credible set for  $\theta$ .

- **High Posterior Density (HPD) intervals**

# Posterior Inference - Interval estimation

How to construct these intervals?

- If the posterior is tractable and belongs to standard distributions (normal, beta, gamma, Laplace, Student's  $t$  etc.), we can use standard tables for quantiles or routines within statistical software. E.g. In R,  
    `qnorm, pnorm`  
    `qbeta, pbeta`  
    `qgamma, pgamma`.
- Otherwise, we can derive them from posterior samples generated.
- Note that credible intervals are not unique, there are a few other variants not discussed here.

# Posterior Inference - Interval estimation

Again, recall Example 1





# Posterior Inference - Density estimation

- Density estimation is particularly useful if we are to construct the density from posterior samples generated.
- Kernel density estimation is a popular option. Suppose posterior samples  $(\theta^{(1)}, \dots, \theta^{(m)})$ , we construct

$$\hat{\pi}_{\lambda}(\theta) = \frac{1}{m\lambda} \sum_{i=1}^m K\left(\frac{\theta - \theta_i}{\lambda}\right),$$

where  $K(\cdot)$  is a non-negative function called kernel, and  $\lambda$  is a smoothing parameter called bandwidth.

- We say  $\hat{\pi}_{\lambda}(\theta)$  is the **Kernel density estimate** of the posterior  $\pi(\theta|\mathbf{y})$ .
- We will look at how to implement some of these using R packages (e.g. using “coda”) later.

# Bayesian Hypothesis Testing

- In classical statistics, hypothesis testing relies heavily on p-value, where

$$\text{p-value} = P(\text{Observed Data or more extreme} | H_0 \text{ true}),$$

$H_0$  being the null hypothesis.

- Criticisms of p-value:
  - Conclusions are made **based on what you DON'T observe**.  
But why should an unobserved region criticize  $H_0$ ?
  - We are not allowed to write  $P(H_0 \text{ true} | \text{Observed Data})$   
because  $\theta$  (and hence  $H_0$ ) is not viewed as a random variable in classical.
  - Different conclusion can be reached from the exact same data.
- Hypothesis testing is much more natural within the Bayesian framework.
- Simply compute  $P(H_0 \text{ true} | \mathbf{y})$  using the posterior distribution or samples.

# Hierarchical modelling

- An extremely powerful modelling tool and closely related to Bayesian statistics.
- Traditionally inaccessible in classical framework but can be **naturally handled** within a Bayesian framework.
- Involves developing a **statistical model in multiple levels**, typically involving multi-dimensional parameters.
- **Easy to integrate** over multiple levels using Bayes theorem, incorporating all sources of uncertainty coherently.
- Very useful when information is available on several levels of observational units (including those with latent variables).
- Hierarchical models also aid in interpretations when problems can be broken down into individual levels to allow analysis of conditional relationships between data and parameters, rather than analysing the more complex marginal model.

# Bayesian Updating

- Simplest version: suppose  $Y_1, Y_2$  independent given  $\theta$ . So joint model is

$$f(y_2|\theta)f(y_1|\theta)\pi(\theta) \propto f(y_2|\theta)\pi(\theta|y_1).$$

- $Y_1$  updates  $\pi(\theta)$  to  $\pi(\theta|y_1)$  before  $Y_2$  arrives.
- Yesterday's posterior becomes today's prior!
- **Sequential** data collection and analysis...
- Easy to generalize to more than two updates, block updating, for dependent/independent data.

# Sampling based Bayesian inference

- In the examples we have seen so far, the posterior distributions are tractable.
- It can be very challenging when they are **analytically intractable** (more common!).
- So what do we do?
- One way may be to formulate an approximation of the posterior, e.g. Laplace approximations, Variational Bayes, INLA etc.
- Alternatively, if we can **draw samples** from the posterior distributions, we can use them **to learn about the posterior** distribution (arbitrarily well).
- This is the first principles of statistics, **use samples from the population/distribution to draw inference from it!**

# Sampling based Bayesian inference

- Importance sampling
- Sampling-Importance-Resampling (SIR)
- Rejection sampling
- Markov chain Monte Carlo (MCMC) methods

Note that independent sampling from  $\pi(\theta|\mathbf{y})$  can be difficult, so consider MCMC methods.

## Section 3: MCMC Methods

[1.30-3 pm]

# MCMC methods

Some basics of Markov chain:

- A Markov chain is generated by sampling

$$\theta^{(t+1)} \sim p(\theta|\theta^{(t)}), t = 0, 1, \dots,$$

where  $p$  is called the transition kernel.

- **Key property:** given the current and past values, the next value depends only on current and not the past,

$$p(\theta|\theta^{(t)}, \theta^{(t-1)}, \dots) = p(\theta|\theta^{(t)})$$

- As  $t \rightarrow \infty$ , the Markov chain converges to its **stationary distribution** (if it exists and is unique).
- **Irreducibility:** any set of states (values) can be reached from any other state (value) in a finite number of moves.



# MCMC methods

Some basics of Markov chain:

- **Aperiodicity:** If the greatest common divisor of return times to any particular state  $i$ , say, is 1, then the Markov chain is aperiodic.
- **Ergodicity:** If a Markov chain is aperiodic and irreducible, then it is said to be ergodic.
- What does this all take us? **An ergodic Markov chain will eventually reach a unique stationary distribution, regardless of the initial state.**

# MCMC methods

Markov chain Monte Carlo (MCMC) methods:

- One of the ingredients that popularizes Bayesian inference, along with high computational speed and power.
- The **main idea** is to construct an ergodic Markov chain with  $\pi(\theta|\mathbf{y})$  as the **stationary distribution**.
- But how? Metropolis et al (1953) showed us how, as generalized by Hastings (1970).
- Briefly, achieved by ensuring that the *detailed balanced equation* is satisfied.

# MCMC methods: Metropolis-Hastings algorithm

- **Step 1:** Start at an arbitrary initial value  $\theta^{(0)}$  and set  $t = 0$ .
- **Step 2:** Generate  $\theta^*$  from the proposal distribution,  $\theta^* \sim p(\theta|\theta^{(t)})$ .
- **Step 3:** Then with probability

$$\alpha(\theta^{(t)}, \theta^*) = \min \left\{ 1, \frac{\pi(\theta^*|\mathbf{y})}{\pi(\theta^{(t)}|\mathbf{y})} \times \frac{p(\theta^{(t)}|\theta^*)}{p(\theta^*|\theta^{(t)})} \right\}$$

set  $\theta^{(t+1)} = \theta^*$  (acceptance).

Otherwise, set  $\theta^{(t+1)} = \theta^{(t)}$  (rejection).

- **Step 4:** Set  $t = t + 1$  and go to **Step 2**, or stop when sufficient samples are obtained.

# MCMC methods: Metropolis-Hastings algorithm

Some remarks:

- The normalising constant of  $\pi(\theta|\mathbf{y})$  is not required (it cancels out in the ratio!).
- The choice of the proposal distribution  $p(\cdot)$  is arbitrary, but defines the resulting algorithm and the efficiency.
- Typically also choose  $p(\cdot)$  such that it is easy to sample from.
- Samples generated are dependent samples from the **target distribution**, here,  $\pi(\theta|\mathbf{y})$ .

# MCMC methods: Metropolis-Hastings algorithm

## 1. Independence sampler

- The proposal does not depend on current value,  $\theta^{(t)}$ :

$$p(\theta|\theta^{(t)}) = p(\theta).$$

- Only useful if:
  - $p$  is a good approximation of to the posterior;
  - Or the tail of  $p$  is heavier than the tail of the posterior for rapid convergence.
- One special case is to choose prior as the proposal,  $p(\theta) = \pi(\theta)$ .
- The acceptance probability reduces to ratio of likelihoods,

$$\alpha(\theta^{(t)}, \theta^*) = \min \left\{ 1, \frac{f(\mathbf{y}|\theta^*)}{f(\mathbf{y}|\theta^{(t)})} \right\}.$$

- Beware:** This can be either very good or very bad!
- Also check acceptance rates (the larger the better for Independence sampler).

# MCMC methods: Metropolis-Hastings algorithm

## 2. Metropolis algorithm

- If the proposal distribution is symmetric:

$$p(\theta^{(t)}|\theta^*) = p(\theta^*|\theta^{(t)})$$

this forms the **Metropolis algorithm**.

- Originally proposed by Metropolis et al. (1953).
- A special case is known as the **random-walk Metropolis algorithm**,

$$p(\theta^*|\theta^{(t)}) = p(|\theta^* - \theta^{(t)}|).$$

- More specifically,  $\theta^*|\theta^{(t)} \sim N(\theta^{(t)}, \sigma_{prop}^2)$
- Then the acceptance probability

$$\alpha(\theta^{(t)}, \theta^*) = \min \left\{ 1, \frac{\pi(\theta^*|\mathbf{y})}{\pi(\theta^{(t)}|\mathbf{y})} \right\}$$

# MCMC methods: Metropolis-Hastings algorithm

## Remarks on **random-walk Metropolis algorithm**:

- Proposal depends on where you are, the current value,  $\theta^{(t)}$ .
- Any proposal with higher posterior density than current value  $\theta^{(t)}$  is automatically accepted.
- The proposal variance  $\sigma_{prop}^2$  determines the trajectory of the posterior samples:
  - $\sigma_{prop}^2$  too large: high rejection rates, trajectory stays at current values too much  $\Rightarrow$  slow exploration.
  - $\sigma_{prop}^2$  too small: high acceptance rates, jump steps very small  $\Rightarrow$  very correlated posterior samples  $\Rightarrow$  slow exploration.
- For optimal performances, tune  $\sigma_{prop}^2$  such that the acceptance rate is between **15 – 40%** (Gelman et al., 1996).

# MCMC methods: Metropolis-Hastings algorithm

A simple example:

- Let  $Y_1, \dots, Y_n \stackrel{iid}{\sim} N(\theta, 1)$  and  $\pi(\theta) = \frac{1}{\pi(1+\theta^2)}$ .
- Posterior (target) density:

$$\pi(\theta|\mathbf{y}) \propto \exp\left\{-\frac{n(\theta - \bar{y})^2}{2}\right\} \times \frac{1}{1 + \theta^2}.$$

- Suppose also  $n = 25$ ,  $\bar{y} = 0.05$ .
- Let's try both independence sampler and random-walk Metropolis algorithms.



# MCMC methods: Metropolis-Hastings algorithm

A simple example: 1. Independence sampler

- We can sample from the prior, Cauchy distribution easily.
- Try Cauchy distribution as the proposal, i.e.  
$$p(\theta|\theta^{(t)}) = \frac{1}{\pi(1+\theta^2)}.$$
- See R.

# MCMC methods: Metropolis-Hastings algorithm

A simple example: 2. Random-walk Metropolis

- Starting with arbitrary initial value, say  $\theta^{(0)} = 5$ .
- We propose  $\theta^*$  from  $\theta|\theta^{(t)} \sim N(\theta^{(t)}, \sigma_{prop}^2)$ .
- Accept with the acceptance probability

$$\alpha(\theta^{(t)}, \theta^*) = \min \left\{ 1, \exp \left[ \frac{25(\theta^{(t)} - 0.05)^2}{2} - \frac{25(\theta^* - 0.05)^2}{2} \right] \times \frac{1 + (\theta^{(t)})^2}{1 + (\theta^*)^2} \right\}$$

- See R.
- Experiment with different values of  $\sigma_{prop}^2$  and monitor the trajectories and acceptance rates.

# MCMC methods: Gibbs algorithm

Gibbs sampling:

- Suppose that  $\theta = (\theta_1, \dots, \theta_k)$  is  $k$  dimensional.
- Gibbs sampler (Gelfand and smith, 1990) uses the **full conditional posterior distributions**: for  $j = 1, \dots, k$ ,

$$\pi(\theta_j | \theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_k, \mathbf{y}) \\ = \frac{\pi(\theta_1, \dots, \theta_{j-1}, \theta_j, \theta_{j+1}, \dots, \theta_k | \mathbf{y})}{\int \pi(\theta_1, \dots, \theta_{j-1}, \theta_j, \theta_{j+1}, \dots, \theta_k | \mathbf{y}) d\theta_j}.$$

- Idea: Explore the full conditional posterior sequentially one-by-one, joining all the samples form the intended joint posterior samples.
- Note that:
  - The full conditional posterior is proportional to the joint posterior since the denominator in the integral above does not depend on  $\theta_j$ .
  - We must be able to sample from

# MCMC methods: Gibbs algorithm

Gibbs sampling algorithm:

- Sample/update in turn:

$$\theta_1^{(t+1)} \sim \pi((\theta_1 | \theta_2^{(t)}, \theta_3^{(t)}, \dots, \theta_k^{(t)}, \mathbf{y}))$$

$$\theta_2^{(t+1)} \sim \pi((\theta_2 | \theta_1^{(t+1)}, \theta_3^{(t)}, \dots, \theta_k^{(t)}, \mathbf{y}))$$

$$\theta_3^{(t+1)} \sim \pi((\theta_3 | \theta_1^{(t+1)}, \theta_2^{(t+1)}, \dots, \theta_k^{(t)}, \mathbf{y}))$$

$$\vdots \quad \vdots \quad \vdots$$

$$\theta_k^{(t+1)} \sim \pi((\theta_k | \theta_1^{(t+1)}, \theta_2^{(t+1)}, \dots, \theta_{k-1}^{(t+1)}, \mathbf{y}))$$

- Always use the most recent values.
- The order of updating matters but often not too much (Roberts and Sahu, 1997).

# MCMC methods: Gibbs algorithm

Example:

- Suppose we have the following joint posterior density:

$$\pi(\theta, \gamma | \mathbf{y}) \propto (\gamma)^{n/2+a-1} \exp\left\{-\frac{\gamma \sum_{i=1}^n (y_i - \theta)^2}{2} - b\gamma\right\}.$$

- This is similar to Example 2, but the precision  $\gamma = \frac{1}{\sigma^2}$  is unknown with the prior  $\gamma \sim \text{Gamma}(a, b)$ , and a flat prior for  $\theta$ ,  $\pi(\theta) \propto 1$ .
- It can be shown that:

$$\gamma | \theta, \mathbf{y} \sim \text{Gamma}\left(n/2 + a, b + \frac{1}{2} \sum_{i=1}^n (y_i - \theta)^2\right)$$

$$\theta | \gamma, \mathbf{y} \sim N(\bar{y}, \frac{1}{n\gamma})$$

# MCMC methods: Gibbs algorithm

Example (cont.):

- So choose an arbitrary starting point  $\theta^{(0)}$ , then update the parameters sequentially, i.e.  
$$\theta^{(0)} \rightarrow \gamma^{(1)} \rightarrow \theta^{(1)} \rightarrow \gamma^{(2)} \rightarrow \theta^{(2)} \rightarrow \gamma^{(3)} \rightarrow \dots$$
- $\{(\gamma^{(1)}, \theta^{(1)}), (\gamma^{(2)}, \theta^{(2)}), \dots, (\gamma^{(m)}, \theta^{(m)})\}$  form  $m$  samples from the target joint posterior distribution.

# MCMC methods: Convergence Diagnostics

- **Burn-in:** First few iterations are usually discarded
  - Remove influence of initial values
  - So if discard  $b$  iterations,  $\{\theta^{b+1}, \dots, \theta^m\}$  forms a size  $m^* = m - b$  posterior sample.
  - The optimal value of  $b$  varies.
- **Trace plots:** Time series plot of the posterior samples
  - Assess if the chain is mixing properly?
  - What should be the value of  $b$ ?
  - Can detect convergence issues, e.g. change-points, algorithm getting “stuck”, chain “drifting” away etc.

# MCMC methods: Convergence Diagnostics

- **Auto-correlation plots:** Plot of auto-correlations
  - MCMC methods generate dependent samples
  - Dependency means we require larger samples to properly explore  $\pi(\theta|\mathbf{y})$ , i.e. reduced **effective sample size**
  - Want to see fast (exponential) decay
  - can consider **thinning** (collecting samples every few iterations) if too correlated
- **Density plots**
  - E.g. Kernel density estimates are useful to visualize the posterior.
  - Can indicate if we have sufficient samples.
- Number of chains many parallel runs/one very long run
- Other diagnostic tools in R package “coda”.



# MCMC methods: Convergence Diagnostics

- Length or chains and number of chains.
  - One very long chain (Geyer, 1992). But how long?
  - Several parallel chains (Gelman and Rubin, 1992)- a sense of statistical security.
- Graphical diagnostics mentioned here are generally enough.
- Other diagnostic tools available in R package “coda”.

Example:

- Gelman and Rubin (1992) convergence diagnostic, *R* (*gelman.diag*)
- Geweke (1992) normality test (*geweke.diag*)

# MCMC methods

## Strengths of MCMC:

- Flexibility in modelling
- Flexibility in inference
- Allows for sensitivity analysis
- Model comparison/criticism/choice
- Opportunities for simultaneous inference

## Weaknesses of MCMC:

- Possibility of slow convergence
- Order  $N^{-1/2}$  precision
- May require tuning/adaptations for higher efficiency

## Section 4: Practical Examples [3.30-5 pm]

# Practical examples



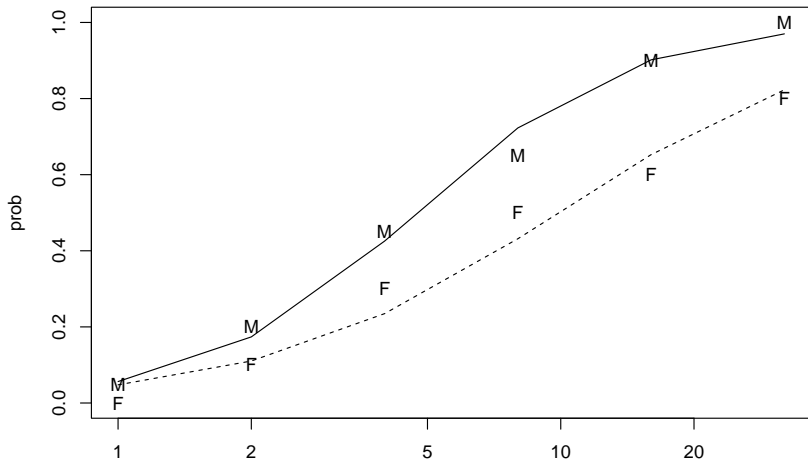
## Practical 1: Budworm data

- Collett (1991, p. 75), an experiment on toxicity to the tobacco budworm *Heliothis virescens* using doses of the pyrethroid, *trans-cypermethrin*.
- Batches of 20 moths of each sex were exposed for three days to the pyrethroid.
- The result were

Sex \ Dose	1	2	4	8	16	32
Male	1	4	9	13	18	20
Female	0	2	6	10	12	16

# Practical examples

## Practical 1: Budworm data



# Practical examples

## Practical 1: Budworm data

- Consider the logistic regression.

$$\begin{cases} Y_i \sim \text{Bernoulli}(p_i) \\ \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_M \times M + \beta_D \times D + \beta_I \times M \times D \end{cases}$$

- Response variable:*
  - $Y_i$  (1=Dead, 0=Alive): Dead/alive indicator
- Explanatory variables:*
  - $M$  (Dummy variable): Sex [1=Male, 0=Female]
  - $D$  (Continuous variable):  $\log_2$  Dose (in  $\mu\text{g}$ )
- Regression coefficients:*
  - $\beta_0$ : Intercept term
  - $\beta_M$ : Individual effect due to sex
  - $\beta_D$ : Individual effect due to dose number
  - $\beta_I$ : Interaction effect

# Practical examples

## Practical 1: Budworm data

- Method 1: Using the R package *MCMCpack*, a function called *MCMClogit*.
- This works for Bernoulli distribution but not binomial, so disaggregate the data.
- The default prior for  $\beta$  is an improper uniform prior, can try others - see *?MCMClogit*.
- Quite a black box...
- *MCMClogit* uses random-walk Metropolis MCMC scheme
- Try playing with the function arguments for sensitivity analysis

# Practical examples

## Practical 1: Budworm data

- Method 2: Using the R package *LearnBayes*, a function called *rwmetrop*.
- This uses random-walk Metropolis MCMC scheme.
- Need to supply log posterior density.
- The prior?
- The function argument *proposal* to be specified. Two things to tune here:
  - var: variance-covariance matrix of the posterior. Usually not easy to compute but can provide an estimate.
  - scale: an arbitrary constant to be adaptively set.
- Try playing with the function arguments for sensitivity analysis



# Practical examples

## Practical 1: Budworm data

- Method 3 (optional): Using the R package *rjags*.
- Need to install JAGS software (<https://sourceforge.net/projects/mcmc-jags/>).
- Also need to write a separate file to specify the model. Syntax very similar to R. We will call this file *budworm.jags*
- The function *jags.model* then compile the model and can be run using *coda.samples* to generate posterior samples.
- This method is very powerful for Bayesian analysis, referred to as **JAGS (Just Another Gibbs Sampling)**.
- More flexibility in model and prior specifications.

# Practical examples



## Practical 2: Puffin data

- A study in Albert (2009), <https://cran.r-project.org/web/packages/LearnBayes/LearnBayes.pdf>
- Measurements on breedings of the common puffin on different habits at Great Island, Newfoundland
- A data frame with 38 observations of 5 variables below:
  - 1 Nest: nesting frequency (burrows per 9 square meters)
  - 2 Grass: grass cover (percentage)
  - 3 Soil: mean soil depth (in centimeters)
  - 4 Angle: angle of slope (in degrees)
  - 5 Distance: distance from cliff edge (in meters)

# Practical examples

## Practical 2: Puffin data

- Consider first the Gaussian linear regression.

$$Y_i = \beta_0 + \beta_G \times G + \beta_S \times S + \beta_A \times A + \beta_D \times D + \epsilon_i$$

- Response variable:*
  - $Y_i$  (Nest): Continuous/discrete variable
- Explanatory variables:*
  - $G$  (Grass): Continuous variable
  - $S$  (Soil): Continuous variable
  - $A$  (Angle): Continuous variable
  - $D$  (Distance): Continuous variable
- Regression coefficients:*
  - $\beta_G, \beta_S, \beta_A, \beta_D$ : Individual effects due to Grass, Soil, Angle, Distance respectively.
- Error term:*
  - $\epsilon_i \sim N(0, \sigma^2)$ :  $\sigma^2$  is the variance.

# Practical examples

## Practical 2: Puffin data

- Now consider the Poisson regression.

$$\begin{cases} Y_i \sim \text{Poisson}(\lambda_i) \\ \log(\lambda_i) = \beta_0 + \beta_G \times G + \beta_S \times S + \beta_A \times A + \beta_D \times D \end{cases}$$

- Response variable:*
  - $Y_i$  (Nest): Discrete variable
- Explanatory variables:*
  - $G$  (Grass): Continuous variable
  - $S$  (Soil): Continuous variable
  - $A$  (Angle): Continuous variable
  - $D$  (Distance): Continuous variable
- Regression coefficients:*
  - $\beta_G, \beta_S, \beta_A, \beta_D$ : Individual effects due to Grass, Soil, Angle, Distance respectively.

# Conclusion

- We have covered basic Bayesian analysis, estimation and computation.
- Also studied how to derive posterior distributions, and how to draw inferences from them.
- Assessed the influence of prior distributions.
- The power of sampling-based inferences!
- How MCMC further enabled posterior sampling.
- Some simple implementations and tools in R.
- Two practical examples.

# What we did not manage to cover?

A lot!

- Model comparison (marginal likelihoods, Bayes factors etc.).
- Bayesian hypothesis testing.
- For maximal flexibility in implementation, consider WinBUGS/OpenBUGS, JAGS, STAN.
- BUGS language: uses Directed Acyclic Graphs to simplify model specification.
- Missing data.
- Big data.
- And more ...

# References



Collett D. (1991). Modelling Binary Data. *Springer*.



Gelfand, A. E., and Adrian F. M. S. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association* 85, no. 410: 398–409.  
<https://doi.org/10.2307/2289776>.







Gelman, A. & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statist. Sci.* 7 (4) 457 - 472.  
<https://doi.org/10.1214/ss/1177011136>



Gelman, A. & Carlin, J. B. & Stern H. S. & Dunson D. B. & Vehtari A. & Rubin D. B. (2013). Bayesian Data Analysis. *Chapman & Hall*.

# References

-  Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* 1 (3) 515 - 534.
-  Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, Volume 57, Issue 1, Pages 97–109,  
<https://doi.org/10.1093/biomet/57.1.97>
-  Metropolis, N. & Rosenbluth, A. W. & Rosenbluth, M. N. & Teller, A. H. & Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*. 21 (6), 1087-1092.
-  Geweke, J. F. (1991). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Staff Report 148, Federal Reserve Bank of Minneapolis*.



# References

-  Peck, R. & Devore, J. & Olsen, C. (2005). Introduction to Statistics And Data Analysis, *Thomson Learning*.
-  Albert, J. (2018). Package “LearnBayes”. <https://cran.r-project.org/web/packages/LearnBayes/LearnBayes.pdf>. Accessed 15 July. 2022
-  Roberts, G.O. and Sahu, S.K. (1997). Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59: 291-317.  
<https://doi.org/10.1111/1467-9868.00070>.
-  Sahu, S.K. (2022). Bayesian Modeling of Spatio-Temporal Data with R (1st ed.). *Chapman & Hall/CRC*.  
<https://doi.org/10.1201/9780429318443>.

Any questions?

Thank You!