Lecture 10: Introduction to Bayesian inference & conjugate Bayesian models

Osvaldo Anacleto
Genetics and Genomics, Roslin Institute
osvaldo.anacleto@roslin.ed.ac.uk

THE UNIVERSITY *of* EDINBURGH

BBSRC
20 Years of Pioneering
Great British Bioscience

Overview

- Key concepts in Bayesian Inference
- Bayesian conjugate models
  - beta-binomial
  - normal-normal
- Conjugate analysis for stochastic SIR models
- Bayesian and Frequentist inference: a comparison

Bayesian inference: the key ideas

In Bayesian inference, all that is known about the possible values of a parameter is represented by a probability distribution: **the prior distribution**

where does prior information come from?

- expert opinion about the likely values a parameter
- previous experiments

After data are observed, the beliefs about a parameter is updated by combining the prior information and the available data (the likelihood): the resulting distribution is called **the posterior distribution**

The posterior combines two sources of information about θ: the subjective prior beliefs about θ, and information about θ contained in the data.

Bayesian inference in a nutshell:

- **data**: $x_1, x_2, \ldots, x_n$ - **i.i.d** observations from a random variable X with probability distribution indexed by parameter θ (usually a **vector** of parameters)
- **likelihood:** $f(\textbf{data}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta})$
- **prior distribution:** initial beliefs about $\boldsymbol{\theta}$: $g(\boldsymbol{\theta})$
- **posterior distribution:** combination of initial beliefs with observed data using **Bayes theorem**

$$g(\boldsymbol{\theta}|\mathbf{x}) = k g(\boldsymbol{\theta}) f(\mathbf{x}|\boldsymbol{\theta})$$

(where k is a constant which doesn't depend on θ)

alternatively, $g(\boldsymbol{\theta}|\mathbf{x}) \propto g(\boldsymbol{\theta}) f(\mathbf{x}|\boldsymbol{\theta})$

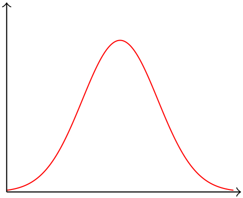
- $g(\boldsymbol{\theta}|\mathbf{x})$ and $\propto g(\boldsymbol{\theta})$ are probability distributions
- inference is done using the posterior distribution $g(\boldsymbol{\theta}|\mathbf{x})$

parameters are random variables in Bayesian inference

Posterior distributions are the key to Bayesian inference

the posterior distribution summarizes all information about parameters after data are observed

posterior distribution $g(\theta|\mathbf{x})$



θ (scalar!!!)

- a point estimate can be the mean or the mode of $g(\theta|\mathbf{x})$
- interval estimates are obtained using the quantiles of the posterior distribution

Crucial task in Bayesian inference: choice of prior

- the prior distribution should reflect the knowledge about the parameters **before data are observed**
- different priors lead to different posteriors (practical)
- priors can also reflect the lack of information about parameters: these are called **non-informative priors** and are extensively used in applications
- depending on the distribution assumed for the data, some posteriors have the same “shape” as the prior distribution - **conjugate priors**

conjugate priors

if posterior $g(\theta|\mathbf{x})$ is from the same family of distributions as the prior $g(\theta)$ - $g(\theta)$ is a conjugate prior

Why are conjugate priors useful?

- As it comes from a standard distribution, the posterior in a conjugate model is easily summarized and understood
- Since the posterior is from the same family of distributions as a conjugate prior, it is very easy evaluate the effects of the observed data on inference (practical).
- Conjugate priors can help defining priors in more complicated inference problems where conjugacy is not possible.

conjugate prior examples (I)
The beta-binomial model

example: Suppose we wish to estimate the prevalence of infected fish in a lake based on a sample of size n

- **parameter:** θ: prevalence (proportion) of infected individuals
- **data:** binary status (infected/healthy) for each fish i in the sample, $i, \dots, n$

practical question: what are the plausible values for θ based on the infection data?

Inference questions

- Is there any preliminary information about the value of θ? How to represent it in terms of probabilities?
- What's the probability model for the data? How to represent the randomness in the sample?

conjugate prior examples
The beta-binomial model

prior for θ

Since prevalence lies between 0 and 1, we can use a beta distribution to define a prior for θ

$$\theta \sim \text{beta}(a, b)$$

choice of **hyperparameters** a and b defines the prior uncertainty about the parameter θ

Probability model for the data

- Suppose X is a random variable representing the number of infected animals in a sample of size n
- X is modelled as a binomial distribution with parameters N and θ - $\theta \sim \text{bin}(N, \theta)$

$$P(X = x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x},$$

P(X=x) is the likelihood function

The beta-binomial model
posterior for the prevalence θ

- **data**: x observed number of infected fish
- **likelihood:** $P(X = x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$
- **prior distribution:** beta(a, b) (a and b must be defined!)
- **posterior distribution:** combination of initial beliefs with observed data using **Bayes theorem:**

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

it can be shown that

$$\theta|x \sim \text{beta}(a + x, b + n - x)$$

$\theta|x$ means distribution of θ **given** the data x

posterior distribution belongs to the same family of distributions as the prior - beta is a **conjugate prior** for the proportion θ

Bayesian conjugate analysis for the parameters of a normal distribution (the normal-normal model)

Example: midge wing length
(Grogan and Wirth, 1981, Hoff, 2009)

goal: learn about of mean and variance of wing length of a midge species based on a sample

- Assume that wing length follows a **normal distribution**
- the normal distribution has **two** parameters:
  - θ: represents the mean wing length of the population (the species)
  - σ^2 represents the wing length variation in the population

Multivariate distributions

- we have only considered univariate distributions so far
- multivariate distributions are required when dealing with random vectors

Examples:

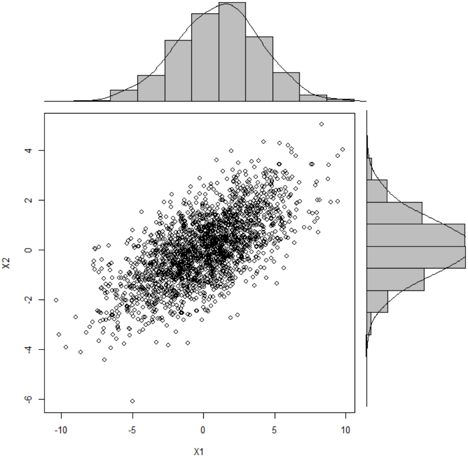
- If (X_1, X_2) is a **discrete** random vector, a bivariate distribution defines a probability for each combination of possible values of (X_1, X_2)
- If (X_1, X_2) is a **continuous** random vector, a bivariate distribution defines a probability for each combination of **ranges** of (X_1, X_2)

In this case,

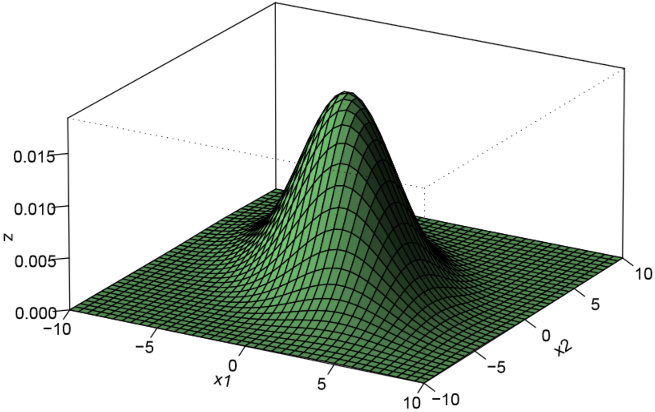
$$P[a_1 \leq X_1 \leq b_1, a_2 \leq X_2 \leq b_2] = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x_1, x_2) dx_1 dx_2$$

Example: the bivariate normal

- these data follows a bivariate normal distribution
- histograms and densities represent **marginal** distributions of X_1 and X_2
- darker regions in the scatterplot represents regions with more frequency (or density)



bivariate density function of a standard normal



$$P[a_1 \leq X_1 \leq b_1, a_2 \leq X_2 \leq b_2] = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x_1, x_2) dx_1 dx_2$$

In this case, probability represents the **volume** under the surface delimited by (a_1, b_1), and (a_2, b_2)

prior distribution for (θ, σ^2)

To define a prior bivariate distribution for (θ, σ^2), we can use the fact that

$$f(\theta, \sigma^2) = f(\theta|\sigma^2)f(\sigma^2),$$

and then set a **conditional** distribution for θ (**given** σ^2) and a **marginal** distribution for σ^2

The normal distribution is a conjugate prior for $\theta|\sigma^2$

For the example, previous studies suggest that midge wing lengths are typically around 1.9mm therefore a conjugate prior for $\theta|\sigma^2$ is

$$\theta|\sigma^2 \sim N(\theta_0 = 1.9, \sigma^2)$$

Prior distribution for σ^2

- σ^2 should be positive, so its prior should consider values on $(0, \infty)$ only.
- A **gamma distribution** is a conjugate prior for the **inverse** of σ^2: $1/\sigma^2$
$$\frac{1}{\sigma^2} \sim \text{gamma}(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0)$$
- $1/\sigma^2$ is called the **precision** of the normal distribution
- the parameters ν_0 and σ_0 represent, respectively, the sample size and sample variance of observations collected **before the sample under study** (prior observations)
- if $1/\sigma^2 \sim \text{gamma}(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0)$, then $\sigma^2 \sim \text{inverse-gamma}(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0)$

for the midge wing length example

- Studies on other population suggest that the the standard deviation of midge wing length is around 0.1 mm

- since the species of interest may be different from other midge species, the prior should be weakly centered around that value.

- This is achieved by using gamma($a = 0.5, b = 0.5 \times 0.01$) as a prior for the precision $1/\sigma^2$. In this case, $\nu_0 = 1$

The likelihood

- $X_1, X_2, ..., X_N$ are **i.i.d** random variables representing the measurements (e.g midge wing length) of a random sample of size N
- the random variable X follows a normal distribution: $X \sim N(\theta, \sigma^2)$ (sampling model)
- therefore, the likelihood is

$$L(\theta, \sigma^2) = f(x_1, \ldots, x_n | \theta, \sigma^2) = \prod_{i=1}^{n} f(x_i | \theta, \sigma^2)$$

in the midge wing length example

- N=9 (9 measurements of wing lengths in the sample)
- measurements (data): 1.64, 1.70, 1.72, 1.74, 1.82, 1.82, 1.82, 1.90, 2.08

Posterior Inference for the mean θ

- **priors:** $\theta|\sigma^2 \sim N(\theta_0, \sigma^2)$ and $\sigma^2 \sim$ inverse-gamma$(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0)$
- **sampling model :** $X_1, X_2, ..., X_N \sim i.i.d\ \ \mathrm{N}(\theta, \sigma^2)$

As done with the prior, the posterior distribution can be decomposed :

$$f(\theta, \sigma^2|x_1, x_2, ..., x_N) = f(\theta|\sigma^2, x_1, x_2, ..., x_N)f(\sigma^2|x_1, x_2, ..., x_N)$$

Using Bayes theorem, it can be shown that the posterior for θ is:

$$\theta|\sigma^2, x_1, x_2, ..., x_N \sim \mathrm{N}(\theta_n, \sigma^2/\kappa_n)$$

where

$$\kappa_n = \nu_0 + n \quad \text{and} \quad \theta_n = (\theta_0 + n\bar{x})/\kappa_n$$

Posterior Inference for the variance σ^2:

- For the posterior distribution of σ^2, we need to calculate $f(\sigma^2|x_1, x_2, ..., x_N)$ (via integration)
- Then, it can be shown that

$$\sigma^2|x_1, x_2, ..., x_N \sim \text{inverse-gamma}(\nu_n/2, \nu_n\sigma_n^2/2)$$

(see Hoff, page 75 for details about ν_n and σ_n^2)

Posterior distributions of mean and variance of wing length

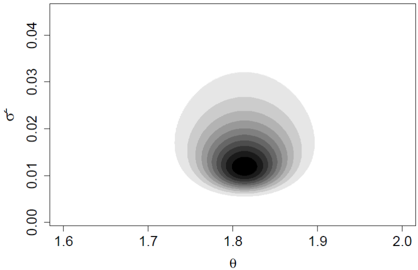
- $\theta|\sigma^2, x_1, \ldots, x_9 \sim \text{N}(1.814, \sigma^2/10)$
- $\sigma^2|x_1, \ldots x_9 \sim \text{inverse-gamma}(10/2, 10x0.015/2)$

Visualising the posterior distribution of θ, σ^2

As the parameter vector has only two dimensions, the posterior for θ, σ^2 can be visualised by

- setting a grid of possible values for θ, σ^2
- calculating $f(\theta, \sigma^2 | x_1, \ldots x_9) = f(\theta | \sigma^2 | x_1, \ldots x_9) f(\sigma^2 | x_1, \ldots x_9)$ for each point of the grid
- plotting $f(\theta, \sigma^2 | x_1, \ldots x_9)$ for the range of values of θ, σ^2 from the grid

contour plot of the posterior:



- darker regions indicate higher probabilities
- contours are more peaked as a function of θ for low values of σ^2 than high values

What if we are interested in the mean only??

- The posterior of the mean depends on the variance: $f(\theta|\sigma^2 x_1, \ldots x_9)$
- different values of σ^2 provides different posteriors for the mean θ
- the marginal distribution of θ can be obtained:
  - analitically (by integration - rarely the case in complex models)
  - by simulation (see Monte Carlo Lecture)
- for the normal-normal model it can be shown that, the marginal of θ follows a **t-distribution**
- in this case, σ^2 is called a **nuisance parameter**

Tutorial 10: Bayesian inference for the beta-binomial model (fish infection data)

Conjugate Bayesian analysis of stochastic SIR models

assumptions

- Infection and removal times are **exactly observed**
- epidemic observed until its end
- i_1 is an artificially infected animal or was infected prior to the start of observation time

data:

- infection times: $\boldsymbol{i} = (i_2, i_3, \ldots i_n)$
- removal times: $\boldsymbol{r} = (r_1, r_2, \ldots r_n)$

likelihood: $L(\boldsymbol{i}, \boldsymbol{r} | \beta, \gamma, i_1)$

Inference problems

- How to calculate the posterior distributions $f(\beta|\boldsymbol{i}, \boldsymbol{r})$ and $f(\gamma|\boldsymbol{i}, \boldsymbol{r})$?
- How to estimate R_0 ?

The gamma distribution is a conjugate prior for Bayesian inference on β and γ when assuming **complete epidemic data** under a SIR model

Conjugate Bayesian analysis of stochastic SIR models

(independent) prior distributions:

$$\beta \sim \text{gamma}(a, b) \qquad \textit{and} \qquad \gamma \sim \text{gamma}(c, d)$$

The hyperparameters a, b, c, d must be defined such that these priors encode subjective beliefs, previous information of ignorance about the parameters

- The likelihood $L(\boldsymbol{i}, \boldsymbol{r} | \beta, \gamma, i_1)$ can be split into infection and removal parts
- It can be shown that the posteriors β and γ also follow gamma distributions, with parameters as functions of hyperparameters and the data (details omitted).
- Therefore, inference for β and γ can be easily done by calculating the mean, medians and quantiles of gamma distributions (using `R`, for example)

How about inference for R_0 ?

Two alternatives for making inference about R_0 assuming complete data and Bayesian conjugate analysis for β and γ
(i) by analytically calculating the posterior distribution of R_0 based on the posteriors of β and γ (using probability theory)
(ii) by obtaining samples from the posterior of R_0 using the following algorithm:

```
do k=1, M
```

- `sample` $\beta^{(k)}$ `from` $\beta|\boldsymbol{i},\boldsymbol{r}$
- `sample` $\gamma^{(k)}$ `from` $\gamma|\boldsymbol{i},\boldsymbol{r}$
- `calculate` $R_0^{(k)} = \frac{\beta^{(k)}}{\gamma^{(k)}}$

```
end do
```

This algorithm gives a sample of size M of the posterior of R_0 based on the (gamma) posteriors of β and γ (M should be large enough to provide a small simulation error)

required ingredients for Bayesian data analysis

1. **model specification:** a probability distribution to represent the data (the sampling model)
2. **prior specification:** a probability distribution to represent someone's information about the parameter values that are likely to describe the sampling distribution
3. **posterior summary:** description of the posterior distribution by using means, medians and quantiles (for credibility intervals / regions)

the big problem: for many models, the posterior distribution is very complicated to deal with (intractable)

solution: simulation methods to approximate the posterior

Bayesian and Frequentist inference: a comparison

" The **frequentist** approach evaluates the accuracy of an estimate of an unknown value in terms of **how different that estimate could have been**. The **Bayesian** approach **updates personal beliefs about the unknown true value**. "

David Hand, Dennis Lindley's Obituary, The Guardian (16/Mar/2014)

Frequentist inference

- parameters are fixed
- inference interpretation depends on the idea of repeatable experiments
- **can be heavily dependent on sample size**

Bayesian Inference

- parameters are random variables
- beliefs about parameters are updated in the light of available data
- **complex models may require complex simulation methods**

References

- Wasserman, L. (2013). All of statistics: a concise course in statistical inference. Springer Science & Business Media.
- Hoff, P. D. (2009). A first course in Bayesian statistical methods. Springer Science & Business Media.
- Berry, D. A. (1996). Statistics: A Bayesian Perspective. Duxbury Press.