[1]A brief introduction to Bayesian inference

1 The basics

The basic idea of Bayesian inference is to setup a full probability model for both observed and unobserved quantities. Inference is then based on the so-called posterior density — that is the conditional density of the unobserved quantity conditional on the observed quantity.

Let y denote the observed quantity (the data) which we assume is a realisation of a random variable Y. Assume futher that the dsitribution of Y depends on an unobserved quantity θ which we assume is a realisation of another random variable Θ. More precisely we assume that Θ is distributed according to the so-called *prior density* $\pi(\theta)$. Given $\Theta = \theta$ we assume that Y is distributed according to the so-called *sampling/data density* $\pi(y|\theta)$ — sometimes also referred to as the *likelihood*. By the definition of conditional densities, these assumption imply that the joint distribution of Y and Θ has density

$$\pi(y, \theta) = \pi(\theta)\pi(y|\theta).$$

The prior density should reflect our prior knowledge (or our prior uncertainty) regarding Θ, i.e. our knowledge about Θ *before* we observe Y. The data density should be chosen so that it is consistent with our knowledge about the problem of interest.

From the definition of conditional densities we obtain the posterior density of Θ:

$$\pi(\theta|y) = \frac{\pi(y, \theta)}{\pi(y)} = \frac{\pi(\theta)\pi(y|\theta)}{\pi(y)}. \tag{1}$$

Notice that given the data $Y = y$ the term $\pi(y)$ is a constant and hence

$$\pi(\theta|y) \propto \pi(\theta)\pi(y|\theta)$$

is an unnormalised posterior density. The posterior density can be interpreted as our updated knowledge about Θ after having observed Y. Inference is typically based on reproducing all or parts of the posterior density graphically (as graphs or contour plots). Another option is to report e.g. posterior mean, mode, and quantiles. Notice that a *central 95% postrior interval* (e.g. the interval between the 2.5% and 97.5% quantiles) can directly be interpreted as containing θ with high probability unlike the classical confidence intervals.

[1]This manuscript is a mild revision and combination of two documents written by prof. Jesper Møller, Dept. for Mathematical Sciences, Aalborg University

It is however not always trivial to obtain the posterior density — or even an approximation of it.

Classical Bayesian inference has been limited by the fact that to make a posterior analysis feasible the prior should be chosen so that the resulting posterior density can be recognised as the density of a known distribution. Such prior distribution are called conjugated priors. This limitation has been drastically reduced in last 15 years by a combination of Markov chain Monte Carlo (MCMC) methods and an increase in available computing power.

2 Examples of Bayesian inference

2.1 Binomial likelihood

Assume that we have perform n independent experiments where each experiment has probability p for success. Here p plays the role of the unkonwn parameter θ in (1). Let $x \in \{0, 1, \ldots, n\}$ denote the random number of successes. The number of successes follows a Binomial distribution

$$x \sim B(n, p)$$

that is

$$\pi(x|n, p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

For a Bayesian analysis we now need to specify our prior distribution. It turns out to be convenient to specify the prior distribution for p by a beta distribution with parameter α and β, ie.

$$\pi \sim Be(\alpha, \beta)$$

so the distribution of p is specified by the probability density function (pdf):

$$\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \text{ for } p \in [0, 1].$$

Recall that $p \sim Be(\alpha, \beta)$ implies $E(p) = \frac{\alpha}{(\alpha+\beta)}$ and $Var(x) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

We obtain the posterior distribution as

$$\begin{aligned}
\pi(p|x) &\propto \pi(x|p)\pi(p) \\
&= \binom{n}{x} p^x (1-p)^{n-x} \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} p^{\alpha-1}(1-p)^{\beta-1} \\
&\propto p^{x+\alpha-1}(1-p)^{n-x+\beta-1}
\end{aligned}$$

which we recognise as the unnormalised density of a beta distribution with parameters $x + \alpha$ and $n - x + \beta$, ie.

$$p|x \sim Be(x + \alpha, n - x + \beta).$$

Consequently, the posterior mean and variances are $E(p|x) = \frac{x+\alpha}{n+\alpha+\beta}$ and $Var(p|x) = \frac{(x+\alpha)(n-x+\beta)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}$.

If we use $\alpha = \beta = 1$ we have a flat prior.

2.2 Binomial likelihood: Placenta Previa data

In this applied example we consider the probability for a female birth given a special condition called placenta previa. The number of female births ($x = 437$) is the observed quantity, and the probability of a female birth is the unobserved quantity p. This leads us to assume that the number of observed female births (given p) is binomially distributed with parameter p, where we assume that the total number of births $n = 980$ is known.

As in Section 2.1 we assume a Beta distribution as the prior for p. Following calculations as in Section 2.1 we obtain a posterior for p which corresponds to a beta distribution with parameters $437 + \alpha$ and $543 + \beta$. Regarding the choice of α and β. We are told the probability of a female birth in the background population is 0.485. One option would be to select α and β so that the prior mean is 0.485, ie. that $\alpha/(\alpha + \beta) = 0.485$.

In Bayesian statistics it is good practice to perform a so-called *sensitivity analysis* to assess how sensitive the posterior distribution is to the choice of prior. Table 1 contains the 2.5%, 50% and 97.5% quantiles for the posterior distribution for a range of α and β values reparameterised as $\alpha/(\alpha + \beta)$ (the prior mean) and $\alpha + \beta$. Further, Figure 1 shows the prior and posterior densities of p for the same values of α and β. Table 1 shows that except for the last row the prior has little influence on the posterior distribution. Note that the prior and posterior densities are quite different, again except the last case, and even here the 95% posterior interval does not contain the prior mean.

2.3 Normal likelihood

Usually a normal distribution is specified by its mean, μ, and variance, σ^2. Working with a normal distribution in a Bayesian setting it is in general easier to more convenient to specify a normal distribution in terms of its mean, μ, and its precision, $\tau = 1/\sigma^2$. A normal distributed random variable

		Quantiles		
$\frac{\alpha}{\alpha+\beta}$	$\alpha+\beta$	2.5%	50%	97.5%
0.5	2	0.415	0.446	0.477
0.485	5	0.415	0.446	0.477
0.485	10	0.415	0.446	0.477
0.485	20	0.416	0.447	0.478
0.485	100	0.420	0.450	0.479
0.485	200	0.424	0.453	0.481

Table 1: Prior parameters and corresponding posterior quantiles.

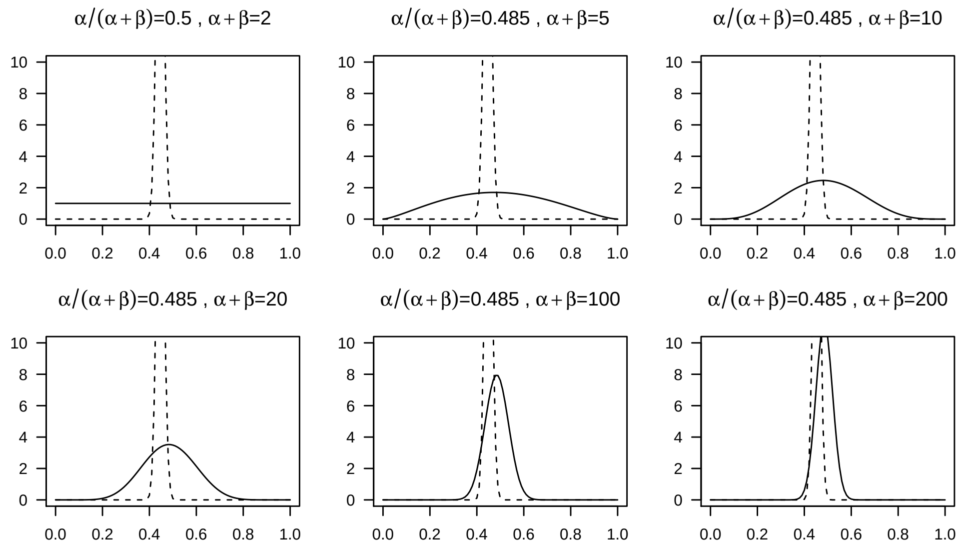

Figure 1: Prior (solid line) and posterior (dashed line) densities for p.

x with mean μ and precision τ has pdf

$$\pi(x|\mu,\tau) = \sqrt{\frac{\tau}{2\pi}} \exp\Big(-\frac{1}{2}\tau(x-\mu)^2\Big)$$
$$\propto \exp\Big(-\frac{1}{2}\tau x^2 + \tau\mu x\Big), \tag{2}$$

which we denote

$$x \sim N(\mu,\tau).$$

Assume we have a single observation $x \sim N(\mu,\tau)$ with known precision but unknown mean. Regarding the unknown mean we assume a priori that the mean is normal, specifically $\mu \sim N(\mu_0,\tau_0)$. Then the posterior distribution of μ is given by

$$\begin{aligned}\pi(\mu|x) &\propto \pi(x|\mu)\pi(\mu)\\ &= \sqrt{\frac{\tau}{2\pi}} \exp\Big(-\frac{1}{2}\tau(x-\mu)^2\Big)\sqrt{\frac{\tau_0}{2\pi}} \exp\Big(-\frac{1}{2}\tau_0(\mu-\mu_0)^2\Big)\\ &\propto \exp\Big(-\frac{1}{2}(\tau+\tau_0)\mu^2 + (\tau x + \tau_0\mu_0)\mu\Big).\end{aligned} \tag{3}$$

Comparing (3) to (2) we see that this implies $\mu|x \sim N(\mu_1,\tau_1)$, where

$$\tau_1 = \tau + \tau_0 \quad \text{and} \quad \mu_1 = \frac{1}{\tau_1}(\tau x + \tau_0\mu_0) = \frac{\tau x + \tau_0\mu_0}{\tau+\tau_0},$$

where we notice that the posterior mean, μ_1, is a weighted average of x and μ_0 with weights $\tau/(\tau+\tau_0)$ and $\tau_0/(\tau+\tau_0)$, respectively.

If we have n independent observations $x_1,\dots,x_n$, where $x_i \sim N(\mu,\tau)$. Again assuming that τ is known and a priori $\mu \sim N(\mu_0,\tau_0)$, it follows that

$$\mu|x_1,\dots,x_n \sim N(\mu_1,\tau_1), \tag{4}$$

where

$$\tau_1 = n\tau + \tau_0 \quad \text{and} \quad \mu_1 = \frac{1}{\tau_1}(\tau\sum_{i=1}^{n} x_i + \tau_0\mu_0) = \frac{n\tau\bar{x} + \tau_0\mu_0}{n\tau+\tau_0},$$

where $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$ is the average of the n observations. Again we see that the posterior mean is a weighted average of the observed mean and the prior mean.

We now consider the situation where the mean is known and the precision is unknown. Assume that the precision is gamma distributed with shape parameter α and scale parameter β, ie.

$$\pi(\tau|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}}\tau^{\alpha-1}e^{-\tau/\beta},$$

which we denote $\tau \sim \text{Gamma}(\alpha, \beta)$. Recall that $E(\tau) = \alpha\beta$ and $Var(\tau) = \alpha\beta^2$. It can then be shown that the posterior distribution of τ is also gamma distributed:

$$\tau|x_1, \ldots, x_n \sim \text{Gamma}\Big(n/2 + \alpha, \big(\frac{1}{2}\sum_i (x_i - \mu)^2 + 1/\beta\big)^{-1}\Big). \qquad (5)$$

The posterior mean is now

$$E(\tau|x_1, \ldots, x_n) = \frac{\frac{n}{2} + \alpha}{\frac{1}{2}\sum_i (x_i - \mu)^2 + 1/\beta} = \frac{\frac{n}{2} + \alpha\beta/\beta}{\frac{n}{2}\hat{\sigma}^2 + 1/\beta},$$

where $\hat{\sigma}^2 = \frac{1}{n}\sum (x_i - \mu)^2$ is the observed variance.

2.4 Conjugated priors

In the examples above the prior and posterior are distributions of the same type. For a given likelihood $l(\theta|x)$ we say that Π is a conjugated family if the posterior belongs to Π whenever the prior does. This definition is too broad in general — in practise we are only interested in conjugated families which consist of well known distributions.

2.5 Semi-conjugate priors

If both mean and variance is unknown we ideally want a joint conjugate prior distribution. One alternative is to use a semi-conjugate prior: We assume a priori that μ and τ are independent and $\mu \sim N(\mu_0, \tau_0)$ and $\tau \sim Gamma(\alpha, \beta)$. It should be clear that the conditional posterior distribution are of a known form, specifically $\mu|\tau, x_1, \ldots, x_n$ $N(\mu_1, \tau_1)$ with μ_1 and τ_1 as above and $\tau|\mu, x_1, \ldots, x_n$ is Gamma distributed as above. It is now straight forward to sample the joint posterior (asymptotically) using a Gibbs sampler.

2.6 Improper priors

Assume we want to perform Bayesian inference for observations from the observation model $\pi(x|\theta)$ specified by a real valued parameter θ. In case we have little or no prior information about the parameter θ we might be tempted to use a flat prior $\pi(\theta) \propto k$. This is an example of an improper prior because $\int_{-\infty}^{\infty} \pi(\theta)\mathrm{d}\theta = \infty$.

3 Exercises

1. Show that posterior distributions (4) and (5) are correct.

2. Suppose your prior beliefs about the probability p of success have mean 1/3 and variance 1/32. What is the posterior distribution after having observed 8 successes in 20 trail?

3. Consider again the posterior distribution in the binomial example. Assume that the prior knowledge comes from previous experience with the same experiment. In this light how could you interpret α and β?

4. A random variable x is said to be Poisson distributed with rate $\lambda > 0$ if it has probability function

$$\pi(x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & \text{if } x \in \mathbf{N}_0 \\ 0 & \text{otherwise,} \end{cases}$$

   where $\mathbf{N}_0 = \{0, 1, 2, 3, \ldots\}$ is the non-negative integers. This is denoted $x \sim Pois(\lambda)$. If $x \sim Pois(\lambda)$ then $E(x) = \lambda$ and $Var(x) = \lambda$.

   (a) Assume a priori that λ follows a gamma distribution with parameters $\alpha > 0$ and $\beta > 0$. Determine the posterior distribution of λ

   (b) In an early draft for a new book on Bayesian statistics, the number of misprints on the first six pages were

$$3, \quad 4, \quad 2, \quad 1, \quad 2, \quad 3.$$

   Assume that these observations are independent and come from a Poisson distribution with rate λ. Based on experience with drafts for other books we want a Gamma prior on λ with mean 3 and variance 4. Find the posterior distribution for λ.

5. Assume we observe data x, which we assume comes from a normal distribution with unknown mean μ and known precision τ. From previous experience a suitable prior has density

$$\pi(\mu) = \frac{1}{3} \times \sqrt{\frac{1}{2\pi}} \exp\left(-\frac{1}{2}\mu^2\right) + \frac{2}{3} \times \sqrt{\frac{1}{2\pi}} \exp\left(-\frac{1}{2}(\mu - 1)^2\right)$$

   (a) What is your interpretation of this prior? Is it even a normalised density?

   (b) Find the posterior distribution of μ.

4 Litterature

There litterature avilable on Bayesian inference is vast and growing fast. Below is list of text that are concerned with introductory Bayesian inference. The list is (obviously) not exhaustive.

- Andrew Gelman, et al. (2014). *Bayesian Data Analysis*, 3rd ed. CRC Press.
- Peter D. Hoff (2009). *A First Course in Bayesian Statistical Methods.* Springer. e
- Peter M. Lee (2004). *Bayesian Statistics: an introduction*, 3rd ed. Arnold.
- Jean-Michel Marin and Christian Robert (2007) *Bayesian Core: A Practical Approach to Computational Bayesian Statistics.* Springer.
- Jean-Michel Marin and Christian Robert (2014) *Bayesian Essentials with R.* Springer.
- Ioannis Ntzoufras (2009). *Bayesian Modeling Using WinBUGS.* Wiley.