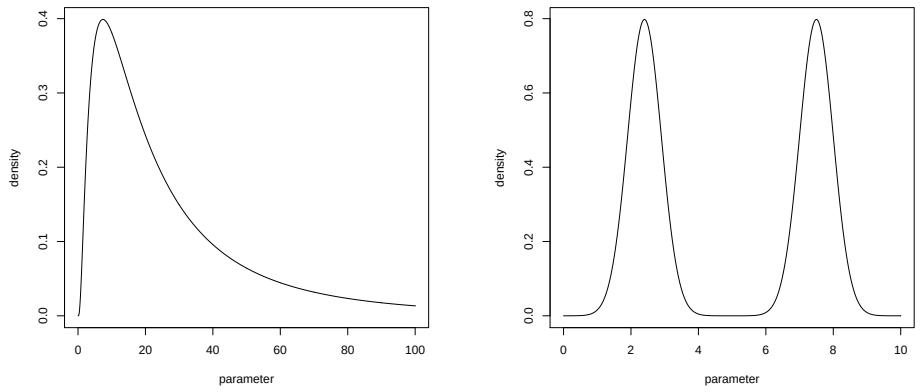
Bayesian Inference: Tutorial 3

1. (From 2009 exam - number in brackets correspond to number of marks - the total exam is out of 50 marks.) Let $X_1, \ldots, X_n$ denote independent random variables that depend on two parameters θ_1 and θ_2. Suppose that the posterior marginal distributions for the two parameters θ_1 and θ_2 are given in Figures 1(a) and (b) respectively.

Figure 1: Marginal posterior densities of parameters θ_1 and θ_2



(a) θ_1 (b) θ_2

Comment on the suitability of summarising these posterior distributions using the following summary statistics:

(a) posterior mean;

(b) posterior median;

(c) 95% symmetric credible interval;

(d) 95% highest posterior density interval.

[6]

Suggest another summary statistic that could be used in the summarising of the joint posterior distribution $\pi(\theta_1, \theta_2|\boldsymbol{x})$, where $\boldsymbol{x}$ denotes the observed data, to provide information on the posterior relationship between θ_1 and θ_2. [1]

2. (From December 2014 exam - number in brackets correspond to number of marks - the total exam is out of 50 marks.) A chemist is interested in the maximum possible yield produced by a certain chemical process. Due to the large variability in the data, he assumes that, given a scalar θ, each yield x_i, $i = 1, ..., n$, is independent of the other yields and follows a uniform distribution $U(0, \theta)$, so that,

$$f(x_i|\theta) = \frac{1}{\theta}, \qquad 0 < x_i < \theta.$$

Before the chemist sees any data, he assumes a Pareto prior distribution for θ, so that,

$$p(\theta) = \begin{cases} \frac{ax_0^a}{\theta^{a+1}} & \theta \geq x_0 \\ 0 & \text{otherwise,} \end{cases}$$

where $a > 0$ and $x_0 > 0$ are known parameters for the prior Pareto distribution. The mean of a Pareto distribution is given by, $\frac{ax_0}{a-1}$, for $a > 1$, whilst the median of a Pareto distribution is given by, $x_0 \times 2^{1/a}$.

a) Calculate the posterior distribution for θ. If the posterior is of standard form, state the distributional form (including the associated parameters of the distribution). [4]

b) Consider a Pareto prior distribution with parameters, $a = 2, x_0 = 0.1$. Consider also observed data $\boldsymbol{x} = \{x_1, x_2, x_3\} = \{3, 10, 17\}$. Obtain the posterior distribution and indicate how the expert's beliefs have changed after observing the data, using point summaries. Briefly discuss your findings. [2]

c) Consider an alternative Uniform prior $p(\theta) = U(0, 10)$. Without performing any calculations, discuss the scenario where $p(\theta) = U(0, 10)$, and the observed data are $\boldsymbol{x} = \{3, 9, 12\}$. [3]

3. We observe data $\boldsymbol{x} = \{x_1, \ldots, x_m\}$, from a Multinomial distribution, such that,

$$\boldsymbol{X} \sim MN(N, \boldsymbol{p}),$$

we wish to make inference on the parameters $\boldsymbol{p} = \{p_1, \ldots, p_m\}$. The prior on the unknown parameters $\boldsymbol{p}$, is specified to be of the form,

$$\boldsymbol{p} \sim Dir(\alpha_1, \ldots, \alpha_m).$$

What is the corresponding posterior distribution for the parameters $\boldsymbol{p}$? What is the posterior mean of p_i, $i = 1, \ldots, m$?
Note that a list of common probability distributions is provided in appendix A of the lecture notes

4. (From May 2009 exam - number in brackets correspond to number of marks - the total was 50 marks.) Let $X_1, \ldots, X_n$ be independent and identically distributed $N(\mu, \sigma^2)$ random variables, where σ^2 is known. We specify the prior,

$$\mu \sim TN(0, \tau^2),$$

with probability density function,

$$p(\mu) = \frac{\sqrt{2}}{\sqrt{\pi\tau^2}} \exp\left(-\frac{\mu^2}{2\tau^2}\right) \qquad 0 \leq \mu < \infty.$$

(a) Calculate the posterior distribution of μ, given observed data $x_1, \ldots, x_n$. [5]

(b) For a given dataset, a practitioner obtains a posterior mean for μ of 3.5, posterior median of 2.96 and 95% highest posterior density interval of $(-1.5, 8.5)$. Without looking at their observed data, describe how the statistician automatically realises that there has been at least one mistake made within the analysis. [2]

(c) Discuss any disadvantages of specifying a $TN(0, \tau^2)$ prior on μ. [2]

Note that the following distributional information was also provided. A random variable X has a (positive) **Truncated-Normal distribution** $TN(\mu, \sigma^2)$, if its probability density function (p.d.f.) is

$$\frac{1}{\left(1 - \Phi\left(-\frac{\mu}{\sigma}\right)\right)\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad 0 \leq x < \infty,$$

where Φ denotes the cumulative distribution function of the standard Normal distribution. This result follows from,

$$\frac{N(\mu,\sigma^2)}{\int_0^\infty N(\mu,\sigma^2)dx} = \frac{N(\mu,\sigma^2)}{P(X>0|\mu,\sigma)} = \frac{N(\mu,\sigma^2)}{P(\frac{X-\mu}{\sigma} > \frac{-\mu}{\sigma}|\mu,\sigma)} = \frac{N(\mu,\sigma^2)}{1-P(Z<\frac{-\mu}{\sigma})},$$

where $Z \sim N(0,1)$.

5. (To discuss in groups if possible) You are presented with the following prior beliefs regarding an unknown parameter of interest $\theta \in \mathcal{R}$:

   (a) Prior information: average of 50 and bounds of [20,80]

   (b) Prior information: average of 50 and bounds of [25,100]

   Use this prior information to propose a sensible prior distribution on θ in each scenario.

Bayesian Inference

Tutorial 3: Solutions

1. For the density of θ_1, the posterior median is usually the chosen point summary, as the distribution is skewed. Any of the two intervals could be used, with little difference in practice (HPDI will be slightly narrower that the symmetric interval). For the density of θ_2, neither point estimate would be helpful as a summary of our beliefs. Conditional medians or modes (conditional, for example on θ_2 being less than or more than 5) would be more useful. For an interval estimator, the HPDI is a much better choice, as it will show that the density is bimodal.

   A higher posterior density credible region (the generalization of the HPDI in 2 dimensions or more) could be used as a summary statistic for the joint posterior distribution. The correlation between θ_1 and θ_2 is another choice of summary statistic.

2. a) When a sample $\boldsymbol{x}$ of size n is observed,

$$f(\boldsymbol{x}|\theta) = \begin{cases} \frac{1}{\theta^n} & \theta \geq \max\{x_1, ..., x_n\} \\ 0 & \text{otherwise.} \end{cases}$$

   The posterior distribution is,

$$\pi(\theta|\boldsymbol{x}) \propto \frac{1}{\theta^n} \frac{a x_0^a}{\theta^{a+1}}, \qquad \theta \geq x_0, \theta \geq \max\{x_1, ..., x_n\}$$

$$\propto \frac{1}{\theta^{n+a+1}}, \qquad \theta \geq \max\{x_0, x_1, ..., x_n\},$$

   which is recognized as a Pareto distribution with parameters $a^{'} = n + a$ and $x_0^{'} = \max\{x_0, x_1, ..., x_n\}$.

   b) For $a = 2, x_0 = 0.1$, and observed data $\boldsymbol{x} = \{x_1, x_2, x_3\} = \{3, 10, 17\}$, the posterior distribution is Pareto with parameters $a^{'} = n + a = 3 + 2 = 5$ and $x_0^{'} = \max\{0.1, 3, 10, 17\} = 17$. The prior mean for θ is $2 \times 0.1/1 = 0.2$ whilst the prior median is $0.1 \times \sqrt{2} = 0.14$. The posterior mean and median are $5 \times 17/4 = 21.25$ and $17 \times 2^{1/5} = 19.52$ respectively. We observe a difference between the mean and median. This indicates a skewed distribution, so we may choose to report prior and posterior medians. In any case, we observe that prior beliefs, at least as described by point estimates, have changed dramatically after observing a small number of observations. This is due to the nature of this problem, where the maximum observation has a drastic effect on the posterior distribution.

   c) Observation $x_3 = 12$ implies that θ is larger than 12, however the prior distribution assigns zero probability to this outcome. There is a direct conflict between the prior and the likelihood that cannot be resolved. In practice, no prior to posterior analysis can take place that will provide with a sensible posterior. The expert should rethink their prior, in an informal manner, given the observations. Then, discard these observations and collect new observations that will be used to formally derive the posterior distribution.

3. When $\boldsymbol{p} \sim Dir(\alpha_1, \ldots, \alpha_m)$,

$$p(\boldsymbol{p}) = \frac{1}{D(\boldsymbol{p})} \prod_{i=1}^{m} p_i^{\alpha_i - 1},$$

with,

$$D(\boldsymbol{p}) = \frac{\prod_{i=1}^{m} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{m} \alpha_i)}.$$

Also, $f(\boldsymbol{x}|\boldsymbol{p}) = \prod_{i=1}^{m} p_i^{x_i}$, $\sum_{i=1}^{m} p_i = 1$, for x_i observations in category i. Using Bayes' Theorem, the posterior distribution is,

$$\begin{aligned} \pi(\boldsymbol{p}|\boldsymbol{x}) &\propto f(\boldsymbol{x}|\boldsymbol{p}) p(\boldsymbol{p}) \\ &\propto \prod_{i=1}^{m} p_i^{x_i} \times \prod_{i=1}^{m} p_i^{\alpha_i - 1} \\ &= \prod_{i=1}^{m} p_i^{x_i + \alpha_i - 1}. \end{aligned}$$

Thus, we have,

$$\boldsymbol{p}|\boldsymbol{x} \sim Dir\left(\alpha_1 + x_1, \ldots, \alpha_m + x_m\right).$$

The posterior mean for $\boldsymbol{p}$ is,

$$\mathbb{E}_{\pi}(p_i) = \frac{\alpha_i + x_i}{\sum_{j=1}^{m} (\alpha_j + x_j)},$$

using the standard result for expectation for a Dirichlet distribution.

4. (a) The posterior distribution, where $\mu \geq 0$, is given by,

$$\begin{aligned} \pi(\mu|\boldsymbol{x}) &\propto f(\boldsymbol{x}|\mu) p(\mu) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \frac{\sqrt{2}}{\sqrt{\pi\tau^2}} \exp\left(-\frac{\mu^2}{2\tau^2}\right) \\ &\propto \exp\left(-\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^2}\right) \exp\left(-\frac{\mu^2}{2\tau^2}\right) \\ &\propto \exp\left(-\frac{(n\mu^2 - 2n\bar{x}\mu)}{2\sigma^2}\right) \exp\left(-\frac{\mu^2}{2\tau^2}\right) \\ &= \exp\left(-\frac{\mu^2(\tau^2 n + \sigma^2) - 2\mu\tau^2 n\bar{x}}{2\sigma^2\tau^2}\right). \end{aligned}$$

Then, by completing the square,

$$\pi(\mu|\boldsymbol{x}) \propto \exp\left(-\frac{\tau^2 n + \sigma^2}{2\sigma^2\tau^2}[\mu - \frac{n\tau^2\bar{x}}{n\tau^2 + \sigma^2}]^2\right)$$

and we can identify the posterior distribution for μ to be,

$$\mu|\boldsymbol{x} \sim TN\left(\frac{\tau^2 n\bar{x}}{\tau^2 n + \sigma^2}, \frac{\sigma^2\tau^2}{\tau^2 n + \sigma^2}\right).$$

for $\mu \geq 0$.

(b) The parameter $\mu \geq 0$ (as specified by the prior). Thus, the 95% HPDI cannot have a lower bound that is negative, so there is at least one error in calculating this CI.

(c) Specifying a $TN\left(0, \sigma^2\right)$ prior necessarily restricts the posterior distribution for μ to be defined on the non-negative values. In other words $\mathbb{P}_p(\mu < 0) \Rightarrow \mathbb{P}_\pi(\mu < 0 \mid \boldsymbol{x})$, irrespective of the information contained within the data. Unless it is not possible (e.g. for structural reasons) that μ cannot be negative, specifying a truncated normal prior is somewhat restrictive. It would be better to specify, for example, a prior probability denoted q_1 that $\mu \geq 0$, so that with probability $1 - q_1, \mu < 0$. We then specify a prior distribution on μ given that $\mu \geq 0$, denoted $p_1(\mu)$ (e.g. positive truncated normal) and a prior given that $\mu < 0$, denoted $p_2(\mu)$ (e.g. negative truncated normal). The prior for μ is then given by,

$$p(\mu) = q_1 p_1(\mu) + (1 - q_1)\, p_2(\mu).$$

Assuming that there is prior information that μ is positive we could specify $q_1 = 0.95$. This is strong prior information that μ is positive but does not preclude the possibility that μ may be negative given the observed data.

5. (a) The prior information can be represented via a symmetric distribution. For example, consider a Gaussian distribution with mean $\mu = 50$. To compute the variance we should consider [20,80] to be a 95% interval. Using the standard properties of the normal distribution we know that a 95% interval for a $N(\mu, \sigma^2)$ is given by:

$$\mu \pm 1.96\sigma.$$

This gives us $\sigma^2 = 15.3^2$

(b) To represent this prior information we need to consider a skewed distribution. For example, we could consider a Gamma distribution. We note that if $\theta \sim \Gamma(\alpha, \beta)$, then $\mathbb{E}(\theta) = \frac{\alpha}{\beta}$. Thus we could set $\frac{\alpha}{\beta} = 50$ or equivalently $\alpha = 50\beta$. This means that we have one more "free" parameter in which to satisfy the prior bound specification. However, we are not able to find any such values of α and β under the restriction that $\alpha = 50\beta$ such that the lower and upper 2.5% quantiles are exactly $[25, 100]$. For example, using trial and error, we may obtain $\alpha = 8$ and $\beta = 0.16$ such that the 95% symmetric credible interval $[21.6, 90.1]$. These quantiles can be obtained in R using:
$>$ qgamma $(c(0.025, 0.5, 0.975), 8, 0.16)$
To get closer to the specified prior information we need to consider a different distribution. For example, we could consider the Inverse Gamma distribution. If $\theta \sim \Gamma^{-1}(\alpha, \beta)$ then $\mathbb{E}(\theta) = \frac{\beta}{(\alpha-1)}$. Thus we set $\beta = 50(\alpha - 1)$. Again, using trial and error we could propose $\alpha = 8$ and $\beta = 350$, giving a symmetric 95% credible interval of $[24.3, 101.3]$. This is again close but does not exactly represent the prior information.
Finally, we could consider a transformation on θ. Consider the log of the parameter, which we denote by $\phi = \log\theta$. This would give an average value of $\log 50 = 3.91$ and bounds of $[3.22, 4.61]$. We can note that the bounds are symmetric about the average value. Thus we could consider a Normal distribution on ϕ. In particular (using the analogous argument as above), we could use $\phi \sim N\left(3.91, 0.35^2\right)$ - or equivalently that $\theta \sim \log N\left(3.91, 0.35^2\right)$. The corresponding median of θ is 50 with 95% symmetric credible interval $[25, 100]$ (note that the mean is 53).