MTH 511a - 2020: Lecture 26

Instructor: Dootika Vats

1 Bayesian models

So far, in data analysis techniques, we have assumed that the data comes from some distribution. This distribution dictates the likelihood, and then having observed the data, the likelihood is maximized.

However, sometimes, we may already have some information about a parameter in the data distribution. We would like to take into account that information. Some examples:

- The mean of the batting average of cricket openers in ODIs. We know that opening batters are likely to have batting averages between 30 - 50.
- The mean GRE quantitative score of MTH students. It is likely that MTH students will have high GRE quant scores, so we should be able to account for this information.
- You want to estimate how much money Akshay Kumar's next movie will make. His latest few movies have been big successes and big failures and everything in between. You have information that his movie can be highly variable in its revenue, and want to take account for that information.

1.1 Prior distribution

Suppose $X_1, \dots, X_n \sim F_\theta$ is data from a distribution where θ is a defining parameter. So far in likelihood-based estimation techniques, we've obtained

$$L(\theta|X_1, \dots, X_n) = \prod_{i=1}^{n} f(X_i \mid \theta)\,,$$

and then we maximized the likelihood. Thus, this likelihood becomes the key quantity of interest.

When we have some information on the parameter θ, we want to be able to include that information into our process. We will assume that θ, the parameter of interest, is random, in that, it has a distribution. The distribution assigned to θ is called the *prior distribution* and is meant to encapsulate the *prior* information about the parameter:

$$\theta \sim \pi(\theta)\,,$$

where $\pi(\theta)$ denotes the density of the prior distribution. For example:

- **Cricket:** $\theta \sim N(40, 20)$
- **GRE Quant:** $\theta \sim N(167, 1)$
- **Akshay Kumar:** $\theta \sim \text{Gamma}(40 \text{ lacs}, 1)$.

1.2 Posterior distribution

Having observed the data $X_1, \dots, X_n \sim F_\theta$, we now have more evidence that helps us understand θ. That is, we can update the prior information using the observed data. As in, we are interested in the distribution of θ *given* $X_1, \dots, X_n$.

$$\pi(\theta|X_1, \dots, X_n)\,.$$

This is called the *posterior distribution.* Since we know $\pi(\theta)$ and we know $f(X_1, \dots, X_n|\theta)$, we can use Bayes' Theorem to the density of the posterior distribution:

$$\begin{aligned}
\pi(\theta \mid X_1, \dots, X_n) &= \frac{f(\theta, X_1, \dots, X_n)}{\int f(\theta, X_1, \dots, X_n) d\theta} \\
&= \frac{f(X_1, \dots, X_n \mid \theta)\, \pi(\theta)}{\int f(\theta, X_1, \dots, X_n) d\theta} \\
&\propto f(X_1, \dots, X_n \mid \theta)\, \pi(\theta)\,,
\end{aligned}$$

where note that the proportionality constant is the inverse of the marginal distribution of the data and is uniquely defined since the right hand side must integrate to 1.

The posterior distribution contains all relevant information about θ having accounted for the observed data. Note that, a key distinguishing feature of Bayesian models, from non-Bayesian (or *frequentist* models) is that the parameter θ is now random and has a distribution. This is a philosophical difference between the two ideologies. We will look at a few examples to understand this better.

Example 1 (Coin probability). Suppose $Y_1, \dots, Y_n \sim \text{Bern}(p)$ are realizations from n series of coin flips of a coin that gives heads with probability p. Suppose you have no reason to believe that the coin is fair, and have no idea what is the probability of heads for the coin. You want to let the model know that p could be anything between $[0, 1]$ and in fact all values seem equally likely to you.

That is

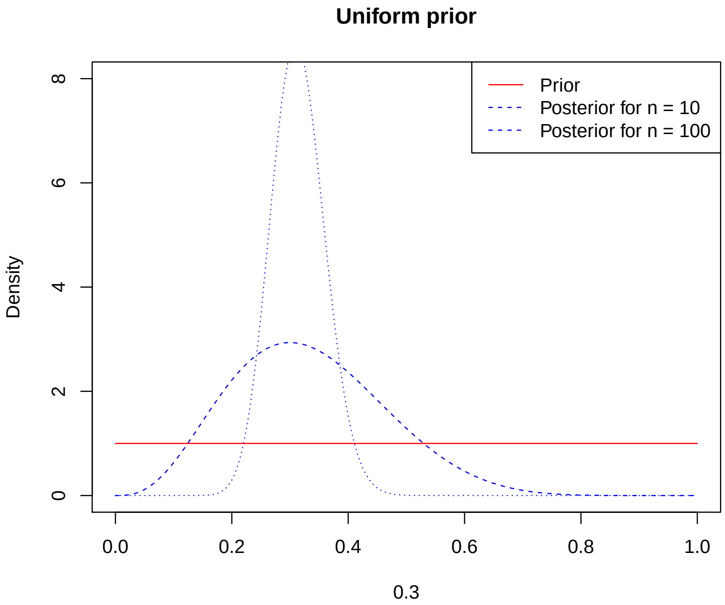
$$\textit{Prior distribution: } \theta \sim U[0, 1]\,.$$

That is, the prior on p is $\pi(p) = 1$ for $0 \leq p \leq 1$. By Bayes' theorem, the posterior distribution is:

$$\begin{aligned}\pi(p \mid y) &\propto \pi(p) \cdot \prod_{i=1}^{n} f(y_i \mid p)\\ &= 1\mathbb{I}(0 \leq p \leq 1) \cdot \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}\\ &= p^{\sum y_i}(1-p)^{n-\sum y_i}\,\mathbb{I}(0 \leq p \leq 1)\,,\end{aligned}$$

Although we haven't found the proportionality constant (the marginal likelihood), we know that the above expression is that of a Beta distribution. Thus, So,

$$\textit{Posterior distribution: } p \mid y \sim \text{Beta}\left(\sum y_i + 1, n - \sum y_i + 1\right)$$



Now, suppose our prior distribution was different, and that we knew some values of p were more favorable than others. We can assume

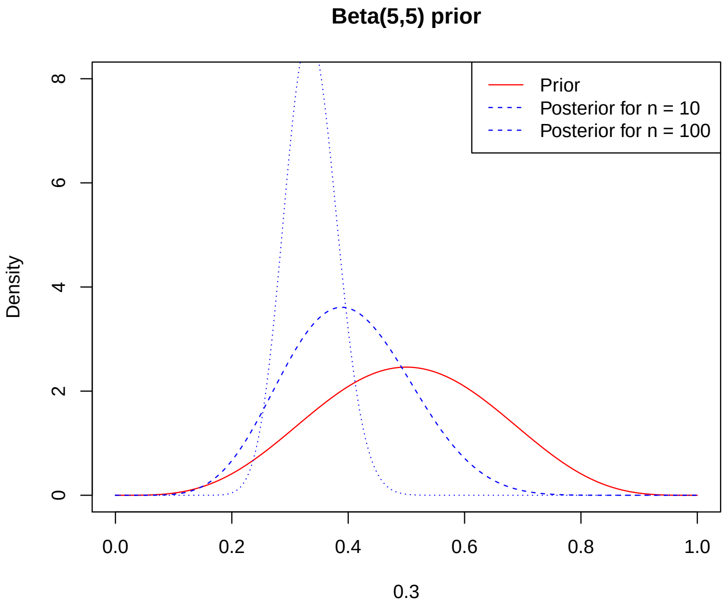
$$\textit{Prior distribution: } \theta \sim \text{Beta}(\alpha, \beta)\,,$$

for $\alpha, \beta > 0$. Note that for $\alpha = \beta = 1$, the prior is the $U[0,1]$ prior. For this prior, the posterior will be different, since

$$\begin{aligned}\pi(p \mid y) &\propto \pi(p) \cdot \prod_{i=1}^{n} f(y_i \mid p) \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\mathbb{I}(0 \le p \le 1) \cdot \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} \\ &\propto p^{\sum y_i + \alpha - 1}(1-p)^{n-\sum y_i + \beta - 1}\, \mathbb{I}(0 \le p \le 1)\,.\end{aligned}$$

Thus, the posterior distribution is

$$\textit{Posterior distribution: } p \mid y \sim \text{Beta}\left(\sum y_i + \alpha, n - \sum y_i + \beta\right)\,.$$



Question: Information about θ is in the form of a distribution. How do we summarize a full distribution:

1. *Maximum a posterior (MAP)*: the mode of the posterior distribution.
2. *Posterior mean*: the mean of the posterior distribution, $\text{E}[\theta|y]$.

Amongst the two, posterior means are used more often that MAP. The above are point estimates and do not give any idea of the variability of the posterior distribution. Thus, we need a way to assess the variability. This is done via *credible intervals*. A 95% credible interval is $[L, U]$ such that

$$\int_L^U \pi(\theta|y) d\theta = .95 \,.$$

Thus, one way to obtain $[L, U]$ is to set $L = .025^{th}$-quantile and $U = .975^{th}$-quantile. Similarly, an 80% quantile will be $L = .10^{th}$-quantile and $U = .90^{th}$-quantile.

Example 2 (Coin probability continued). The MAP estimator requires solving the following optimization problem:

$$\arg\max_p \left\{ p^{\sum y_i + \alpha - 1} (1-p)^{n - \sum y_i + \beta - 1} \, \mathbb{I}(0 \leq p \leq 1) \right\} \,,$$

which, if it exists happens is easy to show, that it happens at

$$\hat{p}_{\text{MAP}} = \frac{\sum y_i + \alpha - 1}{n + \alpha + \beta - 2}$$

The posterior mean estimator is

$$\hat{p}_{\text{PM}} = \frac{\sum y_i + \alpha}{n + \alpha + \beta} \,.$$

Notice also that

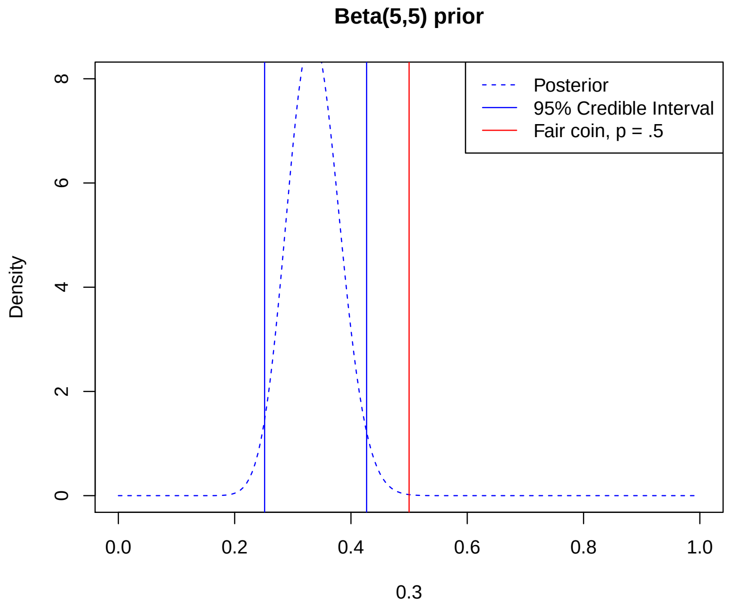
$$\begin{aligned} \hat{p}_{\text{PM}} &= \frac{\sum y_i}{n + \alpha + \beta} + \frac{\alpha}{n + \alpha + \beta} \\ &= \frac{n}{n + \alpha + \beta} \bar{y} + \frac{\alpha + \beta}{n + \alpha + \beta} \cdot \frac{\alpha}{\alpha + \beta} \,. \end{aligned}$$

Notice that the posterior mean is a weighted average between the prior mean and the data mean. As data increases ($n \to \infty$), $n/(n + \alpha + \beta) \to 1$ and thus,

$$\hat{p}_{PM} - \bar{y} \to 0 \,.$$

That is, Bayesian modeling, in many cases, finds a balance between information from the data and information from the prior, and once there is sufficient data, it essentially starts matching the inference made by only the data.

We can also visualize the 95% credible interval and a vertical line for the previous simulated data and we also present a vertical line for $p = .5$. Since this vertical line is outside the credible interval, we can be certain that the coin is not fair.



Example 3 (Normal-normal example). Let $Y_1, Y_2, \ldots, Y_n \sim N(\theta, \sigma^2)$ for $\sigma^2 > 0$ known. Further, let's suppose the prior distribution is

$$\textit{Prior distribution: } \theta \sim N(m_0, s_0^2)\,.$$

The posterior distribution is

$$\textit{Posterior distribution: } \theta \sim \pi(\theta|y)\,.$$

In order to find this

$$
\begin{aligned}
&\pi(\theta|y)\\
&\propto \pi(\theta) \cdot \prod_{i=1}^{n} f(y_i|\theta)\\
&\propto \exp\left\{-\frac{(\theta - m_0)^2}{2s_0^2}\right\} \exp\left\{-\sum_{i=1}^{n} \frac{(y_i - \theta)^2}{2\sigma^2}\right\}\\
&= \exp\left\{-\frac{1}{2}\left(\frac{\theta^2}{s_0^2} + \frac{m_0^2}{s_0^2} - \frac{2\theta m_0}{s_0^2} + \frac{\sum_{i=1}^{n} y_i^2}{\sigma^2} + \frac{n\theta^2}{\sigma^2} - \frac{2\theta \sum_{i=1}^{n} y_i}{\sigma^2}\right)\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left(\theta^2\left(\frac{1}{s_0^2} + \frac{n}{\sigma^2}\right) - 2\theta\left(\frac{m_0}{s_0^2} + \frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right)\right)\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{s_0^2} + \frac{n}{\sigma^2}\right)\left(\theta^2 - 2\theta\left(\frac{m_0}{s_0^2} + \frac{n\bar{y}}{\sigma^2}\right)\left(\frac{1}{s_0^2} + \frac{n}{\sigma^2}\right)^{-1} + \left[\left(\frac{m_0}{s_0^2} + \frac{n\bar{y}}{\sigma^2}\right)\left(\frac{1}{s_0^2} + \frac{n}{\sigma^2}\right)^{-1}\right]^2\right)\right\}
\end{aligned}
$$

So that

$$\theta|y \sim N\left(\left(\frac{m_0}{s_0^2} + \frac{n\bar{y}}{\sigma^2}\right)\left(\frac{1}{s_0^2} + \frac{n}{\sigma^2}\right)^{-1}, \left(\frac{1}{s_0^2} + \frac{n}{\sigma^2}\right)^{-1}\right)$$

How to choose a prior distribution:

- The support of the prior distribution defines the support of the posterior distribution. Thus, one way to restrict the family of distributions is to consider the restrictions on the parameter. For example, for the Bernoulli problem, the prior on p must be between 0,1. If putting a prior on the variance of a normal distribution, the prior distribution must have positive support.
- Within a family of prior distributions, the specific choices depend on prior information available, like in the examples above.
- Sometimes (and this should not happen often), priors may be chosen so as to allow for simpler and computable posterior distributions. We will discuss these next time.

2 Questions to think about

- Show for yourself that the posterior mean in the normal example is a weighted average of the sample mean and the prior mean. What happens at $n \to \infty$?
- Find the posterior distribution for data distribution $\text{Exp}(\lambda)$ and prior $\lambda \sim \text{Gamma}(a, b)$.
- Generate data from Example 3 and pictorial see how the posterior distribution changes as n increases.
- When will MAP and posterior means coincide?