Bayesian Inference
Multiparameters models

UBCO MDS — DATA 582

Introduction

- In previous lectures we considered Bayesian inference in the context of statistical models where only a single scalar parameter θ is to be estimated.
- Of course, many practical problems in statistics involves more than one unknown or unobservable quantity.
- Although a problem can include several parameters of interest, conclusions will often be drawn about one, or only a few, parameters at a time.
- Suppose θ is a vector $\theta = (\theta_1, \theta_2)$. If we are only interested in inferences on θ_1 we call θ_2 a *nuisance parameter*.

- Although we may not be interested in θ_2, we will still need to incorporate the uncertainty about it into our analysis.

- Today we see how the Bayesian framework is well equipped to handle this type of problem.

- In words, our aim now is to find the distribution that quantifies the uncertainty about θ_1 the parameter of interest, conditional *only* on the observed data.

- To accomplish this, we will need to integrate out the unknown nuisance parameter θ_2 from the so-called *joint posterior* distribution of θ_1 and θ_2.

Recall Bayes' rule for a single parameter:

$$\begin{aligned} p(\theta \mid y) &\propto p(y \mid \theta) \times p(\theta) \\ \text{posterior} &\propto \text{likelihood} \times \text{prior} \end{aligned} \tag{1}$$

Moving to the multi-parameter setting we have:

$$\begin{aligned} p(\theta_1, \theta_2 \mid y) &\propto p(y \mid \theta_1, \theta_2) \times p(\theta_1, \theta_2) \\ \textit{joint} \text{ posterior} &\propto \textit{joint} \text{ likelihood} \times \textit{joint} \text{ prior} \end{aligned}$$

The *marginal posterior*, or *posterior marginal*, θ_1 is given by:

$$\begin{aligned} p(\theta_1 \mid y) &= \int p(\theta_1, \theta_2 \mid y) d\theta_2 \\ &= \int p(\theta_1 \mid \theta_2, y) p(\theta_2 \mid y) d\theta_2 \end{aligned} \tag{2}$$

- As we explore in a future lecture, we can avoid explicit evaluation of (2) (which might be impossible) through simulation-based techniques.

- Before exploring this, let's return to the normal model for which an analytic solution of this posterior marginal *is* possible.

- Consider a sample $y_1, \ldots, y_n$ independent and identically drawn from a normal population with both μ and σ^2 unknown.

- Suppose we are only interested in the mean, i.e. the variance is a nuisance parameter.

A note on joint distributions **in general**

Marginal/condition distributions

If X and Y are continuous random variables with joint probability density function $f_{XY}(x, y)$ then: the marginal distribution functions for X and Y are:

$$f_X(x) = \int_y f_{XY}(x, y)dy \quad f_Y(y) = \int_x f_{XY}(x, y)dx$$

and the conditional distribution function of X (resp. Y) given Y (resp. X) is:

$$f_{X|y}(x) = \frac{f_{XY}(x, y)}{f_X(x)} \quad f_{Y|x}(y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

Normal model with μ, σ^2 unknown

Using $\theta_1 = \mu$ and $\theta_2 = \sigma^2$ we have:

$$p(\mu, \sigma^2 \mid y) \propto p(y \mid \mu, \sigma^2) \times p(\mu, \sigma^2)$$

joint posterior $\propto$ likelihood $\times$ *joint* prior

We want the marginal posterior for μ given by:

$$\begin{aligned} p(\mu \mid y) &= \int p(\mu, \sigma^2 \mid y) d\sigma^2 \\ &= \int \boxed{p(\mu \mid \sigma^2, y)} p(\sigma^2 \mid y) d\sigma^2 \end{aligned}$$

Notice how it depends on this which we found last class. We referred to it as $p(\mu \mid y)$ with known σ^2 to keep things cleaner, but it is more generally it is a *conditional posterior distribution*.

Joint Prior

- If we assume μ and σ^2 are independent then the joint prior is the product of their so-called *marginal priors*

$$p(\mu, \sigma^2) = p(\mu)p(\sigma^2)$$

- We have discussed the normal and uniform prior on μ. Popular choices for priors on variance include:

$$p(\sigma^2) \propto \frac{1}{\sigma^2} \quad \text{(improper Jeffrey's prior)} \tag{3}$$

$$p(\sigma^2) \sim \text{Inv-Gamma}(\alpha, \beta) \tag{4}$$

$$p(\sigma^2) \sim \text{Scale-Inv-}\chi^2(\nu, \tau^2) \tag{5}$$

- More generally if μ and σ^2 are not assumed independent we write their joint prior distribution as:

$$p(\mu, \sigma^2) = p(\sigma^2)p(\mu \mid \sigma^2)$$

- For instance, if we assume an inverse gamma for the marginal prior on σ^2 and a Normal(μ_0, σ_0^2) density for μ given σ^2, then $p(\mu, \sigma^2)$ can be expressed as:

$$\begin{aligned} &= p(\sigma^2)p(\mu \mid \sigma^2) \\ &= \mathsf{IG}(\alpha, \beta) \times \mathcal{N}(\mu_0, \sigma_0^2) \end{aligned}$$

Likelihood

The likelihood from observations from a $N(\mu, \sigma^2)$ is:

$$
\begin{aligned}
p(y \mid \mu, \sigma^2) &= \prod_{i=1}^{n} p(y_i \mid \mu, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\} \\
&\propto \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\} \\
&\propto \frac{\exp\left\{-\frac{1}{2\sigma^2}\left[n(\overline{y} - \mu)^2 + SS_y\right]\right\}}{(\sigma^2)^{n/2}}
\end{aligned}
$$

where $SS_y = \sum_{i=1}^{n}(y_i - \overline{y})^2$. Notice that all terms involving μ or σ need to be kept in the expression.

Joint Posterior

To calculate the joint posterior, $p(\mu, \sigma^2 \mid y)$ we must first decide on which joint prior to use. Today we outline two:

1. Assuming μ and σ^2 are independent we'll adopt a uniform prior for μ and the Jeffrey's prior given in (3) for σ^2

$$\begin{aligned} p(\mu, \sigma^2) &= p(\mu) \times p(\sigma^2) \\ &= 1 \times \frac{1}{\sigma^2} \qquad \text{(this is improper)} \end{aligned}$$

2. Assuming μ and σ^2 are *not* independent:

$$\begin{aligned} p(\mu, \sigma^2) &= p(\sigma^2) p(\mu \mid \sigma^2) \\ &= \mathrm{IG}(\alpha, \beta) \times \mathcal{N}(\mu_0, \sigma_0^2) \end{aligned}$$

Example 1.

$$
\begin{aligned}
\textit{joint posterior} &\propto \textit{likelihood} \times \textit{joint prior} \\
p(\mu, \sigma^2 \mid y) &\propto p(y \mid \mu, \sigma^2) \times p(\mu, \sigma^2) \\
&\propto \frac{\exp\left\{-\frac{1}{2\sigma^2}\left[n(\overline{y}-\mu)^2 + SS_y\right]\right\}}{(\sigma^2)^{n/2}} \times \frac{1}{\sigma^2} \\
&\propto \frac{1}{(\sigma^2)^{\frac{n}{2}+1}} \exp\left\{-\frac{1}{2\sigma^2}\left[n(\overline{y}-\mu)^2 + \frac{n-1}{n-1} SS_y\right]\right\} \\
&\propto \sigma^{n-2} \exp\left\{-\frac{1}{2\sigma^2}\left[n(\overline{y}-\mu)^2 + (n-1)s^2\right]\right\}
\end{aligned}
$$

where $s^2 = \frac{SS_y}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$ is the sample variance.

Marginal posterior, $p(\mu \mid y)$

If we only care about μ we extract the marginal posterior of μ:

$$\begin{aligned} p(\mu \mid y) &= \int p(\mu, \sigma^2 \mid y) d\sigma^2 \\ &= \int \boxed{p(\mu \mid \sigma^2, y)} p(\sigma^2 \mid y) d\sigma^2 \end{aligned}$$

From last lecture we know this for known variance and $p(\mu) = c$ for any constant c. The results for the marginal posterior $p(\sigma^2 \mid y)$ are given below (details in supp. material)

$$\mu \mid \sigma^2, y \sim \text{Normal}(\bar{y}, \sigma^2/n) \tag{6}$$

$$\sigma^2 \mid y \sim \text{Inv-Gamma}\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right) \tag{7}$$

The marginal can be obtained by integrating σ^2 out of the joint posterior distribution on the previous page[1]

$$
\begin{aligned}
p(\mu \mid y) &= \int p(\mu, \sigma^2 \mid y) d\sigma^2 \\
&= \int \boxed{p(\mu \mid \sigma^2, y)}\ \boxed{p(\sigma^2 \mid y)}\, d\sigma^2 \\
&= \int \boxed{\text{pdf for } (6)} \times \boxed{\text{pdf for } (7)} \\
&= \ldots
\end{aligned}
$$

which can be recognized as a Generalized student's t distribution with degrees of freedom $n-1$, location $\overline{y}$ and scale s^2/n

$$\mu \mid y \sim t_{n-1}(\overline{y}, s^2/n)$$

[1] see MC Sec 7.2 or AG textbook for details

- For example 2. we still decompose the joint posterior to a product of marginal and conditional probabilities:

$$p(\mu, \sigma^2 \mid y) = p(\mu \mid \sigma^2, y)p(\sigma^2 \mid y)$$

- Again, this is convenient since we know μ has conjugate distributions if we condition on knowing σ^2.
- This phenomenon is known as *semi-conjugacy*.
- The term semi-conjugate is used when the prior specified for each unknown model parameter would have been conjugate if all other model parameters were assumed known, but the entire joint prior is not conjugate.

Joint Prior

- In pdf format, our joint prior, which we'll refer to as the *normal-inverse gamma prior* takes the form:

$$
\begin{aligned}
p(\mu, \sigma^2) &= p(\sigma^2) p(\mu \mid \sigma^2) \\
&= \mathsf{IG}(\alpha, \beta) \times \mathcal{N}(\mu_0, \sigma_0^2)
\end{aligned}
$$

If we let $\sigma_0^2 = \sigma^2/\kappa_0 \implies \kappa_0 = \sigma^2/\sigma_0^2$,

$$
\begin{aligned}
&= \frac{\sqrt{\kappa_0}}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\kappa_0(\mu - \mu_0)^2}{2\sigma^2}\right) \\
&\qquad \cdots \times \frac{\beta^\alpha}{\Gamma(\alpha)} (1/\sigma^2)^{\alpha+1} \exp\left(-\beta/\sigma^2\right)
\end{aligned}
$$

- Here we let $\sigma_0^2 = \sigma^2/\kappa_0$. Notice how σ^2 appears in the conditional prior distribution for μ.

- This is notable since it stresses the fact that we are not assuming there are independent. For example, if σ^2 is large, then it induces a higher prior variance for μ.

- While this dependence is quite restrictive, it simplifies calculations and is largely used for convenience (we'll see how it this prior is conjugate).

- Furthermore, if σ_0^2 is not proportional to σ^2 we do not arrive at a closed-form solution for the posterior.

Joint Posteriors

We now wish to calculate the joint posterior density $p(\mu, \sigma^2 \mid y)$ whichwe again factorize as follows ...

$$p(\mu, \sigma^2 \mid y) = p(\mu \mid \sigma^2, y) p(\sigma^2 \mid y)$$

Recall $p(\mu \mid \sigma^2, y)$ from the one-parameter case:

$$\mu \mid \sigma^2, y \sim \mathcal{N}\left(\frac{\frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}\right)$$

If we let $\sigma_0^2 = \sigma^2/\kappa_0 \implies \kappa_0 = \sigma^2/\sigma_0^2$, :

$$\mu_{\text{post}} = \frac{\frac{\kappa_0}{\sigma^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}} = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0 + n}$$

$$\sigma^2_{\text{post}} = \frac{1}{\frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}} = \frac{\sigma^2}{\kappa_0 + n}.$$

As we have seen previously, μ_{post} is a weighted average of the prior mean μ_0 and sample mean $\bar{y}$

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y}$$
$$= \frac{\kappa_0}{\kappa_n}\mu_0 + \frac{n}{\kappa_n}\bar{y}$$

Additionally, is is dependent on the sample size n and κ_0 which can be viewed as the effective sample size ESS of our prior.

For notational convenience you might let $\kappa_n = \kappa_0 + n$. In addition, κ_n can be viewed as the posterior samples size since it takes the sample size from our data n and adds it to our prior sample size κ_0.

$$\mu \mid \sigma^2, y \sim \mathcal{N}(\mu_{\text{post}}, \sigma^2/\kappa_n) \tag{8}$$

$$\sigma^2 \mid y \sim \text{IG}\left(\alpha + \frac{n}{2}, \beta + \frac{(n-1)s^2}{2} + \frac{\kappa_0 n(\overline{y} - \mu_0)^2}{2\kappa_n}\right) \tag{9}$$

Notice, the inverse gamma prior density was conjugate for the normal likelihood, since the resulting posterior density is in the same family as the prior.

Marginal posterior, $p(\mu \mid y)$

If we only care about μ we extract the marginal posterior of μ:

$$
\begin{aligned}
p(\mu \mid y) &= \int p(\mu, \sigma^2 \mid y) d\sigma^2 \\
&= \int \boxed{p(\mu \mid \sigma^2, y)} \times \boxed{p(\sigma^2 \mid y)} d\sigma^2 \\
&= \int \boxed{\text{pdf for (8)}} \times \boxed{\text{pdf for (9)}} \\
& \dots
\end{aligned}
$$

This can also be recognized as a Generalized student's t distribution with parameter values given on pg 108 of MC (messy)

- The normal distribution is one of the few multiparameter problems for which the posteriors are simple enough to solve in closed form.

- This exercise is a first step toward a more general model which discuss next.

Comments

- Now that simulation-based methods are conveniently available for Bayesian model fitting, this joint prior density is not commonly used for inference in simple models with a normal likelihood.

- However, it is worth knowing about because of both its historical use and its current usefulness as a building block in more complex models.

- Next up, we will see how to perform Bayesian inference for nonconjugate models.

References I