LECTURE 19

§21.3. IMPROPER PRIORS.

For point and interval estimates and to some extent in testing, objective priors are often improper. We discuss a few basic facts about improper priors.

An improper prior density $p(\theta)$ is non-negative for all θ, but

$$\int_\Theta p(\theta)\, d\theta = +\infty.$$

Such an improper prior can be used in the Bayes formula for calculating the posterior, provided the denominator is finite for all $X = (x_1, x_2, ..., x_n)$ (or all, but a set of $(x_1, x_2, ..., x_n)$ with zero probability for all θ), i.e.

$$\int_\Theta p(\theta)\, L(x_1, x_2, ..., x_n|\theta)\, d\theta < \infty.$$

Then the posterior density $p(\theta|x_1, x_2, ..., x_n)$ is a proper probability density function and can be used at least in inference problems.

The most common improper priors are

$$p_1(\mu) = C, \quad \text{for} -\infty < \mu < +\infty,$$
$$p_2(\sigma) = \frac{1}{\sigma}, \quad \text{for} \qquad 0 < \sigma < +\infty,$$

for location and scale parameters. Both the improper priors may be interpreted as a sort of limit of the proper priors:

$$p_{1,L}(\mu) = \begin{cases} \dfrac{1}{2L} & \text{if} \quad -L < \mu < L \\ 0 & \text{otherwise} \end{cases}$$

$$p_{2,L}(\sigma) = \begin{cases} \dfrac{A}{\sigma} & \text{if} \quad 0 < \dfrac{1}{L} < \sigma < L \\ 0 & \text{otherwise} \end{cases}$$

where

$$A = \frac{1}{2 \ln L},$$

in the sense that the posteriors for p_1 and p_2 may be obtained by making $L \to +\infty$ in $p_{i,L}(\theta|x_1, x_2, ..., x_n)$.

If the data comprise a conditionally i.i.d. sample from the Bernoulli distribution with parameter θ, and θ has the beta distribution with parameters α and β, the posterior distribution of θ is the beta distribution with parameters $\alpha + s$, $\beta + n - s$ (n observations, s successes). If we use the improper (and unreal) beta distribution prior with prior parameters $\alpha = \beta = 0$, and we have improper prior

$$p(\theta) \propto \theta^{-1} (1-\theta)^{-1},$$

we obtain a proper posterior proportional to

$$\theta^{s-1} (1-\theta)^{n-s-1},$$

i.e. the posterior density function is the beta distribution with parameters s and $n - s$ i.e. $B(s, n-s)$.

Suppose we can write the likelihood for a given parameter θ and data vector (the sample) $(x_1, x_2, ..., x_n)$ as

$$L(x_1, x_2, ..., x_n|\theta) = g[\theta - t(x_1, x_2, ..., x_n)], \tag{21.3}$$

Here the likelihood is a function $L = g(z)$, where $z = \theta - t(x_1, x_2, ..., x_n)$. If the likelihood is of this form, the data $x_1, x_2, ..., x_n$ only influences θ by a translation on the scale of the function g, i.e., from $g(z)$ to $g(z+a)$. Further, note that $t(x_1, x_2, ..., x_n)$ is the only value of the data that appears, and we call the function t a sufficient statistic. Other data sets with different values of $x_1, x_2, ..., x_n$, but the same value of the sufficient statistic $t(x_1, x_2, ..., x_n)$, have the same likelihood.

When the likelihood can be placed in the form of Equation (21.3), a shift in the data gives rise to the same functional form of the likelihood function except for a shift in location, from $(\theta + t(x_1, x_2, ..., x_n))$ to $(\theta + t(y_1, y_2, ..., y_n))$. Hence, this is a natural scale upon which to measure likelihoods, and on such a scale, a flat prior seems natural.

Example 21.1. Consider n independent samples from a normal distribution with unknown mean μ and known variance σ^2 Here

$$L(x_1, x_2, ..., x_n|\mu) \propto \exp\left(-\frac{(\mu - \overline{x})^2}{2(\sigma^2/n)}\right)$$

Note immediately that $t(x_1, x_2, ..., x_n) = \overline{x}$ is a sufficient statistic for the mean, so that different data sets with the same mean (for n draws) have the same likelihood function for the unknown mean μ. Further note that

$$g(z) = \exp\left(\frac{-z^2}{2(\sigma^2/n)}\right)$$

Hence, a flat prior for μ seems appropriate.

What is the natural scale for a likelihood function that does not satisfy equation (21.3)? Suppose that the likelihood function can be written in data-translated format as

$$L(x_1, x_2, ..., x_n|\theta) = g[h(\theta) - t(x_1, x_2, ..., x_n)] \tag{21.4}$$

When the likelihood function has this format, the natural scale for the unknown parameter is $h(\theta)$. Hence, a prior of the form $p[h(\theta)] = \textit{constant}$ (a flat prior on $h[\theta]$) is suggested.

Using a change of variables to transform $p[h(\theta)]$ back onto the θ scale suggests a prior on θ of the form

$$p(\theta) \propto \left|\frac{\partial h(\theta)}{\partial \theta}\right| \tag{21.5}$$

Indeed, for distribution function we have

$$F_\Theta(\theta) = P(\Theta \leq \theta) = P(h(\Theta) \leq h(\theta)) = F_{h(\Theta)}(h(\theta)).$$

Therefore, for density functions

$$p_\Theta(\theta) = p_{h(\Theta)}(h(x)) \cdot \left|\frac{\partial h(\theta)}{\partial \theta}\right| = c \cdot \left|\frac{\partial h(\theta)}{\partial \theta}\right| \propto \left|\frac{\partial h(\theta)}{\partial \theta}\right|.$$

Example 21.2. Suppose the likelihood function assumes data follow an exponential distribution,

$$L(x_1, x_2, ..., x_n|\theta) = \frac{1}{\theta^n} \exp\left(-\frac{\sum_{i=1}^n x_i}{\theta}\right), \qquad x_1, x_2, ..., x_n > 0.$$

To express this likelihood in a data-translated format, we will make use of the fact that we can multiply any likelihood function by a constant and still have a likelihood function.

In particular, since the data $(x_1, x_2, ..., x_n)$ is known (and hence treated as a constant), we can multiply the likelihood function by any function of the data, e.g.

$$f(x_1, x_2, ..., x_n)\, L(x_1, x_2, ..., x_n|\theta) \propto L(x_1, x_2, ..., x_n|\theta).$$

In this example, we simply multiply the likelihood function by $x_1 \cdot x_2 \cdot ... \cdot x_n$ to give

$$L(x_1, x_2, ..., x_n|\theta) = \frac{\prod x_i}{\theta^n} \exp\left(-\frac{\sum_{i=1}^n x_i}{\theta}\right).$$

Noting that

$$\frac{\prod x_i}{\theta^n} = \exp\left[\ln\left(\frac{\prod x_i}{\theta^n}\right)\right] = \exp\left[\sum_{i=1}^n \ln x_i - n\,\ln\theta\right] = \exp\left[\sum_{i=1}^n (\ln x_i - \ln\theta)\right].$$

and

$$\exp\left\{-\frac{1}{\theta}\sum_{i=1}^n x_i\right\} = \exp\{-x_1/\theta\}\cdot\exp\{-x_2/\theta\}\cdot ...\cdot\exp\{-x_n/\theta\} =$$

$$= \exp\{-\exp\{\ln x_1 - \ln\theta\}\}\ \exp\{-\exp\{\ln x_2 - \ln\theta\}\}\ ...\exp\{-\exp\{\ln x_n - \ln\theta\}\} =$$

$$= \exp\left\{-\exp\left[\sum_{i=1}^n (\ln x_i - \ln\theta)\right]\right\}$$

we can express the likelihood as

$$L(x_1, x_2, ..., x_n|\theta) = \exp\left[\sum_{i=1}^n (\ln x_i - \ln\theta) - \exp\left(\sum_{i=1}^n (\ln x_i - \ln\theta)\right)\right].$$

Hence, in data-translated format the likelihood function becomes

$$g(y) = \exp[y - \exp(y)], \qquad t(x_1, x_2, ..., x_n) = \ln\prod x_i = \sum_{i=1}^n \ln x_i, \qquad h(\theta) = n\,\ln\theta.$$

The natural scale for θ in this likelihood function is thus $\ln\theta$, and a natural prior is

$$F_\Theta(\theta) = P(\Theta \leq \theta) = P(\ln\Theta \leq \ln\theta) = F_{\ln\Theta}(\ln\theta),$$

or relation between density functions is

$$f_\Theta(\theta) = f_{\ln\Theta}(\ln\theta)\,\frac{1}{\theta}.$$

Therefore, for density functions we obtain:

$$p(\ln\theta) = constant,$$

which corresponds to

$$p(\theta) \propto \left|\frac{\partial\ln\theta}{\partial\theta}\right| = \frac{1}{\theta}.$$