Lecture 10

10.1 Bayes estimators.

(Textbook, Sections 6.3 and 6.4)

Once we find the posterior distribution or its p.d.f. or p.f. $\xi(\theta|X_1,\ldots,X_n)$ we turn to constructing the estimate $\hat{\theta}$ of the unknown parameter θ_0. The most common way to do this is simply take the mean of the posterior distribution

$$\hat{\theta} = \hat{\theta}(X_1,\ldots,X_n) = \mathbb{E}(\theta|X_1,\ldots,X_n).$$

This estimate $\hat{\theta}$ is called the *Bayes estimator*. Note that $\hat{\theta}$ depends on the sample $X_1,\ldots,X_n$ since, by definition, the posterior distribution depends on the sample. The obvious motivation for this choice of $\hat{\theta}$ is that it is simply the average of the parameter with respect to posterior distribution that in some sense captures the information contained in the data and our prior intuition about the parameter. Besides this obvious motivation one sometimes gives the following motivation. Let us define the estimator as the parameter a that minimizes

$$\mathbb{E}((\theta-a)^2|X_1,\ldots,X_n),$$

i.e. the posterior average squared deviation of θ from a is as small as possible. To find this a we find the critical point:

$$\frac{\partial}{\partial a}\mathbb{E}((\theta-a)^2|X_1,\ldots,X_n) = 2\mathbb{E}(\theta|X_1,\ldots,X_n) - 2a = 0$$

and it turns out to be the mean

$$a = \hat{\theta} = \mathbb{E}(\theta|X_1,\ldots,X_n).$$

Let us summarize the construction of Bayes estimator.

1. Choose prior distribution of θ, $\xi(\theta)$.

2. Compute posterior distribution $\xi(\theta|X_1,\ldots,X_n)$.

3. Find the expectation of the posterior $\hat{\theta} = \mathbb{E}(\theta|X_1,\ldots,X_n)$.

10.2 Conjugate prior distributions.

Often for many popular families of distributions the prior distribution $\xi(\theta)$ is chosen so that it is easy to compute the posterior distribution. This is done by choosing $\xi(\theta)$ that resembles the likelihood function $f(X_1, \ldots, X_n|\theta)$. We will explain this on the particular examples.

Example. Suppose that the sample comes from Bernoulli distribution $B(p)$ with p.f.

$$f(x|p) = p^x(1-p)^{1-x} \text{ for } x = 0, 1$$

and likelihood function

$$f(X_1, \cdots, X_n|p) = p^{\sum X_i}(1-p)^{n-\sum X_i}.$$

Then the posterior distribution will have the form:

$$\xi(p|X_1, \ldots, X_n) = \frac{f(X_1, \ldots, X_n|p)\xi(p)}{g(X_1, \ldots, X_n)} = \frac{p^{\sum X_i}(1-p)^{n-\sum X_i}\xi(p)}{g(X_1, \ldots, X_n)}.$$

Notice that the likelihood function

$$p^{\sum X_i}(1-p)^{n-\sum X_i}$$

resembles the density of Beta distribution. Therefore, if we let the prior distribution be a Beta distribution $B(\alpha, \beta)$ with some parameters $\alpha, \beta > 0$:

$$\xi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

then the posterior distribution will be

$$\xi(p|X_1, \ldots, X_n) = \frac{1}{g(X_1, \ldots, X_n)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \underbrace{p^{(\alpha+\sum X_i)-1}(1-p)^{(\beta+n-\sum X_i)-1}}_{\textit{resembles Beta distribution}}.$$

We still have to compute $g(X_1, \ldots, X_n)$ but this can be avoided if we notice that $\xi(p|X_1, \ldots, X_n)$ is supposed to be a p.d.f. and it looks like a Beta distribution with parameter $\alpha + \sum X_i$ and $\beta + n - \sum X_i$. Therefore, g has no choice but to be such that

$$\xi(p|X_1, \ldots, X_n) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+\sum X_i)\Gamma(\beta+n-\sum X_i)} p^{(\alpha+\sum X_i)-1}(1-p)^{(\beta+n-\sum X_i)-1}$$

which is the p.d.f. of $B\Big(\alpha+\sum X_i, \beta+n-\sum X_i\Big)$. Since the mean of Beta distribution $B(\alpha, \beta)$ is equal to $\alpha/(\alpha+\beta)$, the Bayes estimator will be

$$\hat{p} = \mathbb{E}(p|X_1, \ldots, X_n) = \frac{\alpha + \sum X_i}{\alpha + \sum X_i + \beta + n - \sum X_i} = \frac{\alpha + \sum X_i}{\alpha + \beta + n}.$$

Let us notice that for large sample size, i.e. when $n \to +\infty$, we have

$$\hat{p} = \frac{\alpha + \sum X_i}{\alpha + \beta + n} = \frac{\frac{\alpha}{n} + \bar{X}}{\frac{\alpha}{n} + \frac{\beta}{n} + 1} \approx \bar{X}.$$

This means that when we have a lot of data our prior intuition becomes irrelevant and the Bayes estimator is approximated by the sample average $\bar{X}$. On the other hand, for $n = 0$

$$\hat{p} = \frac{\alpha}{\alpha + \beta}$$

which is the mean of prior distribution $B(\alpha, \beta)$. If we have no data we simply follow our intuitive guess.

Example. Suppose that the sample comes from the exponential distribution $E(\alpha)$ with p.f.

$$f(x|\alpha) = \alpha e^{-\alpha x} \text{ for } x \geq 0$$

in which case the likelihood function is

$$f(X_1, \ldots, X_n) = \alpha^n e^{-\alpha \sum X_i}.$$

The posterior distribution will have the form:

$$\xi(\alpha | X_1, \ldots, X_n) = \frac{1}{g(X_1, \ldots, X_n)} \alpha^n e^{-\alpha \sum X_i} \xi(\alpha).$$

Notice that the likelihood function resembles the p.d.f. of Gamma distribution and, therefore, we will take prior to be a Gamma distribution $\Gamma(u, v)$ with parameters u and v, i.e.

$$\xi(\alpha) = \frac{v^u}{\Gamma(u)} \alpha^{u-1} e^{-v\alpha}.$$

Then, the posterior will be equal to

$$\xi(\alpha | X_1, \ldots, X_n) = \frac{1}{g} \frac{v^u}{\Gamma(u)} \alpha^{(u+n)-1} e^{-\alpha(\sum X_i + v)}$$

which again looks like a Gamma distribution with parameters $u + n$ and $v + \sum X_i$. Again, $g(X_1, \ldots, X_n)$ will be such that

$$\xi(\alpha | X_1, \ldots, X_n) = \frac{(\sum X_i + v)^{u+n}}{\Gamma(u+n)} \alpha^{(u+n)-1} e^{-\alpha(\sum X_i + v)}$$

which is the p.d.f. of $\Gamma(u + n, v + \sum X_i)$. Since the mean of Gamma distribution $\Gamma(\alpha, \beta)$ with parameters α and β is equal to α/β, the Bayes estimator will be

$$\hat{\alpha} = \mathbb{E}(\alpha | X_1, \ldots, X_n) = \frac{u + n}{v + \sum X_i}.$$

For large sample size n, we get

$$\hat{\alpha} = \frac{\frac{u}{n} + 1}{\frac{v}{n} + \bar{X}} \approx \frac{1}{\bar{X}}.$$

Example. If the sample comes from Poisson distribution $\Pi(\lambda)$ with p.d.f.

$$f(x|\lambda) = \frac{\lambda^x}{x!} e^{-\lambda} \text{ for } x = 0, 1, 2, \ldots$$

then the likelihood function is

$$f(X_1, \ldots, X_n|\lambda) = \frac{\lambda^{\sum X_i}}{\prod X_i!} e^{-n\lambda}$$

and the posterior distribution will have the form

$$\xi(\lambda|X_1, \ldots, X_n) = \frac{1}{g(X_1, \ldots, X_n)} \frac{\lambda^{\sum X_i}}{\prod X_i!} e^{-n\lambda} \xi(\lambda).$$

Since again the likelihood function resembles the Gamma distribution we will take the prior to be a Gamma distribution $\Gamma(u, v)$ in which case

$$\xi(\lambda|X_1, \ldots, X_n) = \frac{1}{g} \frac{v^u}{\Gamma(u)} \lambda^{(\sum X_i + u) - 1} e^{-(n+v)\lambda}.$$

Since this looks like a Gamma distribution $\Gamma(u + \sum X_i, n + v)$ the posterior has no choice but to be equal to this distribution and the Bayes estimator will be:

$$\hat{\lambda} = \frac{\sum X_i + u}{n + v} = \frac{\bar{X} + \frac{u}{n}}{1 + \frac{v}{n}}.$$