

# Stat 710: Mathematical Statistics

## Lecture 3

Jun Shao

Department of Statistics  
University of Wisconsin  
Madison, WI 53706, USA

# Lecture 3: Empirical and hierarchical Bayes actions

## Hyperparameters and empirical Bayes

A Bayes action depends on the chosen prior with a vector  $\xi$  of parameters called *hyperparameters*.

So far, hyperparameters are assumed to be known.

If the hyperparameter  $\xi$  is unknown, one way to solve the problem is to estimate  $\xi$  using some historical data; the resulting Bayes action is called an *empirical Bayes* action.

If there is no historical data, we may estimate  $\xi$  using data  $x$  and the resulting Bayes action is also called an empirical Bayes action.

The simplest empirical Bayes method is to estimate  $\xi$  by viewing  $x$  as a "sample" from the marginal distribution

$$P_{x|\xi}(A) = \int_{\Theta} P_{x|\theta}(A) d\Pi_{\theta|\xi}, \quad A \in \mathcal{B}_x,$$

where  $\Pi_{\theta|\xi}$  is a prior depending on  $\xi$  or from the marginal p.d.f.  $m(x) = \int_{\Theta} f_{\theta}(x) d\Pi$ , if  $P_{x|\theta}$  has a p.d.f.  $f_{\theta}$ .

The method of moments can be applied to estimate  $\xi$ .

## Example 4.4

Let  $X = (X_1, \dots, X_n)$  and  $X_i$ 's be i.i.d. from  $N(\mu, \sigma^2)$  with an unknown  $\mu \in \mathcal{R}$  and a known  $\sigma^2$ .

Consider the prior  $\Pi_{\mu|\xi} = N(\mu_0, \sigma_0^2)$  with  $\xi = (\mu_0, \sigma_0^2)$ .

To obtain a moment estimate of  $\xi$ , we need to calculate

$$\int_{\mathcal{R}^n} x_1 m(x) dx \quad \text{and} \quad \int_{\mathcal{R}^n} x_1^2 m(x) dx, \quad x = (x_1, \dots, x_n).$$

These two integrals can be obtained without calculating  $m(x)$ .

Note that

$$\int_{\mathcal{R}^n} x_1 m(x) dx = \int_{\Theta} \int_{\mathcal{R}^n} x_1 f_{\mu}(x) dx d\Pi_{\mu|\xi} = \int_{\mathcal{R}} \mu d\Pi_{\mu|\xi} = \mu_0$$

and

$$\begin{aligned} \int_{\mathcal{R}^n} x_1^2 m(x) dx &= \int_{\Theta} \int_{\mathcal{R}^n} x_1^2 f_{\mu}(x) dx d\Pi_{\mu|\xi} = \sigma^2 + \int_{\mathcal{R}} \mu^2 d\Pi_{\mu|\xi} \\ &= \sigma^2 + \mu_0^2 + \sigma_0^2 \end{aligned}$$

## Example 4.4: (continued)

Thus, by viewing  $x_1, \dots, x_n$  as a sample from  $m(x)$ , we obtain the moment estimates

$$\hat{\mu}_0 = \bar{x} \quad \text{and} \quad \hat{\sigma}_0^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 - \sigma^2,$$

where  $\bar{x}$  is the sample mean of  $x_i$ 's.

Replacing  $\mu_0$  and  $\sigma_0^2$  in

$$\mu_*(x) = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \bar{x} \quad \text{and} \quad c^2 = \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}$$

(Example 2.25) by  $\hat{\mu}_0$  and  $\hat{\sigma}_0^2$ , respectively, we find that the empirical Bayes action under the squared error loss is simply the sample mean  $\bar{x}$  (which is the generalized Bayes action in Example 4.3).

Note that  $\hat{\sigma}_0^2$  in Example 4.4 can be negative.

Better empirical Bayes methods can be found, for example, in Berger (1985, §4.5)

## Hierarchical Bayes

Instead of estimating hyperparameters, in the *hierarchical Bayes* approach we put a prior on hyperparameters.

Let  $\Pi_{\theta|\xi}$  be a (first-stage) prior with a hyperparameter vector  $\xi$  and let  $\Lambda$  be a prior on  $\Xi$ , the range of  $\xi$ .

Then the "marginal" prior for  $\theta$  is defined by

$$\Pi(B) = \int_{\Xi} \Pi_{\theta|\xi}(B) d\Lambda(\xi), \quad B \in \mathcal{B}_{\Theta}.$$

If the second-stage prior  $\Lambda$  also depends on some unknown hyperparameters, then one can go on to consider a third-stage prior. In most applications, however, two-stage priors are sufficient, since misspecifying a second-stage prior is much less serious than misspecifying a first-stage prior (Berger, 1985, §4.6).

In addition, the second-stage prior can be noninformative (improper). Bayes actions can be obtained in the same way as before.

Thus, the hierarchical Bayes method is simply a Bayes method with a hierarchical prior.

## Remarks

- Empirical Bayes methods deviate from the Bayes method since  $x$  is used to estimate hyperparameters.
- The hierarchical Bayes method is generally better than empirical Bayes methods.

Suppose that  $X$  has a p.d.f.  $f_{\theta}(x)$  w.r.t. a  $\sigma$ -finite measure  $\nu$  and  $\Pi_{\theta|\xi}$  has a p.d.f.  $\pi_{\theta|\xi}(\theta)$  w.r.t. a  $\sigma$ -finite measure  $\kappa$ .

Then the prior  $\Pi$  has a p.d.f. (w.r.t.  $\kappa$ )

$$\pi(\theta) = \int_{\Xi} \pi_{\theta|\xi}(\theta) d\Lambda(\xi)$$

and

$$m(x) = \int_{\Theta} \int_{\Xi} f_{\theta}(x) \pi_{\theta|\xi}(\theta) d\Lambda d\kappa.$$

Let  $P_{\theta|x,\xi}$  be the posterior distribution of  $\vec{\theta}$  given  $x$  and  $\xi$  and

$$m_{x|\xi}(x) = \int_{\Theta} f_{\theta}(x) \pi_{\theta|\xi}(\theta) d\kappa,$$

which is the marginal of  $X$  given  $\xi$ .

## Remarks

- Empirical Bayes methods deviate from the Bayes method since  $x$  is used to estimate hyperparameters.
- The hierarchical Bayes method is generally better than empirical Bayes methods.

Suppose that  $X$  has a p.d.f.  $f_{\theta}(x)$  w.r.t. a  $\sigma$ -finite measure  $\nu$  and  $\Pi_{\theta|\xi}$  has a p.d.f.  $\pi_{\theta|\xi}(\theta)$  w.r.t. a  $\sigma$ -finite measure  $\kappa$ . Then the prior  $\Pi$  has a p.d.f. (w.r.t.  $\kappa$ )

$$\pi(\theta) = \int_{\Xi} \pi_{\theta|\xi}(\theta) d\Lambda(\xi)$$

and

$$m(x) = \int_{\Theta} \int_{\Xi} f_{\theta}(x) \pi_{\theta|\xi}(\theta) d\Lambda d\kappa.$$

Let  $P_{\theta|x,\xi}$  be the posterior distribution of  $\vec{\theta}$  given  $x$  and  $\xi$  and

$$m_{x|\xi}(x) = \int_{\Theta} f_{\theta}(x) \pi_{\theta|\xi}(\theta) d\kappa,$$

which is the marginal of  $X$  given  $\xi$ .

Then the posterior distribution  $P_{\theta|x}$  has a p.d.f.

$$\begin{aligned}\frac{dP_{\theta|x}}{d\kappa} &= \frac{f_{\theta}(x)\pi(\theta)}{m(x)} \\ &= \int_{\Xi} \frac{f_{\theta}(x)\pi_{\theta|\xi}(\theta)}{m(x)} d\Lambda(\xi) \\ &= \int_{\Xi} \frac{f_{\theta}(x)\pi_{\theta|\xi}(\theta)}{m_{x|\xi}(x)} \frac{m_{x|\xi}(x)}{m(x)} d\Lambda(\xi) \\ &= \int_{\Xi} \frac{dP_{\theta|x,\xi}}{d\kappa} dP_{\xi|x},\end{aligned}$$

where  $P_{\xi|x}$  is the posterior distribution of  $\xi$  given  $x$ .

Thus, under the estimation problem considered in Example 4.1, the (hierarchical) Bayes action is

$$\delta(x) = \int_{\Xi} \delta(x, \xi) dP_{\xi|x},$$

where  $\delta(x, \xi)$  is the Bayes action when  $\xi$  is known.

A result similar to this is given in Lemma 4.1.

## Example 4.5

Consider Example 4.4 again.

Suppose that  $\mu_0$  in the first-stage prior  $N(\mu_0, \sigma_0^2)$ , is unknown and  $\sigma_0^2$  is known.

Let the second-stage prior for  $\xi = \mu_0$  be the Lebesgue measure on  $\mathcal{R}$  (improper prior).

From Example 2.25,

$$\delta(x, \xi) = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \xi + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \bar{x}.$$

To obtain the Bayes action  $\delta(x)$ , it suffices to calculate  $E_{\xi|x}(\xi)$ , where the expectation is w.r.t.  $P_{\xi|x}$ .

Note that the p.d.f. of  $P_{\xi|x}$  is proportional to

$$\psi(\xi) = \int_{-\infty}^{\infty} \exp \left\{ -\frac{n(\bar{x} - \mu)^2}{2\sigma^2} - \frac{(\mu - \xi)^2}{2\sigma_0^2} \right\} d\mu.$$

## Example 4.5 (continued)

Using the properties of normal distributions, one can show that

$$\begin{aligned}\psi(\xi) &= C_1 \exp \left\{ \left( \frac{n}{2\sigma^2} + \frac{1}{2\sigma_0^2} \right)^{-1} \left( \frac{n\bar{x}}{2\sigma^2} + \frac{\xi}{2\sigma_0^2} \right)^2 - \frac{\xi^2}{2\sigma_0^2} \right\} \\ &= C_2 \exp \left\{ -\frac{n\xi^2}{2(n\sigma_0^2 + \sigma^2)} + \frac{n\bar{x}\xi}{n\sigma_0^2 + \sigma^2} \right\} \\ &= C_3 \exp \left\{ -\frac{n(\xi - \bar{x})^2}{2(n\sigma_0^2 + \sigma^2)} \right\},\end{aligned}$$

where  $C_1$ ,  $C_2$ , and  $C_3$  are quantities not depending on  $\xi$ .

Hence  $E_{\xi|x}(\xi) = \bar{x}$ .

The (hierarchical) generalized Bayes action is then

$$\delta(x) = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} E_{\xi|x}(\xi) + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \bar{x} = \bar{x}.$$