Lecture 2: Generalized, empirical, and hierarchical Bayes methods

Generalized Bayes action

The minimization in Definition 4.1 is the same as the minimizing

\[ \int_{\Theta} L(\theta, \delta(x)) f_{\theta}(x) d\Pi = \min_{a \in A} \int_{\Theta} L(\theta, a) f_{\theta}(x) d\Pi \]

\( \delta(x) \) is called a generalized Bayes action.

This is still defined even if \( \Pi \) is not a probability measure but a \( \sigma \)-finite measure on \( \Theta \), in which case \( m(x) \) may not be finite.

If \( \Pi(\Theta) \neq 1 \), \( \Pi \) is called an improper prior.

A prior with \( \Pi(\Theta) = 1 \) is then called a proper prior.

The following is a reason why we need to discuss improper priors and generalized Bayes actions.

With no past information, one has to choose a prior subjectively.

In such cases, one would like to select a noninformative prior that tries to treat all parameter values in \( \Theta \) equitably.

A noninformative prior is often improper.
Example 4.3

Suppose that $X = (X_1, \ldots, X_n)$ and $X_i$’s are i.i.d. from $N(\mu, \sigma^2)$, where $\mu \in \Theta \subset \mathbb{R}$ is unknown and $\sigma^2$ is known.

Consider the estimation of $\vartheta = \mu$ under the squared error loss.

If $\Theta = [a, b]$ with $-\infty < a < b < \infty$, then a noninformative prior that treats all parameter values equitably is the uniform distribution on $[a, b]$.

If $\Theta = \mathbb{R}$, however, the corresponding "uniform distribution" is the Lebesgue measure on $\mathbb{R}$, which is an improper prior.

If $\Pi$ is the Lebesgue measure on $\mathbb{R}$, then

$$(2\pi\sigma^2)^{-n/2} \int_{-\infty}^{\infty} \mu^2 \exp \left\{ -\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right\} d\mu < \infty.$$ 

By differentiating $a$ in

$$(2\pi\sigma^2)^{-n/2} \int_{-\infty}^{\infty} (\mu - a)^2 \exp \left\{ -\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right\} d\mu$$
and using the fact that \( \sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \), where \( \bar{x} \) is the sample mean of the observations \( x_1, ..., x_n \), we obtain that

\[
\delta(x) = \frac{\int_{-\infty}^{\infty} \mu \exp \left\{-\frac{n(\bar{x} - \mu)^2}{2\sigma^2}\right\} d\mu}{\int_{-\infty}^{\infty} \exp \left\{-\frac{n(\bar{x} - \mu)^2}{2\sigma^2}\right\} d\mu} = \bar{x}.
\]

Thus, the sample mean is a generalized Bayes action under the squared error loss.

From Example 2.25, if \( \Pi \) is \( N(\mu_0, \sigma_0^2) \), then the Bayes action is

\[
\delta(x) = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \bar{x}
\]

Note that in this case \( \bar{x} \) is a limit of \( \delta(x) \) as \( \sigma_0^2 \to \infty \).

More detailed discussions of the use of improper priors, see Jeffreys (1939, 1948, 1961), Box and Tiao (1973), and Berger (1985).
and using the fact that $\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$, where $\bar{x}$ is the sample mean of the observations $x_1, ..., x_n$, we obtain that

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

Thus, the sample mean is a generalized Bayes action under the squared error loss.

From Example 2.25, if $\Pi$ is $N(\mu_0, \sigma_0^2)$, then the Bayes action is

$$\delta(x) = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \bar{x}$$

Note that in this case $\bar{x}$ is a limit of $\delta(x)$ as $\sigma_0^2 \to \infty$.

More detailed discussions of the use of improper priors, see Jeffreys (1939, 1948, 1961), Box and Tiao (1973), and Berger (1985).
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

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

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 \( X = (X_1, \ldots, X_n) \) and \( X_i \)'s be i.i.d. from \( N(\mu, \sigma^2) \) with an unknown \( \mu \in \mathbb{R} \) 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

\[
\int_{\mathbb{R}^n} x_1 m(x) \, dx \quad \text{and} \quad \int_{\mathbb{R}^n} x_1^2 m(x) \, dx, \quad x = (x_1, \ldots, x_n).
\]

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

Note that

\[
\int_{\mathbb{R}^n} x_1 m(x) \, dx = \int_{\Theta} \int_{\mathbb{R}^n} x_1 f_{\mu}(x) \, dx \, d\Pi_{\mu|\xi} = \int_{\mathbb{R}} \mu \, d\Pi_{\mu|\xi} = \mu_0
\]

and

\[
\int_{\mathbb{R}^n} x_1^2 m(x) \, dx = \int_{\Theta} \int_{\mathbb{R}^n} x_1^2 f_{\mu}(x) \, dx \, d\Pi_{\mu|\xi} = \sigma^2 + \int_{\mathbb{R}} \mu^2 \, d\Pi_{\mu|\xi}
\]

\[
= \sigma^2 + \mu_0^2 + \sigma_0^2
\]
Example 4.4: (continued)

Thus, by viewing \(x_1, \ldots, x_n\) 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

\[
\mu^*_0(x) = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \bar{x} \quad \text{and} \quad c^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}
\]

(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 $\bar{\theta}$ 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 $\bar{\theta}$ 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.

\[
\frac{dP_{\theta|x}}{d\kappa} = f_{\theta}(x)\pi(\theta) \tag{1}
\]

\[
= \int_\Xi \frac{f_{\theta}(x)\pi_{\theta|\xi}(\theta)}{m(x)} d\Lambda(\xi)
\]

\[
= \int_\Xi \frac{f_{\theta}(x)\pi_{\theta|\xi}(\theta)m_{x|\xi}(x)}{m(x)} d\Lambda(\xi)
\]

\[
= \int_\Xi \frac{dP_{\theta|x,\xi}}{d\kappa} dP_{\xi|x},
\]

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 $\mathbb{R}$ (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

\[
\psi(\xi) = C_1 \exp \left\{ \left( \frac{n}{2\sigma^2} + \frac{1}{2\sigma_0^2} \right)^{-1} \left( \frac{n\bar{x}}{2\sigma^2} + \frac{\xi}{2\sigma_0^2} \right)^2 - \frac{\xi^2}{2\sigma_0^2} \right\} 
\]

\[
= C_2 \exp \left\{ -\frac{n\xi^2}{2(n\sigma_0^2 + \sigma^2)} + \frac{n\bar{x}\xi}{n\sigma_0^2 + \sigma^2} \right\} 
\]

\[
= C_3 \exp \left\{ -\frac{n(\xi - \bar{x})^2}{2(n\sigma_0^2 + \sigma^2)} \right\},
\]

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

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