Chapter 4: Estimation in Parametric Models
Lecture 1: Bayesian approach

X is from a population in a parametric family $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$, where $\Theta \subset \mathscr{R}^k$ for a fixed integer $k \geq 1$

Bayes approach

- Optimal rules in the *Bayesian approach*, which is fundamentally different from the classical frequentist approach that we have been adopting
- θ is viewed as a realization of a random vector $\vec{\theta} \in \Theta$ whose *prior* distribution is Π
- Prior distribution: past experience, past data, or a statistician's belief (subjective)
- Sample $X \in \mathscr{X}$: from $P_\theta = P_{x|\theta}$, the conditional distribution of X given $\vec{\theta} = \theta$
- Posterior distribution: updated prior distribution using observed $X = x$

How to construct the posterior?

By Theorem 1.7, the joint distribution of X and $\vec{\theta}$ is a probability measure on $\mathscr{X} \times \Theta$ determined by

$$P(A \times B) = \int_B P_{x|\theta}(A) d\Pi(\theta), \qquad A \in \mathscr{B}_{\mathscr{X}},\ B \in \mathscr{B}_{\Theta}$$

The posterior distribution is the conditional distribution $P_{\theta|x}$ whose existence is guaranteed by Theorem 1.7 a.s. $x \in \mathscr{X}$

Theorem 4.1 (Bayes formula)

Assume $\mathscr{P} = \{P_{x|\theta} : \theta \in \Theta\}$ is dominated by a σ-finite measure ν and $f_\theta(x) = dP_{x|\theta}/d\nu$ is a Borel function on $(\mathscr{X} \times \Theta, \sigma(\mathscr{B}_{\mathscr{X}} \times \mathscr{B}_{\Theta}))$. Let Π be a prior distribution on Θ. Suppose that $m(x) = \int_\Theta f_\theta(x) d\Pi > 0$. ($m(x)$ is called the marginal p.d.f. of X w.r.t. ν)

(i) The posterior distribution $P_{\theta|x} \ll \Pi$ and

$$dP_{\theta|x}/d\Pi = f_\theta(x)/m(x)$$

(ii) If $\Pi \ll \lambda$ and $d\Pi/d\lambda = \pi(\theta)$ for a σ-finite measure λ, then

$$dP_{\theta|x}/d\lambda = f_\theta(x)\pi(\theta)/m(x)$$

If T is a sufficient statistic for θ, then the posterior depends only on T.

Discrete X and $\vec{\theta}$: The Bayes formula in elementary probability

$$P(\vec{\theta}=\theta|X=x)=\frac{P(X=x|\vec{\theta}=\theta)P(\vec{\theta}=\theta)}{\sum_{\theta\in\Theta}P(X=x|\vec{\theta}=\theta)P(\vec{\theta}=\theta)}$$

Remarks on the Bayesian approach

- Without loss of generality we may assume $m(x)>0$
  If $m(x)=0$ for an $x\in\mathscr{X}$, then $f_\theta(x)=0$ a.s. Π
  Either x should be eliminated from $\mathscr{X}$ or the prior Π is incorrect and a new prior should be specified
- The posterior $P_{\theta|x}$ contains all the information we have about θ
- Statistical decisions and inference should be made based on $P_{\theta|x}$, conditional on the observed $X=x$
- In estimating θ, $P_{\theta|x}$ can be viewed as a randomized decision rule under the approach discussed in §2.3
  After $X=x$ is observed, $P_{\theta|x}$ is a randomized rule, which is a probability distribution on the action space $\mathscr{A}=\Theta$
- The Bayesian method can be applied iteratively

Definition 4.1 (Bayes action)

Let $\mathscr{A}$ be an action space in a decision problem and $L(\theta, a) \geq 0$ be a loss function

For any $x \in \mathscr{X}$, a *Bayes action* w.r.t. Π is any $\delta(x) \in \mathscr{A}$ such that

$$E[L(\vec{\theta}, \delta(x))|X = x] = \min_{a \in \mathscr{A}} E[L(\vec{\theta}, a)|X = x]$$

where the expectation is w.r.t. the posterior distribution $P_{\theta|x}$

Remarks

- The Bayes action minimizes the posterior expected loss
- x is fixed, although $\delta(x)$ depends on x
- The Bayes action depends on the prior
- The Bayes action depends on the loss function
- The existence and uniqueness of Bayes actions are discussed in Proposition 4.1
- If $\delta(x)$ is a measurable function of x, then $\delta(X)$ is a nonrandomized decision rule under the frequentist approach

Example 4.1: the estimation of $\vartheta = g(\theta)$

Assume $\int_\Theta [g(\theta)]^2 d\Pi < \infty$, $\mathscr{A}$ = the range of $g(\theta)$, and
$L(\theta, a) = [g(\theta) - a]^2$ (squared error loss).
Using the argument in Example 1.22, we obtain the Bayes action

$$\delta(x) = \frac{\int_\Theta g(\theta) f_\theta(x) d\Pi}{m(x)} = \frac{\int_\Theta g(\theta) f_\theta(x) d\Pi}{\int_\Theta f_\theta(x) d\Pi},$$

which is the posterior expectation of $g(\vec{\theta})$, given $X = x$.

A more specific case

$g(\theta) = \theta^j$ for some integer $j \geq 1$
$f_\theta(x) = e^{-\theta}\theta^x I_{\{0,1,2,\ldots\}}(x)/x!$ (the Poisson distribution) with $\theta > 0$
Π has a Lebesgue p.d.f. $\pi(\theta) = \theta^{\alpha-1} e^{-\theta/\gamma} I_{(0,\infty)}(\theta)/[\Gamma(\alpha)\gamma^\alpha]$
(the gamma distribution $\Gamma(\alpha, \gamma)$ with known $\alpha > 0$ and $\gamma > 0$)
Then, for $x = 0, 1, 2, \ldots$, and some function $c(x)$,

$$f_\theta(x)\pi(\theta)/m(x) = c(x)\theta^{x+\alpha-1} e^{-\theta(\gamma+1)/\gamma} I_{(0,\infty)}(\theta),$$

This is the gamma distribution $\Gamma(x+\alpha, \gamma/(\gamma+1))$.

Without actually working out the integral $m(x)$, we know that

$$c(x) = (1+\gamma^{-1})^{x+\alpha}/\Gamma(x+\alpha),$$

$$\delta(x) = c(x)\int_0^\infty \theta^{j+x+\alpha-1}e^{-\theta(\gamma+1)/\gamma}d\theta.$$

The integrand is proportional to the p.d.f. of the gamma distribution $\Gamma(j+x+\alpha,\gamma/(\gamma+1))$.
Hence

$$\begin{aligned}\delta(x) &= c(x)\Gamma(j+x+\alpha)/(1+\gamma^{-1})^{j+x+\alpha}\\ &= (j+x+\alpha-1)\cdots(x+\alpha)/(1+\gamma^{-1})^{j}.\end{aligned}$$

In particular, $\delta(x) = (x+\alpha)\gamma/(\gamma+1)$ when $j=1$.

Conjugate prior

An interesting phenomenon is that the prior and the posterior are in the same parametric family of distributions.
Such a prior is called a *conjugate* prior.

Remarks

- Whether a prior is conjugate involves a pair of families, the family $\mathscr{P} = \{f_\theta : \theta \in \Theta\}$ and the family from which Π is chosen.
- Example 4.1 shows that the Poisson family and the gamma family produce conjugate priors.
- Many pairs of families in Table 1.1 (page 18) and Table 1.2 (pages 20-21) produce conjugate priors.
- Under a conjugate prior, Bayes actions often have explicit forms (in x) when the loss function is simple.
- Even under a conjugate prior, the integral in $\delta(x)$ in Example 4.1 involving a general g may not have an explicit form.
- In general, numerical methods have to be used in evaluating the integrals in $\delta(x)$ under general loss functions.

Example 2.25/4.8

$X_1, ..., X_n$ i.i.d. $\sim N(\mu, \sigma^2)$, where $\mu \in \mathscr{R}$ is unknown and σ^2 is known.
Let $\pi(\mu)$ be the pdf of $N(\mu_0, \sigma_0^2)$.
Since $\bar{X} \sim N(\mu, \sigma^2/n)$ is sufficient, we treat $\bar{X} = \bar{x}$ as the observation.

$$
\begin{aligned}
f_\mu(\bar{x})\pi(\mu) &= \exp\left(-\frac{(\bar{x}-\mu)^2}{2\sigma^2/n}\right)\exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) \\
&= \exp\left(-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{n\bar{x}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\mu + \frac{n\bar{x}^2}{\sigma^2}+\frac{\mu_0^2}{\sigma_0^2}\right]\right) \\
&= \exp\left(-\frac{1}{2}\left[A\mu^2 - 2B\mu + C\right]\right) = \exp\left(-\frac{1}{2}\left[A(\mu - B/A)^2 - B^2/A + C\right]\right)
\end{aligned}
$$

Integrating out μ, we obtain that the marginal density of $\bar{X}$ is

$$
m(\bar{x}) \propto \exp\left(-\frac{1}{2}\left[C - B^2/A\right]\right) \propto \exp\left(-\frac{(\bar{x}-\mu_0)^2}{2(\sigma^2/n+\sigma_0^2)}\right)
$$

i.e., $m(\bar{x})$ is the density of $N(\mu_0, \sigma^2/n+\sigma_0^2)$.
Also, the posterior of μ given $\bar{x}$ is $N(B/A, A^{-1})$.
Then the Bayes estimate of μ under the squared error loss is

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

Next, assume that both μ and σ^2 are unknown, the prior for $\omega = (2\sigma^2)^{-1}$ is the gamma distribution $\Gamma(\alpha,\gamma)$ with known α and γ, and the prior for μ is $N(\mu_0, \sigma_0^2/\omega)$ (conditional on ω).
Then the posterior p.d.f. of (μ,ω) is proportional to

$$\omega^{(n+1)/2+\alpha-1} \exp\left\{ -\left[\gamma^{-1} + (n-1)s^2 + n(\bar{x}-\mu)^2 + \frac{(\mu-\mu_0)^2}{2\sigma_0^2} \right] \omega \right\},$$

From

$$n(\bar{x}-\mu)^2 + \frac{(\mu-\mu_0)^2}{2\sigma_0^2} = \left(n + \frac{1}{2\sigma_0^2}\right)\mu^2 - 2\left(n\bar{x} + \frac{\mu_0}{2\sigma_0^2}\right)\mu + n\bar{x}^2 + \frac{\mu_0^2}{2\sigma_0^2}.$$

the posterior p.d.f. of (μ,ω) is proportional to

$$\omega^{(n+1)/2+\alpha-1} \exp\left\{ -\left[\gamma^{-1} + W + \left(n + \frac{1}{2\sigma_0^2}\right)\{\mu - \zeta(\bar{x})\}^2 \right] \omega \right\},$$

$$\zeta(\bar{x}) = \frac{n\bar{x} + \frac{\mu_0}{2\sigma_0^2}}{n + \frac{1}{2\sigma_0^2}} \quad \text{and} \quad W = \sum_{i=1}^n x_i^2 + \frac{\mu_0^2}{2\sigma_0^2} - \left(n + \frac{1}{2\sigma_0^2}\right)[\zeta(\bar{x})]^2.$$

Thus, the posterior of ω is $\Gamma(n/2+\alpha, (\gamma^{-1}+W)^{-1})$ and the posterior of μ (given ω and x) is $N\left(\zeta(\bar{x}), [(2n+\sigma_0^{-2})\omega]^{-1}\right)$.
Under the squared error loss, the Bayes estimate of μ is $\zeta(\bar{x})$ and the Bayes estimate of $\sigma^2 = (2\omega)^{-1}$ is $(\gamma^{-1}+W)/(n+2\alpha-2)$, $n+2\alpha > 2$.

Generalized Bayes action

The minimization in Definition 4.1 is the same as the minimizing

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

This is still defined even if Π is not a probability measure but a σ-finite measure on Θ, in which case $m(x)$ may not be finite.

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

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

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 Θ equitably.

A noninformative prior is often improper.

Example 4.3

Suppose that $X = (X_1,...,X_n)$ and X_i's are i.i.d. from $N(\mu,\sigma^2)$, where $\mu \in \Theta \subset \mathscr{R}$ is unknown and σ^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 $\Theta = \mathscr{R}$, however, the corresponding "uniform distribution" is the Lebesgue measure on $\mathscr{R}$, which is an improper prior.
If Π is the Lebesgue measure on $\mathscr{R}$, then

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

By differentiating a and using $\sum_{i=1}^n(x_i-\mu)^2 = \sum_{i=1}^n(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2$, 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 Π 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 can be found in 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 ξ of parameters called *hyperparameters*.
So far, hyperparameters are assumed to be known.
If the hyperparameter ξ is unknown, one way to solve the problem is to estimate ξ using some historical data; the resulting Bayes action is called an *empirical Bayes* action.
If there is no historical data, we may estimate ξ using data x and the resulting Bayes action is also called an empirical Bayes action.
The simplest empirical Bayes method is to estimate ξ by viewing x as a "sample" from the marginal distribution

$$P_{x|\xi}(A) = \int_\Theta P_{x|\theta}(A) d\Pi_{\theta|\xi}, \qquad A \in \mathscr{B}_\mathscr{X},$$

where $\Pi_{\theta|\xi}$ is a prior depending on ξ 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_θ.
The method of moments can be applied to estimate ξ.

Example 4.4

Let $X = (X_1, ..., X_n)$ and X_i's be i.i.d. with an unknown mean $\mu \in \mathscr{R}$ and a known variance σ^2.
Assume the prior $\Pi_{\mu|\xi}$ has mean μ_0 and variance σ_0^2, $\xi = (\mu_0, \sigma_0^2)$.
To obtain a moment estimate of ξ, we need to calculate

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

These two integrals can be obtained without knowing $m(x)$.
Note that

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

and

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

$$= \sigma^2 + \mu_0^2 + \sigma_0^2$$

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

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

where $\bar{x}$ is the sample mean of x_i's.
Replacing μ_0 and σ_0^2 in

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

(Example 2.25) by $\widehat{\mu}_0$ and $\widehat{\sigma}_0^2$, 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 $\widehat{\sigma}_0^2$ 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 ξ and let Λ be a prior on Ξ, the range of ξ.

Then the "marginal" prior for θ is defined by

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

If the second-stage prior Λ 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 $\Pi_{\theta|\xi}$ has a p.d.f. $\pi_{\theta|\xi}(\theta)$ and the prior Λ has a p.d.f. $\lambda(\xi)$ w.r.t. a σ-finite measure κ.
Then the marginal prior of θ has a p.d.f. (w.r.t. κ)

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

Example 2.25.

If $\bar{X} \sim N(\mu, \sigma^2/n)$ with a known σ^2, the prior $\pi(\mu|\xi)$ is the p.d.f. of $N(\xi, \sigma_0^2)$ with a known σ_0^2, and the prior of ξ is $N(\mu_0, \tau^2)$ with known μ_0 and τ^2, then the marginal prior p.d.f. of μ is $N(\mu_0, \sigma_0^2 + \tau^2)$.

This can be derived using the result in Example 2.25 previously discussed with $(\bar{x}, \mu)$ replaced by (μ, ξ).

Because of the hierarchical prior, the prior of μ has more variation.