Lecture 2: Bayes rule and computation

Bayes rule

Under the frequentist approach, a Bayes action $\delta(X)$ as a measurable function of X is a nonrandomized decision rule.

It can be shown that $\delta(X)$ defined in Definition 4.1 (if it exists for $X = x \in A$ with $\int_\Theta P_\theta(A) d\Pi = 1$) also minimizes the Bayes risk

$$r_T(\Pi) = \int_\Theta R_T(\theta) d\Pi$$

over all decision rules T (randomized or nonrandomized), where $R_T(\theta) = E[L(\theta, T(X))]$ is the risk function of T (Chapter 2).

Thus, $\delta(X)$ is a Bayes rule defined in §2.3.2.

In an estimation problem, a Bayes rule is called a *Bayes estimator*.

Generalized Bayes risks and generalized Bayes rules (or estimators) can be defined similarly.

In view of the discussion in §2.3.2, even if we do not adopt the Bayesian approach, the method described in §4.1.1 can be used as a way of generating decision rules.

Frequentist properties of Bayes rules/estimators

Admissibility

Given $R_T(\theta) = E[L(T(X), \theta)]$, T is $\Im$-admissible iff there is no $T_0 \in \Im$ with $R_{T_0}(\theta) \leq R_T(\theta)$ for all θ and $R_{T_0}(\theta) < R_T(\theta)$ for some θ
Admissible = $\Im$-admissible with $\Im = \{$ all rules $\}$

Bayes rules are typically admissible: If T is better than a Bayes rule δ, then T has the same Bayes risk as δ and is itself a Bayes rule: We only need to show that no Bayes rule is worse than another Bayes rule.

Theorem 4.2 (Admissibility of Bayes rules)

In a decision problem, let $\delta(X)$ be a Bayes rule w.r.t. a prior Π.
(i) If $\delta(X)$ is a unique Bayes rule, then $\delta(X)$ is admissible.
(ii) If Θ is a countable set, the Bayes risk $r_\delta(\Pi) < \infty$, and Π gives positive probability to each $\theta \in \Theta$, then $\delta(X)$ is admissible.
(iii) Let $\Im$ be the class of decision rules having continuous risk functions. If $\delta(X) \in \Im$, $r_\delta(\Pi) < \infty$, and Π gives positive probability to any open subset of Θ, then $\delta(X)$ is $\Im$-admissible.

Generalized Bayes rules or estimators are not necessarily admissible.

Many generalized Bayes rules are limits of Bayes rules (Examples 4.3 and 4.7), which are often admissible.

Theorem 4.3

Suppose that Θ is an open set of $\mathscr{R}^k$.
In a decision problem, let $\Im$ be the class of decision rules having continuous risk functions.
A decision rule $T \in \Im$ is $\Im$-admissible if there exists a sequence $\{\Pi_j\}$ of (possibly improper) priors such that
(a) the generalized Bayes risks $r_T(\Pi_j)$ are finite for all j;
(b) for any $\theta_0 \in \Theta$ and $\eta > 0$,

$$\lim_{j\to\infty} \frac{r_T(\Pi_j) - r_j^*(\Pi_j)}{\Pi_j(O_{\theta_0,\eta})} = 0,$$

where $r_j^*(\Pi_j) = \inf_{T\in\Im} r_T(\Pi_j)$ and $O_{\theta_0,\eta} = \{\theta \in \Theta : \|\theta - \theta_0\| < \eta\}$ with $\Pi_j(O_{\theta_0,\eta}) < \infty$ for all j.

Proof

Suppose that T is not $\Im$-admissible.
Then there exists $T_0 \in \Im$ such that $R_{T_0}(\theta) \leq R_T(\theta)$ for all θ and $R_{T_0}(\theta_0) < R_T(\theta_0)$ for a $\theta_0 \in \Theta$.
From the continuity of the risk functions, we conclude that

$$R_{T_0}(\theta) < R_T(\theta) - \varepsilon \quad \theta \in O_{\theta_0,\eta}$$

for some constants $\varepsilon > 0$ and $\eta > 0$.
Then, for any j,

$$\begin{aligned} r_T(\Pi_j) - r_j^*(\Pi_j) &\geq r_T(\Pi_j) - r_{T_0}(\Pi_j) \\ &\geq \int_{O_{\theta_0,\eta}} [R_T(\theta) - R_{T_0}(\theta)] d\Pi_j(\theta) \\ &\geq \varepsilon \Pi_j(O_{\theta_0,\eta}), \end{aligned}$$

which contradicts condition (b). Hence, T is $\Im$-admissible.

While the proof of Theorem 4.3 is easy, the application of Theorem 4.3 is not so easy.

Example 4.6 (An application of Theorem 4.3)

Consider $X_1,...,X_n$ iid from $N(\mu,\sigma^2)$ with unknown μ and known σ^2
Loss = the squared error loss.
By Theorem 2.1, the risk function of any decision rule is continuous in μ if the risk is finite.
Apply Theorem 4.3 to the sample mean $\bar{X}$
Let $\Pi_j = N(0,j)$.
Since $R_{\bar{X}}(\mu) = \sigma^2/n$, $r_{\bar{X}}(\Pi_j) = \sigma^2/n$ for any j.
Hence, condition (a) in Theorem 4.3 is satisfied.
From Example 2.25, the Bayes estimator w.r.t. Π_j is

$$\delta_j(X) = \frac{nj}{nj+\sigma^2}\bar{X}$$

Thus,

$$R_{\delta_j}(\mu) = \frac{\sigma^2 nj^2 + \sigma^4\mu^2}{(nj+\sigma^2)^2}$$

and

$$r_j^*(\Pi_j) = \int R_{\delta_j}(\mu) d\Pi_j = \frac{\sigma^2 j}{nj + \sigma^2}.$$

For any $O_{\mu_0,\eta} = \{\mu : |\mu - \mu_0| < \eta\}$,

$$\Pi_j(O_{\mu_0,\eta}) = \Phi\left(\frac{\mu_0 + \eta}{\sqrt{j}}\right) - \Phi\left(\frac{\mu_0 - \eta}{\sqrt{j}}\right) = \frac{2\eta\Phi'(\xi_j)}{\sqrt{j}}$$

for some ξ_j satisfying $(\mu_0 - \eta)/\sqrt{j} \le \xi_j \le (\mu_0 + \eta)/\sqrt{j}$, where Φ is the standard normal c.d.f. and Φ' is its derivative.
Since $\Phi'(\xi_j) \to \Phi'(0) = (2\pi)^{-1/2}$,

$$\frac{r_{\bar{X}}(\Pi_j) - r_j^*(\Pi_j)}{\Pi_j(O_{\mu_0,\eta})} = \frac{\sigma^4\sqrt{j}}{2\eta\Phi'(\xi_j)n(nj + \sigma^2)} \to 0$$

as $j \to \infty$.
Thus, condition (b) in Theorem 4.3 is satisfied.
Hence, Theorem 4.3 applies and the sample mean $\bar{X}$ is admissible.

For any estimator T of ϑ, its bias is $E(T)-\vartheta$

Proposition 4.2 (Bayes estimators are biased)

If $\delta(X)$ is a Bayes estimator of $\vartheta = g(\theta)$ under the squared error loss, then $\delta(X)$ is not unbiased except in the trivial case where $r_\delta(\Pi) = 0$.

- Proposition 4.2 can be used to check whether an estimator can be a Bayes estimator w.r.t. some prior under the squared error loss.
- However, a generalized Bayes estimator may be unbiased; see, for instance, Examples 4.3 and 4.7.

Proof of Proposition 4.2

Suppose that $\delta(X)$ is unbiased, i.e., $E[\delta(X)|\vec{\theta}] = g(\vec{\theta})$.
Conditioning on $\vec{\theta}$ and using Proposition 1.10, we obtain that

$$E[g(\vec{\theta})\delta(X)] = E\{g(\vec{\theta})E[\delta(X)|\vec{\theta}]\} = E[g(\vec{\theta})]^2.$$

Since $\delta(X) = E[g(\vec{\theta})|X]$, conditioning on X and using Proposition 1.10,

$$E[g(\vec{\theta})\delta(X)] = E\{\delta(X)E[g(\vec{\theta})|X]\} = E[\delta(X)]^2.$$

Then

$$r_\delta(\Pi) = E[\delta(X) - g(\vec{\theta})]^2 = E[\delta(X)]^2 + E[g(\vec{\theta})]^2 - 2E[g(\vec{\theta})\delta(X)] = 0.$$

Bayesian computation

We first consider an example, in which we need the following useful lemma.

Lemma 4.1

Suppose that X has a p.d.f. $f_\theta(x)$ w.r.t. a σ-finite measure ν. Suppose that $\theta = (\theta_1, \theta_2)$, $\theta_j \in \Theta_j$, and that the prior has a p.d.f.

$$\pi(\theta) = \pi_{\theta_1|\theta_2}(\theta_1)\pi_{\theta_2}(\theta_2),$$

where $\pi_{\theta_2}(\theta_2)$ is a p.d.f. w.r.t. a σ-finite measure ν_2 on Θ_2 and for any given θ_2, $\pi_{\theta_1|\theta_2}(\theta_1)$ is a p.d.f. w.r.t. a σ-finite measure ν_1 on Θ_1. Suppose further that if θ_2 is given, the Bayes estimator of $h(\theta_1) = g(\theta_1, \theta_2)$ under the squared error loss is $\delta(X, \theta_2)$. Then the Bayes estimator of $g(\theta_1, \theta_2)$ under the squared error loss is $\delta(X)$ with

$$\delta(x) = \int_{\Theta_2} \delta(x, \theta_2) p_{\theta_2|x}(\theta_2) d\nu_2,$$

where $p_{\theta_2|x}(\theta_2)$ is the posterior p.d.f. of $\vec{\theta}_2$ given $X = x$.

Example 4.9

Consider a linear model

$$X_{ij} = \beta^\tau Z_i + \varepsilon_{ij}, \qquad j = 1, ..., n_i,\ i = 1, ..., k,$$

where $\beta \in \mathscr{R}^p$ is unknown, Z_i's are known vectors, ε_{ij}'s are independent, and ε_{ij} is $N(0, \sigma_i^2)$, $j = 1, ..., n_i$, $i = 1, ..., k$.
Let X be the sample vector containing all X_{ij}'s.
The parameter vector is $\theta = (\beta, \omega)$, $\omega = (\omega_1, ..., \omega_k)$ and $\omega_i = (2\sigma_i^2)^{-1}$.
Assume the prior for θ has the Lebesgue p.d.f.

$$c\,\pi(\beta) \prod_{i=1}^{k} \omega_i^{\alpha} e^{-\omega_i/\gamma},$$

where $\alpha > 0$, $\gamma > 0$, and $c > 0$ are known constants and $\pi(\beta)$ is a known Lebesgue p.d.f. on $\mathscr{R}^p$.
The posterior p.d.f. of θ is then proportional to

$$h(X, \theta) = \pi(\beta) \prod_{i=1}^{k} \omega_i^{n_i/2+\alpha} e^{-[\gamma^{-1} + v_i(\beta)]\omega_i},$$

where $v_i(\beta) = \sum_{j=1}^{n_i} (X_{ij} - \beta^\tau Z_i)^2$.

Example 4.9 (continued)

If β is known, the Bayes estimator of σ_i^2 under the squared error loss is

$$\int \frac{1}{2\omega_i} \frac{h(X,\theta)}{\int h(X,\theta)d\omega} d\omega = \frac{\gamma^{-1} + v_i(\beta)}{2\alpha + n_i}.$$

By Lemma 4.1, the Bayes estimator of σ_i^2 is

$$\widehat{\sigma}_i^2 = \int \frac{\gamma^{-1} + v_i(\beta)}{2\alpha + n_i} f_{\beta|X}(\beta) d\beta,$$

where

$$\begin{aligned} f_{\beta|X}(\beta) &\propto \int h(X,\theta)d\omega \\ &\propto \pi(\beta) \prod_{i=1}^{k} \int \omega_i^{\alpha+n_i/2} e^{-[\gamma^{-1}+v_i(\beta)]\omega_i} d\omega_i \\ &\propto \pi(\beta) \prod_{i=1}^{k} \left[\gamma^{-1} + v_i(\beta)\right]^{-(\alpha+1+n_i/2)} \end{aligned}$$

is the posterior p.d.f. of β.

Example 4.9 (continued)

The Bayes estimator of $l^\tau\beta$ for any $l \in \mathscr{R}^p$ is then the posterior mean of $l^\tau\beta$ w.r.t. the p.d.f. $f_{\beta|X}(\beta)$.

In this problem, Bayes estimators do not have explicit forms.

A numerical method has to be used to evaluate Bayes estimators (see Example 4.10).

Let $\bar{X}_{i\cdot}$ and S_i^2 be the sample mean and variance of X_{ij}, $j = 1,...,n_i$ (S_i^2 is defined to be 0 if $n_i = 1$)

Let $\sigma_0^2 = (2\alpha\gamma)^{-1}$ (the prior mean of σ_i^2).

Then the Bayes estimator $\widehat{\sigma}_i^2$ can be written as

$$\frac{2\alpha}{2\alpha+n_i}\sigma_0^2 + \frac{n_i-1}{2\alpha+n_i}S_i^2 + \frac{n_i}{2\alpha+n_i}\int(\bar{X}_{i\cdot} - \beta^\tau Z_i)^2 f_{\beta|X}(\beta)d\beta.$$

This Bayes estimator is a weighted average of prior information, “within group" variation, and averaged squared “residuals".

Markov chain Monte Carlo (MCMC)

Often, Bayes actions or estimators have to be computed numerically.
Typically we need to compute

$$E_p(g) = \int_\Theta g(\theta)p(\theta)d\nu$$

with some function g, where $p(\theta)$ is a p.d.f. w.r.t. a σ-finite measure ν on $(\Theta, \mathscr{B}_\Theta)$ and $\Theta \subset \mathscr{R}^k$.
If g is an indicator function of $A \in \mathscr{B}_\Theta$ and $p(\theta)$ is the posterior p.d.f. of θ given $X = x$, then $E_p(g)$ is the posterior probability of A.
There are many numerical methods for computing integrals $E_p(g)$.

The simple Monte Carlo method

Generate i.i.d. $\theta^{(1)}, ..., \theta^{(m)}$ from a p.d.f. $h(\theta) > 0$ w.r.t. ν.
By the SLLN, as $m \to \infty$,

$$\widehat{E}_p(g) = \frac{1}{m}\sum_{j=1}^m \frac{g(\theta^{(j)})p(\theta^{(j)})}{h(\theta^{(j)})} \to_{a.s.} \int_\Theta \frac{g(\theta)p(\theta)}{h(\theta)} h(\theta)d\nu = E_p(g).$$

Hence $\widehat{E}_p(g)$ is a numerical approximation to $E_p(g)$.

The simple Monte Carlo method may not work well because

- the convergence of $\widehat{E}_p(g)$ is very slow when k (the dimension of Θ) is large
- generating a random vector from some k-dimensional distribution may be difficult, if not impossible.

More sophisticated MCMC methods

Different from the simple Monte Carlo in two aspects:

- generating random vectors can be done using distributions whose dimensions are much lower than k
- $\theta^{(1)}, ..., \theta^{(m)}$ are not independent, but form a homogeneous Markov chain.

Many MCMC methods were developed in the last 20 years
We only consider one of them as an example

Gibbs sampler

Let $y = (y_1, y_2, ..., y_d)$. (y_j's may be vectors with different dimensions)
At step $t = 1, 2, ...$, given $y^{(t-1)}$, generate
$y_1^{(t)}$ from $P(y_2^{(t-1)}, ..., y_d^{(t-1)} | y_1^{(t-1)})$,...,
$y_j^{(t)}$ from $P(y_1^{(t)}, ..., y_{j-1}^{(t)}, y_{j+1}^{(t-1)}, ..., y_k^{(t-1)} | y_j^{(t-1)})$,...,
$y_k^{(t)}$ from $P_k(y_1^{(t)}, ..., y_{k-1}^{(t)} | y_k^{(t-1)})$.

Example 4.10

Consider Example 4.9 (normal linear model).
Under the given prior for $\theta = (\beta, \omega)$, it is difficult to generate random vectors directly from the posterior p.d.f.

$$p(\theta) \propto \pi(\beta) \prod_{i=1}^{k} \omega_i^{n_i/2+\alpha} e^{-[\gamma^{-1}+v_i(\beta)]\omega_i},$$

which does not have a familiar form.
To apply a Gibbs sampler with $y = \theta$, $y_1 = \beta$, and $y_2 = \omega$, we need to generate random vectors from the posterior of β, given x and ω, and the posterior of ω, given x and β.

Example 4.10 (continued)

Since

$$p(\theta) \propto \pi(\beta) \prod_{i=1}^{k} \omega_i^{n_i/2+\alpha} e^{-[\gamma^{-1}+v_i(\beta)]\omega_i},$$

the posterior of $\omega = (\omega_1, ..., \omega_k)$, given x and β, is a product of marginals of ω_i's that are the gamma distributions $\Gamma(\alpha+1+n_i/2, [\gamma^{-1}+v_i(\beta)]^{-1})$, $i = 1, ..., k$.

Assume now that $\pi(\beta) \equiv 1$ (noninformative prior for β).

The posterior p.d.f. of β, given x and ω, is proportional to

$$\prod_{i=1}^{k} e^{-\omega_i v_i(\beta)} \propto e^{-\|W^{1/2}Z\beta - W^{1/2}X\|^2},$$

where W is the diagonal block matrix whose ith block is $\omega_i I_{n_i}$.

Let $n = \sum_{i=1}^{k} n_i$.

The posterior of $W^{1/2}Z\beta$, given X and ω, is $N_n(W^{1/2}X, 2^{-1}I_n)$ and the posterior of β, given X and ω, is $N_p((Z^\tau WZ)^{-1}Z^\tau WX, 2^{-1}(Z^\tau WZ)^{-1})$ ($Z^\tau WZ$ is assumed of full rank for simplicity), since $\beta = [(Z^\tau WZ)^{-1}Z^\tau W^{1/2}]W^{1/2}Z\beta$.

Random generation using these two posterior distributions is easy.