Introduction to Bayes Statistics

1 Definitions

For a sample X, denote the sample space by $(\mathcal{X}, \mathcal{B}_x)$, the family of distributions by $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ where $\mathcal{B}_\Theta$ is the σ-field generated by subsets of Θ, the action space by $(A, \mathcal{B}_A)$, the loss function by $L(\theta, a) \geq 0$ defined on $\Theta \times A$, which is a $\mathcal{B}_\Theta \times \mathcal{B}_A$-measurable function. For a decision function (or a decision rule) δ, its risk function is denoted by $R(\theta, \delta)$.

Definition 1 *Let Π be a probability measure defined on $(\Theta, \mathcal{B}_\Theta)$, i.e.* a prior distribution. *For a decision function δ, the Bayes risk with respect to a prior distribution Π is defined as follows:*

$$R_\Pi(\delta) = \int_\Theta R(\theta, \delta) d\Pi(\theta). \tag{1}$$

Consider a class of decision functions, denoted by $\mathcal{D}$. For example, $\mathcal{D}$ is a class of all possible decision functions or a class of all deterministic decision functions.

Definition 2 *δ_Π is said to be a Bayes solution (or Bayes rule) with respect to a prior Π if there exists $\delta_\Pi \in \mathcal{D}$ such that*

$$R_\Pi(\delta_\Pi) = \inf\{R_\Pi(\delta) : \delta \in \mathcal{D}\}.$$

In the problem of point estimation, a Bayes solution is often called a Bayes estimation.

Clearly, in the Bayes theory, both conditional distribution of X given θ, P_θ, and prior distribution of θ, Π, are given. In this case, we can calculate the joint distribution of (X, θ), denoted by P^*. For a set $C \in \mathcal{B}_\Theta \times \mathcal{B}_x$,

$$P^*(C) = \int_\Theta P_\theta(C_\theta) d\Pi(\theta), \tag{2}$$

where $C_\theta = \{x : x \in \mathcal{X}, (x, \theta) \in C\}$. Consequently, the marginal distribution of the sample X is given by

$$P(A) = P(X \in A) = P^*(\Theta \times A) = \int_\Theta P_\theta(A) d\Pi(\theta), A \in \mathcal{B}_x. \tag{3}$$

Now it is ready to define the most important concept in the Bayes theory, *posterior distribution* – the conditional distribution of θ given sample X.

Definition 3 *Given sample X, there exists a (regular) conditional distribution function $\Pi(\cdot|\cdot)$ defined on $\mathcal{B}_\Theta \times \mathcal{X}$ satisfying:*

1) *For any fixed $x \in \mathcal{X}$, $\Pi(\cdot|x)$ is a probability measure on $\mathcal{B}_\Theta$; for any given $B \in \mathcal{B}_\Theta$, $\Pi(B|\cdot)$ is $\mathcal{B}_x$-measurable.*

2) *For given $B \in \mathcal{B}_\Theta, C \in \mathcal{B}_x$, the joint probability of the Cartesian product $B \times C$ is*

$$P^*(B \times C) = \int_C \Pi(B|x) dP(x).$$

Then, $\Pi(\cdot|x)$ is called the posterior distribution of θ given x (with respect to prior Π).

In most of applied problems, space $(\Theta, \mathcal{B}_\Theta)$ is Euclidean, so the conditional distribution $\Pi(\cdot|\cdot)$ exists. Moreover, often there exists a σ-finite measure μ on $\mathcal{B}_x$ such that $\mathcal{P} = \{P_\theta, \theta \in \Theta\} \ll \mu$, then denote R-N derivative by $f(x|\theta) = dP_\theta/d\mu$. It follows that

$$\Pi(B|x) = \frac{\int_B f(x|\theta) d\Pi(\theta)}{\int_\Theta f(x|\theta) d\Pi(\theta)}, B \in \mathcal{B}_\Theta. \tag{4}$$

Furthermore, if there is a σ-finite measure ν on $\mathcal{B}_\Theta$ such that $\Pi \ll \nu$, then the resulting R-N derivative is $\pi(\theta) = d\Pi(\theta)/d\nu$. Then, for any $x \in \mathcal{X}$ $H(\cdot|x) \ll \nu$ holds and

$$\frac{d\Pi(\theta|x)}{d\nu} = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\psi)\pi(\psi) d\psi}. \tag{5}$$

Remark: Posterior distribution plays the key role in Bayes theory. However, its derivation is independent of action space and loss function. It represents a transition from prior knowledge about θ to posterior knowledge about θ when data x from model P_θ is given.

How to use posterior distribution to perform parameter estimation and statistical inference?

1) To obtain a point estimation of parameter θ, we may use the mode of posterior distribution $\pi(\theta|x)$ (in a similar spirit to MLE, the probable value given data), the mean of posterior distribution $\pi(\theta|x)$ (called the Bayes estimation), or the median of posterior distribution $\pi(\theta|x)$ (a robust estimation), and so on.

2) To obtain an interval estimation of parameter θ, it is desired to choose an interval of the shortest with a $(1-\alpha)$ posterior probability. In the case of a unimodal posterior distribution, the left and right limits of the interval $[a_x, b_x]$ may be determined by

$$\pi(a_x|x) = \pi(b_x|x), \quad \int_{a_x}^{b_x} \pi(\theta|x) d\theta = 1 - \alpha.$$

Note that here $(1-\alpha)$ cannot be interpreted as a confidence level in the Neyman's sense but it seems to be a reasonable way to calibrate the precision of post-data decision.

3) To perform a hypothesis test, say $H_0 : 0 < \theta \leq \theta_0$, one may calculate the posterior probability for the null hypothesis:

$$P(H_0|x) = \int_0^{\theta_0} \pi(\theta|x)d\theta.$$

The rejection rule may be set as follows: For a prespecified α, reject the null hypothesis H_0 when $P(H_0|x) < \alpha$. Although the posterior probability is in a similar spirit to p-value used in the Neyman-Pearsons theory of hypothesis testing, it is a very different concept (according to Efron, it is called *local false discovery rate.* We will come back to this later.)

1.1 Principle of Minimum Posterior Risk

Definition 4 *Let δ be a deterministic decision function. The posterior risk of δ conditional on an available sample x is defined by*

$$R_\Pi(\delta, x) = \int_\Theta L(\theta, \delta(x))\Pi(d\theta|x). \tag{6}$$

In the case of a randomized decision function $\delta(\cdot|\cdot)$, the posterior risk is defined by

$$R_\Pi(\delta, x) = \int_\Theta \left\{ \int_A L(\theta, a)\delta(da|x) \right\} \Pi(d\theta|x). \tag{7}$$

It is easy to see that for a deterministic decision rule, the Bayes risk is related to the posterior risk in a form given as follows:

$$\begin{aligned} R_\Pi(\delta) &= E_{P^*}\{L(\theta, \delta(X))\} \\ &= \int_{\mathcal{X}} \left\{ \int_\Theta L(\theta, \delta(x))\Pi(d\theta|x) \right\} dP(x) \\ &= \int_{\mathcal{X}} R_\Pi(\delta, x)dP(x). \end{aligned}$$

Likewise, it is easy to show that this relation holds for a randomized decision rule, too.

Theorem 1 *Suppose there exists a set $B \in \mathcal{B}_x$ with $P(B) = 0$ such that for $x \in B^c$ the minimum $\inf\{\int_\Theta L(\theta, a)\Pi(d\theta|x) : a \in A\}$ is attained at a certain $a = a_x \in A$ (a.s.). Define a (deterministic) decision function*

$$\delta(x) = \begin{cases} a_x, & \text{if } x \in B^c; \\ a_0, & \text{if } x \in B, \end{cases}$$

where $a_0 \in A$ is an arbitrary element in A. Suppose that δ is a measurable transformation from $(\mathcal{X}, \mathcal{B}_x)$ to $(A, \mathcal{B}_A)$. Then, the decision function above is the Bayes rule among the class of all decision functions (including both deterministic and randomized).

Proof Let $\mathcal{D}$ be the class of all decision functions, including both deterministic and randomized rules. For an arbitrary rule $\delta^*(\cdot|\cdot) \in \mathcal{D}$, we have for each $x \in B^c$,

$$\begin{aligned} R_\Pi(\delta^*, x) &= \int_\Theta \left\{ \int_A L(\theta, a)\delta^*(da|x) \right\} \Pi(d\theta|x) \\ &= \int_A \left\{ \int_\Theta L(\theta, a)\Pi(d\theta|x) \right\} \delta^*(da|x) \\ &\geq \int_A R_\Pi(\delta, x)\delta^*(da|x) \\ &= R_\Pi(\delta, x). \end{aligned}$$

Because of the arbitrariness of δ^* and $P(B) = 0$, so

$$R_\Pi(\delta^*) = \int_{\mathcal{X}} R_\Pi(\delta^*, x)dP(x) = \int_B + \int_{B^c} R_\Pi(\delta^*, x)dP(x) \geq R_\Pi(\delta).$$

This proves that the δ defined above (the minimizer of the posterior risk) is the Bayes rule. □

The theorem implies that to find the Bayes rule it is sufficient to consider ONLY the class of deterministic decision functions, as long as the extrema above exist. In fact, the attainability of extrema is a necessary condition, as stated in the following theorem.

Attainability Condition: Extremum $\inf\{\int_\Theta L(\theta, a)\Pi(d\theta|x) : a \in A\}$ is attained at almost every $x \in \mathcal{X}$ (with respect to the marginal probability measure $P(\cdot)$).

Theorem 2 *Suppose there exists a decision function δ satisfying $R_\Pi(\delta) < \infty$. Then δ is the Bayes rule if and only if the δ satisfies the attainability condition.*

Proof It is sufficient to prove that if the attainability condition is not satisfied there would not exist Bayes rule. Suppose there is a subset $\mathcal{X}_1 \subset \mathcal{X}$ with $P(\mathcal{X}_1) > 0$ on which the attainability condition does not hold by δ. On the other hand, $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1$ with $\mathcal{X}_0 = \mathcal{X}_1^c \subset \mathcal{X}$ on which the attainability condition holds. We show that it is always possible to construct a new decision function δ^* whose Bayes risk is smaller than that of δ. This construction of δ^* is easily done by two steps. First, for $x \in \mathcal{X}_0$, δ satisfies the attainability condition, so there exists an a_x that achieves $\inf\{\int_\Theta L(\theta, a)\Pi(d\theta|x) : a \in A\}$; that is,

$$\int_\Theta L(\theta, a_x)\Pi(d\theta|x) = R_\Pi(\delta, x) \leq R_\Pi(\delta', x), \forall \delta' \in \mathcal{D}$$

For $x \in \mathcal{X}_1$, δ does not satisfy the attainability condition, so

$$\int_\Theta L(\theta, \delta(x))\Pi(d\theta|x) > \inf\{\int_\Theta L(\theta, a)\Pi(d\theta|x) : a \in A\} \overset{def}{=} m(x),$$

or $R_\Pi(\delta, x) > m(x), x \in \mathcal{X}_1$. For a fixed x, being a continuous function in a, there always exists an a_x^* such that

$$R_\Pi(\delta, x) > \int_\Theta L(\theta, a_x^*)\Pi(d\theta|x) > m(x), x \in \mathcal{X}_1.$$

Then we define a new decision rule:

$$\delta^*(x) = \begin{cases} \delta(x), & \text{if } x \in \mathcal{X}_0; \\ a_x^* & \text{if } x \in \mathcal{X}_1, \end{cases}$$

It is clearly that $R_\Pi(\delta^*, x) \leq R_\Pi(\delta, x)$ where the inequality holds for $x \in \mathcal{X}_1$ with $P(\mathcal{X}_1) > 0$. Thus, $R_\Pi(\delta^*) \leq R_\Pi(\delta)$. This means that there exists a decision rule whose Bayes risk is not larger than that of the given rule δ. In other words, the given rule δ is not the Bayes solution. □

In summary, technically, the derivation of the Bayes solution or Bayes rule is to search for the minimizer of the posterior risk with over the action space A.

1.2 Bayes Method and Sufficient Statistic

Bayes theory provides a different statistical principle to make parameter estimation and inference. It should be compatible with other fundamental statistical principles, such as sufficiency. Sufficiency represents a data reduction principle in the sense that if there is a sufficient statistic $T(x)$ for parameter θ, statistical inference can based completely on $T(x)$.

Theorem 3 *Consider a distribution family $\mathcal{P} = (P_\theta, \theta \in \Theta) \ll \mu$, where μ is a σ-finite measure on $\mathcal{B}_x$. Suppose $T(x)$ is a sufficient statistic. Then the posterior distribution depends on x only through $T(x)$ in a form of $\tilde{\Pi}(\cdot|T(x))$, which is the conditional distribution of θ given $T(x) = t$, i.e. $\tilde{\Pi}(\cdot|t)$.*

Proof Denote the R-N derivative by $f(x|\theta) = dP_\theta(x)/d\mu$. According to the Factorization Theorem, we have

$$f(x|\theta) = g(t, \theta)h(x), \ a.s.P_\theta,$$

where $0 < h(x) < \infty, x \in \mathcal{X}$. Let

$$C = \left\{ t : \int_\Theta g(t, \theta)d\Pi(\theta) = 0 \text{ or } \infty \right\}.$$

It follows from (4) that for $B \in \mathcal{B}_\Theta$,

$$\begin{aligned} \Pi(B|x) &= \begin{cases} \int_B g(t, \theta)d\Pi(\theta) / \int_\Theta g(t, \theta)d\Pi(\theta) & \text{if } t = T(x) \notin C, \\ \Pi(B) & \text{if } t = T(x) \in C \end{cases} \\ &\equiv \tilde{\Pi}(\cdot|t). \end{aligned}$$

Apparently the posterior distribution depends only on $t = T(x)$. $\square$

The implication of this theorem is clear: based on x or sufficient statistic $t = T(x)$ will give the same Bayes solution.

1.3 Bayes Estimation under Squared Error Loss and Absolute Error Loss

As pointed out above, technically, deriving the Bayes solution or Bayes rule is equivalent to searching for the minimizer of the posterior risk over the action space. Obviously, the posterior risk depends on the choice of loss function. Under two importance cases of loss functions, squared error loss and absolute error loss, the Bayes solution may be given explicitly.

1.3.1 Bayes Solution under Squared Error Loss

Let g be a real-value function defined on the parameter space $\Theta \subset R$. Let the action space be $A = (-\infty, \infty)$. Consider the squared error loss function of the form:

$$L(\theta, a) = \lambda(\theta)(g(\theta) - a)^2, \tag{8}$$

where $0 < \lambda(\theta) < \infty$. The objective is to determine a point estimate of $g(\theta)$. The posterior risk is

$$R_\Pi(a, x) = \int_\Theta \lambda(\theta)(g(\theta) - a)^2 \Pi(d\theta|x).$$

The Bayes estimator of $g(\theta)$ is denoted by $\delta_\Pi(x) = \arg\min_{a \in A} R_\Pi(a, x)$. Then

$$\delta_\Pi(x) = \begin{cases} \frac{\int_\Theta \lambda(\theta) g(\theta) \Pi(d\theta|x)}{\int_\Theta \lambda(\theta)\Pi(d\theta|x)}, & \text{if } R_\Pi(a,x) < \infty \text{ for all } a \in A; \\ a(x), & \text{if } \exists \text{ one point } a(x) \text{ such that } R_\Pi(a(x), x) < \infty \\ & \text{but } R_\Pi(a,x) = \infty \text{ for all } a \neq a(x); \\ a^*, & \ R_\Pi(a,x) = \infty \text{ for all } a \in A, \end{cases}$$

where a^* is an arbitrary finite real number. Note that the first term is the expectation of $g(\theta)$ under the following probability measure on $(\Theta, \mathcal{B}_\Theta)$:

$$\eta(B|x) = \frac{\int_B \lambda(\theta)\Pi(d\theta|x)}{\int_\Theta \lambda(\theta)\Pi(d\theta|x)}, \ B \in \mathcal{B}_\Theta, \tag{9}$$

provided that $\int_\Theta \lambda(\theta)\Pi(d\theta|x) < \infty$.

Exercise: Can you give examples for the second and the third cases?

Lemma 1 *(Blackwell-Girshick Lemma) Let L be a squared error loss function given in (8), and let ν be a probability measure on $(\Theta, \mathcal{B}_\Theta)$. Consider a function*

$$f(a) = \int_\Theta L(\theta, a) d\nu(\theta), \ a \in (-\infty, \infty).$$

Suppose there exist $a_1, a_2 \in (-\infty, \infty)$ *such that* $f(a_1) < \infty$ *and* $f(a_2) < \infty$. *Then for all* $a \in (-\infty, \infty), f(a) < \infty$ *and* $\int_\Theta \lambda(\theta) d\nu(\theta) < \infty$.

The proof of this lemma is left for you as an optional exercise. This lemma holds in the case of the absolute error loss.

1.3.2 Bayes Solution under Absolute Error Loss

Consider the absolute error loss function of the form:

$$L(\theta, a) = \lambda(\theta)|g(\theta) - a|, \tag{10}$$

where $0 < \lambda(\theta) < \infty$. The objective is to determine a point estimate of $g(\theta)$. The posterior risk is

$$R_\Pi(a, x) = \int_\Theta \lambda(\theta)|g(\theta) - a|\Pi(d\theta|x).$$

The Bayes estimator of $g(\theta)$ is denoted by $\delta_\Pi(x) = \arg\min_{a \in A} R_\Pi(a, x)$. Then

$$\delta_\Pi(x) = \begin{cases} \text{median of } g(\theta) \text{ under } \theta \sim \eta(\cdot|x) \text{ in (9)}, & \text{if } R_\Pi(a, x) < \infty \text{ for all } a \in A; \\ a(x), & \text{if } \exists \text{ one point } a(x) \text{ such that } R_\Pi(a(x), x) < \infty \\ & \text{ but } R_\Pi(a, x) = \infty \text{ for all } a \neq a(x); \\ a^*, & R_\Pi(a, x) = \infty \text{ for all } a \in A, \end{cases}$$

where a^* is an arbitrary finite real number.

Lemma 2 *Let* F *be the distribution of a one-dimensional random variable* Z, *and* $E(|Z|) < \infty$. *Define an objective function:*

$$G(\rho) = q \int_{-\infty}^{\rho} (\rho - x) dF(x) + p \int_{\rho}^{\infty} (x - \rho) dF(x), \rho \in (-\infty, \infty),$$

where $q = 1 - p, p \in (0, 1)$. *Then, the minimum of* $G(\rho)$ *is attained at the* p*-th percentile for* Z.

1.4 Some Examples

Example 1 Suppose that $X \sim B(n, \theta), 0 \leq \theta \leq 1$ and that the prior distribution of θ is $U(0, 1)$ (a representation of prior ignorance or the same chance for each instance). Consider the squared error loss function $L(\theta, a) = (\theta - a)^2$. We want to find the Bayes estimator of θ.

It is easy to show that the posterior distribution of θ is a beta distribution $Beta(x + 1, n - x + 1)$. Under the squared error loss function, the expectation of the posterior is the Bayes estimator; that is

$$\delta_\Pi(x) = \frac{x + 1}{n + 2}.$$

Its risk function is

$$R(\theta, \delta_\Pi) = E_\theta \left(\frac{x+1}{n+2} - \theta \right)^2 = \frac{n\theta(\theta-1) + (1-2\theta)^2}{(n+2)^2}.$$

It follows that the Bayes risk of δ_Π is

$$R_\Pi(\delta_\Pi) = \int_0^1 R(\theta, \delta_\Pi) d\theta = \frac{1}{6(n+2)}.$$

It is interesting to note that the Bayes risk of the most commonly used estimator $\hat{\theta} = \frac{X}{n}$ is $\frac{1}{6n}$, which is bigger than the Bayes estimator δ_Π above. Another interesting comparison between the two estimators is that when sample $x = 0$ or n, $\hat{\theta} = 0$ or 1, while $\delta_\Pi = 1/(n+2)$ or $(n+1)/(n+2)$, close to 0 or 1, but not on the boundaries of the parameter space. In such extreme cases, the Bayes estimator seems to give a more reasonable solution than the popular estimator $\hat{\theta}$.

Now we generalize the above result using prior distribution $\theta \sim Beta(a, b)$ with the two hyperparameters $a, b > 0$. It is easy to show that in this case the posterior distribution of θ given sample x is $\theta | x \sim Beta(a + x, n + b - x)$. Note that the uniform distribution is a special case of beta distribution with $a = b = 1$. Under the squared error loss function, we obtain the Bayes estimator of θ as

$$\delta_{a,b}(x) = \frac{x+a}{n+a+b}.$$

Also, the risk function of $\delta_{a,b}$ is given by

$$R(\theta, \delta_{a,b}) = \sum_{i=1}^{n} C_n^i \theta^i (1-\theta)^{n-i} \left(\frac{i+a}{n+a+b} - \theta \right)^2 = \frac{n\theta(1-\theta) + \{(a+b)\theta - a\}^2}{(n+a+b)^2}.$$

Moreover the Bayes risk is

$$R_\Pi(\delta_{a,b}) = \int_0^1 R(\theta, \delta_{a,b}) \pi(\theta) d\theta = \frac{ab}{(n+a+b)(a+b+1)(a+b)}.$$

For example, when $a = b = \sqrt{n}/2$ (a beta distribution with mean 1/2 and variance converging to zero when $n \to \infty$), we have

$$\delta_{\sqrt{n}/2, \sqrt{n}/2} = \frac{x}{n+\sqrt{n}} + \frac{\sqrt{n}}{2(n+\sqrt{n})},$$

and the corresponding risk function is

$$R(\theta, \delta_{\sqrt{n}/2, \sqrt{n}/2}) = \frac{n}{4(n+\sqrt{n})^2}.$$

It is easy to see that when n is large, $\delta_{\sqrt{n}/2,\sqrt{n}/2}$ has a smaller Bayes risk than $\delta_{1,1}$ (the one obtained under the uniform prior).

In this example, we observed that the beta distribution has the following property: the posterior distribution and the prior distribution belong to the same distribution family, the beta distribution family; in other words, the posterior distribution falls in the same distribution family that the prior distribution belongs to. Such distribution family for the prior is called *the family of conjugate prior distributions.* This is an important property in the Bayes theory that we will come back to this point later.

Example 2 Let $X = (X_1, \ldots, X_n)$ with $X_1, \ldots, X_n$ i.i.d. $N(\theta, 1)$. Suppose the prior $\theta \sim N(0, \tau^2)$. Consider the squared error loss function $L(\theta, a) = (\theta - a)^2$. We want to derive the Bayes estimator of θ.

Denote $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$. It is easy to express the joint density of $(\theta, X_1, \ldots, X_n)$ as follows:

$$f(\theta, X) = \frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{\theta^2}{2\tau^2}\right)\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - \theta)^2\right).$$

Noting that

$$\frac{\theta^2}{\tau^2} + \sum_{i=1}^n (x_i - \theta)^2 \propto \frac{1 + n\tau^2}{\tau^2}\left(\theta - \frac{n\tau^2\bar{x}}{1 + n\tau^2}\right)^2,$$

we know that the posterior distribution of θ given x is

$$\theta | x \sim N\left(\frac{n\tau^2\bar{x}}{1 + n\tau^2}, \frac{\tau^2}{1 + n\tau^2}\right).$$

Again, here we see in this example that the normal distribution family is also the family of congjugate prior distributions. Under the squared error loss function, the Bayes estimator is

$$\delta_\tau(x) = \frac{n\tau^2\bar{x}}{1 + n\tau^2}.$$

It is easy to calculate its risk function equal to

$$R(\theta, \delta_\tau) = \frac{n\tau^4 + \theta^2}{(1 + n\tau^2)^2}.$$

Moreover the Bayes risk is given by

$$R_\Pi(\delta_\tau) = \frac{\tau^2}{1 + n\tau^2}.$$

It is interesting to notice that the Bayes estimator δ_τ is a shrinkage estimator of the sample mean $\bar{x}$, where the shrinkage factor is $\frac{n\tau^2}{1+n\tau^2}$. It is easy to show that the Bayes risk of the sample mean $\bar{X}$ is $\frac{1}{n}$, which is larger than the above $R_\Pi(\delta_\tau)$.

- The prior distribution $N(0, \tau^2)$ essentially claims *aprior* that the true value of θ is most likely to occur at zero. Thus, it is natural to drag the sample mean $\bar{x}$ towards 0; and the smaller τ^2 the more shrinkage to zero.

- When $\tau^2 \to \infty$, the prior distribution becomes flat, so the prior knowledge about θ is no longer informative (such prior is called *a diffusion prior*). Thus, the sample mean $\bar{x}$ is the solution.

- When the sample size $n \to \infty$, the amount of data information exceeds over the prior knowledge (which is of little influence), so the sample mean $\bar{x}$ is the solution.

- In either case, $\tau^2 \to \infty$ or $n \to \infty$, the Bayes risk of δ_τ converges to $\frac{1}{n}$, the Bayes risk of the sample mean.

Example 3 Let $X = (X_1, \dots, X_n)$ with $X_1, \dots, X_n$ i.i.d. $Exp(\theta) = Ga(\theta, 1)$ with the density function of a gamma distribution $Ga(\alpha, \beta)$

$$\frac{\alpha^\beta}{\Gamma(\beta)} x^{\beta-1} e^{-\alpha x}, \; x > 0; \alpha, \beta > 0.$$

Exponential distribution is one of the simplest distribution to model a life time variable. Suppose the prior $\theta \sim Ga(\tau^{-1}, \nu)$ with two hyperparameters $\tau, \nu > 0$. We like to obtain the Bayes estimator of $g(\theta) = P_\theta(X_1 > a) = e^{-\theta a}$ (the survival probability at a) under the squared error loss function.

Denote $t_n = \sum_{i=1}^n x_i$. The joint density of $(\theta, X_1, \dots, X_n)$ is given by

$$\theta^n e^{-nt_n} \left(\Gamma(\nu)\tau\right)^{-1} e^{-\theta/\tau} \theta^{\nu-1}.$$

Thus, the posterior distribution of θ given sample x is

$$\theta | x \equiv \theta | t_n \sim Ga\left(t_n + \tau^{-1}, n + \nu\right).$$

Then the Bayes estimator of $g(\theta)$ is the expectation of $g(\theta)$ under the above posterior distribution:

$$\delta_{\tau,\nu}(x) = \int_0^\infty g(\theta)\pi(\theta|x) d\theta = \left(1 + \frac{a}{t_n + \tau^{-1}}\right)^{-(n+\nu)}.$$

The further discussion on the Bayes risk and properties is left to exercises.

1.5 Family of Conjugate Prior Distributions

Definition 5 *Let $\mathcal{H}$ be a family of probability distributions on the parameter space $(\Theta, \mathcal{B}_\Theta)$. $\mathcal{H}$ is said to be a family of conjugate prior distributions if for each $\Pi \in \mathcal{H}$ and a given sample x, it holds that $\Pi(\cdot|x) \in \mathcal{H}$.*

Remark: The family of conjugate prior distributions, mathematically speaking, always exists in any problem. For example, $\mathcal{H}$ is the family of all possible distributions on $(\Theta, \mathcal{B}_\Theta)$. However, such $\mathcal{H}$ is useless for a given problem. It is hoped in a practical problem that such family may be more specific and contents relevant. One of the most popular ways to construct the family of conjugate prior distributions is by the means of sufficient statistics.

Let $X_1, \ldots, X_n$ be i.i.d. random variables, where for each X_i the (same) probability space is $(\mathcal{X}, \mathcal{B}_x, \mathcal{P})$ with $\mathcal{P} = (P_\theta, \theta \in \Theta) \ll \mu$, and μ is a σ-finite measure on $\mathcal{B}_x$. Denote $f_\theta(x) = dP_\theta(x)/d\mu$.

Denote $\mathbf{X}_n = (X_1, \ldots, X_n)$. Then the probability space of $\mathbf{X}_n$ is $(\mathcal{X}^{(n)}, \mathcal{B}_x^{(n)}, \mathcal{P}^{(n)})$ where

$$\mathcal{X}^{(n)} = \mathcal{X} \times \cdots \times \mathcal{X},\ \mathcal{B}^{(n)} = \mathcal{B} \times \cdots \times \mathcal{B},\ \mathcal{P}^{(n)} = \mathcal{P} \times \cdots \times \mathcal{P}.$$

The following assumptions are required.

1) For any sample size n, there exists a sufficient statistic $T_n = T_n(\mathbf{X}_n)$ with respect to the probability space $(\mathcal{X}^{(n)}, \mathcal{B}_x^{(n)}, \mathcal{P}^{(n)})$.

2) For all n, all $\theta \in \Theta$ and all $\mathbf{x}_n \in \mathcal{X}^{(n)}$, the factorization theorem holds

$$\prod_{i=1}^n f_\theta(x_i) = g_n(T_n(x_1, \ldots, x_n), \theta) h_n(x_1, \ldots, x_n),$$

where $h_n(x_1, \ldots, x_n) > 0$.

3) There exists a σ-finite measure ν on $\mathcal{B}_\Theta$ such that for all $\mathbf{x}_n \in \mathcal{X}^{(n)}$ and $n = 1, 2, \ldots$, it holds that

$$0 < \int_\Theta g_n(T_n(x_1, \ldots, x_n), \theta) d\nu(\theta) < \infty.$$

Now we construct a family of distributions on $\mathcal{B}_\Theta$: $\mathcal{H} = \{\Pi(\cdot|\mathbf{x}_n) : \mathbf{x}_n \in \mathcal{X}^{(n)}, n = 1, 2, \ldots\}$ where

$$\Pi(B|\mathbf{x}_n) = \frac{\int_B g_n(T_n(x_1, \ldots, x_n), \theta) d\nu(\theta)}{\int_\Theta g_n(T_n(x_1, \ldots, x_n), \theta) d\nu(\theta)}. \tag{11}$$

Theorem 4 *Suppose the the above conditions 1)-3) hold. Then the family $\mathcal{H}$ defined in (11) is a family of conjugate prior distributions.*

Proof Take an arbitrary element from $\mathcal{H}$ as a prior. This prior may be expressed as follows:

$$\pi(\theta) d\nu(\theta) \overset{def}{=} \frac{g_m(T_m(x_1^0, \ldots, x_m^0), \theta) d\nu(\theta)}{\int_\Theta g_m(T_m(x_1^0, \ldots, x_m^0), \psi) d\nu(\psi)},$$

for some m and a certain given sample $\mathbf{x}_m^0 = (x_1^0, \ldots, x_m^0)$.

Now given a sample $\mathbf{x}_n = (x_1, \ldots, x_n)$ at hand, the posterior distribution is given by

$$\pi(\theta|x_1, \ldots, x_n)d\nu(\theta) = C_1(\mathbf{x}_m^0, \mathbf{x}_n)g_m(T_m(x_1^0, \ldots, x_m^0), \theta)\prod_{i=1}^n f_\theta(x_i)d\nu(\theta),$$

where C_1 denotes a certain normalization constant independent of the parameter θ. The same notation applies in the following arguments, too. Applying the Factorization theorem on the $\prod_{i=1}^n f(x_i)$, we have

$$\pi(\theta|\mathbf{x}_n) = C_2(\mathbf{x}_m^0, \mathbf{x}_n)g_m(T_m(x_1^0, \ldots, x_m^0), \theta)g_n(T_n(x_1, \ldots, x_n), \theta).$$

On the other hand, applying the Factorization theorem on the joint sample of $\mathbf{x}_m^0$ and $\mathbf{x}_n$, we have

$$\prod_{i=1}^m f_\theta(x_i^0)\prod_{i=1}^n f_\theta(x_i) = g_{m+n}(T_{m+n}(x_1^0, \ldots, x_m^0, x_1, \ldots, x_n), \theta)h_{m+n}(x_1^0, \ldots, x_m^0, x_1, \ldots, x_n).$$

It follows that

$$\pi(\theta|\mathbf{x}_n)d\nu(\theta) = C_3(\mathbf{x}_m^0, \mathbf{x}_n)g_{m+n}(T_{m+n}(\mathbf{x}_m^0, \mathbf{x}_n), \theta)d\nu(\theta).$$

Because $\int_\Theta \pi(\theta|\mathbf{x}_n)d\nu(\theta) = 1$, we have

$$C_3(\mathbf{x}_m^0, \mathbf{x}_n) = \left\{\int_\Theta g_{m+n}(T_{m+n}(\mathbf{x}_m^0, \mathbf{x}_n), \theta)d\nu(\theta)\right\}^{-1}.$$

Thus, the posterior density is given by

$$\pi(\theta|\mathbf{x}_n)d\nu(\theta) = \frac{g_{m+n}(T_{m+n}(\mathbf{x}_m^0, \mathbf{x}_n), \theta)d\nu(\theta)}{\int_\Theta g_{m+n}(T_{m+n}(\mathbf{x}_m^0, \mathbf{x}_n), \psi)d\nu(\psi)},$$

which belongs to the family $\mathcal{H}$ according to the definition of the family (11). □

Let us look at some examples.

Example 4 Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli distribution with probability of success θ, $\theta \in [0, 1]$. Consider a Lebsegue measure $d\theta$ for $d\nu(\theta)$ on the parameter space $([0, 1], \mathcal{B}_{[0,1]})$. It is easy to see that

$$\prod_{i=1}^n f_\theta(x_i) = \theta^{T_n}(1-\theta)^{n-T_n},$$

where $T_n = \sum_{i=1}^n x_i$ is a sufficient statistic. From the above expression, we identify

$$g_n(T_n(\mathbf{x}_n), \theta) = \theta^{T_n}(1-\theta)^{n-T_n}$$

and $h_n(\mathbf{x}_n) = 1$. According to (11), the family of conjugate prior distributions will be the density generated by g_n with respect to θ. Noting that the above g_n represents a kernel of beta distribution, we can immediately know that the family of beta distributions $Beta(\alpha, \beta), \alpha, \beta > 0$ will be the family of conjugate prior distributions.

Example 5 Let $X_1, \ldots, X_n$ be i.i.d. Poisson distribution with mean parameter θ, $\theta > 0$. Consider a Lebsegue measure $d\theta$ for $d\nu(\theta)$ on the parameter space $((0,\infty), \mathcal{B}_{(0,\infty)})$. First, look at

$$\prod_{i=1}^{n} f_\theta(x_i) = e^{-n\theta}\theta^{T_n}/(x_1! \cdots x_n!),$$

where $T_n = \sum_{i=1}^{n} x_i$ is a sufficient statistic. Second, we identify

$$g_n(T_n(\mathbf{x}_n), \theta) = e^{-n\theta}\theta^{T_n}, \theta > 0$$

and $h_n(\mathbf{x}_n) = 1/(x_1! \cdots x_n!)$. Third, knowing the above g_n represents a kernel of gamma distribution in θ, we conclude that the family of gamma distributions $Ga(\alpha, \beta), \alpha, \beta > 0$ is the family of conjugate prior distributions.

Example 6 Let $X_1, \ldots, X_n$ be i.i.d. normal distribution $N(\theta, \sigma^2)$, with mean parameter θ, $-\infty < \theta < \infty$ and known variance parameter σ^2. Consider a Lebsegue measure $d\theta$ for $d\nu(\theta)$ on the parameter space $((-\infty,\infty), \mathcal{B}_{(-\infty,\infty)})$. First, look at

$$\prod_{i=1}^{n} f_\theta(x_i) = (\sqrt{2\pi}\sigma)^{-n} \exp\left\{-\sum_{i=1}^{n} x_i^2/(2\sigma^2) + nT_n/(2\sigma^2)\right\} \exp\left\{-\frac{n}{2\sigma^2}(\theta - T_n)^2\right\},$$

where $T_n = n^{-1}\sum_{i=1}^{n} x_i$ is a sufficient statistic. Second, we identify

$$\begin{aligned} g_n(T_n(\mathbf{x}_n), \theta) &= \exp\left\{-\frac{n}{2\sigma^2}(\theta - T_n)^2\right\}, \theta \in (-\infty, \infty) \\ h_n(\mathbf{x}_n) &= (\sqrt{2\pi}\sigma)^{-n} \exp\left\{-\sum_{i=1}^{n} x_i^2/(2\sigma^2) + nT_n/(2\sigma^2)\right\}. \end{aligned}$$

Third, knowing the above g_n represents a kernel of normal distribution $N(T_n, \sigma^2/n)$ for θ, and generalizing it we conclude that the family of normal distributions $N(\alpha, \beta^2), \alpha \in (-\infty, \infty), \beta > 0$ is the family of conjugate prior distributions.

Example 7 Let $X_1, \ldots, X_n$ be i.i.d. normal distribution $N(\theta, \sigma^2)$, with known mean parameter θ, and unknown variance parameter σ^2, $\sigma > 0$. Consider a Lebsegue measure $d\sigma$ for $d\nu(\sigma)$ on the parameter space $((0,\infty), \mathcal{B}_{(0,\infty)})$. First, look at

$$\prod_{i=1}^{n} f_\sigma(x_i) = (\sqrt{2\pi}\sigma)^{-n} \exp\left\{-T_n/(2\sigma^2)\right\},$$

where $T_n = \sum_{i=1}^{n}(x_i - \theta)^2$ is a sufficient statistic. Second, we identify

$$\begin{aligned} g_n(T_n(\mathbf{x}_n), \sigma) &= (\sigma^2)^{-n/2} \exp\left\{-\frac{T_n}{2\sigma^2}\right\}, \sigma \in (0, \infty) \\ h_n(\mathbf{x}_n) &= (\sqrt{2\pi})^{-n}. \end{aligned}$$

Third, knowing the above g_n represents a kernel of gamma distribution for $1/\sigma^2$, and generalizing it we conclude that the family of inverse gamma distributions $IG(\alpha, \beta), \alpha, \beta > 0$ is the family of conjugate prior distributions.

1.6 Empirical Bayes Method

Empirical Bayes method is attributed to Herbert Robbins who first proposed the method in 1955. This method was proposed to deal with the specification of prior distribution in the Bayes method when some historical data are available to determine an objective prior, instead of specifying prior by certain subjective assumption.

Suppose the probability space of X is $(\mathcal{X}, \mathcal{B}_x, P_\theta, \theta \in \Theta)$. Let the action space and loss function be $(A, \mathcal{B}_A)$ and $L(\theta, a)$, respectively. Suppose that in the past we had collected historical data of X, say x_i being the sample collected at the i-th experiment conducted in the past according to P_{θ_i}. For example, θ_i may be the rate of plates with flaws in the i-th inspection, and x_i was the observed number of plates with flaws in that inspection. Under the assumption that the prior distribution is Π, $\theta_1, \ldots, \theta_n$ may be regarded as an *i.i.d.* "sample" drawn from Π. Of course, we do not know $\theta_1, \ldots, \theta_n$; instead, we observed $x_1, \ldots, x_n$ that should carry on some information of θ_i's as well as that of Π.

At the current experiment, we observe x from P_θ. Then, it seems natural to establish a decision rule $\delta_n = \delta(x|x_1, \ldots, x_n)$. Conditional on the historical data $x_1, \ldots, x_n$, the Bayes risk is given by

$$\begin{aligned} R_\Pi(\delta_n|x_1, \ldots, x_n) &= \int_{\mathcal{X}\times\Theta} L(\theta, \delta(x|x_1, \ldots, x_n))dP^*(x, \theta) \\ &= \int_\Theta \left\{ \int_{\mathcal{X}} L(\theta, \delta_n(x|x_1, \ldots, x_n))dP_\theta(x) \right\} d\Pi(\theta). \end{aligned}$$

Accounting for the uncertainty from the historical data, the overall Bayes risk is given by

$$R^*_\Pi(\delta_n) = \int_{\mathcal{X}\times\cdots\times\mathcal{X}} R_\Pi(\delta_n|x_1, \ldots, x_n)dP(x_1)\cdots dP(x_n).$$

With given $x_1, \ldots, x_n$, $\delta_n(x|x_1, \ldots, x_n)$ is a decision rule under the prior Π. Thus,

$$R_\Pi(\delta_n|x_1, \ldots, x_n) \geq R_\Pi(\delta_\Pi).$$

This implies that

$$R^*_\Pi(\delta_n) \geq R_\Pi(\delta_\Pi).$$

In other words, the overall Bayes risk of any EB decision rule is not smaller than the Bayes risk of the Bayes rule.

Definition 6 *A decision rule $\delta_n = \delta_n(x|x_1, \ldots, x_n)$ that depends on both current data x and historical data $x_1. \ldots, x_n$ is termed as an empirical Bayes decision function (or rule). If for any prior $\Pi \in \mathcal{H}$,*

$$\lim_{n\to\infty} R^*_\Pi(\delta_n) = R_\Pi(\delta_\Pi),$$

then δ_n is called an asymptotically optimal empirical Bayes decision function.

Example 8 Consider $X \sim N(\theta, 1)$, loss function $L(\theta, a) = (\theta - a)^2$, and the prior distribution of θ is $\mathcal{H} = \{N(0, \tau^2), \tau > 0\}$. Let $x_1, \ldots, x_n$ be the *i.i.d.* historical data. Since the marginal distriubtion of X_i is $N(0, 1 + \tau^2)$, then the estimate of τ^2 is

$$\hat{\tau}^2_n = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 1.$$

Now with given current sample x, under the prior distribution of $N(0, \hat{\tau}^2_n)$, according to Example 2, the empirical Bayes estimator is

$$\delta_n(x|x_1, \ldots, x_n) = \frac{\hat{\tau}^2_n}{1 + \hat{\tau}^2_n} x, \tag{12}$$

and its Bayes risk conditional on $x_1, \ldots, x_n$ is

$$R_\Pi(\delta_n|x_1, \ldots, x_n) = \frac{\hat{\tau}^2_n}{1 + \hat{\tau}^2_n}.$$

By the Law of Large Number, $\tau^2_n \xrightarrow{p} (1 + \tau^2) - 1 = \tau^2$, then

$$\lim_{n\to\infty} R_\Pi(\delta_n|x_1, \ldots, x_n) = \frac{\tau^2}{1 + \tau^2}.$$

This is the Bayes risk of the Bayes solution given in Example 2. This means that the empirical EB rule above is asymptotically optimal.

1.7 James-Stein Estimator

Now we consider many versions of Example 2,

$$X_i|\theta_i \sim N(\theta_i, 1) \text{ and } \theta_i \sim N(0, \tau^2),\ i = 1, \ldots, m,$$

where the (θ_i, X_i) pairs are independent each other. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\top$ and $\mathbf{X} = (X_1, \ldots, X_m)^\top$. Then,

$$\mathbf{X}|\boldsymbol{\theta} \sim N_m(\boldsymbol{\theta}, \mathbf{I}), \text{ and } \boldsymbol{\theta} \sim N_m(0, \tau^2\mathbf{I}),$$

where $\mathbf{I}$ is the $m \times m$ identity matrix. By the results of Example 2 componentwise, it is easy to obtain the posterior distribution

$$\boldsymbol{\theta}|\mathbf{X} \sim N_m\left(\frac{\tau^2}{1+\tau^2}\mathbf{X}, \frac{\tau^2}{1+\tau^2}\mathbf{I}\right).$$

Moreover, under the total squared loss error function $L(\boldsymbol{\theta}, \mathbf{a}) = ||\boldsymbol{\theta} - \mathbf{a}||_2^2 = \sum_{i=1}^m (\theta_i - a_i)^2$, the Bayes estimator is a shrinkage estimator given by

$$\boldsymbol{\delta}_\Pi = \frac{\tau^2}{1+\tau^2}\mathbf{X} = \left(1 - \frac{1}{1+\tau^2}\right)\mathbf{X},$$

and its Bayes risk is

$$R_\Pi(\boldsymbol{\delta}_\Pi) = m\frac{\tau^2}{1+\tau^2}.$$

On the other hand the MLE of $\boldsymbol{\theta}$ is $\boldsymbol{\delta}_{mle} = \mathbf{X}$, and its total risk is

$$R(\boldsymbol{\delta}_{mle}) = m.$$

In conclusion, if the normal prior is correct then the Bayes rule $\boldsymbol{\delta}_\Pi$ offers substantial savings of the risk than the MLE,

$$R(\boldsymbol{\delta}_{mle}) - R_\Pi(\boldsymbol{\delta}_\Pi) = \frac{m}{1+\tau^2},$$

which increases to infinity as the dimension m tends to infinity.

Question: What if the prior is not correct?

The answer was provided by James and Stein in 1961 through the means of empirical Bayes. From Example 8, we know marginally $(X_1, \ldots, X_m) \sim N_m(0, (1+\tau^2)\mathbf{I})$. Then, it is easy to show that

$$S = \sum_{i=1}^m X_i^2 \sim (1+\tau^2)\chi_m^2,$$

so that

$$E\left(\frac{m-2}{S}\right) = \frac{1}{1+\tau^2}$$

The famous James-Stein estimator is defined to be

$$\boldsymbol{\delta}_{JS} = \left(1 - \frac{m-2}{S}\right)\mathbf{X}. \qquad (13)$$

It is not difficult to show that the overall Bayes risk of the above James-Stein estimator is

$$R(\boldsymbol{\delta}_{JS}) = m\frac{\tau^2}{1+\tau^2} + \frac{2}{1+\tau^2}.$$

With no surprise, it is bigger than the Bayes risk, but the difference is modest

$$\frac{R(\boldsymbol{\delta}_{JS})}{R_\Pi(\boldsymbol{\delta}_\Pi)} = 1 + \frac{2}{m\tau^2} \to 1 \text{ if } m \to \infty \text{ and/or } \tau^2 \to \infty.$$

More importantly, when $m > 2$ **it is easy to show that** $R(\boldsymbol{\delta}_{JS}) < R(\boldsymbol{\delta}_{mle})$**.** This implies that the MLE is an inadmissible estimator when the dimension of $\boldsymbol{\theta}$, m, is bigger than 2.

Theorem 5 *For* $m \geq 3$*, the James-Stein estimator everywhere dominates the MLE in terms of expected total squared error; that is*

$$E_{\boldsymbol{\theta}}\{||\boldsymbol{\delta}_{JS} - \boldsymbol{\theta}||_2^2\} < E_{\boldsymbol{\theta}}\{||\boldsymbol{\delta}_{mle} - \boldsymbol{\theta}||_2^2\}, \ \forall \boldsymbol{\theta} \in R^m.$$

Although the theorem is presented using the frequentist terms, it may be viewed by the decision theory point of view from the perspective of the overall risk.

Proof First, for a vector of actions, $\mathbf{a} = (a_1, \ldots, a_m)^\top = \mathbf{a}(\mathbf{x})$,

$$(a_i - \theta_i)^2 = (x_i - a_i)^2 - (x_i - \theta_i)^2 + 2(a_i - \theta_i)(x_i - \theta_i).$$

Summing over $i = 1, \ldots, m$ and taking expectation with respect to the conditional distribution of $\mathbf{X}|\boldsymbol{\theta}$, we have

$$\begin{aligned} E_{\boldsymbol{\theta}}||\mathbf{a} - \boldsymbol{\theta}||_2^2 &= E_{\boldsymbol{\theta}}||\mathbf{X} - \mathbf{a}||_2^2 - \sum_{i=1}^m Var_{\theta_i}(X_i) + 2\sum_{i=1}^m Cov_{\boldsymbol{\theta}}(a_i, X_i) \\ &= E_{\boldsymbol{\theta}}||\mathbf{x} - \mathbf{a}||_2^2 - m + 2\sum_{i=1}^m Cov_{\boldsymbol{\theta}}(a_i, X_i) \end{aligned}$$

It is easy to show that

$$Cov_{\boldsymbol{\theta}}(a_i, X_i) = E_{\boldsymbol{\theta}}\frac{\partial a_i}{\partial X_i}.$$

In effect,

$$\begin{aligned}
Cov_{\boldsymbol{\theta}}(a_i, X_i) &= E_{\boldsymbol{\theta}} a_i(\mathbf{X})(X_i - \theta_i) \\
&= \int_{R^m} a_i(\mathbf{x})(x_i - \theta_i)(2\pi)^{-m/2} \exp\left\{-\frac{1}{2}\sum_{j=1}^{m}(x_j - \theta_j)^2\right\} dx_1 \cdots dx_m \\
&= \int_{R^{m-1}} \left[\int_R a_i(\mathbf{x})(x_i - \theta_i)(2\pi)^{-1/2} \exp\left\{-\frac{1}{2}(x_i - \theta_i)^2\right\} dx_i\right] \times \\
&\qquad (2\pi)^{-(m-1)/2} \exp\left\{-\frac{1}{2}\sum_{j\neq i}(x_j - \theta_j)^2\right\} dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_m \\
&= -\int_{R^{m-1}} \left[\int_R a_i(\mathbf{x})(2\pi)^{-1/2} d \exp\left\{-\frac{1}{2}(x_i - \theta_i)^2\right\}\right] \times \\
&\qquad (2\pi)^{-(m-1)/2} \exp\left\{-\frac{1}{2}\sum_{j\neq i}(x_j - \theta_j)^2\right\} dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_m \\
&= \int_{R^{m-1}} \left[\int_R \frac{\partial a_i(\mathbf{x})}{\partial x_i}(2\pi)^{-1/2} \exp\left\{-\frac{1}{2}(x_i - \theta_i)^2\right\} dx_i\right] \times \\
&\qquad (2\pi)^{-(m-1)/2} \exp\left\{-\frac{1}{2}\sum_{j\neq i}(x_j - \theta_j)^2\right\} dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_m \\
&= \int_{R^m} \frac{\partial a_i(\mathbf{x})}{\partial x_i}(2\pi)^{-m/2} \exp\left\{-\frac{1}{2}\sum_{j=1}^{m}(x_j - \theta_j)^2\right\} dx_1 \cdots dx_m \\
&= E_{\boldsymbol{\theta}} \frac{\partial a_i}{\partial X_i}.
\end{aligned}$$

Now taking $a_i(\mathbf{x}) = \boldsymbol{\delta}_{JS}$ in (13), we have

$$\begin{aligned}
E_{\boldsymbol{\theta}}||\mathbf{X} - \boldsymbol{\delta}_{JS}||_2^2 &= E_{\boldsymbol{\theta}}\left\{\frac{(m-2)^2}{S^2}S\right\} = E_{\boldsymbol{\theta}}\left\{\frac{(m-2)^2}{S}\right\} \\
\sum_{i=1}^{m} Cov_{\boldsymbol{\theta}}(\boldsymbol{\delta}_{JS,i}, X_i) &= E_{\boldsymbol{\theta}}\sum_{i=1}^{m}\frac{\partial \delta_{JS,i}}{\partial X_i} = E_{\boldsymbol{\theta}}\left\{\sum_{i=1}^{m}\frac{2(m-2)X_i^2}{S^2} + m\left(1 - \frac{m-2}{S}\right)\right\} \\
&= = m - E_{\boldsymbol{\theta}}\frac{(m-2)^2}{S}.
\end{aligned}$$

It follows that

$$\begin{aligned}
E_{\boldsymbol{\theta}}||\boldsymbol{\delta}_{JS} - \boldsymbol{\theta}||_2^2 &= E_{\boldsymbol{\theta}}\left\{\frac{(m-2)^2}{S}\right\} - m + 2m - 2E_{\boldsymbol{\theta}}\left(\frac{(m-2)^2}{S}\right) \\
&= E_{\boldsymbol{\theta}}||\boldsymbol{\delta}_{mle} - \boldsymbol{\theta}||_2^2 - E_{\boldsymbol{\theta}}\left\{\frac{(m-2)^2}{S}\right\}
\end{aligned}$$

where $S = \sum_{i=1} X_i^2$. Thus, the second term is always poisitive as long as $m > 2$. This proves that $E_{\boldsymbol{\theta}}||\boldsymbol{\delta}_{JS} - \boldsymbol{\theta}||_2^2 < E_{\boldsymbol{\theta}}||\boldsymbol{\delta}_{mle} - \boldsymbol{\theta}||_2^2$ for all $\boldsymbol{\theta}$. □

1.8 Minimax Estimation

Let $\mathcal{D}$ be the class of decision functions under investigation.

Definition 7 *A decision rule $\delta^* \in \mathcal{D}$ is said to be a minimax solution in $\mathcal{D}$ if*

$$\sup\{R(\theta, \delta^*) : \theta \in \Theta\} = \inf \sup_{\theta \in \Theta}\{R(\theta, \delta) : \delta \in \mathcal{D}\}. \tag{14}$$

Intuitively, such δ^* attempts to minimize the risk of the worst. In the following, we

Theorem 6 *Suppose that there exists a prior distribution Π under which the Bayes solution δ_Π has a constant risk function $R(\theta, \delta_\Pi)$ over Θ. Then this Bayes solution δ_Π is a minimax solution.*

Proof If the Bayes solution δ_Π is not the minimax solution, then there is a decision function δ such that

$$\sup\{R(\theta, \delta) : \theta \in \Theta\} < \sup\{R(\theta, \delta_\Pi) : \theta \in \Theta\} \overset{def}{=} C.$$

It implies that $R(\theta, \delta) < R(\theta, \delta_\Pi) \equiv C$ for all $\theta \in \Theta$. Thus,

$$R_\Pi(\delta) = \int_\Theta R(\theta, \delta) d\Pi(\theta) < \int_\Theta R(\theta, \delta_\Pi) d\Pi(\theta) = R_\Pi(\delta_\Pi).$$

This means that δ_Π is not a Bayes solution. Contradiction! □

Theorem 7 *Suppose that $\{\Pi_n\}$ is a sequence of prior distributions that lead to a corresponding sequence of Bayes solutions $\{\delta_{\Pi_n}\}$. If a decision function $\delta \in \mathcal{D}$ satisfies (i) $\sup\{R(\theta, \delta) : \theta \in \Theta\} < \infty$ and (ii)*

$$\sup\{R(\theta, \delta) : \theta \in \Theta\} \le \lim_{n \to \infty} \sup R_{\Pi_n}(\delta_{\Pi_n}) = C.$$

Then δ is a minimax solution.

Proof If δ is not the minimax solution in $\mathcal{D}$, then there exists a $\delta^* \in \mathcal{D}$ such that

$$\sup\{R(\theta, \delta^*) : \theta \in \Theta\} < \sup\{R(\theta, \delta) : \theta \in \Theta\} \overset{def}{=} \tilde{C} \le C,$$

where $\tilde{C} < \infty$ according to condition (i). Thus, there exists $\epsilon > 0$ such that

$$R(\theta, \delta^*) \le \tilde{C} - 2\epsilon, \forall \theta \in \Theta.$$

On the other hand, according to the definition of constant C, for a sufficiently large N it holds that

$$R_{\Pi_N}(\delta_{\Pi_N}) \ge \tilde{C} - \epsilon.$$

It follows that

$$R_{\Pi_N}(\delta^*) = \int_\Theta R(\theta, \delta^*) d\Pi_N(\theta) \leq \tilde{C} - 2\epsilon < \tilde{C} - \epsilon \leq R_{\Pi_N}(\delta_{\Pi_N}).$$

This means that δ_{Π_N} is not a Bayes solution under the prior Π_N. Contradition! □

Example 9 Consider $X \sim B(n, \theta)$, $\theta \in [0,1]$, We like to find, if possible, the minimax estimator of θ under the squared error loss. From Example 1, we already obtained the following results: Taking $a = b = \sqrt{n}/2$ (a beta distribution with mean 1/2 and variance converging to zero when $n \to \infty$), we have

$$\delta_{\sqrt{n}/2, \sqrt{n}/2} = \frac{x}{n + \sqrt{n}} + \frac{\sqrt{n}}{2(n + \sqrt{n})},$$

and the corresponding risk function is

$$R(\theta, \delta_{\sqrt{n}/2, \sqrt{n}/2}) = \frac{n}{4(n + \sqrt{n})^2}.$$

Because this risk function is a constant over $\theta \in [0,1]$, according to Theorem 6, the corresponding Bayes estimator $\delta_{\sqrt{n}/2, \sqrt{n}/2}$ is the minimax estimator of θ.

Example 10 Let $X_1, \ldots, X_n$ be i.i.d. $N(\theta, 1)$. Under the squared error loss function $L(\theta, a) = (\theta - a)^2$, we like to find the minimax solution for θ.

Consider a sequence of (conjugate) priors $N(0, \tau^2), \tau = 1, 2, \ldots$. We know the following facts from Example 2: under the squared error loss, the Bayes estimator of θ is

$$\delta_\tau(x) = \frac{n\tau^2 \bar{x}}{1 + n\tau^2}.$$

Its Bayes risk equals to

$$R_{\Pi_\tau}(\delta_\tau) = \frac{\tau^2}{1 + n\tau^2} \to \frac{1}{n}, \text{as } \tau \to \infty.$$

Note that $R(\theta, \bar{X}) = E_\theta(\bar{X} - \theta)^2 = \frac{1}{n}$. Appling Theorem 7, the sample mean $\bar{X}$ is the minimax estimator of θ.

Remark: Although $\bar{X}$ has a constant risk function over θ but it has to be a Bayes estimator under a certain prior (which we don't know) in order for the application of Theorem 6.