The Annals of Statistics
1996, Vol. 24, No. 5, 2288–2295

AN INFORMATION INEQUALITY FOR THE BAYES RISK[1]

BY MICHIKAZU SATO AND MASAFUMI AKAHIRA

University of Tsukuba

This paper presents a lower bound, derived from the information inequality for the Bayes risk with respect to truncated priors under quadratic loss. It is discussed in cases where the regularity condition of Brown and Gajek is not always satisfied. A related result for the minimax risk is also given.

1. Introduction. Brown and Gajek (1990) presented lower bounds for the Bayes risk, derived from the information inequality under scaled quadratic loss assuming some regularity conditions including that the prior density is absolutely continuous. So the results cannot be applied to a truncated prior like a proper uniform prior. Sato and Akahira (1995) discuss lower bounds for the minimax risk under quadratic loss, derived from information inequalities for the Bayes risk obtained by Borovkov and Sakhanienko (1980) and Brown and Gajek (1990). Related results can be found in Bobrovsky, Mayer-Wolf and Zakai (1987)

The purpose of this paper is to obtain a lower bound for the Bayes risk of truncated priors under quadratic loss. An information inequality for such priors is given in Theorem 2.1, but the bound is not sharp under regularity conditions as is mentioned in Section 3. In Section 4, we consider what happens for continuous prior densities. In Section 5, we discuss the relation to minimax bounds.

2. A lower bound for the Bayes risk. In this section, we obtain a lower bound for the Bayes risk in cases where the prior density does not satisfy the regularity condition of Brown and Gajek (1990).

Let X be an observable random variable with probability densities p_θ relative to some σ-finite measure ν. Assume $\theta \in \Theta$, where $\Theta \subset \mathbf{R}$ is a (possibly infinite) interval. It is desired to estimate θ by $a \in \Theta$ under loss

$$L(\theta, a) = (a - \theta)^2. \tag{2.1}$$

Let $R(\theta, T) = E_\theta[L(\theta, T)]$ denote the risk of the nonrandomized estimator

Received June 1993; revised August 1995.

[1] This research was supported in part by Grant-in-Aid for Science Research 00001777, Ministry of Education, Science, Sports and Culture, Japan.

AMS 1991 *subject classifications*. Primary 62F10; secondary 62F15.

Key words and phrases. Information inequality, Cramér–Rao inequality, Bayes risk.

$T = T(X)$. Let $\overline{\Theta}$ denote the closure of Θ. Let $g(\cdot)$ be a nonnegative density with respect to the Lebesgue measure on Θ. This is the prior density. For any estimator T, let $B(g,T) = \int R(\theta,T)g(\theta)\,d\theta$ and let $B(g) = \inf_T B(g,T)$. $B(g)$ is the Bayes risk under g.

The setup above is assumed until Section 4.

REMARK. If (2.1) is replaced by $L(\theta, a) = m(\theta)(a - \theta)^2$ where $m > 0$, we may regard $m(\theta)g(\theta)$ as $g(\theta)$ and use (2.1).

We make the following conditions (2a) to (2e):

(2a) There exist $\theta_1, \theta_2 \in \overline{\Theta}$ such that $\theta_1 < \theta_2$ and, for a.e. $\theta \in (\theta_1, \theta_2)$, the amount of Fisher information

$$I(\theta) := E_\theta\left[\left\{\frac{\partial}{\partial\theta}\log p_\theta(X)\right\}^2\right]$$

exists. Define $V(\theta) := 1/I(\theta)$ and assume $0 < V(\theta) \le \infty$ for a.e. $\theta \in (\theta_1, \theta_2)$.

(2b) The prior density g is in C^1, $g(\theta) > 0$ on $[\theta_1, \theta_2]$ and $g(\theta) = 0$ outside $[\theta_1, \theta_2]$.

(2c) Let T_g be the Bayes estimator under g, that is,

$$T_g(x) = \frac{\int_{\theta_1}^{\theta_2}\theta p_\theta(x)g(\theta)\,d\theta}{\int_{\theta_1}^{\theta_2}p_\theta(x)g(\theta)\,d\theta},$$

and assume that $\theta \mapsto E_\theta[T_g]$, $\theta \in (\theta_1, \theta_2)$, can be extended to a C^2 function on $[\theta_1, \theta_2]$.

(2d) For T_g, the Cramér–Rao inequality (or the C–R inequality for short)

$$\operatorname{Var}_\theta T_g \ge V(\theta)\left\{\frac{d}{d\theta}E_\theta[T_g]\right\}^2$$

holds for a.e. $\theta \in (\theta_1, \theta_2)$.

(2e) There is a C^1 function, V_1, on $[\theta_1, \theta_2]$ which satisfies

$$V_1(\theta) \le V(\theta) \quad \text{a.e. } \theta \in (\theta_1, \theta_2),$$
$$0 < V_1(\theta) < \infty \quad \text{for all } \theta \in [\theta_1, \theta_2].$$

In the above, "a.e. θ" means almost all θ with respect to the Lebesgue measure. When we say "C^n on $[\theta_1, \theta_2]$," we consider the right differential coefficient at θ_1 and the left differential coefficient at θ_2.

REMARK. In order to get the best bound of all V_1's, we should let $V_1 = V$ if (2e) is satisfied, but it is often difficult to get the bound concretely.

Define $b(\theta) := E_\theta[T_g] - \theta$; then we get

$$R(\theta,T) \ge V_1(\theta)\{1 + b'(\theta)\}^2 + b^2(\theta),$$

$$B(g) = B(g,T_g) \ge \int\left\{V_1(1+b')^2 + b^2\right\}g\,d\theta.$$

For any C^2 function $y = y(\theta)$ on $[\theta_1, \theta_2]$, define

$$J(y) := \int\left\{V_1(1+y')^2 + y^2\right\}g\,d\theta.$$

Under the boundary conditions $y(\theta_1) = c_1$, $y(\theta_2) = c_2$, we want to minimize $J(y)$. Let η be a C^2 function on $[\theta_1, \theta_2]$ such that $\eta(\theta_1) = \eta(\theta_2) = 0$, let $\eta \not\equiv 0$ and let α be a constant. Then we obtain

$$\begin{aligned} &J(y+\alpha\eta) - J(y) \\ &\quad = 2\alpha\int\{V_1\eta(1+y') + \eta y\}g\,d\theta + \alpha^2\int\left(V_1\eta'^2 + \eta^2\right)g\,d\theta. \end{aligned} \tag{2.2}$$

If $J(y)$ takes its minimum value at y, then we have

$$\int\{V_1\eta'(1+y') + \eta y\}g\,d\theta = 0 \tag{2.3}$$

for all η satisfying the assumptions above.

Conversely, if (2.3) holds at y, then, by letting $\alpha = 1$ in (2.2), we obtain

$$J(y+\eta) > J(y) \quad \text{for all } \eta \text{ satisfying the assumptions above.}$$

Hence $J(y)$ takes its minimum value at y if and only if (2.3) holds at y. By integration by parts, we have

$$\begin{aligned} &\int V_1\eta'(1+y')g\,d\theta \\ &\quad = \int \eta'(1+y')(V_1 g)\,d\theta \\ &\quad = \left[\eta(1+y')(V_1 g)\right]_{\theta_1}^{\theta_2} - \int\eta\{y''(V_1 g) + (1+y')(V_1 g)\}\,d\theta \\ &\quad = \int\eta\{y''(V_1 g) + (1+y')(V_1 g)\}\,d\theta, \end{aligned} \tag{2.4}$$

where the last equality follows from $\eta(\theta_1) = \eta(\theta_2) = 0$. Hence (2.3) is equivalent to

$$\int\eta\{y''(V_1 g) + (1+y')(V_1 g)' - gy\}\,d\theta = 0 \tag{2.5}$$

for all η satisfying the assumptions above.

Then a necessary and sufficient condition for (2.5) is easily shown to be that y is a solution of the differential equation

$$y''(V_1 g) + (1+y')(V_1 g)' - gy = 0,$$

which is equivalent to

$$y'' + \tilde{g}y' - \frac{1}{V_1}y + \tilde{g} = 0, \tag{2.6}$$

where $\tilde{g} = (\log V_1 g)'$.

LEMMA 2.1. *The differential equation* (2.6) *has a unique solution for any choice of* c_1 *and* c_2, *and the solution is* C^2 *on* $[\theta_1, \theta_2]$.

PROOF. Let $y = y_0 + Ay_1 + By_2$ be a general solution of (2.6), where y_1 and y_2 are linearly independent. In order to show that (2.6) has a unique solution, it is enough to prove that the simultaneous equations

$$y_0(\theta_j) + Ay_1(\theta_j) + By_2(\theta_j) = c_j, \qquad j = 1, 2,$$

have a unique solution with respect to A and B. This is equivalent to

$$\begin{vmatrix} y_1(\theta_1) & y_2(\theta_1) \\ y_1(\theta_2) & y_2(\theta_2) \end{vmatrix} \neq 0,$$

which holds if and only if the simultaneous equations

$$y_0(\theta_j) + Ay_1(\theta_j) + By_2(\theta_j) = 0, \qquad j = 1, 2,$$

have the unique solution $A = B = 0$; that is, the differential equation

$$y'' + \tilde{g}y' - \frac{1}{V_1}y = 0, \qquad y(\theta_1) = y(\theta_2) = 0 \tag{2.7}$$

has the only solution $y \equiv 0$. Assume that (2.7) has a solution $y \not\equiv 0$. Then we get to a contradiction by considering the sign of (2.7) at the point θ where $y(\theta)$ takes its maximum (minimum) value. It is easily seen from (2.6) that the solution is C^2. □

THEOREM 2.1. *Assume that conditions* (2a) *to* (2e) *hold. Let* $y = y_0 + Ay_1 + By_2$ *be a general solution of* (2.5), *where* y_1 *and* y_2 *are linearly independent. If* y *moves over all* C^2 *functions on* $[\theta_1, \theta_2]$, *then* $J(y)$ *takes its minimum value* J_0 *and*

$$B(g) \geq J_0 = \frac{a_0 b_0 c_0 + 2 f_0 g_0 h_0 - a_0 f_0^2 - b_0 g_0^2 - c_0 h_0^2}{a_0 b_0 - h_0^2},$$

where

$$a_0 := \int \left(V_1 y_1'^2 + y_1^2\right) g \, d\theta, \qquad b_0 := \int \left(V_1 y_2'^2 + y_2^2\right) g \, d\theta,$$

$$c_0 := \int \left\{V_1 (1 + y_0')^2 + y_0^2\right\} g \, d\theta, \qquad f_0 := \int \{V_1 (1 + y_0') y_2' + y_0 y_2\} g \, d\theta,$$

$$g_0 := \int \{V_1 (1 + y_0') y_1' + y_0 y_1\} g \, d\theta, \quad h_0 := \int (V_1 y_1' y_2' + y_1 y_2) g \, d\theta.$$

PROOF. We need only minimize

$$J := J(y_0 + Ay_1 + By_2) = a_0 A^2 + 2 h_0 AB + b_0 B^2 + 2 g_0 A + 2 f_0 B + c_0,$$

where the coefficients are given in the above. Note that this is a quadratic form in A and B. By definition, $a_0 > 0$, and by applying the Cauchy–Schwarz inequality to

$$\langle y_1, y_2 \rangle := \int (V_1 y_1' y_2' + y_1 y_2) g \, d\theta,$$

we have $h_0^2 < a_0 b_0$. Hence $\lambda_1, \lambda_2 > 0$, where λ_1 and λ_2 are the eigenvalues of $\begin{pmatrix} a_0 & h_0 \\ h_0 & b_0 \end{pmatrix}$. By exchanging coordinates (rotation and moving parallel) (A, B) for (C, D) (say), we get J in the form $J = \lambda_1 C^2 + \lambda_2 D^2 + J_0$ and we obtain

$$J_0 = \frac{\begin{vmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & J_0 \end{vmatrix}}{\begin{vmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{vmatrix}} = \frac{\begin{vmatrix} a_0 & h_0 & g_0 \\ h_0 & b_0 & f_0 \\ g_0 & f_0 & c_0 \end{vmatrix}}{\begin{vmatrix} a_0 & h_0 \\ h_0 & b_0 \end{vmatrix}}$$

as its minimum value J_0, where the second equality follows from the fact that the numerators and the denominators, which are usually used to classify curves of second degree, are invariant with respect to exchanging coordinates. □

REMARK. Another representation of J_0 is pointed out by Shimakura (1993). In a similar way to Lemma 2.1, we get that the differential equation (2.6) has a unique solution under the boundary condition $y'(\theta_1) = y'(\theta_2) = -1$, and the solution is C^2 on $[\theta_1, \theta_2]$. For (and only for) the solution y, $J_0 = J(y)$ holds. The proof is similar to that of Theorem 2.1. [Let η be a C^2 function on $[\theta_1, \theta_2]$ such that $\eta \not\equiv 0$ (not necessarily $\eta(\theta_1) = \eta(\theta_2) = 0$).]

EXAMPLE 2.1. Let $\theta_1 = \theta_0 - \delta$, $\theta_2 = \theta_0 + \delta$, $\delta > 0$. Let $\lambda \in \mathbf{R}$ and define a prior density g by

$$g(\theta) := \frac{e^{\lambda(\theta - \theta_0)}}{\varphi(\lambda)} \quad \text{for } |\theta - \theta_0| < \delta,$$

where

$$\varphi(t) := \int_{-\delta}^{\delta} e^{t\theta} \, d\theta = \begin{cases} \dfrac{e^{\delta t} - e^{-\delta t}}{t}, & \text{for } t \neq 0, \\ 2\delta, & \text{for } t = 0. \end{cases}$$

Assume that (2a), (2c), (2d) and the following (2f) hold.

(2f) $$v_* := \inf_{\theta_1 < \theta < \theta_2} V(\theta), \qquad 0 < v_* < \infty, \qquad V_1 \equiv v_*.$$

Then

$$B(g) \geq J_0 = v_*(v_* \lambda^2 + 1) - \frac{v_* \sum_{j=1}^{2} \nu_j^2 (v_* \nu_j^2 + 1) \varphi^2(\nu_j)}{\rho^2 \varphi(\rho) \varphi(\lambda)},$$

where

$$\rho := \sqrt{\lambda^2 + \frac{4}{v_*}}, \qquad \nu_1 := \frac{-\lambda + \rho}{2}, \qquad \nu_2 := \frac{-\lambda - \rho}{2}.$$

In particular, if $\lambda = 0$, then the prior distribution becomes a uniform one $U(\theta_1, \theta_2)$. Hence

$$B(g) \geq J_0 = v_* \left(1 - \frac{\sqrt{v_*}\,\kappa}{\delta}\right), \tag{2.8}$$

where $\gamma = e^{\delta/\sqrt{v_*}}$, $\kappa = (\gamma - \gamma^{-1})/(\gamma + \gamma^{-1})$, $0 < \kappa < 1$. Indeed, letting $\theta_0 = 0$, we get $y_0 = v_* \lambda$, $y_1 = e^{\nu_1 \theta}$, $y_2 = e^{\nu_2 \theta}$. Substituting them into the inequality of Theorem 2.1, we have the conclusion.

Note that we may replace (2f) by

(2g) $\qquad 0 < v_* \leq \inf_{\theta_1 < \theta < \theta_2} V(\theta), \qquad v_* < \infty, \qquad V_1 \equiv v_*.$

3. On the attainment of the bound. The information inequality $B(g) \geq J_0$ is given in Section 2, but the equality does not hold under regularity conditions. We make the following conditions (3a) to (3d) using T_g given in (2c):

(3a) $\theta_1 \in \Theta$ [cf. (2a)]

(3b) For $\theta(> \theta_1)$ close to θ_1, P_θ is absolutely continuous with respect to P_{θ_1} [cf. (2a)].

(3c) The function $\theta \mapsto E_\theta[T_g]$, $\theta \in [\theta_1, \theta_2)$ *is* C^2 [cf. (2c)].

(3d) The functions V and $\theta \mapsto \mathrm{Var}_\theta T_g$ are right-continuous at θ_1 [cf. (2d) and (2e)].

THEOREM 3.1. *Under* (2a) *to* (2e) *and* (3a) *to* (3d), *the strict inequality* $B(g) > J_0$ *holds.*

PROOF. Assume that $B(g) = J_0$ holds. Then b should be the solution in the remark below Theorem 2.1, and, from (3a) and (3b), it should also be so at θ_1; hence $b'(\theta_1) = -1$. Next, since $B(g) = J_0$, the equality should hold in the C–R inequality for a.e. $\theta \in (\theta_1, \theta_2)$; we get from (3d) that it also holds at θ_1. So $\mathrm{Var}_{\theta_1} T_g = 0$, that is, T_g is a constant (say c) P_{θ_1}-a.e., and, from (3b), it is P_θ-a.e. for $\theta(> \theta_1)$ close to θ_1. For such θ, $E_\theta[T_g] = c$ and $b(\theta) = c - \theta$, but this b is not a solution of (2.6). This is a contradiction. □

REMARK. A similar result holds if we assume the regularity conditions on θ close to θ_2 instead of θ_1.

It is not yet clear whether the case $B(g) = J_0$ exists or not.

4. Cases where the assumption does not hold. Assumption (2b) implies that g is discontinuous at θ_1 and θ_2. We shall consider cases where (2b) does not hold but the following (4a) holds:

(4a) The prior density g is C^1 on $[\theta_1, \theta_2]$, $g(\theta) > 0$ on (θ_1, θ_2) and $g(\theta) = 0$ outside (θ_1, θ_2).

THEOREM 4.1. *Assume that* (2a), (2c) *to* (2e) *and* (4a) *hold. Then the differential equation* (2.6) *has at most one solution on* $[\theta_1, \theta_2]$ without *boundary conditions. In addition,* $J(y)$ *takes its minimum value* (*say* J_0) *in the class of all* C^2 *functions if and only if there is the solution above and for* (*and only for*) *the solution* y, $J_0 = J(y)$ *holds.*

PROOF. In a similar way to the statement above Lemma 2.1, we get that $J(y)$ takes its minimum value at (and only at) the solution of (2.6), where we let η be a C^2 function on $[\theta_1, \theta_2]$ such that $\eta \not\equiv 0$ [not necessarily $\eta(\theta_1) = \eta(\theta_2) = 0$] and the last equality in (2.4) follows from $g(\theta_1) = g(\theta_2) = 0$. The uniqueness follows from (2.3) □

It is not yet clear whether the solution in the theorem above exists or not.

5. Relation to minimax bounds. Let $\theta_1 = \theta_0 - \delta$, $\theta_2 = \theta_0 + \delta$ for $\delta > 0$ and define a prior density g by

$$g(\theta) := \frac{1}{\delta}\cos^2\frac{\pi}{2\delta}(\theta - \theta_0) \quad \text{for } |\theta - \theta_0| < \delta. \tag{5.1}$$

Assume that (2a), (2d), (2f) [or (2g)] and the following condition (5a) hold:

(5a) In (2c), $\theta \mapsto E_\theta[T_g]$ can be extended to an absolutely continuous function on $[\theta_1, \theta_2]$.

Then it follows from Borovkov and Sakhanienko (1980) and Brown and Gajek (1990) that

$$B(g) > v_*\left(1 + \frac{\pi^2 v_*}{\delta^2}\right)^{-1}. \tag{5.2}$$

Define

$$R^*(T) := \sup_\theta R(\theta, T), \qquad r^* := \inf_T R^*(T);$$

then r^* is called the minimax risk and T_0 is said to be minimax if $R^*(T_0) = r^* < \infty$. Since $r^* \geq B(g)$, the bounds for $B(g)$ in Section 2 and Borovkov and Sakhanienko (1980) and Brown and Gajek (1990) are regarded as those for r^*. Now we compare the bounds (2.8) and (5.2) by regarding them as the bounds of r^*. If δ is sufficiently small, then the bound (2.8) is better than (5.2); if δ is sufficiently large, then the bound (5.2) is better than (2.8). Indeed, letting $x = \delta/\sqrt{v_*}$, we have

$$\lim_{x \downarrow 0}(e^{2x} + 1)\left\{\frac{\text{Bound (2.8)}}{\text{Bound (5.2)}} - 1\right\} = \frac{2\pi^2}{3} - 2 > 0,$$

$$\lim_{x \to \infty} x^2\left\{\frac{\text{Bound (2.8)}}{\text{Bound (5.2)}} - 1\right\} = -\infty < 0.$$

In particular, if X is a $N(\theta, 1)$ random variable and $|\theta| \leq \delta$, it is shown in Bickel (1981) that, for large δ (which is equivalent to large n in the independently and identically distributed case), the distribution given by the density (5.1) is an approximate least favorable distribution rather than the uniform distribution on $(-\delta, \delta)$.

Acknowledgments. The authors thank the referee, Professor L. D. Brown and Professor N. Shimakura of Tohoku University for useful comments.

REFERENCES

BICKEL, P. J. (1981). Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Ann. Statist.* **9** 1301–1309.

BOBROVSKY, B. Z., MAYER-WOLF, E. and ZAKAI, M. (1987). Some classes of global Cramér–Rao bounds. *Ann. Statist.* **15** 1421–1438.

BOROVKOV, A. A. and SAKHANIENKO, A. U. (1980). On estimates of the expected quadratic risk (in Russian). *Probab. Math. Statist.* **1** 185–195.

BROWN, L. D. and GAJEK, L. (1990). Information inequalities for the Bayes risk. *Ann. Statist.* **18** 1578–1594.

SATO, M. and AKAHIRA, M. (1995). Information inequalities for the minimax risk. *J. Japan Statist. Soc.* **25** 151–158.

SHIMAKURA, N. (1993). Personal communication.

INSTITUTE OF MATHEMATICS
UNIVERSITY OF TSUKUBA
IBARAKI 305
JAPAN