Ann. Inst. Statist. Math.
Vol. 41, No. 2, 347-363 (1989)

A MINIMUM AVERAGE RISK APPROACH TO SHRINKAGE ESTIMATORS OF THE NORMAL MEAN

D. L. HAWKINS[1] AND CHIEN-PAI HAN[2]

[1]*Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, IA 52242, U.S.A.*
[2]*Department of Mathematics, The University of Texas at Arlington, P. O. Box 19408, Arlington, TX 76019, U.S.A.*

(Received November 18, 1987; revised September 27, 1988)

Abstract. For the problem of estimating the normal mean μ based on a random sample $X_1,\dots, X_n$ when a prior value μ_0 is available, a class of shrinkage estimators $\hat{\mu}_n(k) = k(T_n)\bar{X}_n + (1 - k(T_n))\mu_0$ is considered, where $T_n = n^{1/2}(\bar{X}_n - \mu_0)/\sigma$ and k is a weight function. For certain choices of k, $\hat{\mu}_n(k)$ coincides with previously studied preliminary test and shrinkage estimators. We consider choosing k from a natural non-parametric family of weight functions so as to minimize average risk relative to a specified prior p. We study how, by varying p, the MSE efficiency (relative to $\bar{X}$) properties of $\hat{\mu}_n(k)$ can be controlled. In the process, a certain robustness property of the usual family of posterior mean estimators, corresponding to the conjugate normal priors, is observed.

Key words and phrases: Optimal weight function, Hilbert space, quadratic programming.

1. Introduction

Let $X_1,\dots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ where σ^2 is known. A popular class of estimators of μ in the presence of some prior value μ_0, near which μ is expected to lie, has the form

$$\hat{\mu}_n(k) = k(T_n)\bar{X}_n + (1 - k(T_n))\mu_0\,, \tag{1.1}$$

where $\bar{X}_n$ is the sample mean, $T_n = \sqrt{n}(\bar{X}_n - \mu_0)/\sigma$ tests H_0: $\mu = \mu_0$ and k is a suitable weight function. The popularity of these estimators is due to the fact that if k is chosen properly, $\hat{\mu}_n(k)$ has smaller mean-squared error (MSE) than $\bar{X}_n$ for all μ in the so-called effective interval (EI) of k; i.e., for all μ satisfying $\sqrt{n}|\mu - \mu_0|/\sigma \le C(k)$, a constant depending on k. Thus, if μ_0 is sufficiently close to μ, we gain in MSE efficiency (MSEE) by using $\hat{\mu}_n(k)$ instead of $\bar{X}_n$.

Various choices of k have been considered, the basic idea being that k should give "large" weight to μ_0 and "small" weight to $\bar{X}_n$ if H_0: $\mu = \mu_0$ appears to be true, and just the reverse if H_0 appears to be false. As this loose specification permits great latitude in choice of k, various optimality conditions have been imposed, seeking to define a choice of k optimal in some sense. Hirano (1977) studied a special type of preliminary test estimator (PTE), formed by taking k in (1.1) as $k_1(t) = I(|t| \geq z_{\alpha/2})$ (here and throughout $z_{\alpha/2}$ denotes the $100(1 - \alpha/2)$ percentile of $N(0, 1)$), and searched within the class, obtained by varying α, for the one minimizing Akaike's (1973) information criterion. From a different viewpoint, Thompson (1968) proposed a shrinkage estimator (SE) of the form (1.1) with $k_2(t) = t^2/(1 + t^2)$. k_2 arises by considering estimators of μ of the form $c\bar{X}_n + (1 - c)\mu_0$ (c is a constant), choosing $c = c^*(\mu)$ to minimize the MSE for each μ (giving $c^*(\mu) = \mu^2/(1 + \mu^2)$) and then replacing μ by X_n (giving k_2). Mehta and Srinivasan (1971) (henceforth M & S) considered $k_3(t) = 1 - ae^{-bt^2}$, where a and b are adjustable constants. They attempted to choose a and b so as to simultaneously give maximum MSEE for μ near μ_0 and as long an EI as possible (two antagonistic properties). Although they did not achieve a unique solution, certain values of a and b work quite well. Finally, Inada (1984) combined the idea of PTE's and SE's and proposed (1.1) with $k_4(t) = d^*I(|t| < z_{\alpha/2}) + I(|t| \geq z_{\alpha/2})$, where for fixed α, $d^* \in [0, 1]$ is chosen by a minimax regret criterion.

Looking at these estimators and their motivations, one observes that they derive from a procedure familiar to statisticians: pick a parametric family of candidate functions and search within that family for a member optimal in some sense. As the families of k_1–k_4 are each indexed by a real- or vector-valued parameter, familiar methods (e.g., differentiation) are applicable to solve the corresponding optimization problems.

In this paper we propose and study yet another method of choosing k in (1.1), based on the idea of minimizing average risk of (1.1). To facilitate its introduction, we require some notation. First, reparameterize the parameter space $\{\mu \in \mathbb{R}\}$, for any fixed $n \geq 1$, by $\{\mu_n = \mu_0 + \Delta/\sqrt{n}: \Delta \in \mathbb{R}\}$. Then under μ_n we have $T_n = \sqrt{n}(\bar{X}_n - \mu_n)/\sigma + \Delta/\sigma \sim N(\Delta/\sigma, 1)$, so for any measurable k and $\Delta \in \mathbb{R}$, the MSE (risk) of $\hat{\mu}_n(k)$ at μ_n is

$$(1.2) \qquad R(k, \Delta, \sigma) = E_{\mu_n}\{n^{1/2}(\hat{\mu}_n(k) - \mu_n)/\sigma\}^2 = (2\pi)^{-1/2}\int_{-\infty}^{\infty}\left(xk(x) - \frac{\Delta}{\sigma}\right)^2 \exp\left\{-\frac{1}{2}\left(x - \frac{\Delta}{\sigma}\right)^2\right\} dx\,.$$

We shall for simplicity refer to $R(k, \Delta, \sigma)$ as the MSE of $\hat{\mu}_n(k)$, in spite of the normalization. (Under the same normalization, the MSE of $\bar{X}_n$ is 1.) Since $R(k, \Delta, \sigma)$ depends on Δ and σ only through the ratio Δ/σ, we let $\sigma = 1$ and write it as

$$(1.3) \qquad R(k,\Delta)=(2\pi)^{-1/2}\int_{-\infty}^{\infty}(xk(x)-\Delta)^2 e^{-(x-\Delta)^2/2}dx\,.$$

Now introduce a prior $p(\Delta)$ on $\mathbb{R}$ satisfying

$$(1.4) \quad \int_{-\infty}^{\infty}\Delta^2 p(\Delta)<\infty, \quad p(\Delta)>0, \quad p(\Delta)=p(-\Delta) \quad \text{for all} \quad \Delta\,,$$

and define the average MSE (average risk) of $\hat{\mu}_n(k)$ by

$$(1.5) \qquad R_p(k)=\int_{-\infty}^{\infty}R(k,\Delta)p(\Delta)d\Delta$$
$$=(2\pi)^{-1/2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(xk(x)-\Delta)^2 e^{-(x-\Delta)^2/2}p(\Delta)dxd\Delta\,.$$

Generally, our proposed method is to first choose a prior p, and then choose k in (1.1) from a sensible class K(discussed below) of candidate weight functions so as to minimize $R_p(k)$ over $k \in K$. Our choice of p is made so as to control the EI and MSEE of $\hat{\mu}_n(k)$. This is plausible since, as will be seen later, the dispersion of p may be varied so as to lengthen or shorten the EI, respectively decreasing or increasing the MSEE on the EI. Thus our view of p is merely as a tool for varying the EI and MSEE of $\hat{\mu}_n(k)$, without any direct Bayesian connotations. By varying the shape of p as well as the dispersion, we may hope to generate an even wider variety of EI and MSEE combinations. We remark that our idea of indirectly controlling EI and MSEE by varying p is somewhat similar to the approaches of the authors of k_1, k_2 and k_4, who attempted this by using optimality criteria only indirectly related to EI and MSEE. Of the authors of k_1–k_4, only M & S considered EI and MSEE directly as their goal.

Turning now to our class K of candidate weight functions, we note that the minimization problem $\min\{R_p(k): k \in K\}$ may be legitimately considered over any set K for which $R_p(k)<\infty$ for all $k \in K$, and which contains a minimizing function. However, we sought a class K of weight functions which preserved the spirit of k_1–k_4 above, but which also allowed considerable generalization. Particularly, we noted of k_1–k_4 that

(P1) Each $k(t)$ is an even function of t.

(P2) For $t \geq 0$, each $k(t)$ is non-decreasing with range $[c,1]$ where $0 \leq c < 1$.

(Put another way, (P2) says that these weight functions look like cdf's of non-negative random variables.) This non-parametric family of weight functions contains the parametric families of k_1–k_4 above and obviously many other weight functions. Hence, we decided to consider it as a candidate class, in hopes that its richness of possible weight functions, together with the approach of minimizing $R_p(k)$ over k in this class, would lead to estimators (1.1) superior in terms of MSEE and EI to k_1–k_4.

This turns out not to be the case. The M & S estimator based on k_3 is not dominated in terms of EI and MSEE performance by our estimator $\hat{\mu}_n(k)$ for any of the p we investigated. However, for $p = N(0, \gamma^2)$ for appropriate γ^2, our method produces $\hat{\mu}_n(k)$ with EI and MSEE almost identical to M & S but with k constant and hence simpler than k_3. Also, our class of estimators is clearly more flexible than that of M & S, and is also more intuitive, due to the connection between dispersion of the prior and the behaviour of the resulting MSEE and EI. Possible reasons that our method does not defeat M & S are: (1) Our method does not directly address the MSEE/EI criterion, while M & S does; (2) the average risk approach is too inflexible and (3) the exponential family of weight functions used by M & S is a very rich family within our class. Nonetheless, our method offers another means of obtaining and understanding the choice of weight function k in (1.1). For future reference, define a weight function k^* to be p-optimal in a class K if $R_p(k^*) \le R_p(k)$ for all $k \in K$. Observe that k^* will depend on p, in general.

The rest of the paper is organized as follows. Subsection 2.1 establishes, for p satisfying (1.4), the existence of a p-optimal weight function k^* in a set of weight functions formed from a slight extension of (P1) and (P2). Subsection 2.2 shows how to approximate k^* numerically. Subsection 2.3 illustrates the approximations for a selection of p. Section 3 compares our estimator to some of the above-mentioned estimators in terms of the MSEE and EI. Section 4 contains some conclusions.

2. The p-optimal weight function approximately satisfying (P1) and (P2)

In order to ensure a solution for the minimization problem, min $\{R_p(k): k \in K\}$, care must be taken in defining K. We first replace the class of functions given by (P1) and (P2) by a slightly larger class. For this we need the following

DEFINITION 2.1. If k_n and k are real-valued functions on $\mathbb{R}$, $k_n \xrightarrow{\text{a.u.}} k$ (almost uniform convergence) means that given any $\varepsilon > 0$, there exists a set $A \subset \mathbb{R}$ with $m(A) < \varepsilon$ (m denotes Lebesgue measure on $\mathbb{R}$) such that $\sup\{|k_n(x) - k(x)|: x \in A^c\} \to 0$ as $n \to \infty$.

For some specified $M \in (0, \infty)$, let

$$K_I = \{k: \mathbb{R} \to \mathbb{R} \mid k \text{ satisfies (i) and (ii) below}\}$$

(i) $k(-x) = k(x)$

(2.1) (ii) $k(x) = \int_0^x f(t)dt$, $0 \le x \le M$, $k(x) = 1$, $x > M$ for some continuous function f satisfying $f(t) \geq 0$,

$$0 \leq t \leq M \text{ and } \int_0^M f(t)dt = 1 .$$

The functions in K_I are all even and non-decreasing on $[0, M]$. They are also absolutely continuous, a restriction we do not wish to impose. To remove this restriction, we extend K_I to the bigger set K_1 by attaching a.u. limits. Specifically, define

$$K_1 = \{k\colon \mathbb{R} \to \mathbb{R} \mid \text{there exist } k_n \in K_I \text{ such that } k_n \xrightarrow{\text{a.u.}} k\} . \tag{2.2}$$

The class K_1 provides a practically adequate approximation to the class of functions specified by (P1) and (P2). The slight flaw in K_1 is that, since M is finite, it does not contain cdf's (like k_2 or k_3 in the introduction) which are supported on all of $[0, \infty)$. However, since we may make M arbitrarily large, there are functions in K_1 which approximate such functions arbitrarily closely. Otherwise, K_1 contains all finitely-supported cdf's on $[0, M]$, including the step functions which are a.u. limits of continuous cdf's. It also contains functions which are constant (between 0 and 1) everywhere on $(0, M)$ (again as a.u. limits). Although such functions are not of the form given by (P2), they turn out to be important in this investigation.

With K_1 as defined in (2.2), we now turn, for a specified p satisfying (1.4), to the problem of minimizing R_p over $k \in K_1$.

2.1 *Existence of a solution*

Define $w(\Delta, x) = p(\Delta)e^{-(x-\Delta)^2/2}$, $(\Delta, x) \in \mathbb{R}^2$, and let

$$L^2(\mathbb{R}^2, dw) = \left\{h\colon \mathbb{R}^2 \to \mathbb{R} \text{ measurable} \;\middle|\; \int_{\mathbb{R}^2} h^2(\Delta, x)w(\Delta, x)d\Delta dx < \infty\right\} .$$

Then, $L^2(\mathbb{R}^2, dw)$ is a Hilbert space. Let

$$K' = \{\phi\colon \mathbb{R}^2 \to \mathbb{R} \mid \text{for every } \Delta,\ \phi(\Delta, x) = xk(x) \text{ a.e. } (x) \text{ for some } k \in K_1\} .$$

THEOREM 2.1. *Under* (1.4),

(i) $K' \subset L^2(\mathbb{R}^2, dw)$.

(ii) *K' is convex and closed in $L^2(\mathbb{R}^2, dw)$.*

(iii) *$\phi_0(\Delta, x) = \Delta$ for all $x \in \mathbb{R}$ is in $L^2(\mathbb{R}^2, dw)$.*

(iv) *The problem,* min $R_p(k)$, $k \in K_1$ *has a unique solution k_1^*.*

PROOF. (i) Let $\phi \in K'$. Then for every Δ, $\phi(\Delta, x) = xk(x)$ a.e. where, for some $\{k_n\} \subset K_I$, $k_n \xrightarrow{\text{a.u.}} k$. By Theorem 2.5.2 in Ash (1972), $k_n \to k$ a.e. (M). Thus, $k(x) = 1$ a.e., $|x| > M$ and $0 \leq k(x) \leq 1$ a.e., $|x| \leq M$. Hence

$$\int_{\mathbb{R}^2} \phi^2(\Delta, x) w(\Delta, x) d\Delta dx = \int_{-\infty}^{\infty} p(\Delta) \left[\int_{-\infty}^{\infty} x^2 k^2(x) e^{-(x-\Delta)^2/2} dx \right] d\Delta ,$$

and the inner integral may be written

$$\begin{aligned} &\int_{|x|>M} x^2 e^{-(x-\Delta)^2/2} dx + \int_{|x|<M} x^2 k^2(x) e^{-(x-\Delta)^2/2} dx \\ &\leq \int_{-\infty}^{\infty} x^2 e^{-(x-\Delta)^2/2} dx + M^2 \int_{-\infty}^{\infty} e^{-(x-\Delta)^2/2} dx \\ &\leq C_1 + C_2 \Delta^2, \end{aligned}$$

where C_1, C_2 are constants. That $\phi \in L^2(\mathbb{R}^2, dw)$ follows by (1.4).

(ii) First consider convexity. Let $\phi_1, \phi_2 \in K'$ and $0 < \lambda < 1$. There exist $k_1, k_2 \in K_1$ such that for every Δ, $\phi_1(\Delta, x) = xk_1(x)$ a.e., $\phi_2(\Delta, x) = xk_2(x)$ a.e. Further, since $k_1, k_2 \in K_1$, there exist sequences $\{k_n^{(1)}\}$, $\{k_n^{(2)}\}$ of K_I-functions such that $k_n^{(1)} \xrightarrow{\text{a.u.}} k_1$, $k_n^{(2)} \xrightarrow{\text{a.u.}} k_2$. Let $g_n = \lambda k_n^{(1)} + (1-\lambda)k_n^{(2)}$. Then $g_n \in K_I$ for all n, and it is easily checked that $g_n \xrightarrow{\text{a.u.}} \lambda k_1 + (1-\lambda)k_2$. Thus, $\lambda k_1 + (1-\lambda)k_2 \in K_1$. This implies that $\lambda\phi_1 + (1-\lambda)\phi_2 \in K'$, so K' is convex.

To see that K' is closed in the $L^2(\mathbb{R}^2, dw)$ norm, suppose that $\{\phi_n\} \subset K'$ and $\phi_n \to \phi$ in the $L^2(\mathbb{R}^2, dw)$ norm. We shall show that this implies that $\phi \in K'$. First, by Theorems 2.5.1 and 2.5.3 in Ash (1972), $\phi_n \xrightarrow{L^2} \phi$ implies the existence of a subsequence $\{n_\nu\}$ such that $\phi_{n_\nu} \to \phi$ a.e. ($\mathbb{R}^2$). Since $\phi_{n_\nu} \in K'$, there exists a function $k_{n_\nu} \in K_1$ such that for every Δ,

$$\phi_{n_\nu}(\Delta, x) = xk_{n_\nu}(x) \quad \text{for all} \quad x \in A_\nu^c, \quad m(A_\nu) = 0. \tag{2.3}$$

Since $\phi_{n_\nu}(\Delta, x) \to \phi(\Delta, x)$ for all $x \in B^c$, with $m(B) = 0$, (2.3) implies that for all x outside $E \equiv \bigcup_{\nu=1}^{\infty} A_\nu \cup B \cup \{0\}$, which has m-measure zero,

$$k(x) \equiv \lim_{\nu\to\infty} k_{n_\nu}(x) = \lim_{\nu\to\infty} \frac{\phi_{n_\nu}(\Delta, x)}{x} = \frac{\phi(\Delta, x)}{x} \quad \text{exists} . \tag{2.4}$$

Thus, for every Δ, $\phi(\Delta, x) = xk(x)$ a.e., and it remains to show that $k \in K_1$. For this we shall construct a sequence $\{g_\alpha\} \subset K_I$ with $g_\alpha \xrightarrow{\text{a.u.}} k$ as $\alpha \to \infty$. This will give the result.

In this direction, we first claim that there exists a subsequence $\{n_{\nu(\alpha)}\}$ of $\{n_\nu\}$ such that

$$k_{n_{\nu(\alpha)}} \xrightarrow{\text{a.u.}} k \quad \text{as} \quad \alpha \to \infty . \tag{2.5}$$

To see this, note that since $\phi_{n_\nu} \to \phi$ in $L^2(\mathbb{R}^2, dw)$,

$$(2.6)\qquad \iint_{\mathbb{R}^2} [xk_{n_\nu}(x) - xk(x)]^2 e^{-(x-\Delta)^2/2} p(\Delta) d\Delta\, dx$$
$$= \int_{-\infty}^{\infty} [k_{n_\nu}(x) - k(x)]^2 \left[x^2 \int_{-\infty}^{\infty} e^{-(x-\Delta)^2/2} p(\Delta) d\Delta \right] dx \to 0 .$$

Hence,

$$(2.7)\qquad k_{n_\nu} \to k \quad \text{in} \quad L^2(\mathbb{R}^1, dg),$$

where $g(x) = x^2 \int_{-\infty}^{\infty} e^{-(x-\Delta)^2/2} p(\Delta) d\Delta$ and

$$L^2(\mathbb{R}^1, dg) = \left\{ f\colon \mathbb{R}^1 \to \mathbb{R}^1 \text{ measurable} \,\middle|\, \int_{-\infty}^{\infty} f^2(x) g(x) dx < \infty \right\}$$

(another Hilbert space). (2.7) implies (by Theorems 2.5.1 and 2.5.3 in Ash (1972)) the existence of a subsequence $\{n_{\nu(\alpha)}\}$ of $\{n_\nu\}$ such that $k_{n_{\nu(\alpha)}} \to k$ a.u. (dg). This implies (2.5) since m is absolutely continuous with respect to dg under (1.4).

Secondly, note that since $k_{n_{\nu(\alpha)}} \in K_1$, for each fixed $\alpha \geq 1$ there exists a sequence $\{k_{n_{\nu(\alpha)}, l}\} \subset K_I$ such that

$$(2.8)\qquad k_{n_{\nu(\alpha)}, l} \xrightarrow{\text{a.u.}} k_{n_{\nu(\alpha)}} \quad \text{as} \quad l \to \infty.$$

Using (2.5) and (2.8), we construct the sequence $\{g_\alpha\}$. Let $\varepsilon > 0$, $\delta > 0$ be given. By (2.5) pick $N_1(\varepsilon, \delta)$ and a set $A \subset \mathbb{R}$ with $m(A) < \varepsilon/2$ so that

$$(2.9)\qquad \sup_{x \in A^c} |k_{n_{\nu(\alpha)}}(x) - k(x)| < \frac{\delta}{2}, \qquad \alpha \geq N_1(\varepsilon, \delta) .$$

Fix $\alpha \geq N_1(\varepsilon, \delta)$. By (2.8) pick $N_2(\alpha, \varepsilon, \delta)$ and a set $B_\alpha \subset \mathbb{R}$ with $m(B_\alpha) < 2^{-\alpha} \cdot \varepsilon/2$ so that

$$(2.10)\qquad \sup_{x \in B_\alpha^c} |k_{n_{\nu(\alpha)}, l}(x) - k_{n_{\nu(\alpha)}}(x)| < \frac{\delta}{2}, \qquad l \geq N_2(\alpha, \varepsilon, \delta) .$$

Set $g_\alpha(x) = k_{n_{\nu(\alpha)}, l(\alpha)}(x)$, where $l(\alpha) = N_2(\alpha, \varepsilon, \delta)$, $\alpha \geq 1$. By (2.9) and (2.10), $C = A \cup \left(\bigcup_{\alpha=1}^{\infty} B_\alpha \right)$ satisfies $m(C) < \varepsilon$ and, for any $\alpha_0 \geq N_1$,

$$\sup_{x \in C^c} |g_{\alpha_0}(x) - k(x)| = \sup_{x \in C^c} |k_{n_{\nu(\alpha_0)}, l(\alpha_0)}(x) - k(x)|$$
$$\leq \sup_{x \in B_{\alpha_0}^c} |k_{n_{\nu(\alpha_0)}, l(\alpha_0)}(x) - k_{n_{\nu(\alpha_0)}}(x)| + \sup_{x \in A^c} |k_{n_{\nu(\alpha_0)}}(x) - k(x)|$$
$$< \frac{\delta}{2} + \frac{\delta}{2} = \delta .$$

Thus, $g_\alpha \overset{\text{a.u.}}{\to} k$.

(iii) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \phi_0^2(\Delta, x)w(\Delta, x)d\Delta dx = \int_{-\infty}^{\infty} \Delta^2 p(\Delta)\left[\int_{-\infty}^{\infty} e^{-(x-\Delta)^2/2}dx\right]d\Delta < \infty$

since the bracketed term equals $(2\pi)^{1/2}$.

(iv) By (i) and (ii), K' is a closed convex set in the Hilbert space $L^2(\mathbb{R}^2, dw)$. Thus, by a well-known result (see Ash (1972), Theorem 3.2.9), there exists a unique $\phi^* \in K'$ such that

$$\|\phi^* - \phi_0\|_{L^2(\mathbb{R}^2, dw)} = \min\{\|\phi - \phi_0\|_{L^2(\mathbb{R}^2, dw)} : \phi \in K'\}\,. \tag{2.11}$$

But $\|\phi - \phi_0\|^2_{L^2(\mathbb{R}^2, dw)} = R_p(\phi/x)$, $\phi \in K'$. Thus, if $k_1^*(x) = \phi^*(x)/x$, (2.11) gives, since $\phi \in K'$ iff $k(x) = \phi(x)/x \in K_1$,

$$\begin{aligned}\min\{R_p(k): k \in K_1\} &= \min\{R_p(\phi/x): \phi \in K'\}\\ &= \min\{\|\phi - \phi_0\|^2_{L^2(\mathbb{R}^2, dw)} : \phi \in K'\}\\ &= \|\phi^* - \phi_0\|^2_{L^2(\mathbb{R}^2, dw)} = R_p(k_1^*)\,.\end{aligned}$$

Thus, k_1^* is a solution. Since ϕ^* is unique in K' (up to a.e. (m) equivalence), so is k_1^* in K_1. □

2.2 *Numerical approximation of k_1^**

For computational purposes we restrict our search to the set K_I, and rewrite $R_p(k)$ in (1.5) (with M as in the definition of K_I) as

$$\begin{aligned} R_p(f) = (2\pi)^{-1/2}\Bigg\{\int_{-M}^{M}\int_{-\infty}^{\infty}\left[y\int_0^{|y|} f(x)dx - \Delta\right]^2 e^{-(y-\Delta)^2/2}p(\Delta)d\Delta dy \\ + \int_{|y|>M}\int_{-\infty}^{\infty}[y-\Delta]^2 e^{-(y-\Delta)^2/2}p(\Delta)d\Delta dy\Bigg\}\,, \end{aligned} \tag{2.12}$$

where

$$\begin{aligned} f \in K_f \\ = \left\{f: [0, M] \to [0, \infty)\,\middle|\, f \geq 0 \text{ is continuous and} \int_0^M f(t)dt = 1\right\}. \end{aligned} \tag{2.13}$$

Clearly, K_I is one-to-one with K_f, so we may consider the problem

$$\min\{R_p(f): f \in K_f\}\,, \tag{2.14}$$

as a reasonable approximation to our problem $\min\{R_p(k): k \in K_1\}$. Also, only the first term on the RHS of (2.12) is involved in the minimization; henceforth we redefine $R_p(f)$ as equal to this first term only.

For $N \geq 1$ we discretize the feasible set K_f by taking $K_f^{(N)} = \{f_N: f_N$ is piecewise linear connecting points (x_{Ni}, f_{Ni}) satisfying (i*)–(iii*)$\}$:

$$\text{(i*)} \quad x_{Ni} = \frac{i}{N}: \quad 0 \leq i \leq MN \quad \text{partitions the interval} \quad [0, M].$$

$$\text{(2.15)} \quad \text{(ii*)} \quad f_{Ni} = f_N(x_{Ni}) \quad \text{satisfies} \quad f_{Ni} \geq 0, \quad 0 \leq i \leq M.$$

$$\text{(iii*)} \quad \frac{1}{2N}(f_{N0} + f_{N,MN}) + \frac{1}{N}\sum_{i=1}^{MN-1} f_{Ni} = 1.$$

Since any $f_N \in K_f^{(N)}$ may be written in the form

$$\text{(2.16)} \quad f_N(x) = \sum_{i=1}^{MN} [f_{N,i-1} + N(x - x_{N,i-1})(f_{Ni} - f_{N,i-1})]\, I_{[x_{N,i-1}, x_{Ni})}(x)\,, \qquad 0 \leq x < M\,,$$

an easy calculation shows that the condition $\int_0^M f_N(x)dx = 1$ is equivalent to (iii*). Thus, $K_f^{(N)} \subset K_f$. As $N \to \infty$, the set $K_f^{(N)}$ approximates the set K_f arbitrarily closely in the sense that for any fixed $f \in K_f$, there exists $f_N \in K_f^{(N)}$ such that

$$\sup\{|f_N(x) - f(x)|: 0 \leq x \leq M\} \to 0 \quad \text{as} \quad N \to \infty.$$

For $f_N \in K^{(N)}$, we approximate the functional $R_p(f)$ as follows. Choose a number $P > 0$, so that the support of the prior pdf $p(\cdot)$ is (exactly or approximately) $[-P, P]$. Partition this interval as $\{a_{Nj} = -P + ((N+j)/N)P: -N \leq j \leq N\}$, so that $a_{N0} = 0$. Define $p_{Nj} = p(a_{Nj})$, $-N \leq j \leq N$. Partition the interval $[-M, M]$ of the y-integration as $\{y_{Nk} = k/N: -MN \leq k \leq MN\}$, so that $y_{N0} = 0$. Replacing integrals by Riemann sums, we thus approximate $R_p(f)$ in (2.12) (remember: first term only) by

$$\text{(2.17)} \quad R_p^{(N)}(f_N) = (2\pi)^{-1/2}\frac{P}{N}\sum_{j=-N}^{N} p_{Nj} \times \left\{\frac{1}{N}\sum_{k=-MN}^{MN}\left[\frac{y_{Nk}}{N}\left(\frac{f_{N0} + f_{N,|k|}}{2} + \sum_{i=1}^{|k|-1} f_{Ni}\right) - a_{Nj}\right]^2 \times \exp\left[-\frac{1}{2}(y_{Nk} - a_{Nj})^2\right]\right\}.$$

In (2.12) the integral $\int_0^{|k/N|} f(x)dx$ has been replaced in (2.17) by

$$\int_0^{|y_{Nk}|} f_N(x)dx = \frac{1}{N}\left[\frac{f_{N0}+f_{N,|k|}}{2} + \sum_{i=1}^{|k|-1} f_{Ni}\right],$$

the equality following from (2.16). The approximation of $R_p(f)$ by (2.17) thus involves two components of error: that due to truncating the support of p to $[-P, P]$, and that due to approximating integrals by Riemann sums. The latter tends to zero as $N \to \infty$; the former does not. (However, if p is finitely supported, there is no error due to truncation.)

With the approximations to $R_p(f)$ and K_f as above, we replace the problem (2.14) by the problem

$$\min\{R_p^{(N)}(f_N): f_N \in K_f^{(N)}\}, \tag{2.18}$$

for some N sufficiently large. (2.18) may be converted to a quadratic programming problem as follows. Define

$$f_N^* = (f_{N0}, f_{N1}, \ldots, f_{N,MN})' \in \mathbb{R}^{MN+1}, \quad \text{and let}$$
$$K_f^{(N)^*} = \left\{ f_N^*: f_{Ni} \geq 0,\ 0 \leq i \leq MN \text{ and } \sum_{i=0}^{MN} \alpha_{Ni} f_{Ni} = 1 \right\}, \tag{2.19}$$

where $\alpha_{Ni} = 1/(2N)$ if $i \in \{0, MN\}$ and $\alpha_{Ni} = 1/N$ if $i \in \{1, 2, \ldots, MN-1\}$. Then $K_f^{(N)^*}$ is one-to-one with $K_f^{(N)}$, so we may clearly write $R_p^{(N)}(f_N)$ in (2.17) as $R_p^{(N)}(f_N^*)$. Therefore, the problem (2.18) is equivalent to

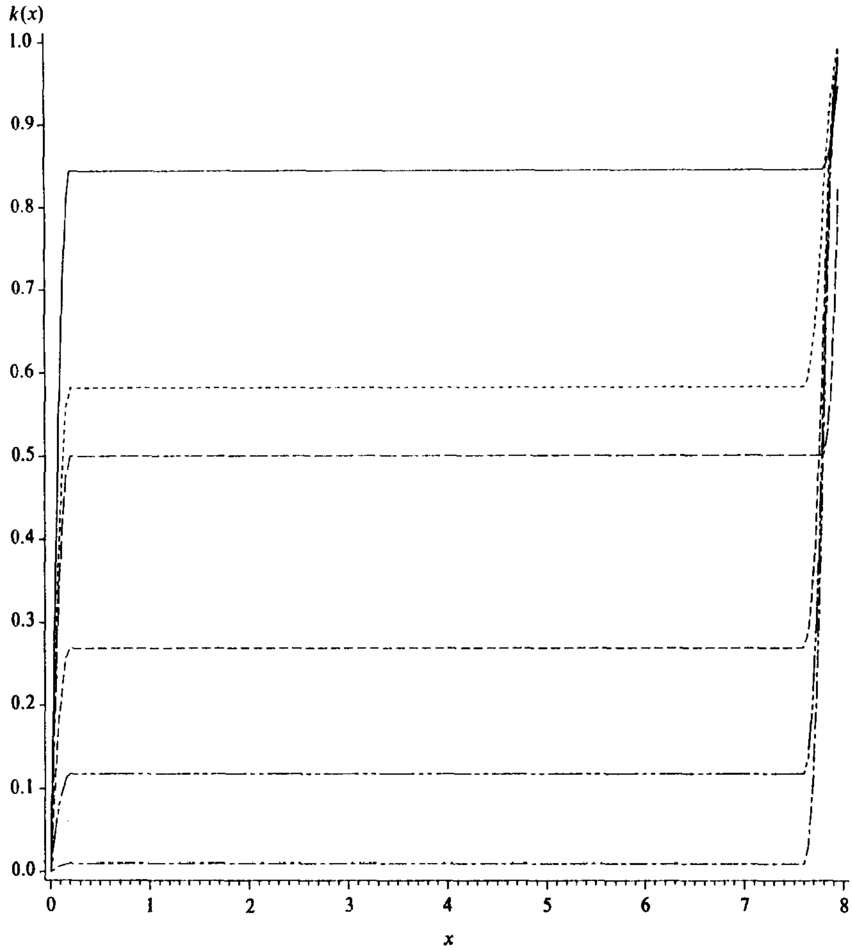
$$\min\{R_p^{(N)}(f_N^*): f_N^* \in K^{(N)^*}\}. \tag{2.20}$$

But (2.20) is a quadratic programming problem since $R_p^{(N)}(f_N^*)$ is clearly a convex quadratic objective function and the feasible set is a closed convex set in $\mathbb{R}^{MN+1}$ (defined by $MN+1$ inequality constraints and a single equality constraint) in the $MN+1$ variables f_{Ni}, $0 \leq i \leq MN$.

The solution $f_N^* = (f_{N0}, \ldots, f_{N,MN})$ to (2.20) gives the values of the optimal $f_N \in K_f^{(N)}$ at the points x_{Ni}, and the piecewise linearity of f_N defines it elsewhere. If N is sufficiently large, we may regard this solution f_N as corresponding to some $k \in K_I$ which is near the optimal $k_1^* \in K_1$.

2.3 *Approximations to k_1^* for selected p*

Before embarking on this discussion, we dispense with some special cases. First, if $p = N(0, \gamma^2)$ for some $\gamma^2 > 0$, it may be shown analytically that $R_p(k)$ in (1.5) is minimized over all measurable k (not just over $k \in K_1$) by $k_1^*(t) = \gamma^2/(1+\gamma^2)$, $t > 0$. Since this k_1^* happens also to lie in K_1 (as an a.u. limit), it is also the constrained (over K_1) minimizer. Thus, for normal p, no approximation is required. (We include a couple of normal priors in Fig. 1 below to check the computational algorithm.) The resulting estimator



1. ——— $U(-4,4)$: $v(p) = 16/3$, $c_p = .8446$
2. ------- $U(-2,2)$: $v(p) = 4/3$, $c_p = .5828$
3. ----- $U(-1,1)$: $v(p) = 1/3$, $c_p = .2683$
4. ——— $N(0,1)$: $v(p) = 1$, $c_p = .5000$
5. —-— $N(0,.01)$: $v(p) = .01$, $c_p = .0099$
6. --——· BETA (.5,.5): $v(p) = 1/8$, $c_p = .1172$

Fig. 1. Approximate optimal weight functions.

(1.1) is the usual Bayesian posterior mean.

We also note that for non-normal priors p, the constrained (over K_1) and unconstrained minimizers do not typically agree. In fact, the unconstrained minimizer is generally not nondecreasing.

This and the next section are, respectively, devoted to studying how the p-optimal k_1^* depends on p, and how the MSEE/EI properties of $\mu_n(k_1^*)$ vary with p. Therefore, for a variety of p, we used the quadratic programming technique described in Subsection 2.2 to compute approximations to the optimal weight function k_1^*. M was set equal to 8 and the grid size parameter N was set to 10, yielding a grid-width of 0.1 and a quadratic programming problem with $MN + 1 = 81$ variables. (When the prior was very concentrated—e.g., the $N(0, .01)$—a suitably smaller grid was used.) The reader should note that the priors with finite support do not satisfy the positivity restriction in (1.4) required for Theorem 2.1. However, we can alter these slightly by adding "super-light" tails, without significantly affecting k_1^*. For approximation of k_1^*, this is of no import.

Shown in Fig. 1 are the approximations for the $U(-4,4)$, $U(-2,2)$, $U(-1,1)$, $N(0,1)$ and Beta $(.5,.5)$ (shifted to have mean zero) priors. These priors are grouped together because the corresponding optimal weight functions k_1^* are apparently constant. The non-constant portions of

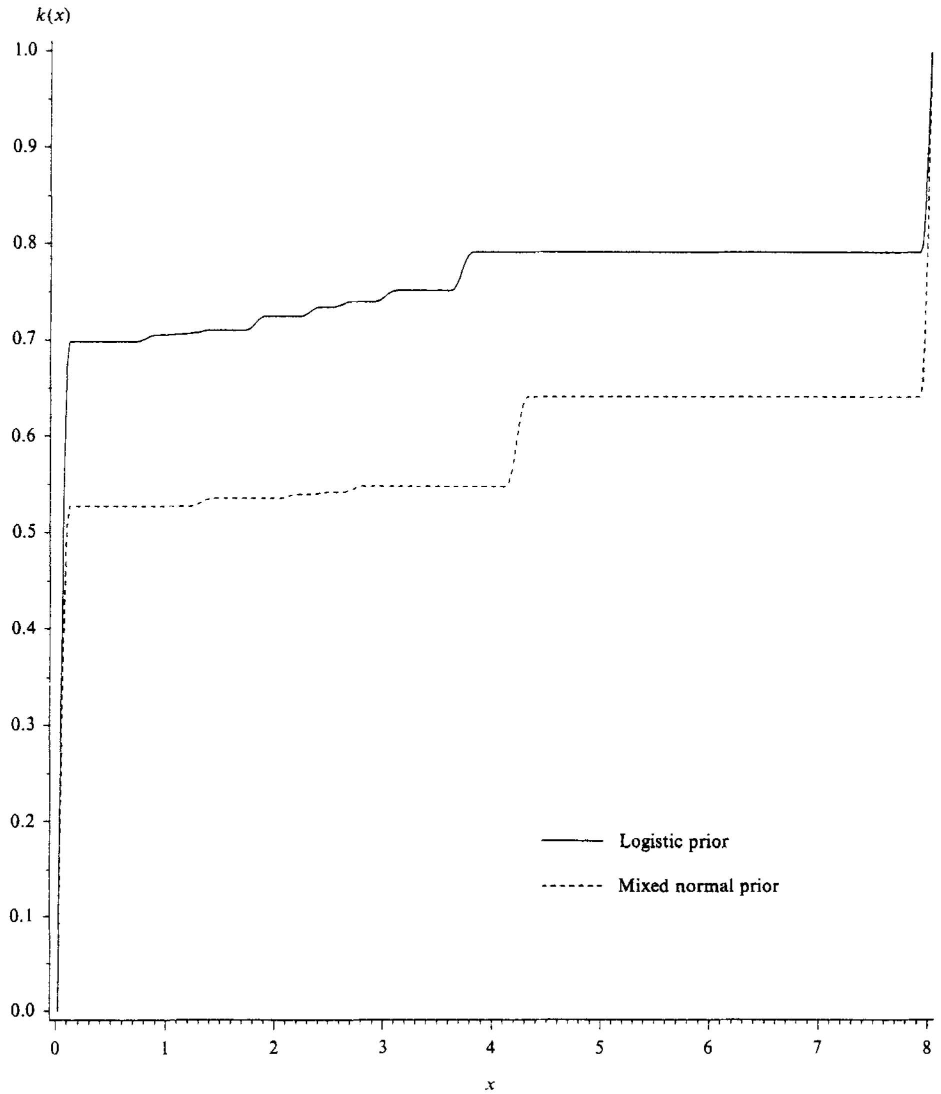


Fig. 2. Approximate optimal weight functions for logistic and mixed normal priors.

the functions, near $x = 0$ and $x = 8$, are due to constraint (ii)—recall (2.1)—on k_1. The x-values at which these functions become non-constant depend on the grid-width and are thus artificial. (The functions rise from height zero to a positive value at $x = 1/N$, and increase to one near $x = 8 - 1/N$, regardless of N.) The normal priors, as noted above, are included as test cases. The uniform priors are included because they provide precise concentration of the averaging process in (1.5) to a specific region of the parameter space, and hence allow intuitive control of the EI. The U-shaped Beta distribution is included for a look at the effects on k_1^* of a bimodal prior. We note that we also computed k_1^* for p a truncated normal, and k_1^* was essentially constant, as for the above priors.

For these p for which k_1^* is apparently a constant, say c_p, we see that c_p increases with $\upsilon(p) = \int p^2(\Delta)d\Delta$, the dispersion of p. In view of (1.1), this is intuitively reasonable. c_p apparently does not, however, depend on p only through $\upsilon(p)$, since $N(0,1)$ and $U(-\sqrt{3}, \sqrt{3})$ (not shown) both have $\upsilon(p) = 1$, but have $c_p = .500$ and $.513$, respectively.

Figure 2 shows the approximate optimal weight functions for the logistic and mixed normal ($.8 * N(0,1) + .2 * N(0,2)$) priors. For these priors, and also for the double exponential (not shown), k_1^* is apparently not constant. The question of which priors p have constant k and which do not is open to the authors.

3. Comparison of $\hat{\mu}_n(k_1^*)$ with other estimators

The estimator $\hat{\mu}_n(k_1^*)$ claims only to minimize the AMSE (average risk) with respect to the specified p, which as noted earlier only indirectly addresses EI and MSEE. Therefore, we now compare its EI and MSEE for the p in Subsection 2.3 with that of some of the estimators in the introduction.

First, consider the priors in $P' = \{p\colon p \text{ satisfies (1.4) and } k_1^*(t) = c_p(\text{constant})\ \forall\, t \in (0,\ M)\}$, which is apparently non-empty from Fig. 1. For these p, $\hat{\mu}_n(k_1^*)$ has a very simple form:

$$\hat{\mu}_n(k_1^*) = c_p \overline{X}_n + (1 - c_p)\mu_0\,, \tag{3.1}$$

with c_p depending on p (recall Fig. 1). We consider its MSEE at $\mu_n = \mu_0 + \Delta/\sqrt{n}$:

$$e_p(\Delta) = \mathrm{MSE}_\Delta(\overline{X}_n)/\mathrm{MSE}_\Delta(\hat{\mu}_n(k_1^*)) = [c_p^2 + (1 - c_p)^2\Delta^2]^{-1}\,, \tag{3.2}$$

where $\Delta = n^{1/2}(\mu - \mu_0)/\sigma$. (Note: Δ is in units of standard errors of $\overline{X}_n$.) From (3.2) it follows that the MSEE is a decreasing function of $|\Delta|$, with maximum value $e_p(0) = 1/c_p^2$. Also,

$$e_p(\Delta) \geq 1 \quad \text{iff} \quad |\Delta| \leq [(1 + c_p)/(1 - c_p)]^{1/2}, \tag{3.3}$$

which gives the EI in terms of Δ. The EI in terms of μ can be easily obtained.

We note, in view of (3.2), that for $p \in P'$ the EI and MSEE of $\hat{\mu}_n(k_1^*)$ depend on p only through c_p. This means that if we restrict p to P', the set of possible EI/MSEE combinations is indexed by $c_p \in (0, 1)$. Further, any family of priors $P_1' \subset P'$ such that $\{c_p: p \in P_1'\} = (0, 1)$ (call this property S) will generate all possible constant (between 0 and 1) weight functions, and hence all possible (for $p \in P'$) EI/MSEE combinations. The normal family $P_1' = \{N(0, \gamma^2): \gamma^2 > 0\}$ is such a family, and is convenient because it requires no approximation of k_1^*. The uniform priors (apparently in P' from Fig. 1) would do as well, but are not so convenient.

These results are interesting for several reasons. First, relative to our method of using the prior p as a tool for selecting k, they show that a large class (namely P') of priors (containing $N(0, \gamma^2)$ and apparently uniform and some betas) lead to constant k_1^*. In fact, the family of constant k's can apparently be generated by any of a number of families of priors. Although some priors lead to non-constant k_1^*, this seems to indicate an unattractive inflexibility of this method of choosing k to control EI and MSEE, since the EI/MSEE combinations seem restricted.

To the Bayesian, however, these results are very interesting. Since $\hat{\mu}_n(k)$ becomes the posterior mean estimator when p is $N(0, \gamma^2)$ (and $k_1^*(t) = \gamma^2/(1 + \gamma^2)$), he is pleased to note that any family $P_1' \subset P'$ with property S generates this same family of posterior mean estimators. In other words, the family (1.1) with k constant is in a sense "prior robust"! This could be made precise if P' were known. However, the characterization of P' is an open question to the authors.

To get some idea of the possible EI/MSEE combinations corresponding to P', Fig. 3 shows these functions for a selection of $N(0, \gamma^2)$ priors. For comparison the MSEE functions of the M & S estimator (with the recommended values $a = .302$, $b = .01$) and the Thompson estimator are also shown. We see that $\hat{\mu}_n(k_1^*)$, with $p = N(0, \gamma^2)$ and γ^2 between 2 and 3, will very closely approximate the performance of the M & S estimator. However, choosing γ^2 so that $\hat{\mu}_n(k_1^*)$ dominates the M & S estimator in terms of EI and MSEE does not appear possible.

The next question is whether any of the priors leading to non-constant k_1^* lead to $\hat{\mu}_n(k_1^*)$ with superior MSEE/EI. This is a formidable mathematical question which we did not attempt in general. Instead we computed the MSEE functions for the logistic and mixed normal to illustrate what can happen in this case.

Table 1 gives MSEE of $\hat{\mu}_n(k)$ for these priors and the p-optimal weight functions pictured in Fig. 2. By comparing these figures to Fig. 3, we observe the following. First, except for $\Delta \geq 3.5$, the estimator based on the

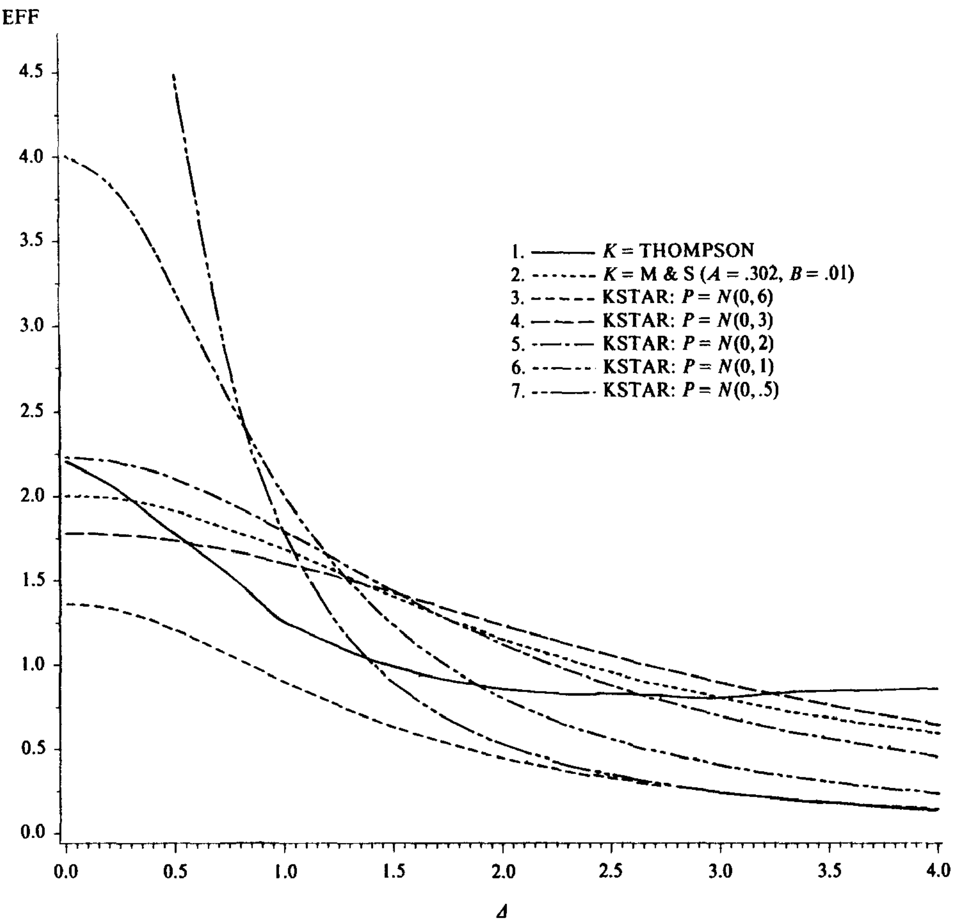


Fig. 3. Efficiency for normal priors versus other shrinkage estimators.

Table 1. MSEE for logistic and mixed normal priors.

	MSEE	
Δ	Logistic	Mixed normal
0.00	1.95	3.50
0.50	1.86	2.93
1.00	1.63	1.97
1.50	1.37	1.28
2.00	1.14	0.87
2.50	0.96	0.62
3.00	0.82	0.47
3.50	0.73	0.37
4.00	0.67	0.31

logistic prior is dominated by the M & S estimator which, as noted above, has approximately the MSEE of $\hat{\mu}_n(k)$ with prior $N(0, \gamma^2)$ for some $\gamma^2 \in (2, 3)$. The estimator based on the mixed normal prior is dominated by the one based on $N(0, 1)$ for $0 \leq \Delta < 1.5$, but slightly dominates this one for $\Delta \geq 1.5$. (This result is intuitive.)

These observations, though scanty, suggest that priors in the class P' yield estimators $\hat{\mu}_n(k)$ which are not uniformly dominated in MSEE by estimators corresponding to priors outside P', but may be dominated over some sets of Δ. We conjecture that this is the case, but as far as we know it is an open question.

4. Conclusions

We have studied an indirect way of controlling the MSEE function of $\hat{\mu}_n(k)$ by minimizing its average risk over $k \in K_1$ for a given prior p. With a normal prior, this method chooses a constant k in K_1, so that (1.1) becomes the usual Bayesian posterior mean estimator. For a suitable choice of the normal prior variance, this estimator compares well in terms of MSEE and EI with the M & S estimator, which was aimed directly at optimizing these criteria.

For an apparently large class of non-normal priors (including uniform, Beta, truncated normal) the p-optimal k_1^* is still constant, although the unconstrained (Bayes) minimizer of (1.5) is not constant. (Typically, it is decreasing in $|t|$, which is contrary to intuition.) These priors lead to the same class of posterior mean estimators as do the normal priors.

Other non-normal priors (e.g., the logistic and mixed normal) have non-constant p-optimal weight functions, but apparently do not give estimators (1.1) which uniformly dominate those estimators (1.1) generated by priors with constant weight functions.

Finally, we note a potentially fruitful alternative to (1.1), due to Srivastava and Ramkaran (1982), of the form

$$\hat{\mu}_n = \left(1 - \frac{g}{t_n^2 + k}\right)\overline{X}_n + \left(\frac{g}{t_n^2 + k}\right)\mu_0, \qquad t_n = \sqrt{n}(\overline{X}_n - \mu_0)/S_n\,,$$

where S_n is the sample standard deviation, and g and k are constants to be determined. These authors attempt to show that this estimator dominates $\overline{X}_n$ uniformly in MSEE over all μ, which of course contradicts the admissibility of $\overline{X}_n$ in the one-dimensional case. (The authors have been notified.) Nonetheless, estimators of this form, or generalized to depend on t_n in the spirit of (1.1) , may be worthy of further study.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *Second International Symposium on Information Theory*, (eds. B. N. Petrov and F. Csaki), 267–281, Akademiai, Budapest.

Ash, R. B. (1972). *Real Analysis and Probability*, Academic Press, New York.

Hirano, K. (1977). Estimation procedures based on preliminary test, shrinkage technique and information criterion, *Ann. Inst. Statist. Math.*, **29**, 21–34.

Inada, K. (1984). A minimax regret estimator of a normal mean after preliminary test, *Ann. Inst. Statist. Math.*, **36**, 207–215.

Mehta, J. S. and Srinivasan, R. (1971). Estimation of the mean by shrinkage to a point, *J. Amer. Statist. Assoc.*, **66**, 86–90.

Srivastava, V. K. and Ramkaran (1982). A class of shrunken estimates of the normal mean, *J. Statist. Plann. Inference*, **6**, 297–299.

Thompson, J. R. (1968). Some shrinkage techniques for estimating the mean, *J. Amer. Statist. Assoc.*, **63**, 113–122.