The Annals of Statistics
1991, Vol. 19, No. 4, 2259–2265

CONSTRAINED MINIMAX ESTIMATION OF THE MEAN OF THE NORMAL DISTRIBUTION WITH KNOWN VARIANCE

BY ISRAEL FELDMAN

Hebrew University

In this paper we shall discuss the estimation of the mean of a normal distribution with variance 1. The main question in this work is the existence and computation of a least favorable distribution among all the prior distributions satisfying a given set of constraints.

In the following we show that if this distribution is bounded from above on some even moment, then the least favorable distribution exists and it is either normal or discrete. The support of the discrete distribution function does not have any accumulation point. The least favorable distribution is normal if and only if the second moment is bounded from above, without any other relevant constraint.

These theorems shed light on the James–Stein estimator as the minimax estimator for a prior with unknown bounded variance.

1. Introduction. Let θ be the unobserved parameter and let $X = \theta + \varepsilon$ such that

$$\varepsilon \sim N(0,1).$$

We assume that θ is a random variable with unknown distribution. We also assume that there is some partial knowledge about this distribution, as specified below.

Let τ be the distribution function of θ. For a given observation x, let $\delta_\tau(x)$ be an estimate of θ.

Consider square error loss and define the risk function

$$R(\theta, \delta_\tau) = E\left[(\theta - \delta_\tau(X))^2 \middle| \theta\right].$$

Define the Bayes risk by $r(\tau, \delta_\tau) = E_\tau R(\theta, \delta_\tau)$. Assume that τ satisfies a set of moment constraints. Thus τ belongs to some set P'_k of distribution functions:

$$P'_k = \left\{\tau: l_n \le E\theta^{2n} \le m_n, \lambda_n \le E\theta^{2n-1} \le \mu_n, n = 1, \dots, k, E\theta = 0\right\}.$$

For a given distribution function τ, let δ_τ be the Bayes estimator with respect to τ. Thus $r(\tau, \delta_\tau) = \inf_\delta r(\tau, \delta)$.

Our problem is to find $\tau^* \in P'_k$ and $\delta^*_{\tau^*}$ such that

$$r(\tau^*, \delta^*_{\tau^*}) = \inf_\delta \sup_{\tau \in P'_k} r(\tau, \delta).$$

Received August 1989; revised October 1990.

AMS 1980 *subject classifications*. Primary 62F10; secondary 62F15, 62C99, 60E15.

Key words and phrases. Minimax risk, normal distribution, quadratic risk, estimating a bounded normal mean.

If such τ^* exists, it is the least favorable distribution among all distribution functions that belong to P_k' and $\delta^*_{\tau^*}$ is then the desired minimax estimator.

2. Background. Let τ be the prior distribution function of θ. Define

$$h(x) = \int_{-\infty}^{\infty} \phi(x-\theta)\, d\tau(\theta).$$

$h(x)$ is the marginal density of X and ϕ is the density function of a standard normal distribution. Then $\delta_\tau(x) = x + h'(x)/h(x)$ and $r(\tau, \delta_\tau) = 1 - \int_{-\infty}^{\infty} h'^2(u)/h(u)\, du$. The last equality is known as Brown's equality [see Bickel (1981)].

It is well known that if there are no moment constraints on the prior, the least favorable distribution does not exist and $\delta(x) = x$ is minimax [see Ferguson (1967)].

The problem of bounded normal mean became interesting because of two reasons: (a) x is no longer minimax as it is in the unrestricted problem. (b) The minimax estimator is not obvious.

Casella and Strawderman (1981) have found the exact minimax estimator and least favorable prior for problems with small bounds, but have not been able to determine the solution for larger values. Other convenient priors, such as uniform prior, have been tried by some people for problems with larger bounds, which can be used to improve the approximation of the minimax risk from the two-point and three-point Bayes risks. Levit (1980) studied the behavior of the minimax risk when the bound is large. Bickel (1981) claimed that, for a bound sufficiently large, the minimax risk can be well approximated by the Bayes risk with a "cos" prior. Ibragimov and Haminskii (1984) proved that the risk induced by the best linear estimator is within a finite factor to the minimax risk over all possible bounds. And this factor is later determined in a precise way by Donoho and Liu (1988) to be less than 1.25, with applications to nonparametric estimations. Another relevant problem was studied by Donoho and Johnstone (1988), when θ is known to lie in an n-dimensional l_p ball.

3. The main theorems. Let $P_k = \{\tau\colon\ E\theta^{2n} \le m_n,\ n = 1, \dots, k\}$, $0 < m_n \le \infty$, $n = 1, \dots, k-1$ and $0 < m_k < \infty$.

P_k is a set of distribution functions satisfying given constraints on the k first even moments.

Define a problem π_k as follows.

Find $\tau^* \in P_k$ and $\delta^*_{\tau^*}$ such that $r(\tau^*, \delta^*_{\tau^*}) = \inf_\delta \sup_\tau r(\tau, \delta)$.

By standard methods we can prove that there exists a solution $(\tau^*, \delta^*_{\tau^*})$ to the problem π_k for $k \ge 1$, such that $\tau^* \in P_k$ and $\delta^*_{\tau^*}$ is Bayes with respect to τ^*, and also

$$r(\tau^*, \delta^*_{\tau^*}) = \sup_\tau \inf_\delta r(\tau, \delta) = \inf_\delta \sup_\tau r(\tau, \delta).$$

THEOREM 1. *Let P_k be the set of all distribution functions that are constrained on their first k even moments* ($k \geq 1$).

$$P_k = \{\tau: E_\tau \theta^{2n} \leq m_n, n = 1, \dots, k\},$$

when $0 < m_n \leq \infty$, $n = 1, \dots, k-1$ *and* $0 < m_k < \infty$.

The least favorable distribution function is either normal with mean 0 *and variance* m_1 (*this happens if and only if the normal distribution with mean* 0 *and variance* m_1 *belongs to* P_k) *or it is discrete with support that does not have any accumulation point.*

PROOF. Assume that there is an interval $[a, b]$ in the support of τ_k^* that contains an accumulation point. Let δ_k^* be the minimax estimator that is Bayes with respect to τ_k^*, then

$$R(\theta, \delta_k^*) = \sum_{i=0}^{k} a_i \theta^{2i} \quad \text{a.e. } (\tau_k^*), \theta \in [a, b].$$

To see that, suppose that there is no real $a_0, \dots, a_k$ such that

$$R(\theta, \delta_k^*) = \sum_{i=0}^{k} a_i \theta^{2i} \quad \text{a.e. } (\tau_k^*), \theta \in [a, b].$$

Thus $R(\theta, \delta_k^*) \notin \mathrm{Span}(1, \theta^2, \dots, \theta^{2k})$ in the Hilbert space $L_2(\tau_k^*)$.

There are $c_0, \dots, c_k$ (obtained by the Gram–Schmidt orthogonalization procedure) such that

$$G(\theta) = R(\theta, \delta_k^*) - \sum_{i=0}^{k} c_i \theta^{2i}.$$

And also:

$$\int_a^b R(\theta, \delta_k^*) G(\theta)\, d\tau_k^*(\theta) > 0, \tag{1.1}$$

$$\int_a^b G(\theta) \theta^{2i}\, d\tau_k^*(\theta) = 0 \quad \text{for } i = 0, \dots, k. \tag{1.2}$$

Define

$$d\lambda_k(\theta)/d\tau^* = \begin{cases} 1, & \theta \notin [a, b], \\ 1 + \varepsilon G(\theta), & \theta \in [a, b]. \end{cases}$$

Choose $\varepsilon > 0$ such that $1 + \varepsilon G(\theta) > 0$ for every θ, $\theta \in [a, b]$.

It is obvious from (1.2) that

$$\int_{-\infty}^{\infty} \theta^{2i}\, d\lambda_k(\theta) = \int_{-\infty}^{\infty} \theta^{2i}\, d\tau_k^*(\theta), \qquad i = 0, \dots, k.$$

And hence $\lambda_k \in P_k$.

Compute the risk of δ_k^* with respect to λ_k:

$$r(\lambda_k, \delta_k^*) = r(\tau_k^*, \delta_k^*) + \varepsilon \int_a^b R(\theta, \delta_k^*) G(\theta)\, d\tau_k^*(\theta) > \tau(\tau_k^*, \delta_k^*).$$

The last inequality contradicts the fact that (τ_k^*, δ_k^*) is a saddle point such that

$$r(\tau_k^*, \delta_k^*) \geq r(\tau, \delta_k^*) \quad \text{for all } \tau.$$

Therefore $R(\theta, \delta^*) \in \mathrm{Span}(1, \theta^2, \dots, \theta^{2k})$ in the Hilbert space $L_2(\tau_k^*)$.

Let $\psi(x) = x - \delta_k^*(x)$.

$$|\psi(x)| = |x - E(\theta|x)| = \left| \int_{-\infty}^{\infty} (x - \theta) \phi(x - \theta)\, d\tau_k^*(\theta) \Big/ \int_{-\infty}^{\infty} \phi(x - \theta)\, d\tau_k^* \right|$$

$$\leq \int_{-\infty}^{\infty} |x - \theta| \phi(x - \theta)\, d\tau_k^*(\theta) \Big/ \int_{-\infty}^{\infty} \phi(x - \theta)\, d\tau_k^*.$$

Notice that $\phi(x - \theta)$ is a decreasing function in $|\theta - x|$, while $|x - \theta|$ is an increasing function in $|\theta - x|$. From Hardy, Littlewood and Polya (1952), it follows that

$$E\{|x - \theta| \phi(x - \theta)\} \leq E|x - \theta| E\phi(x - \theta).$$

Thus for a fixed x we get

$$(1.3) \quad 0 \leq |\psi(x)| \leq E|x - \theta| E\phi(x - \theta)/E(\phi - \theta) = E|x - \theta| \leq |x| + E|\theta|.$$

Using Stein's identity [see Bickel (1981), page 1303], we get

$$R(\theta, \delta_k^*) = 1 - E_\theta(2\psi'(x) - \psi^2(x)) = E_\theta(\psi^2(x) - 2\psi'(x) + 1).$$

Thus

$$R(\theta, \delta_k^*) = \int_{-\infty}^{\infty} (\psi^2(x) - 2\psi'^2(x) + 1) \phi(x - \theta)\, dx.$$

Since $\int_{-\infty}^{\infty} x^i \phi(x - \theta)\, dx$ is a polynomial in θ of degree i, there are $c_0, \dots, c_{2k}$ such that

$$\int_{-\infty}^{\infty} \sum_{i=0}^{2k} c_i x^i \phi(x - \theta)\, dx = \sum_{i=0}^{k} a_i \theta^{2i} = R(\theta, \delta_k^*) \quad \text{a.e. } (\tau_k^*).$$

Thus, from analyticity of $R(\theta, \delta_k^*)$ and the assumption of an accumulation point in the support of τ_k^*, it follows that the risk equals the degree $2k$ polynomial everywhere. However using (1.3) and analyticity of $\psi(x)$, it follows that the risk function can grow at most quadratically in θ. It also must be symmetric, so that

$$R(\theta, \delta_k^*) = a_0 + a_1 \theta^2.$$

Now we shall show that the risk function $R(\theta, \delta_k^*) = a_0 + a_1\theta^2$ is obtained by the linear estimator $\delta_b(x) = bx$.

For this estimator the risk is given by

$$R(\theta, \delta_b) = b^2 + (1-b)^2\theta^2.$$

It is clear that $a_0 \ge 0$ otherwise $R(\theta, \delta_k^*) < 0$ for $\theta = 0$ and also $a_1 \ge 0$ otherwise $R(\theta, \delta_k^*) < 0$ for some values of θ. If $a_1 > 1$, then δ_k^* is dominated by δ_0 and hence is not Bayes, a contradiction. If $a_1 = 1$ and $a_0 > 0$, then δ_k^* is dominated by δ_0, a contradiction. If $a_1 = 1$, $a_0 = 0$, then δ_k^* has the same risk function as δ_0, but δ_0 is not Bayes, a contradiction. If $0 \le a_1 < 1$, let $a_1 = (1-b)^2$. Then $a_0 < b^2$ contradicts the known admissibility of δ_b and $a_0 > b^2$ contradicts the assumed admissibility of δ_k^*.

Hence $a_0 = b^2$ and $\delta_k^* = \delta_b$ (a.e.).

In our problem there is one-to-one correspondence between the a priori distribution function and its generalized Bayes procedures [see Brown (1971)].

Thus, τ_k^* must be normal with mean 0 and variance $b/(1-b)$.

This means that when there is an accumulation point in some interval that belongs to the support of τ_k^*, τ_k^* must be normal. Suppose that τ_k^* is normal with mean 0 and variance μ such that $\mu < m_1$. Since $\tau_k^* \in P_k$,

$$E_{\tau_k^*}\theta^{2i} \le m_i, \qquad i = 1, \dots, k,$$

$$r(\delta_k^*, \tau_k^*) = E_{\tau_k^*}R(\theta, \delta_k^*) = E_{\tau_k^*}(a_0 + a_1\theta^2) = a_0 + a_1\mu.$$

Define distribution function λ as follows:

$$\lambda(\theta) = \begin{cases} (1-\varepsilon)\tau_k^*(\theta), & -\infty < \theta < -\sqrt{m_1}, \\ \frac{1}{2}\varepsilon + (1-\varepsilon)\tau_k^*(\theta), & -\sqrt{m_1} \le \theta < \sqrt{m_1}, \\ \varepsilon + (1-\varepsilon)\tau_k^*(\theta), & \sqrt{m_1} \le \theta < \infty, \end{cases}$$

$$E_\lambda\theta^{2i} \le (1-\varepsilon)m_i + \varepsilon m_1^i \le m_i.$$

Thus $\lambda \in P_k$.

$$r(\delta_k^*, \lambda) = E_\lambda R(\theta, \delta_k^*) = a_0 + a_1[(1-\varepsilon)\mu + \varepsilon m_1] > r(\delta_k^*, \tau_k^*),$$

a contradiction to the assumption that τ_k^* is a least favorable distribution function. Thus if τ_k^* is normal, it must be with mean 0 and variance m_1. In the case that the normal distribution does not belong to P_k, τ_k^* is discrete without any accumulation point in its support. □

Remarks. (a) In a special case when only the second moment of the distribution function is bounded from above, there is a short proof that a least favorable distribution is normal.

(b) Ghosh (1964) proved that when the mean is restricted in a given interval, the least favorable distribution puts mass on a finite number of points. This result can also be easily derived from the proof of the above Theorem 1. Since the least favourable prior cannot be normal in this case, it is discrete with finite support.

(c) Efron and Morris (1973) have showed that the James–Stein (1961) estimator can be derived by an empirical Bayes approach, assuming normality of the prior. Using our Theorem 1, this assumption can be replaced by a weaker one. The normality is derived by assuming that the prior has a bounded unknown variance and by taking the least favourable distribution function having this property, as a prior.

Theorem 2 is an extension of Theorem 1 when the constraints are given also on the odd moments. In this case the least favourable distribution must be nonsymmetric.

Let

$$P'_k = \{\tau\colon l_i \le E_\tau \theta^{2i} \le m_i,\, 0 < \lambda_i \le E_\tau \theta^{2i-1} \le \mu_i,\\ i = 1, \dots, k,\, k \ge 2,\, m_k < \infty,\, E\theta = 0\}.$$

Assume $P'_k \neq \varnothing$.

Define the problem π'_k in the same manner as π_k.

THEOREM 2. *There is a solution to the problem* π'_k *for* $k \ge 2$. *The least favourable distribution function* τ^* *is discrete with support that does not have any accumulation point.*

The proof is similar to the previous one and is omitted.

Acknowledgments. This paper is a part of a Ph.D. dissertation that was carried out under the supervision of Professor J. Yahav and Dr. Y. Ritov. Some important remarks were made by Professor L. D. Brown, who also helped in its preparation.

REFERENCES

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

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

BILLINGSLEY, P. (1978). *Probability and Measure*. Wiley, New York.

BROWN, L. D. (1971). Admissible estimators, recurrent diffusions and insoluble boundary problems. *Ann. Math. Statist.* **42** 855–903.

CASELLA, G. and STRAWDERMAN, W. E. (1981). Estimating a bounded normal mean. *Ann. Statist.* **9** 870–878.

DONOHO, D. L. and JOHNSTONE, I. (1988). Minimax risk over l_p balls. Technical Report, Univ. California, Berkeley.

DONOHO, D. L. and LIU R. C. (1988). Geometrizing rates of convergence, III. Technical Report 138, Univ. California, Berkeley.

EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors—an empirical bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130.

FERGUSON, T. S. (1967). *Mathematical Statistics*: *A Decision Theoretic Approach*. Academic, New York.

GHOSH, M. N. (1964). Uniform approximation of minimax point estimates. *Ann. Math. Statist.* **35** 1031–1047.

HARDY, G. H., LITTLEWOOD, J. G. and POLYA, G. (1952). *Inequalities*, 2nd ed. Cambridge Univ. Press.

IBRAGIMOV, I. A. and HASMINSKII, R. Z. (1984). Nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theory Probab. Appl.* **29** 1–32.

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 361–379. Univ. California Press, Berkeley.

LEVIT, B. YA. (1980). On the second order asymptotically minimax estimates. *Theory Probab. Appl.* **25** 561–576.

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 197–206. Univ. California Press, Berkeley.

DEPARTMENT OF STATISTICS
HEBREW UNIVERSITY OF JERUSALEM
MOUNT SCOPUS, JERUSALEM
ISRAEL 91905