The Annals of Statistics
1987, Vol. 15, No. 4, 1389–1400

ESTIMATING THE MEAN OF A NORMAL DISTRIBUTION WITH LOSS EQUAL TO SQUARED ERROR PLUS COMPLEXITY COST[1]

BY PETER J. KEMPTHORNE

Massachusetts Institute of Technology and Harvard University

Estimating the mean of a p-variate normal distribution is considered when the loss is squared error plus a complexity cost. The complexity of estimates is defined using a partition of the parameter space into sets corresponding to models of different complexity. The model implied by the use of an estimate determines the estimate's complexity cost. Complete classes of estimators are developed which consist of preliminary-test estimators. As is the case when loss is just squared error, the maximum-likelihood estimator is minimax. However, unlike the no-complexity-cost case, the maximum-likelihood estimator is inadmissible even in the case when $p = 1$ or 2.

1. Introduction. For estimating the mean of a multivariate normal distribution, most interest has focused on the performance of estimators under squared-error loss. Stein (1956) proved the inadmissibility of the maximum-likelihood estimator for this loss and the study of estimators which dominate the maximum-likelihood estimator has been extensive. See for example, Strawderman (1971), Efron and Morris (1976), Berger (1976, 1980, 1982a, 1982b) and George (1986). Many of the proposed estimators are also inadmissible and the characterization of admissible dominating estimators is addressed in Kempthorne (1986). An important result is that any discontinuous estimator is inadmissible. Preliminary-test estimators are of this form. See, for example, Sclove, Morris and Radhakrishnan (1972).

We address the estimation of a multivariate normal mean when the loss is equal to squared error *plus* a cost depending on the complexity of estimates. In this case, preliminary-test estimators which use maximum-likelihood estimates can be admissible. Meeden and Arnold (1979) prove the inadmissibility of a single such preliminary-test estimator in a one-dimensional problem. Stone (1982) extends their result to the case of two dimensions. In three or more dimensions, these "maximum-likelihood" preliminary test estimators can be inadmissible. See, for example. Bock (1980) and Ghosh and Dey (1984). These papers characterize dominating procedures in special cases but the admissibility of dominating estimators is not proven.

We characterize complete classes of estimators for this problem which necessarily include all admissible estimators. Any admissible estimator can be

Received June 1985; revised April 1987.

[1]Research supported in part by Office of Naval Research Contract N00014-75-0444, and National Science Foundation grants MCS-80-02535, MCS-78-25301, SES-84-01422 and DMS-86-05819.

AMS 1980 *subject classifications*. Primary 62C07; secondary 62C20.

Key words and phrases. Admissibility, complete class, generalized Bayes, minimax, preliminary-test estimators.

interpreted as a preliminary-test estimator. The stigma of discontinuous estimators vanishes when the loss incorporates complexity costs. Interestingly, only one preliminary-test estimator in the complete class uses maximum-likelihood estimates. Consequently, only one maximum-likelihood preliminary-test estimator could be admissible for a given loss.

For a common specification of the complexity costs, we also consider the performance of the maximum-likelihood estimator (which uses no preliminary test). We show that it is minimax, but it is inadmissible no matter what the dimension of the mean.

2. Decision-theoretic preliminaries. Let X be a p-variate normal random variable with unknown mean $\theta \in R^p$ and known $p \times p$ covariance matrix Σ. Consider estimating θ given X in a decision-theoretic framework with loss equal to squared error plus a cost depending on the complexity of an estimate. For $t \in R^p$, an estimate of θ, this loss is defined to be

$$L(t, \theta) = \|t - \theta\|^2 + C(t),$$

where $\|\cdot\|^2$ is the norm induced by an inner product $(\cdot, \cdot)$, specified by a positive-definite matrix M, such that for $u, v \in R^p$, $(u, v) = u^T M v$; $C(t)$ is the cost for using estimate t.

To define the complexity-cost function $C(\cdot)$, let $\{S_k, k = 1, 2, \ldots, K\}$ be a collection of closed convex subsets of $\Theta = R^p$, which is closed under intersections and contains Θ. For $t \in R^p$, set $C(t) = C_k$ if $t \in S_k$ and $t \notin S_j$ for all $S_j \subsetneq S_k$, that is, the cost associated with using an estimate from S_k. So that the costs assess complexity, assume that $C_j \geq 0$ for all j and $C_j \leq C_k$ for all j, k such that $S_j \subset S_k$.

For any estimator δ of θ given X, let $\delta(\cdot|x)$ denote the distribution of estimates of θ given $X = x$. When $\delta(\cdot|x)$ is degenerate for every $x \in R^p$, let $\delta(x)$ denote the estimate of the nonrandomized estimator. Assume that for all $x \in R^p$, $\delta(\cdot|x)$ is a probability measure on the Borel sets of $\bar{R}^p$ (the compactification of R^p) and for any fixed Borel set $A \subset \bar{R}^p$, $\delta(A|x)$ is Lebesgue measurable in x. Let D denote the collection of all such estimators.

For any estimator δ, the probability measures $\delta(\cdot|x)$, for $x \in R^p$, can be reexpressed in a form which incorporates two steps: First, a model S_k of Θ is chosen depending on the data; second, the parameter is estimated subject to the constraint that it lie in S_k. Define the partition $\{S_k', k = 1, 2, \ldots, K\}$ of R^p in terms of the S_j as follows. For each k, let

$$S_k' = S_k - \bigcup_{j \neq k:\, S_j \subset S_k} S_j.$$

It is easily verified that $\bigcup_k S_k' = R^p$ and $S_k' \cap S_j' = \varnothing$ for $j \neq k$. Each set in the partition consists of points which have the same complexity cost when considered as estimates of θ.

For an estimator δ, let

$$g_\delta(k|x) = \delta(S_k'|x), \qquad k = 1, 2, \ldots, K \text{ and } x \in R^p,$$

that is, the probability given x that the estimator δ assigns to that part of S_k which does not lie in a proper subset S_j. Define the measure $\delta_k(\cdot|x)$ as follows: For any Borel $A \subset R^p$,

$$\delta_k(A|x) = \begin{cases} \delta(A \cap S_k'|x)/\delta(S_k'|x), & \text{if } g_\delta(k|x) > 0, \\ \text{arbitrary}, & \text{if } g_\delta(k|x) = 0. \end{cases}$$

The probability measure given x of the estimator δ can now be expressed as

$$\delta(\cdot|x) = \sum_{k=1}^{K} g_\delta(k|x)\delta_k(\cdot|x),$$

The distribution $g_\delta(\cdot|x)$ on the indices of the sets $\{S_k', k = 1, 2, \ldots, K\}$ partitioning R^p can be interpreted as the critical function of a multiple-alternative hypothesis test used by estimator δ to decide which model to use before estimating the parameter. After deciding on the model, say k, θ is estimated with an estimate whose distribution given x is specified by $\delta_k(\cdot|x)$, a probability measure over S_k', the range of θ corresponding to model k and no proper submodel. The estimator δ can thus be interpreted as a preliminary-test estimator.

The performance of an estimator δ is completely characterized by its risk function

$$\begin{aligned} R(\delta, \theta) &= E_\theta L(\delta, \theta) \\ &= \int_X \left[\int_{\overline{R}_p} L(t, \theta)\delta(dt|x) \right] p(x|\theta)\, dx, \end{aligned}$$

where $p(\cdot|\theta)$ is the density of X given θ. An estimator is admissible if no other estimator has risk function which is everywhere as small as that of δ and smaller for some values of θ.

In the next section we address the characterization of complete classes of estimators. Theorem 3.3 applies the theory of Wald (1950) and Le Cam (1955): If D_B denotes the class of Bayes procedures, then under appropriate assumptions $\overline{D}_B$, the closure of D_B, is an essentially complete class. This closure is in the topology of regular convergence, which is defined as follows: A sequence $\{\delta_n, n = 1, 2 \ldots\}$ converges regularly to δ if for every Lebesgue-measurable function f satisfying $\int_{R^p} |f(x)|\, dx < \infty$ and every bounded continuous real-valued function u on R^p, which is zero outside a compact set, then

$$\lim_{n \to \infty} \int_{R^p} \int_{\overline{R}^p} f(x)u(t)\delta_n(dt|x)\, dx = \int_{R^p} \int_{\overline{R}^p} f(x)u(t)\delta(dt|x)\, dx.$$

3. Complete class results. Our first result is that a (strictly) complete class consists of all estimators δ, whose probability measures given x are discrete on the values (at x) of K nonrandomized estimators of θ: $\{t_k = t_k(x), k = 1, 2, \ldots, K\}$, where each estimator t_k is such that $t_k(x') \in S_k$ for all $x' \in R^p$. Theorem 3.2 characterizes the form of the Bayes procedures for this problem. Lindley (1968) proved a special case of this result in a purely Bayesian treatment

of the variable-selection problem in regression. Theorem 3.3 characterizes an essentially complete class of procedures which includes all procedures which are a limit of a sequence of Bayes rules whose risks are not everywhere infinite. This is the collection of generalized Bayes procedures which is a subclass of the complete class in Theorem 3.1. A version of Stein's (1955) necessary condition for admissibility given by Farrell (1968) is used to prove that the class in Theorem 3.3 is, in fact, strictly complete. Theorem 3.5 demonstrates that the class of generalized Bayes procedures coincides with the class of limits of Bayes procedures in the case of one dimension.

THEOREM 3.1. *A complete class of estimators consists of all δ whose probability measures*

$$\delta(\cdot|x) = \sum_{k=1}^{K} g_\delta(k|x)\delta_k(\cdot|x), \qquad x \in R^p,$$

are such that δ_k is nonrandomized on $\{x\colon g_\delta(k|x) > 0\}$, $k = 1, 2, \ldots, K$.

PROOF. Since each S_k is convex and $\{C_k\}$ is nondecreasing, this follows in the usual way from Jensen's inequality; see, for example, the proof of the Rao–Blackwell theorem in Berger [(1985), page 41]. □

The procedures in this complete class are almost nonrandomized. For the one-dimensional case, the class of nonrandomized procedures is complete; see Theorem 4.1 in Section 4. The argument is particular to the one-dimensional case, however, and the truth of the proposition in the multidimensional case remains an open question.

The complete class of Theorem 3.1 suggests that estimating θ proceeds in two stages. First, a possibly random preliminary test is applied to choose a model Θ_k among the K alternatives. Second, the best nonrandom estimator t_k for model Θ_k is used to estimate θ. This structure of admissible estimators will have an alternative interpretation for the complete class based upon the closure of the class of Bayes estimators; see the discussion following Theorem 3.4.

The Bayes procedures for this decision problem are addressed in Lindley (1968) for the special case of variable selection in regression. Their general characterization is provided by

THEOREM 3.2. *An estimator δ^π is Bayes with respect to the prior distribution π on Θ if, given $x \in R^p$, the corresponding probability measure $\delta^\pi(\cdot|x) = \Sigma_k g_{\delta^\pi}(k|x)\delta_k^\pi(\cdot|x)$ satisfies*

(i) $\delta_k^\pi(x) \equiv P_k m_\pi(x)$ *when* $g_{\delta^\pi}(k|x) > 0$ *and*
(ii) $g_{\delta^\pi}(k|x) > 0$ *only if* k *minimizes* $\|P_k m_\pi(x) - m_\pi(x)\|^2 + C_k$,

where P_k is the projection of R^p onto the subset S_k, orthogonal with respect to

the inner product $(\cdot, \cdot)$ *and*

$$m_\pi(x) = \frac{\int_{R^p} \theta p(x|\theta)\pi(d\theta)}{\int_{R^p} p(x|\theta)\pi(d\theta)},$$

the mean of the posterior distribution of θ *given* x.

We omit the proof since the argument is straightforward and parallels that given by Lindley (1968).

Let D_B denote the class of proper Bayes procedures for this decision problem, with $\bar{D}_B$ denoting its closure in the topology induced by regular convergence. Let D_0 be the subclass of procedures in $\bar{D}_B$ whose risks are not everywhere infinite. The essentially complete class D_0 is characterized in

THEOREM 3.3. *If* $\delta \in D_0$, *then there exists a measure* F *on* Θ *such that* $\int_{R^p} p(x|\theta)F(d\theta) < \infty$ *for all* x *and the measure* $\delta(\cdot|x) = \sum_{k=1}^K g_\delta(k|x)\delta_k(\cdot|x)$ *satisfies conditions* (*i*) *and* (*ii*) *of Theorem* 3.2 *with* π *replaced by* F.

PROOF. When the complexity-cost function is constant, the minimizing k in condition (ii) is that for which $S_k = R^p$ and δ always uses the nonrandomized estimate $m_F(x)$. The one-dimensional case was proven by Sacks (1963) and Brown (1971) proved the extension to $p > 1$ dimensions. We will use this result to prove the theorem when the complexity-cost function is not constant.

For a procedure $\delta^0 \in D_0$, let $\{\delta^n\}_{n=1}^\infty$ denote a sequence of Bayes procedures which converges regularly to δ^0 and let $\{F^n\}_{n=1}^\infty$ and $\{m_{F^n}\}_{n=1}^\infty$ denote the corresponding sequences of (proper) prior distributions and posterior mean functions.

First, we note that there exists a subsequence of $\{m_{F^n}\}_{n=1}^\infty$ which converges almost everywhere to a finite-valued function $m(x)$. To show this, let $k^n = k^n(x)$ denote the index of the subset selected by the Bayes procedure δ^n given x and observe that its loss satisfies

$$\begin{aligned} L(P_{k^n} m_{F^n}(x), \theta) &\geq \|P_{k^n} m_{F^n}(x)\|^2 - \|\theta\|^2 \\ &\geq \|m_{F^n}(x)\|^2 - \bar{C} - \|\theta\|^2, \end{aligned} \tag{3.1}$$

where $\bar{C} = \max_k C_k$ $(< \infty)$. The first inequality follows by ignoring the nonnegative complexity costs and the second by comparing the objective function of (ii) for k^n and the k for which $S_k = R^p$. The proposition is easily shown by assuming the contrary and using (3.1) to deduce the contradiction that the risks of the δ^n diverge for any θ.

By replacing sequences with subsequences, we may assume, without loss of generality, that the sequence of Bayes procedures $\{\delta_n\}_{n=1}^\infty$ converges regularly to δ^0 and the corresponding sequence of posterior mean functions $\{m_{F^n}\}_{n=1}^\infty$ converges almost everywhere to a finite-valued function $m(\cdot)$. Since the projections P_k, $k = 1, 2, \ldots, K$, are continuous bounded functions on R^p, $P_k m_{F^n}(x) \to P_k m(x)$ almost everywhere. Hence, the sequence of conditional distributions

defining the Bayes procedures $\{\delta_k^n(\cdot|x)\}_{n=1}^{\infty}$ converges weakly to $\delta_k(\cdot|x)$ almost everywhere which is degenerate at $P_k m(\cdot)$, for each k. Conditions (i) and (ii) of the theorem easily follow then with $m(\cdot)$ in place of $m_F(\cdot)$.

To complete the proof of the theorem, we must show that there exists a prior measure F such that

$$\int_{R^p} p(x|\theta)F(d\theta) < \infty \quad \text{and} \quad m_F(\cdot) = m(\cdot) \quad \text{a.e.} \tag{3.2}$$

The sequence of estimators $\{m_{F^n}(\cdot)\}$, which are Bayes with respect to the prior probability measures in the sequence $\{F^n\}$ for loss equal to squared error only, converge regularly to the estimator $m(\cdot)$. Note that the squared-error risk of $m(\cdot)$ has the following bound in terms of the risk of δ^0:

$$\int_{\chi} \|m(x) - \theta\|^2 p(x|\theta)\, dx \le R(\delta^0, \theta) + \bar{C}.$$

Since the risk of δ^0 is not everywhere infinite, so is the squared-error risk of $m(\cdot)$. Because the conditions of the theorem with no complexity costs are satisfied, (3.2) follows; see the arguments in the proofs of Theorems 2.2.1 and 3.1.1 in Brown (1971). □

The essential completeness of D_0 can be strengthened to show that D_0 contains all admissible procedures. Consider

THEOREM 3.4. *The class D_0 is complete.*

PROOF. We use Farrell's (1968) formulation of Stein's (1955) necessary and sufficient condition for admissibility; see also Berger [(1985), pages 546–547]. Because D_0 is essentially complete, a procedure δ^0 is admissible if and only if it is admissible relative to D_0. It is easily verified that the set of risk functions of procedures in D_0 is a sequentially weakly subcompact convex set of continuous real-valued functions on Θ, a σ-compact locally compact metric space. So, a procedure δ^0 is admissible if and only if there exists a sequence $\{F^n\}$ of (generalized) prior distributions such that

(a) each F^n has finite mass concentrating in a compact set $\Theta_n \subset \Theta$ with $\Theta_n \uparrow \Theta$,

(b) there is a compact set $C \subset \Theta$ such that $F^n(C) = 1$ for all n and

(c) $\lim_{n\to\infty}[\int_{\Theta} R(\delta^0, \theta)F^n(d\theta) - \int_{\Theta} R(\delta^n, \theta)F^n(d\theta)] = 0$ where δ^n is the Bayes procedure with respect to F^n.

Suppose δ^0 is an admissible procedure and let $\{F^n\}$ and $\{\delta^n\}$ be as given in (a)–(c). That $\delta^0 \in D_0$ will follow if we show that $\{\delta^n\}$ converges regularly to δ^0.

In (c), we can interchange the order of integration over Θ and χ because the risks of δ^n and δ^0 are continuous, finite and each F^n has finite mass concentrated on a compact set. After the first integration over Θ, we have that (c) is

equivalent to

$$\lim_{n\to\infty} \int_{\chi}\Big\{\int_{R^p}\big[\|t_0 - m_{F^n}(x)\|^2 + C(t_0)\big]\delta^0(dt_0|x) - \int_{R^p}\big[\|t_n - m_{F^n}(x)\|^2 + C(t_n)\big]\delta^n(dt_n|x)\Big\}p^n(x)\,dx = 0, \tag{3.3}$$

where $p^n(x) = \int_\Theta p(x|\theta)F^n(d\theta)$ is the (generalized) marginal density of X.

Since δ^n is Bayes, it concentrates on those estimates $t_n = P_{k^n}m_{F^n}(x)$ which minimize $\|t - m_{F^n}(x)\|^2 + C(t)$ almost everywhere. So the part of (3.3) in braces can be reexpressed as

$$\int_{R^p}\big[\|t_0 - m_{F^n}(x)\|^2 + C(t_0) - \|P_{k^n}m_{F^n}(x) - m_{F^n}(x)\|^2 - C_{k^n}\big]\delta^0(dt_0|x), \tag{3.4}$$

which is nonnegative.

By (a) and (b) for any compact set $A \subset \chi$ of positive Lebesgue measure, there exists an $\varepsilon = \varepsilon(A) > 0$ such that $p^n(x) > \varepsilon$ for all n and $x \in A$. Thus (3.3), equivalently (c), holds only if (3.4) converges to zero almost everywhere, i.e.,

$$\lim_{n\to\infty}\big[\|\delta^0(x) - m_{F^n}(x)\|^2 + C(\delta^0(x)) - \|P_{k^n}m_{F^n}(x) - m_{F^n}(x)\|^2 - C_{k^n}\big] = 0, \quad \text{a.e.}, \tag{3.5}$$

where $\delta^0(x)$ denotes the possibly random estimate of θ given x for δ^0.

By an argument analogous to that used in the proof of Theorem 3.3, it must be the case that the sequence $\{m_{F^n}\}$ converges a.e. to a finite-valued function $m(\cdot)$. Otherwise, the risk of the Bayes procedures would diverge and, thus, so would the limit in (c). It follows then that $\delta^0(\cdot|x)$ concentrates almost everywhere on those estimates $t = P_k m(x)$, where $k = k(x)$ is chosen to minimize

$$\|P_k m(x) - m(x)\|^2 + C_k.$$

As argued in the proof of Theorem 3.3, this is the regular limit of $\{\delta^n\}$. □

The complete class D_0 is a subclass of the almost nonrandom preliminary-test estimators in Theorem 3.1. The two-stage operation of estimators given earlier can now be revised to: First, the most accurate estimator $m_F(x)$ is determined, incorporating any prior beliefs into the specification of F and, second, when $X = x$ is observed, the most accurate approximation $P_k m_F(x)$ to $m_F(x)$ is chosen which balances the costs of complexity and squared-error inaccuracy. So, separate analyses for the K models are combined into one.

As a curiosity, we include a result concerning the converse of Theorem 3.3 for the case of one dimension.

THEOREM 3.5. *For $\Theta = R^1$, suppose that F is a measure on Θ such that $\int_\Theta p(x|\theta)F(d\theta) < \infty$ for all x and that δ_F is the generalized Bayes procedure*

with respect to the prior measure F, satisfying conditions (i) *and* (ii) *of Theorem* 3.2. *Then,* δ_F *is a limit of a sequence of Bayes procedures.*

To prove this result we will need

LEMMA 3.1. *Let F be any measure on* R^1 *satisfying the condition of Theorem* 3.5 *and let* $m_F(\cdot)$ *denote the mean of the (formal) posterior distribution with respect to F. For any* $m_0 \in R^1$, *let* $H = \{x: m_F(x) = m_0\}$. *Then H is either* R^1 *(if and only if F is concentrated on* m_0*) or at most one point [if and only if* $F((-\infty, m_0)) > 0$ *and* $F((m_0, \infty)) > 0$*].*

PROOF. If F is not concentrated on a single point, then $m_F(x)$ is strictly increasing because of the strict monotone-likelihood-ratio property of $p(x|\theta)$. See, e.g., Karlin (1956) or Karlin and Rubin (1956). □

PROOF OF THEOREM 3.5. The nontrivial case is when F does not concentrate on one point. For $n = 1, 2, \ldots$, define F^n to be the normalized measure F restricted to the interval $[-n, n]$. It then follows that $m_{F^n}(\cdot)$ converges almost everywhere to $m_F(\cdot)$. This implies that $g_{F^n}(\cdot|x) \to g_F(\cdot|x)$ and $\delta_{F^n}(\cdot|x) \to \delta_F(\cdot|x)$ except possibly on the set $\{x: m_F(x) \in E\}$, where

$$E = \left\{m: \text{there exists } k, k' \text{ such that } \|P_k m - m\|^2 + C_k = \|P_{k'} m - m\|^2 + C_{k'}\right\}.$$

However, E is a finite set since $\Theta = R^1$ and, hence by Lemma 3.1, so is $\{x: m_F(x) \in E\}$. □

The extension of Theorem 3.5 to the multidimensional case has been elusive. But this does not affect the utility of the theory for identifying inadmissible procedures. Theorems 3.3 and 3.4 provide us with a necessary condition for admissibility. The converse of Theorem 3.3 would not provide a sufficient condition for admissibility, but rather a sufficient condition for lying in the closure of the class of Bayes procedures. In fact, there are procedures satisfying the conditions of Theorem 3.3 which are inadmissible.

In the next section, we apply Theorem 3.5 to prove that in the one-dimensional case, the class of monotone nonrandomized generalized Bayes procedures is complete. In such problems, nonmonotone Bayes procedures can be shown to be inadmissible. Hence, the complete class D_0 is not minimal complete.

4. Related results. The general conclusion of Theorem 3.1 remains true if the loss function is the sum of two components: a function convex in t for each θ and a function which is discrete, depending only on the subsets S_k in which the estimate t and the parameter θ lie. Cohen (1965) addresses a problem of this form: Estimation of the parameter of a general exponential distribution under loss equal to squared error plus a cost for misspecification; i.e., if the true parameter is zero but a nonzero estimate is used, a cost is imposed, whereas if the true parameter had been nonzero, no cost would be incurred. He proves results similar to Theorem 3.1 and to Theorem 4.1 for the one-dimensional case.

Meeden and Arnold (1979) treat the special one-dimensional case of our problem with $K = 2$, $S_1 = R^1$, $S_2 = \{0\}$ and cost function $C(\cdot)$: $C(t) = C\ (> 0)$ if $t \neq 0$ and $C(0) = 0$. They suggest that the class of monotone estimators is essentially complete and claim that minor modifications of theory presented in Karlin and Rubin (1956) yield a proof. We present a simple proof of the completeness of the monotone nonrandomized procedures in the general one-dimensional problem.

THEOREM 4.1. *For the one-dimensional problem, the class of monotone nonrandomized generalized Bayes estimators is complete.*

PROOF. If δ is admissible, then δ satisfies conditions (i) and (ii) of Theorem 3.2 for some prior measure F. If F is concentrated on more than one point, then conditions (i) and (ii) uniquely determine δ almost everywhere and δ is nonrandomized as in Theorem 3.5. Hence, the only case that need be considered is where F concentrates on one point, say θ_0, and

$$H = \left\{k: (P_k\theta_0 - \theta_0)^2 + C_k = \min_j \left[(P_j\theta_0 - \theta_0)^2 + C_j\right]\right\}$$

contains at least two points.

There is then no loss of generality in assuming $H = \{1, 2, \ldots, j\}$; $P_i\theta_0 = \theta_i$, $i = 1, \ldots, j$, and $\theta_1 < \theta_2 < \cdots < \theta_j$. Define δ' as the monotone procedure

$$\delta'(\{\theta_1\}|x) = 1 - \delta'(\{\theta_j\}|x) = \begin{cases} 1, & \text{if } x \leq K', \\ 0, & \text{if } x > K', \end{cases}$$

where K' is determined so that

$$E_{X|\theta_0}\left[E_{\delta(\cdot|X)}(t|X)\right] = E_{X|\theta_0}\left[E_{\delta'(\cdot|X)}(t|X)\right].$$

Assume that δ is not of this form. Then $E_{\delta'}(t|x) - E_{\delta}(t|x)$ crosses zero once, at $x = K'$. So

$$e'(\theta) - e(\theta) \equiv E_{X|\theta}\left[E_{\delta(\cdot|X)}(t|X)\right] - E_{X|\theta}\left[E_{\delta'(\cdot|X)}(t|X)\right]$$

is less than zero for $\theta < \theta_0$ and is greater than zero for $\theta > \theta_0$ as a consequence of the strict monotone-likelihood-ratio property of the normal distribution.

Now, for $1 \leq i \leq k \leq j$,

$$\begin{aligned} L(\theta, \theta_i) - L(\theta, \theta_k) &= (\theta - \theta_0 + \theta_0 - \theta_i)^2 + C_i - \left[(\theta - \theta_0 + \theta_0 - \theta_k)^2 + C_k\right] \\ &= (\theta - \theta_0)(\theta_k - \theta_i), \end{aligned}$$

since $(\theta_0 - \theta_i)^2 + C_i = (\theta_0 - \theta_k)^2 + C_k$. Hence,

$$R(\delta, \theta) - R(\delta', \theta) = 2(\theta - \theta_0)[e'(\theta) - e(\theta)] \geq 0,$$

with strict inequality whenever $\delta \neq \delta'$ and $\theta \neq \theta_0$. It follows that the monotone procedures $\{\delta'\}$ dominate every other Bayes procedure relative to the given F. □

A common specification has the complexity of an estimate $t \in R^p$ depend only on which components of t are nonzero. If t is the estimate of a regression

parameter, then the nonzero components correspond to the subset of the p possible explanatory variables in the regression model which are selected for inclusion in the fitted model. See, for example, Kempthorne [(1982), Chapter 4].

An interesting consequence of Theorem 3.4 is that if there is a "null model," that is, a k^* with $S_{k^*} = \{0\}$ and $C_{k^*} = 0$, then $\delta_0(X) = X$, the maximum-likelihood estimator, is inadmissible no matter what the dimension of θ. The only estimator in D_0 which ever uses the estimate $\delta_0(x) = x$ is $\delta_* = \delta_F$, where F is Lebesgue measure on R^p. But δ_* does not use the estimate $t = x$ when $\|x\|^2 < C_1$, because the estimate $t = 0$ (with zero complexity cost) has smaller posterior risk. So δ_0 is not in the complete class D_0. This is contrary to the case of zero complexity cost when the maximum-likelihood estimator is always in the complete class D_0 and, for the cases of one and two dimensions, it is admissible. Despite its inadmissibility, however, the maximum-likelihood estimator is minimax. Consider

THEOREM 4.2. *Suppose that* $\{S_k,\ k = 1, 2, \ldots, K\}$ *with* $S_1 = R^p$ *and for each* $k > 1$, S_k *lies in a subspace of dimension less than* p. *The estimator* $\delta_0(X) = X$ *is minimax.*

PROOF. It is easy to verify that δ_0 has constant risk equal to $\operatorname{trace}(\Sigma M) + C_1$, where Σ is the covariance matrix of X, M is the defining matrix of the norm $\|\cdot\|$ and C_1 is the cost for complexity of an estimate t with full complexity. The minimaxity of δ_0 follows if we show that the risk of δ is the limit of Bayes risks of a sequence of Bayes procedures. See, for example, Theorem 2 of Ferguson [(1967), page 90] or Proposition 10.4.2 of Bickel and Doksum [(1977), page 426].

Consider the sequence of prior distributions $\{F^n,\ n = 1, 2, \ldots\}$, where F^n is the p-variate normal distribution with mean 0, covariance matrix $n\Sigma$ and density $f^n(\theta)$. Let $f^n(\theta|x)$ denote the density of the posterior distribution given x, which is normal with mean $\gamma_n x$ and covariance matrix $\gamma_n\Sigma$, where $\gamma_n = n/(n+1)$. The marginal distribution P^n of X is normal with mean 0 and covariance matrix $(n+1)\Sigma$. Denote its density by $p^n(x)$.

The Bayes risk of δ^n, the Bayes procedure for F^n, is

$$r(\delta^n, F^n) = \int_\Theta R(\delta^n, \theta) F^n(d\theta).$$

Interchanging the order of integration, the Bayes risk can be reexpressed as the average normalized posterior variance (which is independent of the Bayes estimates) plus the sum of the average normalized squared bias and the complexity of the Bayes estimates,

$$r(\delta^n, F^n) = \gamma_n \operatorname{trace}(\Sigma M) + \sum_{k=1}^{K} \int_{A_k^n} \left[\gamma_n^2 \|P_k x - x\|^2 + C_k\right] p^n(x)\, dx,$$

where $A_k^n = \{x\colon \delta^n(P_k\gamma_n X|x) = 1\}$, $k = 1, \ldots, K$.

Since the complexity costs are nonnegative, we can bound the Bayes risk from below,

$$r(\delta^n, F^n) \geq \gamma_n \operatorname{trace}(\Sigma M) + C_1 P^n(A_1^n).$$

For $k \neq 1$, it is clear that

$$\|P_k m_{F^n}(x) - m_{F^n}(x)\|^2 = \gamma_n^2 \|P_k X - P_k X\|^2 \to \infty \quad .$$

in probability as $n \to \infty$ under P^n. Consequently, $P^n(A_1^n) = P^n(\delta^n(X) = \gamma_n X) \to 1$. Thus, we have $r(\delta^n, F^n) \to \text{trace}(\Sigma M) + C_1 \equiv R(\delta_0, \theta)$. □

Perhaps surprisingly, the generalized Bayes estimator for the uniform prior, δ_* is not minimax. It is straightforward to show that its risk is smaller than $\delta_0(X) = X$ near $\theta = 0$, but it is larger when $\|\theta\|$ is large. This raises the problem of finding estimators which dominate the maximum-likelihood estimator when the loss incorporates a complexity cost. Kempthorne (1985) provides partial characterizations of admissible estimators which dominate the maximum-likelihood estimator under squared-error loss when $p \geq 3$. The approach is generalized in Kempthorne (1987) to the problem of dominating arbitrary inadmissible decision procedures. The theory and methods developed there can be applied to the case when the loss includes complexity costs.

Acknowledgments. I am indebted to the late Jack Kiefer for helpful discussions and advice during the research for this paper. The extremely helpful comments of an Associate Editor led to significant improvements in the original manuscript. Also, I thank Larry Brown for a discussion concerning how to strengthen Theorem 3.3 from an essentially complete class to a complete class result.

REFERENCES

BERGER, J. (1976). Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. *Ann. Statist.* **4** 223–226.

BERGER, J. (1980). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *Ann. Statist.* **8** 716–761.

BERGER, J. (1982a). Bayesian robustness and the Stein effect. *J. Amer. Statist. Assoc.* **77** 358–368.

BERGER, J. (1982b). Estimation in continuous exponential families: Bayesian estimation subject to risk restrictions and inadmissibility results. In *Statistical Decision Theory and Related Topics III* (S. S. Gupta and J. Berger, eds.) **1** 109–141. Academic, New York.

BERGER, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, New York.

BICKEL, P. J. and DOKSUM, K. A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day, San Francisco.

BOCK, M. E. (1980). Multiple subspace selection, in estimation of multivariate normal means. Mimeograph Series No. 80-31, Dept. Statistics, Purdue Univ.

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

COHEN, A. (1965). A hybrid problem on the exponential family. *Ann. Math. Statist.* **36** 1185–1206.

EFRON, B. and MORRIS, C. (1976). Families of minimax estimators of the mean of a multivariate normal distribution. *Ann. Statist.* **4** 11–21.

FARRELL, R. H. (1968). Towards a theory of generalized Bayes tests. *Ann. Math. Statist.* **39** 1–22.

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

GEORGE, E. (1986). A formal Bayes multiple shrinkage estimator. *Comm. Statist. A—Theory Methods* **15** 2099–2114.

GHOSH, M. and DEY, D. K. (1984). On the inadmissibility of preliminary-test estimators when the loss involves a complexity cost. Technical Report No. 230, Dept. Statistics, Univ. Florida.

KARLIN, S. (1956). Decision theory of Polya type distributions; case of two actions. I. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 115–120. Univ. California Press.

KARLIN, S. and RUBIN, H. (1956). The theory of decision procedures for distributions with monotone likelihood ratio. *Ann. Math. Statist.* **27** 272–299.

KEMPTHORNE, P. J. (1982). Variable selection and parameter estimation for normal linear regression models. Ph.D. dissertation, Dept. Statistics, Univ. California, Berkeley.

KEMPTHORNE, P. J. (1985). Controlling risks under different loss functions: The compromise decision problem. Technical Report NS-520, Dept. Statistics, Harvard Univ.

KEMPTHORNE, P. J. (1986). Optimal minimax squared error risk estimation of a multivariate normal distribution. *Comm. Statist. A—Theory Methods* **15** 2145–2158.

KEMPTHORNE, P. J. (1987). Dominating inadmissible procedures using compromise decision theory. In *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. Berger, eds.). To appear.

LE CAM, L. (1955). An extension of Wald's theory of statistical decision functions. *Ann. Math. Statist.* **26** 31–53.

LINDLEY, D. V. (1968). The choice of variables in multiple regression. *J. Roy. Statist. Soc. Ser. B* **30** 31–53.

MEEDEN, G. and ARNOLD, B. C. (1979). The admissibility of a preliminary test estimator when the loss incorporates a complexity cost. *J. Amer. Statist. Assoc.* **74** 872–874.

SACKS, J. (1963). Generalized Bayes solutions in estimation problems. *Ann. Math. Statist.* **34** 751–768.

SCLOVE, S. L., MORRIS, C. and RADHAKRISHNAN, R. (1972). Non-optimality of preliminary-test estimators for the multinormal mean. *Ann. Math. Statist.* **43** 1481–1490.

STEIN, C. (1955). A necessary and sufficient condition for admissibility. *Ann. Math. Statist.* **26** 518–522.

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.

STONE, C. J. (1982). Admissibility and local asymptotic admissibility of procedures which combine estimation and model selection. In *Statistical Decision Theory and Related Topics III* (S. S. Gupta and J. Berger, eds.) **2** 317–333. Academic, New York.

STRAWDERMAN, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math Statist.* **42** 385–388.

WALD, A. (1950). *Statistical Decision Functions*. Wiley, New York.

SLOAN SCHOOL OF MANAGEMENT
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02139