The Annals of Statistics
1980, Vol. 8, No. 5, 1036–1056

MINIMAX RIDGE REGRESSION ESTIMATION[1]

BY GEORGE CASELLA

Rutgers University

The technique of ridge regression, first proposed by Hoerl and Kennard, has become a popular tool for data analysts faced with a high degree of multicollinearity in their data. By using a ridge estimator, one hopes to both stabilize one's estimates (lower the condition number of the design matrix) and improve upon the squared error loss of the least squares estimator.

Recently, much attention has been focused on the latter objective. Building on the work of Stein and others, Strawderman and Thisted have developed classes of ridge regression estimators which dominate the usual estimator in risk, and hence are minimax. The unwieldy form of the risk function, however, has led these authors to minimax conditions which are stronger than needed.

In this paper, using an entirely new method of proof, we derive conditions that are necessary and sufficient for minimaxity of a large class of ridge regression estimators. The conditions derived here are very similar to those derived for minimaxity of some Stein-type estimators.

We also show, however, that if one forces a ridge regression estimator to satisfy the minimax conditions, it is quite likely that the other goal of Hoerl and Kennard (stability of the estimates) cannot be realized.

1. Introduction. Beginning with the work of Stein (1955), who showed that in higher dimensional problems, the sample mean of a multivariate normal distribution is inadmissible against squared error loss, much research has been aimed at developing estimators whose risk functions dominate that of the sample mean. More recently, a new estimation procedure, ridge regression, has been developed to improve upon the numerical stability of the least squares estimator in linear regression. Although it was not the original purpose of the ridge regression estimator to dominate the risk of the least squares estimator, recent research has gone in that direction.

In the present paper we develop a class of ridge regression estimators and, utilizing a new method of proof, derive necessary and sufficient conditions for these estimators to be minimax and thus dominate the least squares estimator in risk. We also point out that "forcing" ridge regression estimators to be minimax makes it difficult for them to provide the numerical stability for which they were originally intended.

We start with the familiar linear model

$$Y = Z\beta + \varepsilon, \tag{1.1}$$

Received August 1977; revised September 1978.

[1]This research is based on the author's Ph.D. dissertation submitted at Purdue University, May 1977, supported by the Air Force Office of Scientific Research, Air Force Systems Command under Grant AFOSR-72-2350C at Purdue University.

AMS 1970 *subject classifications.* Primary 62C99; secondary 62F10, 62H99, 62J05.

Key words and phrases. Minimax, ridge regression, normal distribution, mean, quadratic loss, risk function.

where Y is an $n \times 1$ vector of observations, Z is the known $n \times p$ design matrix of rank p, β is the $p \times 1$ vector of unknown regression coefficients, and ε is $n \times 1$ vector of experimental errors. We assume that ε has a multivariate normal distribution with mean vector zero and covariance matrix $\sigma^2 I_n$. (I_n denotes the $n \times n$ identity matrix.)

The usual estimator of β in (1.1) is the least squares estimator

$$\hat{\beta} = (Z'Z)^{-1}Z'Y. \tag{1.2}$$

$\hat{\beta}$ minimizes the residual sum of squares of the regression, i.e.,

$$\min_{\beta}(Y - Z\beta)'(Y - Z\beta) = (Y - Z\hat{\beta})'(Y - Z\hat{\beta}), \tag{1.3}$$

and is thus the estimate which best "fits" the data. Two different lines of research, however, pointed out deficiencies in $\hat{\beta}$.

The first deficiency in $\hat{\beta}$ is its inadmissibility. If we measure the loss of an estimator δ of β by

$$L(\delta, \beta, \sigma^2) = \frac{1}{\sigma^2}(\delta - \beta)'Q(\delta - \beta) \tag{1.4}$$

where Q is an arbitrary positive definite matrix, and let the risk of δ be given by

$$R(\delta, \beta, \sigma^2) = EL(\delta, \beta, \sigma^2), \tag{1.5}$$

then the results of Brown (1966) show that $\hat{\beta}$ is inadmissible. Several authors (e.g., Bhattacharya (1966), Berger (1976b)) have exhibited large classes of estimators whose risk functions dominate that of $\hat{\beta}$. Since $\hat{\beta}$ is a minimax estimator of β with constant risk

$$R(\hat{\beta}, \beta, \sigma^2) = \operatorname{tr} Q(Z'Z)^{-1}, \tag{1.6}$$

where "tr" denotes the trace operator, this search for estimators better than $\hat{\beta}$ is a search for minimax estimators.

A second deficiency in $\hat{\beta}$ was first noted by Hoerl and Kennard (1970). If the matrix Z arises from observation rather than from a designed experiment, it is possible that there will be high correlation among the Z variables. This will lead to a $Z'Z$ matrix that is "nearly singular", i.e., $Z'Z$ will have a wide eigenvalue spectrum. Hoerl and Kennard point out that, if this is the case, the least squares estimator $\hat{\beta}$ will be "unstable" in the sense that a nearly singular $Z'Z$ will produce an inverse with inflated diagonal values, and (see (1.2)) small changes in the observations might produce large changes in $\hat{\beta}$. To correct this problem, they proposed the ridge estimator

$$\hat{\beta}(k) = (Z'Z + kI_p)^{-1}Z'Y \tag{1.7}$$

where k is a positive number. Adding the number k before inverting amounts to increasing each eigenvalue of $Z'Z$ by k. This can be made clear as follows: Let P be the matrix of orthonormal eigenvectors of $Z'Z$, and let $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$ be

its eigenvalues. It follows that

$$P'Z'ZP = D_\lambda, \qquad P'P = I_p, \tag{1.8}$$

where $D_\lambda = \text{diag}(\lambda_1, \cdots, \lambda_p)$. Then (1.7) can be written as

$$\hat{\beta}(k) = \left(P'(D_\lambda + kI_p)P\right)^{-1}Z'Y. \tag{1.9}$$

To see that the ridge estimator is more stable than $\hat{\beta}$, we note that the condition number of the matrix being inverted in (1.9) is decreased. The condition number of a matrix is a measure of its ill-conditioning and is given by

$$\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}, \tag{1.10}$$

where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the largest and smallest roots of a matrix. Large values of $\kappa(A)$ imply that A is ill conditioned. Since

$$\frac{\lambda_1 + k}{\lambda_p + k} < \frac{\lambda_1}{\lambda_p} \tag{1.11}$$

for $k > 0$, the ridge estimator is relieving the ill-conditioning problem of $Z'Z$. A straightforward generalization of (1.9) is the generalized ridge estimator

$$\hat{\beta}(K) = \left(P'(D_\lambda + K)P\right)^{-1}Z'Y \tag{1.12}$$

where $K = \text{diag}(k_1, \cdots, k_p)$. Here, we allow each eigenvalue of $Z'Z$ to be increased by differing amounts.

Hoerl and Kennard list many properties of the ridge estimator, and prove the "ridge existence theorem". This theorem asserts that for a fixed parameter point β_0, there exists a value of k (or values of k_i, $i = 1, 2, \cdots, p$) depending on β_0, for which the risk of $\hat{\beta}(k)$ is smaller than the risk of $\hat{\beta}$. This theorem, together with results arising from the work of Stein, has lead to the search for minimax ridge estimators.

In Section 2, we discuss the canonical form of the problem, and develop the necessary notation. Section 3 contains the asymptotic (as the parameter value increases) results needed as a preliminary step in developing the main theorem. Section 4 contains the main theorem, the sufficient conditions for minimaxity of the estimators, while in Section 5 we show that for a smaller class of estimators these conditions are necessary and sufficient. Section 6 contains a discussion of the relationship between minimaxity and the conditioning problem.

2. The canonical problem. The technique of simultaneous diagonalization has found frequent use in proving minimaxity of classes of estimators (see, for example, Berger (1976b) or Strawderman (1978)). The problem is rotated into a space where both the covariance matrix and the loss matrix are diagonal, which greatly simplifies calculations while preserving minimaxity. However, with estimators of the form (1.12) it is necessary to simultaneously diagonalize three matrices $(Z'Z, P'KP, Q)$ which, in general, is not possible. A sufficient condition for the simultaneous diagonalization of these three matrices is that Q and $Z'Z$ have

common eigenvectors. In the absence of any prior knowledge, an experimenter will usually choose $Q = I$ or $Q = (Z'Z)^{-1}$ and the simultaneous diagonalization can be carried out. However, it is often the case that an experimenter has some knowledge of the losses he is willing to incur in the individual components, possibly from cost considerations or prior knowledge. In such cases, it is worthwhile for the estimator to perform well against an arbitrary choice of Q.

Since Hoerl and Kennard's estimator was proposed only with the choice $Q = I$ in mind, we cannot expect it to perform well when Q is arbitrary. A slight generalization, however, will handle any choice of Q. As an extension of (1.12) we define

$$\hat{\beta}(K, Q) = (Z'Z + M'KM)^{-1}Z'Y, \tag{2.1}$$

where M is a nonsingular matrix which simultaneously diagonalizes $Z'Z$ and Q. If Q and $Z'Z$ have common eigenvectors, (2.1) is the original ridge estimator. If D is the diagonal matrix of eigenvalues of $(Q^{-\frac{1}{2}}(Z'Z)Q^{-\frac{1}{2}})^{-1}$, M satisfies

$$\begin{aligned} M'D^{-1}M &= Z'Z \\ M'M &= Q, \end{aligned} \tag{2.2}$$

and showing that $\hat{\beta}(K, Q)$ is minimax against the loss

$$L(B, \beta, \sigma^2) = \frac{1}{\sigma^2}(B - \beta)'Q(B - \beta) \tag{2.3}$$

can be reduced as follows. $\hat{\beta}(K, Q)$ can be written

$$\begin{aligned} \hat{\beta}(K, Q) &= \left(M'(D^{-1} + K)M\right)^{-1}M'D^{-1}M\hat{\beta} \\ &= M^{-1}(D^{-1} + K)^{-1}D^{-1}M\hat{\beta}. \end{aligned} \tag{2.4}$$

Let $X = M\hat{\beta}$, $\theta = M\beta$. Since $\hat{\beta} \sim N(\beta, \sigma^2(Z'Z)^{-1})$, it follows that $X \sim N(\theta, \sigma^2 D)$. Also, from (2.2),

$$\begin{aligned} L(B, \beta, \sigma^2) &= \frac{1}{\sigma^2}(MB - M\beta)'(MB - M\beta) \\ &= \frac{1}{\sigma^2}(MB - \theta)'(MB - \theta). \end{aligned}$$

If we let $\delta(K, Q) = M\hat{\beta}(K, Q)$, we have

$$\delta(K, Q) = (D^{-1} + K)^{-1}D^{-1}X,$$

where the ith component can be written

$$\delta_i(K, Q) = \left(1 - \frac{k_i d_i}{k_i d_i + 1}\right)X_i, \tag{2.5}$$

and the loss of (2.3) becomes

$$L(\delta(K, Q), \theta, \sigma^2) = \frac{1}{\sigma^2}(\delta(K, Q) - \theta)'(\delta(K, Q) - \theta). \tag{2.6}$$

It then follows that $\hat{\beta}(K, Q)$ is minimax against loss (2.3) if and only if $\delta(K, Q)$ is minimax against the loss (2.6).

In the following we will suppress the dependence of the estimator on Q, and since K will be a function of X and s, the variance estimate, we will denote the ridge estimators by $\delta^R(X, s)$.

Finally, we note that since X is minimax with constant risk

$$R(X, \theta, \sigma^2) = EL(X, \theta, \sigma^2) = \operatorname{tr} D,$$

an estimator $\delta(X, s)$ is minimax if and only if

$$\Delta(\delta, \theta, \sigma^2) = R(\delta, \theta, \sigma^2) - R(X, \theta, \sigma^2) \leqslant 0, \qquad \forall \theta.$$

3. Tail minimax conditions. The form of Hoerl and Kennard's ridge estimator, while intuitively pleasing, leads to a rather complicated risk function. If one tries to apply Stein's integration-by-parts technique (Efron and Morris (1976)) in which an unbiased estimate of the risk is obtained and bounded above for all X, it seems that one is led either to bounds that, in some cases, are not sharp (Thisted (1976)) or to additional conditions on the estimator (Strawderman (1978)). The proof in this paper avoids these complications by obtaining an upper bound on the risk of $\delta^R(X, s)$ by an indirect method.

We begin with the concept of tail minimaxity, first introduced by Berger (1976a) to deal with losses other than quadratic. We use tail minimaxity here to obtain a simplified expression for the risk of $\delta^R(X, s)$.

DEFINITION 3.1. An estimator $\delta(X, s)$ is *tail minimax* if $\exists M > 0$ such that $\forall \theta$ satisfying $\theta'\theta > M, \Delta(\delta, \theta, \sigma^2) \leqslant 0$.

Since $\delta^R(X, s)$ shrinks X toward zero, (as can be seen from (2.5)), it should perform well against quadratic loss for small values of θ. Thus, we begin our investigation for minimax ridge estimators by examining conditions under which the risk of the ridge estimators dominates that of X for large values of θ, i.e., those that are tail minimax. We first develop conditions under which, for large values of θ, the quantity $Ef(X)$ can be approximated by $f(\theta)$ with error small enough to be ignored. We then use this approximation on the risk function of $\delta^R(X, s)$ to derive conditions for tail minimaxity.

From the work of Brown (1971) and Berger (1976a), it is reasonable to choose k_i so that the quantity

$$\gamma(X, s) = X - \delta(X, s), \tag{3.1}$$

is, for large values of $X'X$, approximately $c/X'X$ for some constant c, i.e.,

$$\gamma(X, s) \sim c/X'X. \tag{3.2}$$

To this end, we consider k_i of the form

$$k_i = \frac{a_i s r(X'D^{-1}X/s)}{X'D^{-1}X}, \tag{3.3}$$

where a_i is a positive constant and $r(\cdot)$ is a bounded function satisfying certain

regularity conditions. While the quadratic form in the denominator may contain any positive definite matrix and still satisfy (3.2), it will be important later in this paper for the quadratic form to have a noncentral chi-square distribution.

For k_i as in (3.3), the ridge estimator of (2.5) can be written componentwise as

$$\delta_i^R(X,s) = \left(1 - \frac{a_i d_i r(X'D^{-1}X/s)}{a_i d_i r(X'D^{-1}X/s) + X'D^{-1}X/s}\right)X_i, \quad 1 \leqslant i \leqslant p. \tag{3.4}$$

We start with the following lemma, which gives conditions on a function $f(x)$ under which, for large values of θ, $Ef(X)$ can be approximated by $f(\theta)$ with sufficiently small error.

LEMMA 3.1. *Let $X \sim N(\theta, I)$, and let the function $f: \mathbb{R}^p \to \mathbb{R}$ satisfy*

(i) *f has all second order partial derivatives;*

(ii) *$E(f(X) - f(\theta))^2 \leqslant K|\theta|^q$ for some constants q and K;*

(iii) *$\sup_{y>|\theta|/2}|f^{ij}(y) - f^{ij}(\theta)| = o(|\theta|^{-2})$, $1 \leqslant i,j \leqslant p$, where $f^{ij}(X) = (\partial^2/\partial X_i \partial X_j)f(X)$.*

Then as $|\theta| \to \infty$,

$$|Ef(X) - f(\theta)| = o(|\theta|^{-2}).$$

PROOF. The technique of proof is very similar to that used in Theorem 1 of Berger (1976a). Thus, we will only sketch the essential details.

Define the regions W and W^c by

$$W = \{X: |X - \theta| \leqslant |\theta|/2\}, \qquad W^c = \{X: |X - \theta| > |\theta|/2\}$$

and expand $f(X)$ in a Taylor series around θ (up to second order terms). It can then be verified that

$$|Ef(X) - f(\theta)| \leqslant \int_W |\rho(X,\theta)|d\Phi(X - \theta) + \int_{W^c}|f(X) - f(\theta)|d\Phi(X - \theta), \tag{3.5}$$

where $\Phi(\cdot)$ denotes the cumulative normal distribution with mean 0 and covariance I, and $\rho(X,\theta)$ is the remainder in the Taylor expansion. Then, using conditions (i)–(iii) it can be shown that the right-hand side of (3.5) is $o(|\theta|^{-2})$. □

We now derive the asymptotic expression for the risk of the estimator $\delta^R(X,s)$, given by (3.4), and the conditions under which it is tail minimax.

THEOREM 3.1. *Let $X \sim N(\theta, \sigma^2 D)$, $D = \operatorname{diag}(d_1, \cdots, d_p)$, and let $s \sim \sigma^2\chi_m^2$ be independent of X. Let the loss of an estimator $\delta(X,s)$ of θ be given by* (2.6), *and let $\delta^R(X,s)$ be the ridge estimator given by* (3.4) *where $r(t): \mathbb{R} \to [0,\infty)$ satisfies*

(i) *$r'(t) = o(t^{-\frac{1}{2}})$ as $t \to \infty$;*

(ii) *$r''(t) = o(t^{-\frac{3}{2}})$ as $t \to \infty$;*

(iii) *$r(t)$ is bounded and nondecreasing;*

(iv) *$r(t)/t$ is nonincreasing.*

If $\exists \varepsilon_1 > 0$ and $\varepsilon_2 > 0$ such that

$$\varepsilon_1 < r(t) < \left[2(m+2)^{-1}\left(\operatorname{tr} AD^2 - 2\lambda_{\max} AD^2\right)/\lambda_{\max} A^2D^3\right] - \varepsilon_2,$$

where $A = \text{diag}(a_1, \cdots, a_p), a_i \geqslant 0$, $1 \leqslant i \leqslant p$, *then* $\exists K > 0$ *such that* $\forall \theta' \theta > K, R(\delta^R, \theta, \sigma^2) \leqslant R(X, \theta, \sigma^2)$.

PROOF. Define $\Delta(\delta^R, \theta, \sigma^2) = R(\delta^R, \theta, \sigma^2) - R(X, \theta, \sigma^2)$. From (2.6) and (3.4) straightforward calculation yields

$$\Delta(\delta^R, \theta, \sigma^2) = (1/\sigma^2)\Sigma_{i=1}^p E\left\{ \frac{(a_i d_i r(t) X_i)^2}{(a_i d_i r(t) + t)^2} - \frac{2X_i(X_i - \theta_i) a_i d_i r(t)}{a_i d_i r(t) + t} \right\}, \tag{3.6}$$

where $t = X'D^{-1}X/s$. Integrating the last term in (3.6) by parts and defining

$$w_m = s/\sigma^2, \quad Z_i = X_i/\sigma, \quad v = Z'D^{-1}Z,$$
$$h_i(w_m, v) = (a_i d_i r(v/w_m) w_m + v)^{-1}$$

yields

$$\begin{aligned} \Delta(\delta^R, \theta, \sigma^2) = \Sigma_{i=1}^p E\{ h_i^2(w_m, v) [& a_i d_i r(v/w_m) w_m^2 Z_i^2 \\ & - 2h_i(w_m, v) a_i d_i^2 r(v/w_m) w_m \\ & + 4 a_i d_i r(v/w_m) w_m Z_i^2 \\ & - 4\sigma^{-2} w_m^{-2} a_i d_i Z_i^2 v r'(v/w_m)] \}. \end{aligned} \tag{3.7}$$

Since r is nondecreasing, the last term is bounded above by zero. Noting that $t = X'D^{-1}X/s = Z'D^{-1}Z/w_m$, and applying Lemma 4 of the Appendix to the function $q(t) = t^{-1}s(t)$, we have

$$E\{\chi_m^2 s(Z'D^{-1}Z, \chi_m^2)\} = mE\{s(Z'D^{-1}Z, \chi_{m+2}^2\}. \tag{3.8}$$

Using (3.8) on each of the first three terms of (3.7), bounding the last by zero, and rearranging terms gives

$$\begin{aligned} \Delta(\delta^R, \theta, \sigma^2) \leqslant m\Sigma_{i=1}^p E\{ & h_i^2(w_{m+2}, v) a_i d_i r(v/w_{m+2}) (a_i d_i r(v/w_{m+2}) w_{m+2} + 4) Z_i^2 \\ & - 2h_i(w_{m+2}, v) a_i d_i^2 r(v/w_{m+2}) \}. \end{aligned} \tag{3.9}$$

It follows from conditions (i) and (ii) that $r(v/w)$ is nonincreasing in w, and $wr(v/w)$ is nondecreasing in w, and hence the function

$$q_i(w) = \left(\frac{a_i d_i r(v/w) Z_i}{a_i d_i w r(v/w) + v} \right)^2$$

is nonincreasing in w. Applying Lemma 5, Appendix shows

$$E\{q_i(\chi_{m+2}^2)(\chi_{m+2}^2 - (m+2))\} \leqslant 0,$$

so that (3.9) is bounded above by

$$\begin{aligned} \Delta(\delta^R, \theta, \sigma^2) \leqslant m\Sigma_{i=1}^p E\{ & h_i^2(w, v) a_i d_i r(v/w) (a_i d_i r(v/w)(m+2) + 4) Z_i^2 \\ & - 2h_i(w, v) a_i d_i^2 r(v/w) \} \end{aligned} \tag{3.10}$$

where, from here on, $w = w_{m+2} \sim \chi^2_{m+2}$. Divide the region of integration of w into the two intervals

$$W_0 = \{w : w \leqslant M\},$$
$$W_1 = \{w : w > M\},$$

where M is a positive constant. The exact method of choosing M will be detailed later in the proof. Let $G_i(w, Z)$ denote the quantity in braces in expression (3.10) and let $F(\cdot)$ denote the cumulative χ^2 distribution with $m + 2$ degrees of freedom. Then

$$\Delta(\delta^R, \theta, \sigma^2) \leqslant m\int_{W_0}\Sigma_{i=1}^{p} E_Z G_i(w, Z)\, dF(w) + m\int_{W_1}\Sigma_{i=1}^{p} E_Z G_i(w, Z)\, dF(w).$$

Consider first the integral over W_1. Since $a_i d_i r(v/w)w \geqslant 0$ and $Z'D^{-1}Z > Z_i^2/d_i$,

$$\begin{aligned}(3.11)\quad &\int_{W_1}\Sigma_{i=1}^{p} E_Z(G_i(w, Z))\, dF(w)\\ &\leqslant \int_{W_1}\Sigma_{i=1}^{p} E_Z\left\{\left(\frac{a_i d_i r(v/w)}{v}\right)\right.\\ &\qquad \left.\times\left(a_i d_i^2 r(v/w)(m + 2) + 2d_i\right)\right\} dF(w)\\ &\leqslant \int_{W_1}\Sigma_{i=1}^{p} E_Z\left\{\left(\frac{a_i d_i r^*}{v}\right)\left(a_i d_i^2 r^*(m + 2) + 2d_i\right)\right\} dF(w)\\ &= \left[E_Z\left\{v^{-1}\left(\operatorname{tr}(m + 2)r^{*2}A^2D^3 + 2r^*AD^2\right)\right\}\right] P(w > M)\end{aligned}$$

where $r^* = \sup_t r(t)$. Since

$$E_Z(v^{-1}) = E_Z(Z'D^{-1}Z)^{-1} = \sigma^2/\theta'D^{-1}\theta + o(\sigma^2/\theta'\theta)$$

the last expression in (3.11) is equal to

$$(3.12)\quad (\sigma^2/\theta'D^{-1}\theta)\left(\operatorname{tr}\left[(m + 2)r^{*2}A^2D^3 + 2r^*AD^2\right]\right)P(w > M) + o(\sigma^2/\theta'\theta).$$

Consider next the integral over W_0. It is straightforward to verify that, for fixed w, $G_i(w, Z)$ satisfies the conditions of Lemma 3.1. Thus

$$\begin{aligned}\int_{W_0}\Sigma_{i=1}^{p} E_Z G_i(w, Z)\, dF(w) &= \int_{W_0}\Sigma_{i=1}^{p} G_i(w, \theta/\sigma)\, dF(w)\\ &\quad + \int_{W_0} o(\sigma^2/\theta'\theta)\, dF(w).\end{aligned} \tag{3.13}$$

Straightforward calculation will show that the individual terms comprising the $o(\sigma^2/\theta'\theta)$ term in (3.13), which are the higher order derivatives of $G_i(w, Z)$, can each be bounded by a function which is independent of w and of order $o(\sigma^2/\theta'\theta)$. Now define

$$\begin{aligned}\nu &= \theta'D^{-1}\theta/\sigma^2,\\ s_i(\nu/w, \theta) &= a_i d_i(a_i d_i r(\nu/w)(m + 2) + 4)\theta_i^2/\sigma^2,\\ \gamma_i(\nu, w) &= a_i d_i r(\nu/w)w/\left(a_i d_i r(\nu/w)w + \nu\right).\end{aligned}$$

Supressing the arguments of s_i and γ_i we can write (3.13) as

$$\begin{aligned}
&\int_{W_0}\Sigma_{i=1}^{p}E_Z G_i(w,Z)\,dF(w)\\
&\quad=\int_{W_0}\frac{r(v/w)}{v}\Sigma_{i=1}^{p}\left[(s_i/v)-2a_id_i^2\right]dF(w)\\
&\quad-\int_{W_0}\frac{r(v/w)}{v}\Sigma_{i=1}^{p}\left[\gamma_i\left((s_i/v)-2a_id_i^2\right)-\gamma_i^2 s_i/v\right]dF(w) \qquad (3.14)\\
&\quad+o(\sigma^2/\theta'\theta).
\end{aligned}$$

Recall $r^*=\sup_t r(t)$. Then for $w\in W_0$

$$\begin{aligned}
\gamma_i &\leqslant a_id_ir^*M(a_id_ir^*M+v)^{-1}, && 1\leqslant i\leqslant p,\\
s_i/v &\leqslant a_id_i^2\sigma^2(a_id_ir^*(m+2)+4), && 1\leqslant i\leqslant p,
\end{aligned}$$

and thus it is clear that the second integral in (3.14) is $o(v^{-1})=o(\sigma^2/\theta'\theta)$. Hence, summing the first term in (3.14) yields

$$\begin{aligned}
&\int_{W_0}\Sigma_{i=1}^{p}E_Z G_i(w,Z)\,dF(w)\\
&\quad\leqslant\int_{W_0}\frac{r(v/w)}{v}\left[\frac{r(v/w)(m+2)\theta'A^2D^2\theta+\theta'AD\theta}{\theta'D^{-1}\theta}-2\operatorname{tr}AD^2\right]dF(w)\\
&\quad+o(\sigma^2/\theta'\theta) \qquad (3.15)\\
&\quad\leqslant\int_{W_0}\frac{r(v/w)}{v}(\lambda_{\max}A^2D^3)(m+2)\\
&\qquad\times\left[r^*-\frac{2(m+2)^{-1}(\operatorname{tr}AD^2-2\lambda_{\max}AD^2)}{\lambda_{\max}A^2D^3}\right]dF(w)+o(\sigma^2/\theta'\theta),
\end{aligned}$$

since

$$\frac{\theta'A^2D^2\theta}{\theta'D^{-1}\theta}\leqslant\lambda_{\max}A^2D^3,\frac{\theta'AD\theta}{\theta'D^{-1}\theta}\leqslant\lambda_{\max}AD^2.$$

By assumption, the quantity in square brackets in (3.15) is bounded above by $-\varepsilon_2, r(v/w)\geqslant\varepsilon_1$, and since $\lambda_{\max}A^2D^3>0$,

$$\begin{aligned}
\int_{W_0}\Sigma_{i=1}^{p}E_Z G_i(w,Z)\,dF(w) &\leqslant\frac{\sigma^2}{\theta'D^{-1}\theta}\left[-\varepsilon_1\varepsilon_2\lambda_{\max}A^2D^3(m+2)P(w\leqslant M)\right]\\
&\quad+o(\sigma^2/\theta'\theta). \qquad (3.16)
\end{aligned}$$

Combining (3.12) and (3.16) yields

$$\begin{aligned}
\Delta(\delta^R,\theta,\sigma^2) &\leqslant\frac{\sigma^2 m}{\theta'D^{-1}\theta}\left\{-\varepsilon_1\varepsilon_2\lambda_{\max}A^2D^3(m+2)P(w\leqslant M)\right.\\
&\quad\left.+\left(\operatorname{tr}\left[(m+2)r^{*2}A^2D^3+2r^*AD^2\right]\right)P(w>M)\right\} \qquad (3.17)\\
&\quad+o(\sigma^2/\theta'\theta).
\end{aligned}$$

Now M is chosen large enough that

$$-\varepsilon_1\varepsilon_2\lambda_{\max}A^2D^3(m+2)P(w \leqslant M)$$
$$+\operatorname{tr}\left[(m+2)r^{*2}A^2D^3+2r^*AD^2\right]P(w>M) \leqslant -\varepsilon_3 < 0,$$

for some $\varepsilon_3 > 0$, and thus from (3.17),

$$\Delta(\delta^R,\theta,\sigma^2) \leqslant -\left(\varepsilon_3 m\sigma^2/\theta'D^{-1}\theta\right)+o(\sigma^2/\theta'\theta)$$
$$\leqslant -\left(\varepsilon_3 m\lambda_{\min}D\sigma^2/\theta'\theta\right)+o(\sigma^2/\theta'\theta),$$

and for sufficiently large $\theta'\theta, \Delta(\delta^R,\theta,\sigma^2) \leqslant 0$ and so $\delta^R(X,s)$ is tail minimax. □

While Theorem 3.1 does not guarantee that the risk of $\delta^R(X,s)$ will lie below that of X for any specified values of θ, it does provide a bound on the tail behavior of the risk function of $\delta^R(X,s)$. In the next section we show that this bound is, in fact, a global bound.

4. Sufficient conditions for minimaxity. The main theorem of this section, Theorem 4.1, extends the tail minimax bound of Theorem 3.1 to a global bound. We introduce a new method of proof, which differs sharply from the techniques previously used to prove minimaxity. Rather than bounding the risk function pointwise by a function which lies below $R(X,\theta,\sigma^2)$ we identify the extrema of $R(\delta^R,\theta,\sigma^2)$ and show that at these points the risk function of $\delta^R(X,s)$ is below that of X.

THEOREM 4.1. *Let* $\delta^R(X,s)$ *be the ridge estimator of* (3.4) *where* $r(t):\mathbb{R}\to[0,\infty)$ *satisfies conditions* (i)–(iv) *of Theorem* 3.1. *If*

$$0 \leqslant r(t) \leqslant 2(m+2)^{-1}\left[\operatorname{tr}AD^2-2\lambda_{\max}AD^2\right]/\lambda_{\max}A^2D^3, \tag{4.1}$$

$\forall t \geqslant 0$, *then* $\delta^R(X,s)$ *is minimax against the loss* (2.6).

PROOF. Assume that the bound in (4.1) is strict, i.e., $\exists\varepsilon_1$ and ε_2, both positive, such that $\forall t \geqslant 0$

$$\varepsilon_1 < r(t) < \left(2(m+2)^{-1}\left[\operatorname{tr}AD^2-2\lambda_{\max}AD^2\right]/\lambda_{\max}A^2D^3\right)-\varepsilon_2. \tag{4.2}$$

Then from Theorem 3.1 $\exists M>0$ such that $\forall\theta'\theta \geqslant M\ \Delta(\delta^R,\theta,\sigma) \leqslant 0$. Consider the set $\mathbb{S}=\{\theta:\theta'\theta\leqslant M\}$, a compact sphere in $\mathbb{R}^p$. We will bound $\Delta(\delta^R,\theta,\sigma^2)$ by a continuous function $\gamma(\delta^R,\theta,\sigma^2)$, which must have a maximum on $\mathbb{S}$. We will then show that if $\theta_0\in\mathbb{S}$ is an extremum of γ, then $\gamma(\delta^R,\theta_0,\sigma^2)\leqslant 0$. Thus, if M is taken sufficiently large it will follow from Theorem 3.1 that $\gamma(\delta^R,\theta,\sigma^2)\leqslant 0\ \forall\theta$. A simple argument, using Fatou's lemma, then allows the result to be extended to the case where the inequality in condition (4.1) is not strict.

Using the notation of Theorem 3.1, from (3.10) we have

$$\Delta(\delta^R,\theta,\sigma^2) \leqslant \gamma(\delta^R,\theta,\sigma^2)$$
$$= m\Sigma_{i=1}^p Eh_i^2(w,v)g_i(w,v)Z_i^2-2h_i(w,v)a_id_i^2r(v/w) \tag{4.3}$$

where $Z_i \sim n(\theta_i/\sigma, d_i)$, $v = Z'D^{-1}Z$, $w \sim \chi^2_{m+2}$ independent of Z and

$$g_i(w,v) = a_i d_i r(v/w)(a_i d_i r(v/w)(m+2)+4). \tag{4.4}$$

Letting $\chi^2_p(\alpha)$ denote a chi-square random variable with p degrees of freedom and noncentrality parameter $\alpha/2$, we have from Lemma 2, Appendix,

$$\begin{aligned}\gamma(\delta^R,\eta,\sigma^2) = m\Sigma_{i=1}^p E\{&d_i h_i^2(w,\chi^2_{p+2}(v))g_i(w,\chi^2_{p+2}(v))\\ &+\eta_i^2 h_i^2(w,\chi^2_{p+4}(v))g_i(w,\chi^2_{p+4}(v))\\ &-2a_i d_i^2 r(\chi^2_p(v)/w)h_i(w,\chi^2_p(v))\}\end{aligned} \tag{4.5}$$

where $\eta_i = \theta_i/\sigma$ and $v = \eta' D^{-1}\eta$. From (4.5) it can be seen that $\gamma(\delta^R,\eta,\sigma^2)$ is a function of η only through η_i^2. Thus, with the possible exception of $\eta = 0$, a point η_0 is an extreme point of $\gamma(\delta^R,\eta\sigma^2)$ only if

$$\frac{\partial}{\partial\eta_i^2}\gamma(\delta^R,\eta,\sigma^2)\Big|_{\eta=\eta_0} = 0, \qquad 1 \leqslant i \leqslant p. \tag{4.6}$$

We now show that if η_0 satisfies (4.6), then $\gamma(\delta^R,\eta_0,\sigma^2) \leqslant 0$. For the sake of clarity, we take some liberty with notation and define

$$\begin{aligned} b_i(p) &= Eh_i^2(w,\chi^2_p(v))g_i(w,\chi^2_p(v))\\ t_i(p) &= Eh_i(w,\chi^2_p(v))r(\chi^2_p(v)/w).\end{aligned} \tag{4.7}$$

Then

$$\gamma(\delta^R,\eta,\sigma^2) = m\Sigma_{i=1}^p\{d_i b_i(p+2) + \eta_i^2 b_i(p+4) - 2a_i d_i^2 t_i(p)\}. \tag{4.8}$$

From Lemma 6, Appendix, differentiating (4.8) with respect to η_i^2 yields

$$\begin{aligned}\frac{\partial}{\partial\eta_k^2}\gamma(\delta^R,\eta,\sigma^2) = \frac{m}{2d_k}\Sigma_{i=1}^p\{&d_i(b_i(p+4)-b_i(p+2))\\ &+\eta_i^2(b_i(p+6)-b_i(p+4))\\ &-2a_i d_i^2(t_i(p+2)-t_i(p))\}\\ &+mb_k(p+4).\end{aligned} \tag{4.9}$$

Notice that, with the exception of the multiplier d_k^{-1}, the sum in (4.9) does not depend on k, the index of differentiation. Denoting this sum by $\mathcal{D}(\eta)$,

$$\frac{\partial}{\partial\eta_k^2}\gamma(\delta^R,\eta,\sigma^2) = \frac{m}{2d_k}\mathcal{D}(\eta) + mb_k(p+4), \qquad k = 1,\cdots,p. \tag{4.10}$$

Thus, in order for (4.6) to be satisfied it must be the case that

$$d_i b_i(p+4) = d_j b_j(p+4), \qquad 1 \leqslant i,j \leqslant p. \tag{4.11}$$

Assume, without loss of generality, that $d_1 \geqslant d_2 \geqslant \cdots \geqslant d_p$. Lemma 7, Appendix, and the definition of the b_i's show that each b_i is strictly increasing in the quantity $a_i d_i$. Therefore, if $a_i d_i > a_j d_j$ for some $i < j$ (i.e., when $d_i \geqslant d_j$) then (4.11) cannot

be satisfied and $\gamma(\delta^R, \eta, \sigma^2)$ will have no extrema in $\mathcal{S}$. Therefore we only need consider the case $a_1 d_1 \leqslant a_2 d_2 \leqslant \cdots \leqslant a_p d_p$.

Note that the solution of (4.11), and hence of (4.6), depends only on $\nu_0 = \eta_0' D^{-1} \eta_0$, not on the particular η_0. Now suppose $\exists \nu_0$ such that (4.6) is satisfied, and define $\mathcal{H} = \{\eta : \eta' D^{-1} \eta = \nu_0\}$. We will show that $\gamma(\delta^R, \eta, \sigma^2) \leqslant 0 \ \forall \eta \in \mathcal{H}$. Combining (4.8) and (4.9) we have for $\eta \in \mathcal{H}$

$$\begin{aligned} \frac{\partial}{\partial \eta_k^2} \gamma(\delta^R, \eta, \sigma^2) &= 2 d_k^{-1} \gamma(\delta^R, \eta, \sigma^2) \\ &\quad + 2 m d_k^{-1} \Sigma_{i=1}^p \{ d_i b_i (p+4) + \eta_i^2 b_i (p+6) \\ &\quad - 2 a_i d_i^2 t_i (p+2) \} \\ &\quad + m b_k (p+4). \end{aligned} \tag{4.12}$$

Setting (4.12) equal to zero and using the identity $d_i b_i (p+4) = d_j b_j (p+4) \ \forall i,j$ yields

$$\begin{aligned} \gamma(\delta^R, \eta, \sigma^2) &= m \{ (p+2) d_k b_k (p+4) + \Sigma_{i=1}^p \eta_i^2 b_i (p+6) \\ &\quad - 2 \Sigma_{i=1}^p a_i d_i^2 t_i (p+2) \} \ \forall k = 1, \cdots, p. \end{aligned} \tag{4.13}$$

Set $k = p$ and recall that $a_p d_p \geqslant a_i d_i \forall i$. From the definition of t_i it is clear that

$$t_i (p+2) \leqslant t_p (p+2) \ \forall i.$$

Also, from Lemma 8, Appendix,

$$d_i b_i (p+6) \leqslant d_p b_p (p+6) \ \forall i.$$

Using these bounds in (4.13) yields for $\eta \in \mathcal{H}$

$$\begin{aligned} \gamma(\delta^R, \eta, \sigma^2) &\leqslant m \{ (p+2) d_p b_p (p+4) + \Sigma_{i=1}^p (\eta_i^2 / d_i) d_p b_p (p+6) \\ &\quad - 2 \Sigma_{i=1}^p a_i d_i^2 t_p (p+2) \} \\ &= m \{ d_p ((p+2) b_p (p+4) + \nu_0 b_p (p+6)) - 2 \operatorname{tr} A D^2 t_p (p+2) \}, \end{aligned}$$

where $\nu_0 = \Sigma \eta_i^2 / d_i$ and $\operatorname{tr} A D^2 = \Sigma a_i d_i^2$. Applying Lemma 3, Appendix to the first two terms on the right hand side, and recalling the definition of b_i, we have for $\eta \in \mathcal{H}$

$$\gamma(\delta^R, \eta, \sigma^2) \leqslant m E \left\{ \frac{a_p d_p^2 r(u^*) (a_p d_p r(u^*)(m+2) + 4) u}{(a_p d_p r(u^*) w + u)^2} - \frac{2 \operatorname{tr} A D^2 r(u^*)}{a_p d_p r(u^*) w + u} \right\},$$

where $u = \chi_{p+2}^2 (\nu_0)$ and $u^* = u/w$. Collecting terms, and using the fact that $u (a_p d_p r(u^*) w + u)^{-1} \leqslant 1$, we have for $\eta \in \mathcal{H}$,

$$\begin{aligned} &\gamma(\delta^R, \eta, \sigma^2) \\ &\leqslant m E \left\{ \frac{r(u^*)}{(a_p d_p r(u^*) w + u)} \left[a_p d_p^2 (a_p d_p r(u^*)(m+2) + 4) - 2 \operatorname{tr} A D^2 \right] \right\}. \end{aligned}$$

The right-hand side will be bounded above by zero if

$$r(t) \leqslant 2(m+2)^{-1}\left(\operatorname{tr} AD^2 - 2a_p d_p^2\right)/a_p^2 d_p^3 \ \forall t. \tag{4.14}$$

But (4.14) is implied by assumption (4.2). Hence $\gamma(\delta^R, \eta, \sigma^2) \leqslant 0 \ \forall \eta \in \mathfrak{N}$. If $\eta = 0$, or equivalently $\theta = 0$, it is obvious that

$$\Delta(\delta^R, 0, \sigma^2) \leqslant 0,$$

since $\delta^R(X,s)$ is always closer to zero than X. Thus, if (4.1) is replaced by (4.2), $\delta^R(X,s)$ is a minimax estimator of θ. If we define

$$r_\varepsilon(t) = (1-\varepsilon)r(t) + c\varepsilon, \tag{4.15}$$

where $0 < \varepsilon < 1$ and $c > 0$ satisfies

$$0 < c < 2(m+2)^{-1}\left(\operatorname{tr} AD^2 - 2\lambda_{\max} AD^2\right)/\lambda_{\max} A^2 D^3,$$

then the ridge estimator $\delta_\varepsilon^R(X,s)$ given componentwise by

$$\delta_{\varepsilon i}^R(X,s) = \left(1 - \frac{a_i d_i r_\varepsilon(X'D^{-1}X/s)}{a_i d_i r_\varepsilon(X'D^{-1}X/s) + X'D^{-1}X/s}\right)X_i,$$

satisfies the theorem with (4.1) replaced by (4.2), and hence is minimax $\forall \varepsilon$, $0 < \varepsilon < 1$. It is clear that $\lim_{\varepsilon \to 0} \delta_\varepsilon^R(X,s) = \delta^R(X,s)$, and thus from Fatou's lemma

$$\begin{aligned} R(X,\theta,\sigma^2) &\geqslant R\left(\delta_\varepsilon^R(X,s),\theta,\sigma^2\right) \\ &\geqslant \lim_{\varepsilon \to 0} \inf R\left(\delta_\varepsilon^R(X,s),\theta,\sigma^2\right) \\ &\geqslant E\left\{\lim_{\varepsilon \to 0} \inf L\left(\delta_\varepsilon^R(X,s),\theta,\sigma^2\right)\right. \\ &= EL\left(\delta^R(X,s),\theta,\sigma^2\right) \\ &= R\left(\delta^R(X,s),\theta,\sigma^2\right), \end{aligned}$$

and hence $\delta^R(X,s)$ is minimax. □

Condition (4.1) is essentially the same condition derived by other authors working with certain Stein-type estimators. For example, Bock (1975) showed that the spherically symmetric Stein-type estimator

$$\delta^B(X,s) = \left(1 - \frac{ar(X'D^{-1}X/s)}{X'D^{-1}X/s}\right)X$$

is minimax provided

$$0 \leqslant ar(t) \leqslant 2(m+2)^{-1}(\operatorname{tr} D - 2\lambda_{\max} D)/\lambda_{\max} D,$$

which is exactly the condition of Theorem 4.1 if we choose $a_i d_i = c$ to make $\delta^R(X,s)$ spherically symmetric. If $D = I, A = aI$, then (4.1) reduces to the familiar

$$0 \leqslant ar(t) \leqslant 2(p-2)(m+2)^{-1}.$$

Theorem 4.1 has an immediate extension to a wider class of functions. We state this in the following corollary.

COROLLARY 4.1. *Let $\delta^R(X,s)$ be given componentwise by*

$$\delta_i^R(X,s) = \left(1 - \frac{a_i d_i r(X'D^{-1}X,s)}{a_i d_i r(X'D^{-1}X,s) + X'D^{-1}X/s}\right)X_i, \tag{4.16}$$

where $r: \mathbb{R}^2 \to [0,\infty)$ satisfies

(i) $(\partial/\partial t_1)r(t_1,t_2) = o(t_1^{-\frac{1}{2}})$ *as* $t_1 \to \infty$;

(ii) $(\partial^2/\partial t_1^2)r(t_1,t_2) = o(t^{-\frac{3}{2}})$ *as* $t_1 \to \infty$;

(iii) $r(t_1,t_2)$ *is nondecreasing in* t_1 *and nonincreasing in* t_2;

(iv) $r(t_1,t_2)/t_1$ *is nonincreasing in* t_1;

(v) $r(t_1,t_2)t_2$ *is nondecreasing in* t_2.

If

$$0 \leqslant r(t_1,t_2) \leqslant 2(m+2)^{-1}(\operatorname{tr}AD^2 - 2\lambda_{\max}AD^2)/\lambda_{\max}A^2D^3, \tag{4.17}$$

for all $t_1, t_2 \geqslant 0$, then $\delta^R(X,s)$ is minimax against the loss (2.6).

The class of functions of Corollary 4.1 includes the ridge estimator $\delta^s(X,s)$, given componentwise by

$$\delta_i^s(X,s) = \left(1 - \frac{ad_i^{-1}}{ad_i^{-1} + X'D^{-1}X/s + g + h/s}\right)X_i \tag{4.18}$$

where a, g and h are positive constants. Strawderman (1978) showed $\delta^s(X,s)$ is minimax if

(i) $h \geqslant 0$;

(ii) $g \geqslant 2(p-2)(m+2)^{-1}$;

(iii) $a \leqslant (\min_i d_i)2(p-2)(m+2)^{-1}$.

If we define

$$r(X'D^{-1}X,s) = \frac{X'D^{-1}X/s}{X'D^{-1}X/s + g + h/s} \tag{4.19}$$

$$a_i = ad_i^{-2},$$

we can write (4.18) in the form given by (4.16). It is easy to verify that the function r in (4.19) satisfies the conditions of Corollary 4.1, and that the minimax bound (4.17) can be written

$$a \leqslant (\min d_i)2(p-2)(m+2)^{-1},$$

and that the restriction $g \geqslant 2(m+2)^{-1}(p-2)$ is not necessary.

5. Necessary and sufficient conditions. In this section we treat the case of known variance (i.e., $X \sim N(\theta, D)$), and show that condition (5.3) is, in fact, necessary and sufficient for minimaxity of the class of estimators developed here. The main theorem of this section is the following.

THEOREM 5.1. *Let $X \sim N(\theta, D)$, $D = \operatorname{diag}(d_1, \cdots, d_p)$, and let the ridge estimator $\delta^R(X)$ be given componentwise by*

$$\delta_i^R(X) = \left(1 - \frac{a_i d_i r(X'D^{-1}X)}{a_i d_i r(X'D^{-1}X) + X'D^{-1}X}\right)X_i, \qquad 1 \leqslant i \leqslant p, \tag{5.1}$$

where a_i are positive constants and $r: \mathbb{R} \to [0, \infty)$ satisfies

(i) $r'(t) = o(t^{-1})$ *as* $t \to \infty$;

(ii) $r''(t) = o(t^{-\frac{3}{2}})$ *as* $t \to \infty$;

(iii) $r(t)$ *is bounded and nondecreasing*;

(iv) $r(t)/t$ *is nonincreasing.*

$\delta^R(X)$ *is minimax against the loss*

$$L(\delta^R(X), \theta) = (\delta^R(X) - \theta)'(\delta^R(X) - \theta) \tag{5.2}$$

if and only if

$$0 \leqslant r(t) \leqslant 2(\operatorname{tr} AD^2 - 2\lambda_{\max} AD^2)/\lambda_{\max} A^2D^3, \tag{5.3}$$

for all $t \geqslant 0$, *where* $A = \operatorname{diag}(a_1, \cdots, a_p)$.

REMARK. Condition (i) is a slightly stronger requirement on the first derivative of r than was previously needed, and is only needed for the necessity of the theorem. The sufficiency of the theorem holds if $r'(t) = o(t^{-\frac{1}{2}})$. It should be noted, however, that the strengthening of this condition merely eliminates the more pathological choices of the function r.

PROOF. The sufficiency will follow from Theorem 4.1. Define $\delta^R(X, s)$ componentwise by

$$\delta_i^R(X, s) = \left(1 - \frac{a_i d_i r(t)}{a_i d_i r(t) + t}\right) X_i, \qquad 1 \leqslant i \leqslant p,$$

where $t = (m + 2)X'D^{-1}X/s$, r satisfies conditions (i)–(iv) and $s \sim \chi_m^2$ independent of X. From Theorem 4.1, $\delta_i^R(X, s)$ is minimax if

$$0 \leqslant r(t) \leqslant 2(\operatorname{tr} AD^2 - 2\lambda_{\max} AD^2)/\lambda_{\max} A^2D^3, \qquad \forall t \geqslant 0.$$

Since $\lim_{m\to\infty} s(m + 2)^{-1} = 1$ a.e., it follows that $\lim_{m\to\infty} \delta^R(X, s) = \delta^R(X)$. Also, from Lebesgue's dominated convergence theorem it is easy to verify that

$$\lim_{m\to\infty} R(\delta^R(X, s), \theta) = R(\delta^R(X), \theta),$$

and hence the sufficiency is proved.

For the necessity, we again define $\Delta(\delta^R, \theta) = R(\delta^R, \theta) - R(X(X), \theta)$. We proceed as in Theorem 3.1, integrating by parts and applying Lemma 3.1. We note that condition (i) insures that the term involving $r'(t)$ is $o(|\theta|^{-2})$. After collecting terms we have for sufficiently large θ,

$$\Delta(\delta^R, \theta) = \frac{r(\tau)}{\tau}\left\{\frac{r(\tau)\theta' A^2D^2\theta + 4\theta' AD\theta}{\theta' D^{-1}\theta} - 2\operatorname{tr} AD^2\right\} + o(|\theta|^{-2}), \tag{5.4}$$

where $\tau = \theta' D^{-1}\theta$. Define a sequence of vectors θ_n^* as follows. Note that the matrices A^2D^3 and AD^2 have common eigenvectors, and let α^* be the normed eigenvector of A^2D^3 corresponding to its largest root. α^* is then also the normed eigenvector of AD^2 corresponding to its largest root. Define θ_n^* by

$$\theta_n^* = n^{\frac{1}{2}}\alpha^* / (\alpha^{*\prime} D^{-1}\alpha^*)^{\frac{1}{2}}.$$

Then $\theta^{*\prime}D^{-1}\theta^* = n$ and

$$\frac{\theta_n^{*\prime}A^2D^2\theta_n^*}{\theta_n^{*\prime}D^{-1}\theta_n^*} = \frac{\alpha^{*\prime}A^2D^2\alpha^*}{\alpha^{*\prime}D^{-1}\alpha^*} = \lambda_{\max}A^2D^3.$$

Similarly, $\theta_n^{*\prime}AD\theta_n^*/\theta_n^*D^{-1}\theta_n^* = \lambda_{\max}AD^2$. Thus (5.4) becomes, for $\theta_i = \theta_n^*$,

$$\begin{aligned}\Delta(\delta^R, \theta_n^*) &= \frac{r(n)}{n}\{r(n)\lambda_{\max}A^2D^3 + 4\lambda_{\max}AD^2 - 2\operatorname{tr}AD^2\} + o(|\theta|^{-2}) \\ &= \frac{r(n)}{n}\lambda_{\max}A^2D^3\{r(n) - 2(\operatorname{tr}AD^2 - \lambda_{\max}AD^2)/\lambda_{\max}A^2D^3\} \\ &\quad + o(n^{-1}).\end{aligned}$$

Now suppose (5.3) is violated, i.e., $\exists T > 0$ and $\varepsilon > 0$ such that $\forall t > T$,

$$r(t) > \left(2(\operatorname{tr}AD^2 - 2\lambda_{\max}AD^2)/\lambda_{\max}A^2D^3\right) - \varepsilon > 0. \tag{5.5}$$

It then follows that for sufficiently large n

$$\Delta(\delta^R, \theta_n^*) > \varepsilon\frac{r(n)}{n}\lambda_{\max}A^2D^3 + o(n^{-1}) \tag{5.6}$$

and, since (5.5) bounds $r(t)$ from below, for sufficiently large n (5.6) is positive and $\delta^R(X)$ is not minimax. Therefore, the contrapositive and hence the theorem is proved. □

The proof of necessity in Theorem 5.1 did not require condition (iii) on $r(\cdot)$. We state this in the following corollary.

COROLLARY 5.1. *Let $\delta^R(X)$ be the ridge estimator of* (5.1) *where $r: \mathbb{R} \to [0, \infty)$ is bounded and satisfies*

(i) $tr'(t) = o(1)$;

(ii) $t^{\frac{3}{2}}r''(t) = o(1)$.

If $\delta^R(X)$ is minimax against the loss (5.2), then

$$\lim_{t\to\infty} \inf r(t) \leqslant 2(\operatorname{tr}AD^2 - \lambda_{\max}AD^2)/\lambda_{\max}A^2D^3.$$

Thisted (1976) derived necessary conditions similar to the above for a different class of ridge estimators.

6. Minimaxity and conditioning. The crucial condition for the minimaxity of $\delta^R(X)$ is that

$$0 \leqslant r(t) \leqslant 2(\operatorname{tr}AD^2 - 2\lambda_{\max}AD^2)/\lambda_{\max}A^2D^3, \tag{6.1}$$

and hence, it must necessarily be the case that

$$\operatorname{tr}AD^2 > 2\lambda_{\max}AD^2. \tag{6.2}$$

We wish to point out the following inconsistency between the original goal of ridge regression estimators and the performance of minimax ridge regression estimators. Hoerl and Kennard saw ridge regression as a solution to the "ill-conditioning"

problem mentioned earlier, which means, in particular, that the a_i's should be chosen such that

$$a_i d_i \leqslant a_j d_j \text{ when } d_i \leqslant d_j, \qquad 1 \leqslant i,j \leqslant p, \tag{6.3}$$

which will lower the condition number of the matrix inverted in the regression situation, and lead to what Hoerl and Kennard refer to as a more "stable" estimator.

Choosing the a_i's to satisfy (6.3) is also intuitively appealing since it seems sensible to add only small amounts of bias to directions which are providing good information (small d_i's). One can also view the ridge estimator as a combination of an estimator based on sample information and a prior guess that the mean is zero, with the a_i's being the weights given to the prior guess. It is well known that if we assume $\theta \sim N(0, \sigma^2 K^{-1}), K = \operatorname{diag}(k_1, \cdots, k_p)$, then the estimator given componentwise by

$$\delta_i^B(x) = \left(1 - \frac{k_i}{k_i + d_i^{-1}}\right) X_i \tag{6.4}$$

is Bayes against squared error loss. The ridge estimator (5.1) is in the form of (6.4) with $k_i = a_i r(X'D^{-1}X)/X'D^{-1}X$. Although this argument is not a formal justification, it lends credence to the interpretation of the a_i's as weights for a prior guess of the mean vector. Thus if the sampling information is good (in the form of a small values of d_i) it is reasonable to down weight the prior guess (and choose a smaller value of a_i).

An inconsistency arises, however, when the condition of minimaxity is forced into the estimator. If the d_i's are very unequal (as will occur in an ill-conditioned problem), the matrix D is likely to satisfy

$$\operatorname{tr} D^2 \leqslant 2\lambda_{\max} D^2. \tag{6.5}$$

As the number of dimensions, p, increases, it is more likely that the inequality in (6.5) will reverse, but in general one would expect (6.5) to be the case. If the ridge estimator is to be minimax, (6.2) must hold, so the a_i's must be chosen to "reverse" the inequality in (6.5), and this cannot be done if the a_i's satisfy (6.3).

The resulting is an incompatibility between minimaxity and the conditioning problem. Most minimax estimators will have the constants a_i satisfying

$$a_i d_i \geqslant a_j d_j \quad \text{when } d_i \leqslant d_j, \qquad 1 \leqslant i,j \leqslant p, \tag{6.6}$$

(see, e.g., Strawderman (1978)). Choosing the a_i's to satisfy (6.6), however, is not only intuitively unappealing but, in many cases, will aggravate the conditioning problem. The solution seems to lie in a compromise between the two criteria, possibly resulting in an estimator with bounded risk which will improve the conditioning problem. This idea is developed more fully in Casella (1977).

APPENDIX

Computational lemmas. Let X have a p-variate normal distribution with mean θ and covariance matrix D. Let $\chi_p^2(j)$ denote a chi-square random variable with p degrees of freedom and noncentrality parameter $j/2$.

LEMMA 0. *If $K \sim$ Poisson $(\alpha/2)$ and $Z|K \sim \chi_{p+2K}^2$, then $Z \sim \chi_p^2(\alpha)$. In particular, if $Eh(\chi_p^2(\alpha))$ exists,*

$$E\big[h\big(\chi_p^2(\alpha)\big)\big] = E_K E_{\chi^2}\big[h\big(\chi_{p+2K}^2\big)|K\big].$$

PROOF. This is a relatively well-known result, stated here simply for completeness (see, e.g., James and Stein (1961)). □

The next five lemmas are from Bock (1975), and are stated without proof.

LEMMA 1. *Let $h:[0,\infty)\to(-\infty,\infty)$. Then*

$$E\{h(X'D^{-1}X)X\} = \theta E\{h\big(\chi_{p+2}^2(\theta'D^{-1}\theta)\big)\}.$$

LEMMA 2. *If $D = diagonal(d_1,\cdots,d_p)$, and $h:[0,\infty)\to(-\infty,\infty)$, then*

$$E\{h(X'D^{-1}X)X_i^2\} = d_i E\{h\big(\chi_{p+2}^2(\theta'D^{-1}\theta)\big)\} + \theta_i^2 E\{h\big(\chi_{p+4}^2(\theta'D^{-1}\theta)\big)\}.$$

LEMMA 3. *Let $W_{p\times p}$ be symmetric positive definite, and let $h:[0,\infty)\to(-\infty,\infty)$. Then*

$$E\{h(X'D^{-1}X)X'WX\} = \operatorname{tr} WDE\{h\big(\chi_{p+2}^2(\theta'D^{-1}\theta)\big)\} + \theta'W\theta E\{h\big(\chi_{p+4}^2(\theta'D^{-1}\theta)\big)\}.$$

LEMMA 4. *Let $s:[0,\infty)\to(-\infty,\infty)$. Then, if the expected values on both sides exist,*

$$E\{s\big(\chi_p^2\big)\} = E\left\{\frac{ps\big(\chi_{p+2}^2\big)}{\chi_{p+2}^2}\right\}.$$

LEMMA 5. *Let $s:[0,\infty)\to[0,\infty)$ and $t:[0,\infty)\to[0,\infty)$ be monotone nondecreasing and nonincreasing functions, respectively. Let W be a nonnegative random variable. Assume $E(W)$, $E(s(W))$, $E(Ws(W))$, $E(t(W))$ and $E(Wt(W))$ exist and are finite. Then*

$$E\{s(W)(E(W)-W)\} \leqslant 0 \leqslant E\{t(W)(E(W)-W)\}.$$

LEMMA 6. *Let $h:[0,\infty)\to(-\infty,\infty)$, $D = diagonal(d_1,\cdots,d_p)$. If $E\{h(\chi_p^2(\theta'D^{-1}\theta))\}$ exists then*

$$\frac{\partial}{\partial\theta_i^2}E\{h\big(\chi_p^2(\theta'D^{-1}\theta)\big)\} = \frac{1}{2d_i}\big[E\{h\big(\chi_{p+2}^2(\theta'D^{-1}\theta)\big)\} - E\{h\big(\chi_p^2(\theta'D^{-1}\theta)\big)\}\big],$$

for $1 \leqslant i \leqslant p$.

PROOF. Straightforward calculation. □

LEMMA 7. *Let* $p \geq 3$ *and* $r: \mathbb{R} \to [0, \infty)$ *satisfy*

(i) $r(t)$ *is nondecreasing;*

(ii) $r(t)/t$ *is nonincreasing.*

Let $v = \chi^2_{p+4}(\theta'\theta)/\chi^2_m$, *where* $\chi^2_{p+4}(\theta'\theta)$ *and* χ^2_m *are independent. The function*

$$f(a) = E\left\{\frac{ar(v)(ar(v)m + 4)}{\left(ar(v)\chi^2_m + \chi^2_{p+4}(\theta'\theta)\right)^2}\right\}$$

is strictly increasing in a *if either*

(a) $0 \leq r(t) \leq 2(p-2)/ma$,

or

(b) $p \geq 4$.

PROOF. After differentiating and collecting terms we obtain

$$\frac{\partial}{\partial a} f(a) = E\left\{\frac{2r(v)(ar(v)m + 2)}{\left(ar(v)\chi^2_m + \chi^2_{p+4}(\theta'\theta)\right)^3}\left(\chi^2_{p+4}(\theta'\theta) - \frac{2ar(v)\chi^2_m}{amr(v) + 2}\right)\right\}.$$

Adding $\pm 2amr(v)(amr(v) + 2)^{-1}$ inside the parentheses yields

$$\frac{\partial}{\partial a} f(a) = E\left\{\frac{2r(v)(ar(v)m + 2)}{\left(ar(v)\chi^2_m + \chi^2_{p+4}(\theta'\theta)\right)^3}\left(\chi^2_{p+4}(\theta'\theta) - \frac{2amr(v)}{amr(v) + 2}\right)\right\}$$

$$+ E\left\{\frac{4ar^2(v)}{\left(ar(v)\chi^2_m + \chi^2_{p+4}(\theta'\theta)\right)^3}(m - \chi^2_m)\right\}.$$

From condition (ii), the definition of v, and Lemma 5 it follows that the second expectation above is nonnegative. Now from Lemma 1, the first expectation is equal to

$$E_K E\left\{\frac{2r(w)(ar(w)m + 2)}{\left(ar(w)\chi^2_m + \chi^2_{p+4+2K}\right)^3}\left(\chi^2_{p+r+2K} - \frac{2amr(w)}{amr(w) + 2}\right)|K\right\}, \tag{1}$$

where $K \sim$ Poisson $(\theta'\theta/2)$ and $w = \chi^2_{p+4+2K}/\chi^2_m$. Now applying Lemma 4 three times shows that (1) is equal to

(2)

$$E_K E\left\{\frac{s(K)r(u)(ar(u)m + 2)}{\left(ar(u)\chi^2_m + \chi^2_{p-2+2K}\right)^3}\left(\chi^2_{p-2+2K}\right)^3 \times \left(\chi^2_{p-2+2K} - \frac{2amr(u)}{amr(u) + 2}\right)|K\right\},$$

where $s(K) = 2(p + 2 + 2K)^{-1}(p + 2K)^{-1}(p - 2 + 2K)^{-1} \geq 0$, and $u = \chi^2_{p-2+2K}/\chi^2_m$. Define

$$q\left(\chi^2_{p-2+2K}, \chi^2_m\right) = \frac{s(K)r(u)(ar(u)m + 2)\left(\chi^2_{p-2+2K}\right)^3}{\left(ar(u)\chi^2_m + \chi^2_{p-2+2K}\right)^3},$$

which is nondecreasing in χ^2_{p-2+2K} from the conditions on r. Adding $\pm(p-2+2K)$ inside the parentheses shows that (2) is equal to

$$E_K Eq\left(\chi^2_{p-2+2K}, \chi^2_m\right)\left(\chi^2_{p-2+2K} - (p-2+2K)\right)$$

$$+E_k Eq\left(\chi^2_{p-2+2K}, \chi^2_m\right)\left(p-2+2K - \frac{2amr(u)}{amr(u)+2}\right). \tag{3}$$

The first expectation is nonnegative from Lemma 5, and if $p \geqslant 4$, the second expectation is strictly positive since

$$p-2+2K > p-2 \geqslant 2 > 2amr(u)(amr(u)+2)^{-1}.$$

If $p=3$, since $0 \leqslant r(t) \leqslant 2(p-2)/ma$, the only concern is if $r(t_0) = 2(p-2)/ma = 2/ma$, for some t_0. But then it follows from condition (i) that $r(t) = 2/ma$ $\forall t > t_0$, and a simple argument will show that the first expectation in (3) is positive. Hence the derivative of $f(a)$ is always positive so $f(a)$ is strictly increasing. □

LEMMA 8. *Let $f(a)$ be defined as in Lemma 7, and define*

$$g(a) = E\left\{\frac{ar(v)(ar(v)m+4)}{\left(ar(v)\chi^2_m + \chi^2_{p+6}(\theta'\theta)\right)^2}\right\},$$

i.e., $g(a)$ *is obtained by replacing* $\chi^2_{p+4}(\theta'\theta)$ *in* $f(a)$ *with* $\chi^2_{p+6}(\theta'\theta)$. *Suppose that there exist positive constants* $a_1 < a_2, d_1 > d_2$ *such that* $d_1 f(a_1) = d_2 f(a_2)$. *Then* $d_1 g(a_1) \leqslant d_2 g(a_2)$.

PROOF. After some algebra we obtain

$$\begin{aligned} 0 &= d_1 f(a_1) - d_2 f(a_2) \\ &= E\left[\frac{r(v)}{\left(a_1 r(v)\chi^2_m + \chi^2_{p+4}(\theta'\theta)\right)^2}\right. \\ &\quad \left.\times\left\{\frac{d_1 a_1}{d_2 a_2} - \left[\left(\frac{a_1 r(v)m + \chi^2_{p+4}(\theta'\theta)}{a_2 r(v)m + \chi^2_{p+4}(\theta'\theta)}\right)^2\left(\frac{a_2 r(v)m+4}{a_1 r(v)m+4}\right)\right]\right\}\right]. \end{aligned}$$

From the restrictions on the function r and the constants a_1, a_2, d_1, d_2 it is easily shown that the function in square brackets is the product of two nonnegative, nondecreasing functions of $\chi^2_{p+4}(\theta'\theta)$, and hence is nonnegative and nondecreasing. Therefore the function in braces has only one sign change, from positive to negative values. Using the fact that the noncentral chi-square distribution has monotone likelihood ratio in its degrees of freedom, it follows that if we replace $\chi^2_{p+4}(\theta'\theta)$ with $\chi^2_{p+6}(\theta'\theta)$, the expectation becomes nonpositive. Hence the lemma is proved. □

Acknowledgments. The author wishes to thank his major professor, Leon Jay Gleser, whose advice and encouragement was invaluable in the preparation of this manuscript. Discussions with Jim Berger and William Strawderman were also of great help. The author also wishes to thank the referee for pointing out an error in an earlier version of Theorem 4.1, and the Associate Editor for helpful suggestions that led to a more readable paper.

REFERENCES

[1] BERGER, JAMES O. (1976a). Tail minimaxity in location vector problems and its applications. *Ann. Statist.* **4** 33–50.

[2] BERGER, JAMES O. (1976b). Minimax estimation of a multivariate normal mean with arbitrary quadratic loss. *J. Multivariate Anal.* **6** 1–9.

[3] BHATTACHARYA, P. K. (1966). Estimating the mean of a multivariate normal population with general quadratic loss function. *Ann. Math. Statist.* **37** 1819–1824.

[4] BOCK, M. E. (1975). Minimax estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **3** 209–218.

[5] BROWN, L. D. (1966). On the admissibility of invariant estimators of one or more location parameters. *Ann. Math. Statist.* **37** 1087–1136.

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

[7] CASELLA, G. (1977). Minimax ridge regression estimation. Ph.D. Thesis, Purdue Univ.

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

[9] HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12** 55–68.

[10] JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Probability* **1** 361–379.

[11] STEIN, C. (1955). Inadmissibility of the usual estimator of the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Probability* **1** 197–206.

[12] STRAWDERMAN, W. E. (1978). Minimax adaptive generalized ridge regression estimators. *J. Amer. Statist. Assoc.* **73** 623–627.

[13] THISTED, RONALD A. (1976). Ridge regression, minimax estimation, and empirical Bayes methods. Ph.D. Thesis, Stanford Univ.

RUTGERS UNIVERSITY
STATISTICS AND COMPUTER SCIENCE
P.O. BOX 231
NEW BRUNSWICK, N.J. 08903