Minimax Ridge Regression Estimation*

by

George Casella
Purdue University

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #497

May 1977

*This research was supported by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant AFOSR-72-2350C. Reproduction in whole or part is permitted for any purpose of the United States Government.

Abstract

The technique of ridge regression, first proposed by Hoerl and Kennard (1970), has become a popular tool for data analysts faced with a high degree of multicollinearity in their data. By using a ridge estimator, it was hoped that one could both stabilize his 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 (1955) and others, Strawderman (1976) and Thisted (1976) 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 lead 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), which 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 the original purpose of the ridge regression estimator was not 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 nearly impossible for them to provide the numerical stability for which they were originally intended.

We start with the familiar linear model

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

where Y is an nx1 vector of observations, Z is the known nxp design matrix of rank p, β is the px1 vector of unknown regression coefficients, and ε is nx1 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 nxn 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 thus is 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

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

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 function dominates that of $\hat{\beta}$. Since $\hat{\beta}$ is a minimax estimator of β with constant risk

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

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. If this is the case, Hoerl and Kennard point out that 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 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ be its eigenvalues. It follows that

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

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

$$\hat{\beta}_k = (P'(D\lambda + kI_p)P)^{-1}Z'Y. \tag{1.9}$$

To see how 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, 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 $\kappa(A)$ mean 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

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

where $K = \text{diag}(k_1,\ldots,k_p)$. Here, we allow each eigenvalue of $Z'Z$ to be increased by a different amount.

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, $i=1,2,\ldots,p$) 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 about 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 (1976)). 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 experimentor has some knowledge of the losses he is willing to incur in the individual components, possibly from cost considerations or prior knowledge. For this purpose, 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 extention of (1.12) we define

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

where M is a non-singular 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

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

and showing that $\hat{\beta}_Q(K)$ 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. $\hat{\beta}_Q(K)$ can be written

$$\hat{\beta}_Q(K) = (M'(D^{-1}+K)M)^{-1}M'D^{-1}M\hat{\beta}$$
$$= M^{-1}(D^{-1}+K)^{-1}D^{-1}M\hat{\beta}. \tag{2.4}$$

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),

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

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

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

where the ith component can be written

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

and the loss of (2.3) becomes

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

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

In the following we will surpress 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

$$R(X,\theta,\sigma^2) = E\;L(X,\theta,\sigma^2) = \mathrm{tr}D,$$

where "tr" denotes the trace operator, an estimator $\delta(X,s)$ is minimax if and only if

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

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 lead to either bounds that are not sharp (Thisted (1976)) or additional conditions on the estimator (Strawderman (1976)). 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 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 $\theta'\theta > M$, $\Delta(\delta(X,s),\theta,\sigma^2) \leq 0$.

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 follow a non-central chi-square distribution.

For k_i as in (3.3), the ridge estimator of (2.9) can be written componentwise as

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

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 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) $E(f(X) - f(\theta))^2 \leq K|\theta|^q$ for some constants q and K

iii) $\sup_{y>|\theta|/2} |f^{ij}(y) - f^{ij}(\theta)| = o(|\theta|^{-2}) \quad 1 \leq i, j \leq p$

where $f^{ij}(X) = \frac{\partial^2}{\partial X_i \partial X_j} f(X)$

Then

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

Proof: Define the regions W and W^c by

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

The Taylor expansion of f about θ (up to second order terms) is

$$f(X) = f(\theta) + \sum_{i=1}^{p} (X_i - \theta_i) f^i(\theta) + \rho(X,\theta) \tag{3.5}$$

where

$$f^i(\theta) = \frac{\partial}{\partial X_i} f(X)\Big|_{X=\theta}, \tag{3.6}$$

$$\rho(X,\theta) = \frac{(1-t)^2}{2!} \sum_{i,j} (X_i - \theta_i)(f^{ij}(\theta + t(X-\theta)) - f^{ij}(\theta))$$

for some t, $0 < t < 1$. Letting $\Phi(\cdot)$ denote the cumulative normal distribution with mean 0 and covariance matrix I we have

$$Ef(X) = \int_W \{f(\theta) + \sum_{i=1}^{p} (X_i-\theta_i)f^i(\theta) + \rho(X,\theta)\} \, d\Phi(X-\theta)$$

$$+ \int_{W^c} f(X)d\Phi(X-\theta)$$

From the definition of W, a simple sign invariance argument will show

$$\int_W (X_i-\theta_i)\Phi(X-\theta) = 0, \; i = 1,\ldots,p,$$

therefore,

$$\begin{aligned} Ef(X) = f(\theta) + &\int_W \rho(X,\theta)d\Phi(X-\theta) \\ &+ \int_{W^c} (f(X) - f(\theta))d\Phi(X-\theta), \end{aligned} \tag{3.7}$$

and hence,

$$\begin{aligned} |Ef(X) - f(\theta)| \le &\int_W |\rho(X,\theta)|d\Phi(X-\theta) \\ &+ \int_{W^c} |f(X) - f(\theta)|d\Phi(X-\theta) \end{aligned} \tag{3.8}$$

Noting that $X \in W \Rightarrow |\theta + t(X-\theta)| \ge |\theta|/2$ for $0 \le t \le 1$, we have

$$\sup_{X\in W} |f^{ij}(\theta+t(X-\theta)) - f^{ij}(\theta)| \le \sup_{\neg|y|>|\theta|/2} |f^{ij}(y) - f^{ij}(\theta)|.$$

It then follows from (3.6) and condition (iii) that

$$\int_W |\rho(X,\theta)|d\Phi(X-\theta) \le \int_W \frac{(1-t)^2}{2} \{\sum_{i,j} |X_i-\theta_i||X_j-\theta_j|$$

$$\times \sup_{|y|>|\theta|/2} |f^{ij}(y)-f^{ij}(\theta)|\} \, d\Phi(X-\theta) \tag{3.9}$$

$$\le \frac{1}{6} \max_{i,j} \sup_{|y|>|\theta|/2} |f^{ij}(y)-f^{ij}(\theta)| \sum_{i,j} E|X_i-\theta_i||X_j-\theta_j|$$

$$= N \max_{i,j} \sup_{|y|>|\theta|/2} |f^{ij}(y)-f^{ij}(\theta)|$$

$$= o(|\theta|^{-2}),$$

where $N = (1/6) \sum_{i,j} E|X_i-\theta_i||X_j-\theta_j| < \infty$. Also, from the definition of W^c,

$$\int_{W^c} |f(X) - f(\theta)|d\Phi(X-\theta) = E|f(X) - f(\theta)||I_{(|\theta|/2,\infty)}(|X-\theta|)$$

$$\le \{E(f(X) - f(\theta))^2 EI_{(|\theta|/2,\infty)}(|X-\theta|)\}^{1/2}$$

by Holder's inequality. Using the well known fact (see, e.g., Chung (1968)) that if $a > 0$ then

$$\int_a^\infty \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}} \, dz \le \frac{e^{-\frac{1}{2}a^2}}{a},$$

we have

$$EI_{(|\theta|/2,\infty)}(|X-\theta|) = P(|X-\theta| > |\theta|/2)$$

$$\le P(\bigcup_{i=1}^{p} \{|X_i-\theta_i| \ge |\theta|/2p^{\frac{1}{2}}\})$$

$$\leq 2p\, P(n(0,1) > |\theta|/2p^{\frac{1}{2}})$$

$$\leq \frac{4p^{3/2}}{|\theta|} \exp\{-|\theta|^2/8p\},$$

and combining this with condition (ii) and the fact that $\lim_{y\to\infty} y^n e^{-y^2} = 0\ \forall n$ we have

$$\int_{W^c} |f(X) - f(\theta)|\, d\Phi(X-\theta) = o(|\theta|^{-n})$$

$\forall n$, and hence the result follows.

The extension of Lemma 3.1 to the case $X \sim N(\theta,\Sigma)$, Σ a known positive definite matrix, proceeds in the usual manner (i.e., diagonalizing Σ), and is stated without proof.

Lemma 3.2: Let $X \sim N(\theta,\Sigma)$, and let $f\colon\ \mathbb{R}^p \to \mathbb{R}$ satisfy conditions i) - iii) of Lemma 3.1. Then

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

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 $X \sim N(\theta,\sigma^2 D)$, $D = \text{diag}(d_1,\ldots,d_p)$, and let $s \sim \sigma^2\chi^2_m$ be independent of X. Let the loss of an estimator $\delta(X,s)$ of θ be given by (2.5), and let $\delta^R(X,s)$ be the ridge estimator given by (3.4) where $r(t)\colon\ \mathbb{R} \to [0,\infty)$ satisfies

i) $t^{\frac{1}{2}}r'(t) = o(1)$

ii) $t^{3/2}r''(t) = o(1)$

iii) $r(t)$ is bounded and non-decreasing

iv) $r(t)/t$ is non-increasing.

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

$$\varepsilon_1 < r(t) < [2(m+2)^{-1}(\mathrm{tr}AD^2 - 2\lambda\max AD^2)/\lambda\max A^2D^3] - \varepsilon_2, \qquad (3.10)$$

where $A = \mathrm{diag}(a_1,\ldots,a_p)$, $a_i \geq 0$, $1 \leq i \leq p$, then $\exists\ K > 0$ such that $\forall\ \theta'\theta > K$,

$$R(\delta^R(X,s),\theta,\sigma^2) \leq R(X,\theta,\sigma^2).$$

Proof: Define

$$\Delta(\delta^R,\theta,\sigma^2) = R(\delta^R(X,s),\theta,\sigma^2) - R(X,\theta,\sigma^2).$$

From (2.5) and (3.4) straightforward calculation yields

$$\Delta(\delta^R,\theta,\sigma^2) = (1/\sigma^2) \sum_{i=1}^{p} E\{ \frac{(a_i d_i r(t) X_i)^2}{(a_i d_i r(t)+t)^2} - \frac{2X_i(X_i-\theta_i)a_i d_i r(t)}{a_i d_i r(t) + t}\}, \qquad (3.11)$$

Where $t = X'D^{-1}X/s$. Integrating the last term in (3.11) by parts and defining

$$w_m = s/\sigma^2, \quad Z_i = X_i/\sigma, \quad v = Z'D^{-1}Z,$$

yields

$$\Delta(\delta^R,\theta,\sigma^2) = \sum_{i=1}^{p} E\{ \frac{(a_i d_i r(v/w_m))^2 w_m^2 Z_i^2}{(a_i d_i r(v/w_m)w_m+v)^2} - \frac{2a_i d_i^2 r(v/w_m)w_m}{a_i d_i r(v/w_m)w_m+v}$$
$$+ \frac{4a_i d_i r(v/w_m)w_m Z_i^2}{(a_i d_i r(v/w_m)w_m+v)^2} - \frac{4a_i d_i Z_i^2(v/w_m)r'(v/w_m)}{\sigma^2 w_m(a_i d_i r(v/w_m)w_m+v)^2} \} .$$

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

$$E\{\chi_m^2 h(Z'D^{-1}Z,\chi_m^2)\} = mE\{h(Z'D^{-1}Z,\chi_{m+2}^2\} \tag{3.13}$$

Using (3.13)on each of the first three terms of (3.12), bounding the last by zero, and rearranging terms gives

$$\Delta(\delta^R(X,s),\theta,\sigma^2) \le m \sum_{i=1}^{p} E\{\frac{a_i d_i r(v/w_{m+2})(a_i d_i r(v/w_{m+2})w_{m+2}+4)Z_i^2}{(a_i d_i r(v/w_{m+2})w_{m+2}+v)^2} - \frac{2a_i d_i^2 r(v/w_{m+2})}{(a_i d_i r(v/w_{m+2})w_{m+2}+v)}\} \tag{3.14}$$

It follows from conditions (i) and (ii) that r(v/w) is non-increasing in w, and wr(v/w) is non-decreasing 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 non-increasing in w. Applying Lemma 5, Appendix shows

$$E\{q_i(\chi^2_{m+2})(\chi^2_{m+2} - m+2)\} \leq 0,$$

so that (3.14) is bounded above by

$$\Delta(\delta^R(X,s),\theta,\sigma^2) \leq m \sum_{i=1}^{p} E\{ \frac{a_i d_i r(v/w)(a_i d_i r(v/w)(m+2)+4) Z_i^2}{(a_i d_i r(v/w) w+v)^2} - \frac{2 a_i d_i^2 r(v/w)}{(a_i d_i r(v/w) w+v)} \}, \tag{3.15}$$

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

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

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 (3.15) and let $F(\cdot)$ denote the cumulative χ^2 distribution with m+2 degrees of freedom. Then

$$\Delta(\delta^R(X,s),\theta,\sigma^2) \leq m \int_{W_0} \sum_{i=1}^{p} E_Z(g_i(w,Z))dF(w)$$

$$+ m \int_{W_1} \sum_{i=1}^{p} E_Z(g_i(w,Z))dF(w)$$

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

$$\int_{W_1} \sum_{i=1}^{p} E_Z(g_i(w,Z))dF(w) \leq \int_{W_1} \sum_{i=1}^{p} E_Z\{(\frac{a_i d_i r(v/w)}{v})$$

$$\times(a_i d_i^2 r(v/w)(m+2)+2d_i)\}dF(w)$$

$$\leq \int_{W_1} \sum_{i=1}^{p} E_Z\{(\frac{a_i d_i r^*}{v})(a_i d_i^2 r^*(m+2)+2d_i)\}dF(w) \tag{3.16}$$

$$= [E_Z\{v^{-1}tr(m+2)r^{*2}A^2D^3+2r^*AD^2)\}]\ P(w > M)$$

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 (3.16) is equal to

$$(\sigma^2/\theta'D^{-1}\theta)(tr[(m+2)r^{*2}A^2D^3 + 2r^*AD^2])P(w > M) + o(\sigma^2/\theta'\theta). \tag{3.17}$$

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

$$\int_{W_0} \sum_{i=1}^{p} E_Z g_i(w,Z)dF(w) = \int_{W_0} \sum_{i=1}^{p} g_i(w,\theta/\sigma)dF(w)$$

$$+ \int_{W_0} o(\sigma^2/\theta'\theta)dF(w)$$

$$= \int_{W_0} \sum_{i=1}^{p} \{ \frac{a_i d_i r(\nu/w)(a_i d_i r(\nu/w)(m+2)+4)\theta_i^2}{\sigma^2 (a_i d_i r(\nu/w)w + \nu)^2} - \frac{2a_i d_i^2 r(\nu/w)}{(a_i d_i r(\nu/w)w+\nu)} \} dF(w) + \int_{W_0} o(\sigma^2/\theta'\theta) dF(w), \tag{3.18}$$

where $\nu = \theta' D^{-1}\theta/\sigma^2$. Straightforward calculation will show that the individual terms comprising the $o(\sigma^2/\theta'\theta)$ term in (3.18), 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 writing

$$\begin{aligned}(a_i d_i r(\nu/w)w+\nu)^{-1} &= \nu^{-1}(1-(a_i d_i r(\nu/w)w/(a_i d_i r(\nu/w)w+\nu)) \\ &= \nu^{-1}(1-\gamma_i(\nu,w)),\end{aligned}$$

and

$$s_i(\nu,w) = a_i d_i (a_i d_i r(\nu/w)(m+2)+4)\theta_i^2/\sigma^2,$$

we can write (3.18) as

$$\begin{aligned}\int_{W_0} \sum_{i=1}^{p} E_Z g_i(w,Z) dF(w) &= \int_{W_0} \{ \frac{r(\nu/w)}{\nu} \sum_{i=1}^{p} (1-\gamma_i(\nu,w))[(1-\gamma_i(\nu,w))s_i(\nu,w)/\nu \\ &\qquad - 2a_i d_i^2]\} dF(w) \\ &\qquad + o(\sigma^2/\theta'\theta) \\ &= \int_{W_0} \{ \frac{r(\nu/w)}{\nu} \sum_{i=1}^{p} (s_i(\nu,w)/\nu - 2a_i d_i^2)\} dF(w)\end{aligned}$$

$$- \int_{W_0} \{\frac{r(\nu/w)}{\nu} \sum_{i=1}^{p} [\gamma_i(\nu,w)(s_i(\nu,w)/\nu - 2a_i d_i^2) - \gamma_i^2(\nu,w)s_i(\nu,w)/\nu]dF(w) \tag{3.19}$$

$$+ o(\sigma^2/\theta'\theta)$$

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

$$\gamma_i(\nu,w) \leq a_i d_i r^* M(a_i d_i r^* M+\nu)^{-1}, \ 1 \leq i \leq p$$

$$s_i(\nu,w)/\nu \leq a_i d_i^2 \sigma^2 (a_i d_i r^*(m+2)+4), \ 1 \leq i \leq p$$

and thus it is clear that the second integral in (3.19) is $o(\nu^{-1}) = o(\sigma^2/\theta'\theta)$. Hence, summing the first term in (3.19) yields

$$\int_{W_0} \sum_{i=1}^{p} E_Z g_i(w,Z)dF(w) \leq \int_{W_0} \frac{r(\nu/w)}{\nu} [\frac{r(\nu/w)(m+2)\theta'A^2D^2\theta + r\theta'AD}{\theta'D^{-1}\theta} - 2trAD^2]dF(w)$$

$$+ o(\sigma^2/\theta'\theta)$$

$$\leq \int_{W_0} \{ \frac{r(\nu/w)}{\nu} (\lambda maxA^2D^3)(m+2) \times [r^* - \frac{2(m+2)^{-1}(trAD^2 - 2\lambda maxAD^2)}{\lambda maxA^2D^3}] dF(w) \tag{3.20}$$

$$+ o(\sigma^2/\theta'\theta),$$

since

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

By assumption, the quantity in square brackets in (3.20) is bounded above by $-\varepsilon_2$, $r(\nu/w) \ge \varepsilon_1$, and since $\lambda_{max} A^2 D^3 > 0$,

$$\int_{W_0} \sum_{i=1}^{p} E_Z g_i(w,Z)\, dF(w) \le \frac{\sigma^2}{\theta' D^{-1} \theta} [-\varepsilon_1 \varepsilon_2 \lambda_{max} A^2 D^3 (m+2) P(w \le M)] + o(\sigma^2/\theta'\theta) \quad (3.21)$$

Combining (3.17) and (3.21) yields

$$\Delta(\delta^R, \theta, \sigma^2) \le \frac{\sigma^2 m}{\theta' D^{-1} \theta} \{- \varepsilon_1 \varepsilon_2 \lambda_{max} A^2 D^3 (m+2) P(w \le M) + (tr[(m+2) r^{*2} A^2 D^3 + 2r^* A D^2]) P(w > M)\} + o(\sigma^2/\theta'\theta). \quad (3.22)$$

Now M is chosen large enough so that

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

for some $\varepsilon_3 > 0$, and thus from (3.22),

$$\Delta(\delta^R, \theta, \sigma^2) \le - \varepsilon_3 m \sigma^2 / \theta' D^{-1} \theta + o(\sigma^2/\theta'\theta)$$

$$\le - \varepsilon_3 m \lambda_{min} D \sigma^2 / \theta'\theta + o(\sigma^2/\theta'\theta),$$

and for sufficiently large $\theta'\theta$, $\Delta(\delta^R, \theta, \sigma^2) \le 0$ 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 $r(t)$: $\mathbf{R} \to [0,\infty)$ satisfies conditions i) - iv) of Theorem 3.1. If

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

$\forall\, t \ge 0$, then $\delta^R(X,s)$ is minimax against the loss (2.5).

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

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

Then from Theorem 3.1 $\exists\ M > 0$ such that $\forall\ \theta'\theta \ge M$ $\Delta(\delta^R,\theta,\sigma) \le 0$. Consider the set $\mathcal{S} = \{\theta:\ \theta'\theta \le M\}$, a compact sphere in $\mathbf{R}^p$. 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 $\mathcal{S}$. We will then show that, with the exception of the point $\theta = 0$, $\gamma(\delta^R(X),\theta)$ does not have an extreme point in the interior of $\mathcal{S}$ and thus achieves its maximum either at $\theta = 0$ or $\theta'\theta = M$.

If $\theta'\theta = M$, it will follow from Theorem 3.1 that $\gamma(\delta^R,\theta,\sigma^2) \leq 0$. We then show that $\gamma(\delta^R,0,\sigma^2) \leq 0$. A simple argument, using Fatou's Lemma, then allows the result to be extended to the case when the inequality in condition (4.1) is not strict.

Using the notation of Theorem 3.1, from (3.15) we have

$$\Delta(\delta^R,\theta,\sigma^2) \leq \gamma(\delta^R,\theta,\sigma^2)$$

$$= m \sum_{i=1}^{p} E\{\frac{a_i d_i r(v/w)(a_i d_i r(v/w)(m+2)+4)Z_i^2}{(a_i d_i r(v/w)w+v)^2} \tag{4.3}$$

$$- \frac{2a_i d_i^2 r(v/w)}{(a_i d_i r(v/w)w+v} \},$$

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

$$\eta = \theta/\sigma,$$

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

$$h_i(w,Z) = (a_i d_i r(v/w)w+v)^{-1},$$

then

$$\gamma(\delta^R,\theta,\sigma^2) = m \sum_{i=1}^{p} E\{g_i(w,Z)h_i^2(w,Z)Z_i^2$$

$$- 2a_i d_i^2 r(v/w)h_i(w,Z)\} \tag{4.5}$$

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

$$\gamma(\delta^R,\eta,\sigma^2) = m \sum_{i=1}^{p} E\{g_i(w\,,\,\chi^2_{p+2}(\nu))h_i^2(w,\chi^2_{p+2}(\nu))$$

$$+ \eta_i^2 g_i(w,\chi^2_{p+4}(\nu))h_i^2(w,\chi^2_{p+4}(\nu)) \tag{4.6}$$

$$- 2a_i d_i^2 r(\chi_p^2(\nu)/w)h_i(w,\chi_p^2(\nu))\},$$

where $\nu = \eta' D^{-1}\eta$. We note the following: if $f(x)$ is a function of x only through x^2, then with the possible exception of $x = 0$,

$$\frac{d}{dx}\; f(x)\Big|_{x=x_0} = 0 \Leftrightarrow \frac{d}{dx^2}\; f(x)\Big|_{x=x_0} = 0.$$

From (4.6) 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 $\gamma(\delta^R,\gamma,\sigma^2)$ only if

$$\frac{\partial}{\partial\eta_i^2}\; \gamma(\delta^R,\gamma,\sigma^2)\Big|_{\eta=\eta_0} = 0,\ 1 \le i \le p \tag{4.7}$$

We now show that such a point does not exist. From Lemma 6, Appendix,

$$\frac{\partial}{\partial\eta_k^2}\; \gamma(\delta^R,\gamma,\sigma^2) = \sum_{i=1}^{p} E\{\tfrac{1}{2}[g_i(w,\chi^2_{p+4}(\nu))h_i^2(w,\chi^2_{p+4}(\nu))$$

$$- g_i(w,\chi^2_{p+2}(\nu))h_i^2(w,\chi^2_{p+2}(\nu))]$$

$$+ (\eta_i^2/2d_i)[g_i(w,\chi^2_{p+6}(\nu))h_i^2(w,\chi^2_{p+6}(\nu))$$

$$- g_i(w,\chi^2_{p+4}(\nu))h_i^2(w,\chi^2_{p+4}(\nu))] \tag{4.8}$$

$$+ (a_i d_i^2)[r(\chi^2_{p+2}(\nu)/w)h_i(w,\chi^2_{p+2}(\nu))$$

$$- r(\chi^2_p(\nu)/w)h_i(w,\chi^2_p(\nu))]\}$$

$$+ E\{g_k(w,\chi^2_{p+4}(\nu))h_k^2(w,\chi^2_{p+4}(\nu))\}$$

Notice that the sum in (4.8) does not depend on k, the index of differentiation. Therefore, denoting the sum by $\mathcal{S}(w,\eta)$

$$\frac{\partial}{\partial \eta_k^2} \gamma(\delta^R,\eta,\sigma^2) = E\mathcal{S}(w,\eta) + Eg_k(w,\chi^2_{p+4}(\nu))h_k^2(w,\chi^2_{p+4}(\nu)) \tag{4.9}$$

for all k, $1 \leq k \leq p$. Thus, in order for (4.7) to be satisfied at some point $\eta_0 \neq 0$, it must be the case that

$$Eg_i(w,\chi^2_{p+4}(\eta))h_i^2(w,\chi^2_{p+4}(\eta)) = Eg_j(w,\chi^2_{p+4}(\eta))h_j(w,\chi^2_{\mu+4}(\nu))$$

for all i, j, $1 \leq i, j \leq p$. From (4.4),

$$Eg_i(w,\chi^2_{p+4}(\nu))h_i^2(w,\chi^2_{p+4}(\nu))$$

$$= E\{\frac{a_i d_i r(\chi^2_{p+4}(\nu)/w)(a_i d_i r(\chi^2_{p+4}(\nu)/w)(m+2)+4)}{(a_i d_i r(\chi^2_{p+4}(\nu)/w)w + \chi^2_{p+4}(\nu))^2},$$

and from Lemma 9, Appendix, this is a strictly increasing function of $a_i d_i$. Therefore, (4.7) can be satisfied only if $a_i d_i = c \ \forall_i$, but if this is the case, from (4.3),

$$\Delta(\delta^R,\theta,\sigma^2) \leq m \sum_{i=1}^{p} E\{\frac{cr(v/w)(cr(v/w)(m+2)+4)Z_i^2}{(cr(v/w)w+v)^2} - \frac{2cd_i^2 r(v/w)}{cr(v/w)w+v}\}$$

$$\leq m\, E(\frac{cr(v/w)}{cr(v/w)w+v})(\frac{(cr(v/w)(m+2)+4)Z'Z}{Z'D^{-1}Z} - 2trD) \qquad (4.10)$$

since $Z'D^{-1}Z = v \leq cr(v/w)w+v$. Since $Z'Z/Z'D^{-1}Z \leq \lambda_{max}D$, rearranging terms in (4.10) shows

$$\Delta(\delta^R,\theta,\sigma^2) \leq cm(m+2)E(\frac{cr(v/w)}{cr(v/w)w+v})(r(v/w) - \frac{2(trD-\lambda_{max}D)}{c(m+2)\lambda_{max}D}). \qquad (4.11)$$

Under the restrictions $a_i d_i = c$, (4.2) can be written

$$\varepsilon_1 \leq r(t) \leq 2(trD-2\lambda_{max}D)/(c(m+2)\lambda_{max}D) - \varepsilon_2,$$

and hence the right hand side of (4.11) is negative. If $\eta = 0$, or equivalently $\theta = 0$, it is obvious that

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

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

$$r_\varepsilon(t) = (1-\varepsilon)r(t) + c\varepsilon, \qquad (4.12)$$

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

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

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

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

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

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

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

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

is minimax provided

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

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

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

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

<u>Corollary 4.1</u>: Let $\delta^R(X,s)$ be given componentwise by

$$\delta_i^R(X,s) = \left(1 - \frac{a_i d_i r(X'D^{-1}X,s)}{a_i d_i r(X'D^{-1}X,s) + X'D^{-1}X/s}\right) X_i, \qquad (4.13)$$

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

i) $\frac{\partial}{\partial t_1} r(t_1,t_2) = o(t_1^{-\frac{1}{2}})$

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

iii) $r(t_1,t_2)$ is non-decreasing in t_1 and non-increasing in t_2

iv) $r(t_1,t_2)/t_1$ is non-increasing in t_1

V) $r(t_1,t_2)t_2$ is non-decreasing in t_2.

If

$$0 \le r(t_1,t_2) \le 2(m+2)^{-1}(trAD^2 - 2\lambda_{max}AD^2)/\lambda_{max}A^2D^3, \qquad (4.14)$$

for all t_1, $t_2 \ge 0$, then $\delta^R(X,s)$ is minimax against the loss (2.5).

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

$$\delta_i^S(X,s) = (1 - \frac{ad_i^{-1}}{ad_i^{-1}+X'D^{-1}X/s+g+h/s})X_i \qquad (4.15)$$

where a, g and h are positive constants. Strawderman (1976) showed $\delta^S(X,s)$ is minimax if

i) $h \geq 0$

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

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

If we define

$$r(X'D^{-1}X,s) = \frac{X'D^{-1}X/s}{X'D^{-1}X/s+g+h/s} \qquad (4.16)$$

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

we can write (4.15) in the form given by (4.13). It is easy to check that the function r in (4.16) satisfies the conditions of Corollary 4.1 , and that the minimax bound (4.14) can be written

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

and that the restriction $g \geq 2(m+2)^{-1}(p-2)$ 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 (4.1) is, in fact, necessary and sufficient for minimaxity of the ridge estimator. The main theorem of this section is the following.

<u>Theorem 5.1</u>: Let $X \sim N(\theta, D)$, $D = \text{diag}(d_1,\ldots,d_p)$, and let the ridge estimator $\delta^R(X)$ be given componentwise by

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

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

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

ii) $t^{3/2}r''(t) = o(1)$,

iii) $r(t)$ is bounded and non-decreasing,

iv) $r(t)/t$ is non-increasing.

$\delta^R(X)$ 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

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

for all $t \ge 0$, where $A = \text{diag}(a_1,\ldots,a_p)$.

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

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

$$\delta_i^R(X,s) = \left(1 - \frac{a_i d_i (m+2)^{-1} r(X'D^{-1}X/s)}{a_i d_i (m+2)^{-1} r(X'D^{-1}X/s) + X'D^{-1}X/s}\right) X_i, \quad 1 \le i \le p$$

where 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

$$0 \le r(t) \le 2(\mathrm{tr}AD^2 - 2\lambda_{max}AD^2)/\lambda_{max}A^2D^3, \quad \forall t \ge 0.$$

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 check 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 first define $\Delta(\delta^R,\theta) = R(X,\theta) - R(\delta^R(X),\theta)$, and from (5.1) and (5.2) we have

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

where $t = X'D^{-1}X$. As in Theorem 3.1, we integrate the last term by parts and rearrange terms to get

$$\Delta(\delta^R,\theta) = \sum_{i=1}^{p} E\left\{\frac{a_i d_i r(t)(a_i d_i r(t)+4)X_i^2}{(a_i d_i r(t)+t)^2} - \frac{2a_i d_i^2 r(t)}{a_i d_i r(t)+t} - \frac{4a_i d_i X_i^2 t r'(t)}{(a_i d_i r(t)+t)^2}\right\}.$$

Now we apply Lemma 3.1, and noting that condition (i) insures that the term involving $r'(t)$ is $o(|\theta|^{-2})$, we have for sufficiently large θ,

$$\Delta(\delta^R,\theta) = \sum_{i=1}^{p} \left\{\frac{a_i d_i r(\tau)(a_i d_i r(\tau)+4)\theta_i^2}{(a_i d_i r(\tau)+\tau)^2} - \frac{2a_i d_i^2 r(\tau)}{a_i d_i r(\tau)+\tau}\right\} + o(|\theta|^{-2}),$$

where $\tau = \theta' D^{-1}\theta$. Now applying an argument similar to that used in Theorem 3.1 in going from (3.18) to (3.20), we have for sufficiently large θ,

$$\Delta(\delta^R,\theta) = \frac{r(\tau)}{\tau}\left\{\frac{r(\tau)\theta' A^2 D^3\theta + 4\theta' AD\theta}{\theta' D^{-1}\theta} - 2\mathrm{tr}AD^2\right\} + o(|\theta|^{-2}). \tag{5.4}$$

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

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

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

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

$$= \alpha^{*\prime} A^2 D^3 \alpha^*$$

$$= \lambda_{max} A^2 D^3.$$

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

$$\Delta(\delta^R(X), \theta_n^*) = \frac{r(n)}{n} \{r(n)\lambda_{max} A^2 D^3 + 4\lambda_{max} AD^2 - 2trAD^2\}$$

$$+ o(|\theta|^{-2})$$

$$= \frac{r(n)}{n} \lambda_{max} A^2 D^3 \{r(n) - 2(trAD^2 - \lambda_{max} AD^2)/\lambda_{max} A^2 D^3\}$$

$$+ o(n^{-1}).$$

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

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

It then follows that for sufficiently large n

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

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 conditions (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 $r\colon\ \mathbb{R} \to [0,\infty)$ is bounded and satisfies

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

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

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

$$\liminf_{t\to\infty} r(t) \leq 2(trAD^2-\lambda_{max}AD^2)/\lambda_{max}A^2D^3.$$

Thisted (1976) derived necessary conditions similar to the above for the case $r(t)$ = constant.

6. Minimaxity and Conditioning

The crucial condition for the minimaxity of $\delta^R(X)$ is that

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

and hence, it must necessarily be the case that

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

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 that was mentioned earlier, which means, in particular, that the a_i's should be chosen so that

$$a_i \geq a_j \text{ when } d_i \geq d_j, \quad 1 \leq i, j \leq p \tag{6.3}$$

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 for two reasons. One, it is Bayesian in nature, and two, it is sensible to add only small amounts of bias to directions with good information (small d_i's). An inconsistency arises, however, when the condition of minimaxity is forced into the estimator. If the d_i's are very spread out (as will occur in an ill-conditioned problem), the matrix D is likely to satisfy

$$\text{tr}D^2 \leq 2\lambda_{max}D^2. \tag{6.4}$$

As the number of dimensions, p, increases, it is more likely that the inequality in (6.4) will reverse, but in general one would expect (6.4) 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 (6.4), and this cannot be done if the a_i's satisfy (6.3).

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

$$a_i \geq a_j \text{ when } d_i \leq d_j, \quad 1 \leq i, j \leq p, \tag{6.5}$$

(see, e.g., Strawderman (1976)). Choosing the a_i's to satisfy (6.5), 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 non-centrality parameter j/2.

Lemma 0. If $K \sim \text{Poisson}(\alpha/2)$ and $Z|K \sim \chi^2_{p+2K}$, then $Z \sim \chi_p^2(\alpha)$. In particular, if $E\, h(\chi_p^2(\alpha))$ exists,

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

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

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

Lemma 2: If D = diagonal $(d_1,\ldots,d_p)$, and $h: [0,\infty) \to (-\infty,\infty)$, then

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

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

$$E\{h(X'D^{-1}X)X'WX\} = \text{tr } WDE\{h(\chi^2_{p+2}(\theta'D^{-1}\theta))\}$$

$$+ \theta'W\theta E\{h(\chi^2_{p+4}(\theta'D^{-1}\theta))\}.$$

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

$$E\{h(\chi^2_p)\} = E\{\frac{p\ h(\chi^2_{p+2})}{\chi^2_{p+2}}\}.$$

Lemma 5: Let $S: [0,\infty) \to [0,\infty)$ and $t: [0,\infty) \to [0,\infty)$ be monotone non-decreasing and non-increasing functions, respectively. Let W be a non-negative random variable. Assume E(W), E(S(W)), E(WS(W)), E(t(W)) and E(Wt(W)) exist and are finite. Then

$$E\{S(W)(E(W)-W)\} \leq 0 \leq E\{t(W)(E(W)-W)\}.$$

Lemma 6: Let $h: [0,\infty) \to (-\infty,\infty)$. If $E\{h(\chi^2_p(\theta'\theta))\}$ exists, then

$$\frac{\partial}{\partial\theta_i^2} E\{h(\chi^2_p(\theta'\theta))\} = \frac{1}{2}\ [E\{h(\chi^2_{p+2}(\theta'\theta))\} - E\{h(\chi^2_p(\theta'\theta))\}]$$

for $1 \leq i \leq p$.

Proof:

$$E\{h(\chi^2_p(\theta'\theta))\} = \int_0^\infty \sum_{k=0}^\infty h(y)(\frac{\theta'\theta}{2})^k \frac{e^{-\frac{1}{2}\theta'\theta}}{k!} C_{p+2k}^{\frac{1}{2}} y^{\frac{p+2k}{2}-1} e^{-\frac{1}{2}y}\ dy,$$

where $C_{p+2k} = (\Gamma(\frac{p+2k}{2})2^{\frac{p+2k}{2}})^{-1}$. Interchanging the order of summation and integration yields

$$E\{h(\chi_p^2(\theta'\theta))\} = \sum_{k=0}^{\infty} (\frac{\theta'\theta}{2})^k \frac{-e^{\frac{1}{2}\theta'\theta}}{k!} Eh(\chi^2_{p+2k})$$

$$= \sum_{k=0}^{\infty} e^{-\frac{1}{2}\theta'\theta} e^{k(\log \theta'\theta)} \frac{e^{-k \log 2}}{k!} Eh(\chi^2_{p+2k}).$$

From Lehmann (1959), Theorem 9, page 52, we can differentiate the above expression, with respect to log $\theta'\theta$, inside the summation. Thus,

$$\frac{\partial}{\partial \log \theta'\theta} E\{h(\chi_p^2(\theta'\theta))\}$$

$$= \sum_{k=0}^{\infty} \frac{d}{d \log \theta'\theta} (e^{-\frac{1}{2}\theta'\theta} e^{k \log \theta'\theta}) \frac{e^{-k \log 2}}{k!} Eh(\chi^2_{p+2k})$$

$$= \sum_{k=0}^{\infty} (e^{-\frac{1}{2}\theta'\theta} e^{k \log \theta'\theta} (k - \frac{1}{2} \frac{\partial \theta'\theta}{\partial \log \theta'\theta})) \frac{e^{k \log 2}}{k!}$$

$$Eh(\chi^2_{p+2k}).$$

Since

$$\frac{\partial \theta'\theta}{\partial \log \theta'\theta} = \theta'\theta,$$

rearranging terms yields

$$\frac{\partial}{\partial \log \theta'\theta} E\{h(\chi_p^2(\theta'\theta))\}] = \frac{\theta'\theta}{2} \sum_{k=1}^{\infty} \frac{e^{-\frac{1}{2}\theta'\theta}}{(k-1)!} (\frac{\theta'\theta}{2})^{k-1} Eh(\chi^2_{p+2k})$$

$$- \frac{\theta'\theta}{2} \sum_{k=0}^{\infty} \frac{e^{-\frac{1}{2}\theta'\theta}}{k!} (\frac{\theta'\theta}{2})^k Eh(\chi^2_{p+2k}).$$

$$= \frac{\theta'\theta}{2} \left[\sum_{j=0}^{\infty} \frac{e^{-\frac{1}{2}\theta'\theta}}{j!} \left(\frac{\theta'\theta}{2}\right)^j Eh(\chi^2_{p+2+2j}) \right.$$

$$\left. - \sum_{k=0}^{\infty} \frac{e^{-\frac{1}{2}\theta'\theta}}{k!} \left(\frac{\theta'\theta}{2}\right)^k Eh(\chi^2_{p+2k}) \right]$$

$$= \frac{\theta'\theta}{2} [Eh(\chi^2_{p+2}(\theta'\theta)) - Eh(\chi^2_p(\theta'\theta))].$$

From the chain rule,

$$\frac{\partial}{\partial\theta_i^2} E\{h(\chi_p^2(\theta'\theta))\} = \left(\frac{\partial \log \theta'\theta}{\partial \theta_i^2}\right)\left(\frac{\partial\, Eh(\chi_p^2(\theta'\theta))}{\partial \log \theta'\theta}\right),$$

and since

$$\frac{\partial \log \theta'\theta}{\partial\theta_i^2} = \frac{1}{\theta'\theta}, \qquad 1 \leq i \leq p,$$

the result is proved. ||

<u>Lemma 8</u>: Let D = diagonal $(d_1,\ldots,d_p)$. If $E\{h(\chi_p^2(\theta' D^{-1}\theta))\}$ exists, then

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

$$E\{h(\chi_p^2(\theta' D^{-1}\theta))\}],$$

for $1 \leq i \leq p$.

<u>Proof</u>: Similar to that of Lemma 7. ||

Lemma 9: Let $p \geq 3$ and $r: \mathbb{R} \to [0,\infty)$ satisfy

i) $r(t)$ is non-decreasing

ii) $r(t)/t$ is non-increasing.

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

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

is stricly increasing in a if either $0 \leq ar(t) \leq 2(p-2)/m$, $\forall\, t > 0$, or $p \geq 4$.

Proof: By an argument similar to that in Lemma 7 we can differentiate inside the expectation, and after some algebra we obtain

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

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

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

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

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

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

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

$$E_K E\left\{\frac{s(K)r(u)(ar(u)m+4)}{(ar(u)\chi_m^2+\chi_{p-2+2K}^2)^3}(\chi_{p-2+2K}^2)^3 \times\left(\chi_{p-2+2K}^2-\frac{2amr(u)}{amr(u)+2}\right)\Big|K\right\}, \tag{2}$$

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

$$q(\chi_{p-2+2K}^2,\chi_m^2) = \frac{s(K)r(u)(ar(u)m+4)(\chi_{p-2+2K}^2)^3}{(ar(u)\chi_m^2+\chi_{p-2+2K}^2)^3},$$

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

$$E_K E\, q(\chi_{p-2+2K}^2,\chi_m^2)(\chi_{p-2+2K}^2-(p-2+2K)) + E_K E\, q(\chi_{p-2+2K}^2,\chi_m^2)\left(p-2+2K-\frac{2amr(u)}{amr(u)+2}\right) \tag{3}$$

The first expectation is non-negative from Lemma 5, and if $p \geq 4$, the second expectation is strictly positive since

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

If $p = 3$, since $0 \leq ar(t) \leq 2(p-2)/m$, the only concern is if $ar(t_0) = 2(p-2)/m = 2/m$, for some t_0. But then it follows from condition (i) that $ar(t) = 2/m$, $\forall\, t > t_0$, 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.||

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. ____________(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 University.

8. Chung, Kai Lai. (1968). A Course in Probability Theory. Harcourt, Brace & World, Inc., New York.

9. Efron, B., and Morris, C. (1976). Families of minimax estimators of the mean of a multivariate normal distribution. Ann. Statist. 4, 11-21.

10. Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12, 55-68.

11. James, W., and Stein, C. (1961). Estimation with quadratic loss. Proc. Fourth Berkeley Symp. Math. Statist. Probl. 1, 361-379.

12. Lehmann, E. L. (1959). Testing Statistical Hypotheses. John Wiley and Sons, New York.

13. Stein, C. (1955). Inadmissibility of the usual estimator of the mean of a multivariate normal distribution. Proc. Third Berkeley Symp. Math. Statist. Prob. 1, 197-206.

14. Strawderman, W. E. (1976). Minimax adaptive generalized ridge regression estimators. Technical Report, Department of Statistics, Rutgers University.

15. Thisted, Ronald A. (1976). Ridge regression, minimax estimation, and empirical Bayes methods. Ph. D. Thesis, Stanford University.

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Mimeo. Series No. 497

2. GOVT ACCESSION NO.:

3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): Minimax Ridge Regression Estimation

5. TYPE OF REPORT & PERIOD COVERED: Scientific Interim

6. PERFORMING ORG. REPORT NUMBER: Mimeo. Series #497

7. AUTHOR(s): Goerge Casella

8. CONTRACT OR GRANT NUMBER(s): AFOSR-72-2350E

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
Purdue University
Department of Statistics
West Lafayette, IN 47907

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F 9769-05

11. CONTROLLING OFFICE NAME AND ADDRESS:
Air Force Office of Scientific Research
Bldg. 410, Bolling Air Force Base
Washington, DC 20332

12. REPORT DATE: May, 1977

13. NUMBER OF PAGES:

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):

15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report):
Approved for public release, distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report):

18. SUPPLEMENTARY NOTES:

19. KEY WORDS (Continue on reverse side if necessary and identify by block number):
Ridge regression, minimax, normal, mean, quadratic loss, risk function.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number):

The technique of ridge regression, first proposed by Hoerl and Kennard (1970), has become a popular tool for data analysts faced with a high degree of multicollinearity in their data. By using a ridge estimator, it was hoped that one could both stabilize his 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 (1955) and others, Strawderman (1976) and Thisted (1976)

DD FORM 1473, 1 JAN 73 — EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

have developed classes of ridge regression estimators which dominate the usual estimator in risk, and hence are minimax. The unwieldly form of the risk function, however, has lead 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.