ON THE USE OF RIDGE AND STEIN-TYPE ESTIMATORS IN PREDICTION

BY

ALAN E. GELFAND

TECHNICAL REPORT NO. 374
MAY 21, 1986

PREPARED UNDER CONTRACT
N00014-86-K-0156 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

Reproduction in Whole or in Part is Permitted for any purpose of the United States Government

Approved for public release; distribution unlimited.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ON THE USE OF RIDGE AND STEIN-TYPE ESTIMATORS IN PREDICTION

BY

ALAN E. GELFAND

TECHNICAL REPORT NO. 374
MAY 21, 1986

Prepared Under Contract
N00014–86–K–0156 (NR–042–267)
For the Office of Naval Research

Herbert Solomon, Project Director

Reproduction in Whole or in Part is Permitted
for any purpose of the United States Government

Approved for public release; distribution unlimited.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ON THE USE OF RIDGE AND STEIN-TYPE ESTIMATORS IN PREDICTION

Alan E. Gelfand

1. Introduction

For the usual regression model with fixed regressors, \( Y = X\beta + \epsilon, \) \( Y_{n\times1}, X_{n\times p} \) full rank, \( \beta_{p\times1} \) and \( \epsilon_{n\times1} \sim (0, \sigma^2 I) \), there is considerable literature devoted to alternatives to the ordinary least squares estimator, \( \hat{\beta}_{OLS} \) of \( \beta \). From work originally dating to Stein (1956) and James and Stein (1961) when \( \epsilon \) is normally distributed and \( p \geq 3 \), \( \hat{\beta}_{OLS} \) is inadmissible under loss \( (\hat{\beta} - \beta)^T Q (\hat{\beta} - \beta) \), \( Q \) an unrestricted positive definite matrix. Thus, much of this extant discussion focuses on the development of biased estimators with small "variances" which achieve a smaller expected loss either uniformly over \( p \)-dimensional Euclidean space or at least in the vicinity of some specified \( \beta^* \). Two "classes" of such reduced variance regression estimators are particularly well discussed - ridge estimators and Stein-type estimators. Either directly or upon orthogonal transformation these estimators take the form

\[
\hat{\beta}_C = C\hat{\beta}_{OLS} + (I-C)\beta^*
\]

where \( C \) is a diagonal matrix, usually data dependent. They may also be seen to be Bayes or "Empirical" Bayes procedures as well. The review paper by Draper and Van Nostrand (1979) provides an
excellent summary of both the theoretical and simulated effort in this area. In the context of cross-validation, i.e. of examining the performance of an estimator obtained in one sample in prediction in a second independent sample, the work of Stone (1974) leads to estimators of the form in (1) as well.

Herein we consider the simplest such cross-validation problem. At a new vector of predictor values, \( X_0 \), we seek to estimate \( X_0^T \beta \). We take as loss function \((\delta(Y) - X_0^T \beta)^2\) for an estimator \( \delta(Y) \) and we assume henceforth that \( \epsilon \) is normally distributed with \( \sigma^2 \) unknown. Our problem differs from that of estimating the vector \( \beta \) since the results of Cohen (1965) show that \( \alpha X_0^T \beta_{OLS} \) is an admissible estimator of \( X_0^T \beta \) for \( 0 \leq \alpha \leq 1 \), i.e. the UMVU estimator is admissible. (In fact, \( \delta(Y) \) of the form \( \gamma^T Y \) is admissible for \( X_0^T \beta \) i.f.f.
\[(2 \gamma - X(X^T X)^{-1} X_0)^T (2 \gamma - X(X^T X)^{-1} X_0) \leq X_0^T (X^T X)^{-1} X_0.\]) Nonetheless, if we have some confidence in \( \beta^* \), i.e. that \( \beta^* \) is near the true value \( \beta \), then it makes sense to attempt to improve upon \( X_0^T \beta_{OLS} \) in the "vicinity of \( \beta^* \)" using estimators of the form (1). More specifically, how well do the "classes" of ridge estimators and of Stein-type estimators perform in this prediction? Can we make a "best" choice within these classes for a particular prediction?

The problem of prediction of an independent observation \( Y_0 \) at \( X_0 \) using the loss \((\delta(Y) - Y_0)^2\) is equivalent to that of predicting \( X_0^T \beta \), i.e. \( E\beta (\delta(Y) - Y_0)^2 = \sigma^2 + E\beta (\delta(Y) - X_0^T \beta)^2. \)
For an estimator of the form $X_0^T \hat{\beta}$, the expected loss becomes

\begin{equation}
E_\beta (\hat{\beta} - \beta)^T X_0 X_0^T (\hat{\beta} - \beta) = E_\beta [\Sigma X_0 (\hat{\beta}_1 - \beta_1)]^2.
\end{equation}

In the sequel we take the generalized ridge estimator $\hat{\beta}_R$ to be

\begin{equation}
\hat{\beta}_R = (X^T X + A)^{-1} (X^T Y + A \beta^*).
\end{equation}

where $A$ is p.d. symmetric and possibly dependent on $Y$. We take the general Stein-type estimator $\hat{\beta}_S$ to be

\begin{equation}
\hat{\beta}_S = (1 - c/Q) \hat{\beta}_\text{OLS} + c/Q \beta^*
\end{equation}

where $Q = (\hat{\beta}_\text{OLS} - \beta^*)^T X^T X (\hat{\beta}_\text{OLS} - \beta^*)$ and $c$ may depend on $Y$. In practice $c/Q$ is usually replaced by min($c/Q$, 1).

In section 2 we calculate the risk (2), of the estimators (3) and (4) when $A$, $c$ are constant. We then investigate "best" choices for $A$, $c$. Since these choices will be functions of $\beta$ and $\sigma^2$ as well as $X_0$, $A$ and $c$ must be estimated from $Y$. In section 3 we summarize a simulation study which compares the performance of versions of (3) and (4) which are discussed for the estimation of $\beta$ along with others motivated by work in section 2. In section 4 we offer concluding remarks in particular with regard to multiple prediction.
2. Theoretical Results

We first note that for \( \hat{\beta}_{\text{OLS}} \) (2) becomes

\[
\sigma^2 \mathbf{X}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_0.
\]

We now claim that

Theorem 1: For \( \hat{\beta}_R \) as in (3), (2) becomes

\[
\sigma^2 \mathbf{X}_0^T (\mathbf{X}^T \mathbf{X}+\mathbf{A})^{-1} \mathbf{X} (\mathbf{X}^T \mathbf{X}+\mathbf{A})^{-1} \\
+ \mathbf{X}_0^T (\mathbf{X}^T \mathbf{X}+\mathbf{A})^{-1} \mathbf{A} (\mathbf{\beta}-\mathbf{\beta}^*) (\mathbf{\beta}-\mathbf{\beta}^*)^T \mathbf{A} (\mathbf{X}^T \mathbf{X}+\mathbf{A})^{-1} \mathbf{X}.
\]

Proof: We transform to principal components form. Let \( \mathbf{R} \) be nonsingular such that \( \mathbf{R} \mathbf{X} \mathbf{R}^T = \mathbf{I}, \mathbf{R} \mathbf{A} \mathbf{R}^T = \mathbf{D}, \mathbf{D} \) a diagonal matrix with diagonal entries \( \mathbf{d}_1 \). For any point \( \mathbf{\beta} \) in \( p \)-dimensional Euclidean space, let \( \mathbf{\alpha} = (\mathbf{R}^{-1})^T \mathbf{\beta} \). Then

\[
\hat{\mathbf{\alpha}}_R = (\mathbf{R}^{-1})^T \hat{\mathbf{\beta}}_R = (\mathbf{I}+\mathbf{D})^{-1} \hat{\mathbf{\alpha}}_{\text{OLS}} + (\mathbf{I}+\mathbf{D})^{-1} \mathbf{D} \mathbf{\alpha}^* ,
\]

i.e. of the form (1) with \( C = (\mathbf{I}+\mathbf{D})^{-1} \). In terms of \( \mathbf{\alpha} \), (2) becomes \( \mathbf{E}_\mathbf{\alpha} (\mathbf{\hat{\alpha}}-\mathbf{\alpha}) \mathbf{w}_0^T \mathbf{w}_0 (\mathbf{\hat{\alpha}}-\mathbf{\alpha}) \), \( \mathbf{w}_0 = \mathbf{R} \mathbf{X} \). Since \( \mathbf{\hat{\alpha}} \sim \mathbf{N} (\mathbf{\alpha}, \sigma^2 \mathbf{I}) \), this expectation is readily calculated to be

\[
\sigma^2 \mathbf{w}_0^T \mathbf{C}^2 \mathbf{w}_0 + \mathbf{w}_0^T (\mathbf{I}-\mathbf{C}) (\mathbf{\alpha}-\mathbf{\alpha}^*) (\mathbf{\alpha}-\mathbf{\alpha}^*)^T (\mathbf{I}-\mathbf{C}) \mathbf{w}_0 .
\]

Substitution for \( \mathbf{\alpha}, \mathbf{w}_0 \) and \( \mathbf{C} \) yields (6). \( \square \)
Note: Normality is not employed in this calculation.

In (3), $A_i$ is usually taken to be diagonal and, in fact, the class of ridge (as opposed to generalized ridge) estimators sets $A = aI$, $a > 0$. The case where either by design or transformation $X^T X = I$ reduces (5) to $\sigma^2 \Sigma x_{0i}^2$ and reduces (6), for generalized ridge estimators ($a_i$ are the diagonal elements of the diagonal matrix $A$) to

\begin{equation}
\sigma^2 \Sigma x_{0i}^2 \frac{1}{(1+a_i)^2} + [\Sigma x_{0i}(\beta_i - \beta^*) \frac{a_i}{1+a_i}]^2.
\end{equation}

Investigation of this expression reveals that an optimal choice for the $a_i$ to minimize (7) needn't exist although local minima can be found. In the case of ridge estimation, i.e. all $a_i = a$, a unique minimum can be found. This occurs at

\begin{equation}
a_0 = \sigma^2 \gamma^{-2} X_{0}^T X_0
\end{equation}

where $\gamma = X_0^T (\beta - \beta^*)$. Note that $a_0 > 0$ and finite provided $\beta - \beta^*$ isn't orthogonal to $X_0$. The associated minimum equals

\begin{equation}
\frac{\sigma^2 \gamma^2 X_{0}^T X_0}{\gamma^2 + \sigma^2 X_{0}^T X_0}
\end{equation}

When $\beta$ is such that $\beta - \beta^*$ is orthogonal to $X_0$, then $X_0^T \beta^*$ predicts perfectly. For such $\beta$'s we can obtain zero expected loss and
would want no weight attached to $\hat{\beta}_{OLS}$, i.e. would want $a = \infty$. In fact, it is clear that for $X_0, \beta$ fixed there will be a set of $\beta' \ 's which predict $X_0^T$ perfectly and that $\beta' \ 's needn't be close to $\beta$ in Euclidean distance. Thus the appropriate pseudometric for the prediction problem is $(\beta_1-\beta_2)^T X_0 X_0^T (\beta_1-\beta_2)$. This pseudometric clarifies the earlier notion of "vicinity of $\beta^*$" and under this distance the further $\beta^*$ is from $\beta$ the closer $a$ is to 0, i.e. the more weight is placed on $\hat{\beta}_{OLS}$, the closer $\beta^*$ is to $\beta$ the larger $a$ becomes, i.e. the more weight is placed on $\beta^*$. As would be expected, $a_0$ is invariant to scaling of $X_0$, although the risk clearly isn't.

Using (8) our estimator of $X_0^T \beta$ is

$$T_{a_0} = (1+a_0)^{-1} X_0^T \hat{\beta}_{OLS} + (1+a_0)^{-1} a_0 X_0^T \beta^*$$

and, in fact, for any fixed $a > 0$, $T_{a}$ improves upon $X_0^T \beta_{OLS}$ whenever $\gamma^2 < \sigma^2 a^{-1}(2+a)$.

From (8) a convenient estimator of $a_0$ is:

(9) $$\hat{a}_0 = \hat{\sigma}^2 \gamma^2 - 2 T_{X TX} X_0$$

when $\sigma^2$ is the usual UNIVU estimator of $\sigma^2$ and $\gamma = X_0^T (\hat{\beta}_{OLS} - \beta^*)$. The fact that $E \gamma^{-2}$ doesn't exist suggests that $\hat{a}_0$ will be very unstable and that $T_{a_0}$ will perform poorly. We return to this point in the discussion of the simulation study. Since $\hat{\sigma}^2$ is
independent of \( \hat{\beta}_{OLS} \) and \( \hat{a}_0 \) depends on \( \hat{\beta}_{OLS} \) only through \( X_0^T \hat{\beta}_{OLS} \), we may compute the expected loss for \( T_{\hat{a}_0} \). If \( \tau^2 = \sigma^2 X_0^T X_0 \), then \( \gamma \sim N(\gamma, \tau^2) \) and, with \( \tau^2 = \sigma^2 X_0^T X_0 \), (2) for \( T_{\hat{a}_0} \) becomes

\[
(10) \quad E_{\gamma} \left( \frac{\gamma^2}{\gamma^2 + \tau^2} \right) \gamma - \gamma)^2 = \tau^2 + E_{\gamma} \left( \frac{\gamma^2(\tau^2)^2 + 2\tau^2 \gamma^2 - 2\tau^2 (\gamma^2)^2}{\gamma^2 + \tau^2 (\gamma^2)^2} \right).
\]

The equality (10) is seen using the identity \( E_{\gamma} f(\gamma)(\gamma - \gamma) = \tau^2 E_{\gamma} f'(\gamma) \) (Stein (1973)) valid provided \( E_{\gamma} |f'(\gamma)| < \infty \) which, as the following calculations show, is the case. Now \( \gamma^2 / \tau^2 - X_1^2 \), \( \gamma^2 / 2 \tau^2 \) independent of \( \gamma^2 / \tau^2 - X_{n-p}^2 \). Hence

\[
(\gamma^2 + \tau^2)^{-1} \gamma^2 \sim L - B(\frac{n-p}{2}, \frac{2L+1}{2}) \text{ where } L \sim Po(\gamma^2 / 2 \tau^2).
\]

The expectation of each term in (10) can thus be evaluated and (10) becomes

\[
(11) \quad \tau^2 \{L + (n-p)E(n-p+2L+3)^{-1}[(n-p+2)(2L+1)] - \frac{n-p-2L+1}{n-p+2L+1}\}.
\]

If we divide (11) by \( \tau^2 \), i.e. consider the risk relative to that of \( X_0^T \hat{\beta}_{OLS} \), then this relative risk is a function of \( \gamma^2 / \tau^2 \). Hence we set \( \tau^2 = 1 \) and examine the simpler estimator \( (\gamma^2 + 1)^{-1} \gamma^3 \) which may be thought of as an "empirical" Bayes estimator against a normal prior centered at 0, adjusted to have no singularities in \( \mathbb{R}^1 \). The risk of this estimator is readily obtained to be

\[
1 + E_{\gamma} (\gamma^2 + 1)^{-2} (3\gamma^2 - 2) \text{ by an argument similar to that leading to (10). This risk (symmetric about 0) is graphed in Figure 1 against } \gamma > 0 \text{ to illustrate what may be expected, up to scaling, if (11) is evaluated. Note that the risk is bounded and considerably}
\]
less than 1 for \( \gamma \) small. Because \((\gamma^2 + 1)^{-2}\gamma^3\) has singularities in the complex plane it is not admissible.

If we restore \(X^TX\), not necessarily diagonal, our estimator in (3) has \(A = a_0X^TX\) or \(a_0^TX^TX\) according to (8) or (9).

Theorem 2: For \(\beta_S\) as in (4) with \(p > 2\), (2) becomes

\[\sigma^2x_0^T(X^TX)^{-1}x_0 + x_0^T(X^TX)^{-1}x_0[(c^2 + 4c\gamma^2)\Gamma_1/\sigma^2 - 2c\Gamma_2]\]

where

\[\Gamma_1 = \frac{2L+1}{p+2(L+M)/(p+2(L+M)-2)}, \quad \Gamma_2 = \frac{1}{p+2(L+M)-2}\]

with

\[L \sim \text{Po}(\lambda), \quad \lambda = \gamma^2/2\sigma^2x_0^T(X^TX)^{-1}x_0\]

\[M \sim \text{Po}(\delta), \quad \delta = (\Delta x_0^T(X^TX)^{-1}x_0 - \gamma^2)/2\sigma^2x_0^T(X^TX)^{-1}x_0\]

where \(L, M\) independent and \(\Delta = (\hat{\beta} - \beta_*)^T X^TX (\hat{\beta} - \beta_*)\).

Proof: As in Theorem 1, let \(R\) be nonsingular such that \(RX^TXR^T = I\) and let \(\hat{\alpha} = (R^{-1})^T(\hat{\beta}_{OLS} - \beta_*)\). Then \(\hat{\alpha} \sim N(\alpha, \sigma^2I)\) with \(\alpha = (R^{-1})^T(\beta - \beta_*)\), \(Q = \hat{\alpha}^T\hat{\alpha}\), and (2) becomes
(14) \[ E_\alpha \left[ (1 - \frac{c}{q}) w_0^T \alpha - w_0^T \alpha \right]^2 \]

where \( w = RX_0 \) (and \( w_0^T = X_0^T (X_0^T X_0^{-1}) X_0 \)).

If we expand (14) we obtain

(15) \[ E_\alpha (w_0^T \alpha - \alpha \alpha_0)^2 + c^2 E(w_0^T \alpha)^2 / (\alpha_0^T \alpha)^2 - 2c E \frac{w_0^T \alpha}{\alpha_0^T \alpha} \ w_0^T \alpha (\alpha - \alpha) . \]

The last term may be written as \(-2c \sum w_0^T \alpha f(\alpha) (\alpha_0 - \alpha_0)\) where \( f(\alpha) = (\alpha_0^T \alpha)^{-1} (w_0^T \alpha) \). Using the Stein identity,

\[ (\sigma^2 E \frac{\partial f(\alpha)}{\partial \alpha} = E_\alpha f(\alpha) (\alpha_0 - \alpha_0) \), which is valid here, on this expression, after manipulation (15) becomes

(16) \[ \sigma^2 w_0^T \alpha + (c^2 + 4c\sigma^2) E_\alpha (w_0^T \alpha)^2 / (\alpha_0^T \alpha)^2 - 2c \sigma^2 w_0^T \alpha E_\alpha (1 / \alpha_0^T \alpha) . \]

Finally if we let \( U = \frac{w_0^T \alpha}{w_0^T \alpha}, V = \alpha_0^T \alpha - \frac{(w_0^T \alpha)^2}{w_0^T \alpha} \), then

\[ \frac{U}{\sigma^2} | L - \frac{X_1^2}{1 + 2L} \quad \text{with L as in (13)} \]

\[ \frac{V}{\sigma^2} | M - \frac{X_2}{p - 1 + 2M} \quad \text{with M as in (13)} \]

and given L and M, \( \frac{U}{U+V} \sim \text{Be} \left( \frac{1 + 2L}{2}, \frac{p - 1 + 2M}{2} \right) \) independent of \( \frac{U+V}{\sigma^2} \sim \frac{X_2}{p + 2(L + M)} \). Hence
\[ E_{\alpha\left(\frac{w^T\hat{\alpha}}{\alpha^{\top}\alpha}\right)^2} = w^T w E_{\alpha\left(\frac{U}{U+V}\right) | L,M} = w^T w E_{\alpha\left(\frac{U}{U+V}| L,M\right) E_{\left(\frac{1}{U+V}| L,M\right)} \]
\[ = \frac{w^T w}{\sigma^2} E_{\left(\frac{2L+1}{p+2(L+M)}\right) \frac{1}{p+2(L+M)-2}} = \frac{w^T w}{\sigma^2} \Gamma_1 \]

and similarly

\[ E_{\alpha\left(\frac{1}{\alpha^{\top}\alpha}\right)} = \frac{1}{\sigma^2} \Gamma_2. \]

Making these substitutions into (16) and restoring \(X_0\) we obtain (12). \(\Box\)

Note: The proof reveals that the expected loss, (2), for more general estimators of the form \((1-h(Q))\hat{\beta}_{OLS}+h(Q)\beta^*\) can be developed. In fact, if

\[ E_{\beta} \left| \frac{\text{ah}(Q)\hat{\alpha}}{\hat{\beta}_{OLS,1}} \right| < \infty \]

the loss is

\[ \sigma^2 X_0^T (X^T X)^{-1} X_0 + E_{\beta} h^2(Q) \hat{\gamma}^2 - 2\sigma^2 X_0^T (X^T X)^{-1} X_0 E_{\beta} h(Q) - 4\sigma^2 E_{\beta} h'(Q) \hat{\gamma}^2 \]

Inspection of (12) reveals that the unique best \(c\) is

\[ c_0 = \sigma^2 \left( \frac{\Gamma_2}{\Gamma_1} - 2 \right). \]

Since \(\lambda,\delta\) are invariant to scaling of \(X_0\) so is \(c_0\). It is
apparent that for $X_0$ fixed as $\Delta = 0$, $\Gamma_2/\Gamma_1 + p$, i.e.
$c_0 = \sigma^2(p-2)$. Hence if we believe $\beta^*$ is close to $\beta$ the
"usual" constant, $\sigma^2(p-2)$, may be employed. Using this constant,
if $\Delta$ is near 0, the relative risk of $X_0^T\hat{\beta}_S$ to $X_0^T\hat{\beta}_{OLS}$ will, from
(12), be near $2/p$, as it is in the case of estimating $\beta$. As in
remarks after (10), if $\sigma^2$ is the usual estimator of $\sigma^2$ independent
of $\beta_{OLS}$ we may compute (2) for $\hat{\beta}_S$ as in (4) with $c = \sigma^2(p-2)$.¹
We obtain

$$\sigma^2 X_0^T (X^T X)^{-1} X_0 \{1 + \Gamma_1 (p^2 - 4 + 2(n-p)^{-1}(p-2)^2) - 2\Gamma_2 (p-2)\}. \tag{19}$$

Expressions similar to (19) can be obtained for instances
of the more general estimators mentioned above (17). This suggests
that the risk (2) may be calculated for commonly used (adaptive)
ridge estimators, e.g. those considered in section 3. However,
without restrictions on the design matrix $X$, these estimators
often fail to either provide $\hat{\alpha}$ in closed form or to define
$\hat{\alpha}$ as a function of $Q$.

Since to a first order approximation $c_0 \approx \sigma^2(2\lambda + 1)^{-1}(p-2 + 2(\delta - \lambda))$
we may estimate $c_0$ by

$$\hat{c}_0 = \sigma^2(2\lambda + 1)^{-1} (p-2 + 2(\delta - \hat{\lambda})) \tag{20}$$

where $\lambda, \delta$ are the expressions in (13) with $\beta$ replaced by $\hat{\beta}_{OLS}$. We would truncate $c_0$ to the interval $[0, Q]$. Since $E\lambda^{-1}$ doesn't
exist $c_0$ will be very unstable (as with $\hat{a}_0$ in (9)) suggesting that the resulting predictor will perform poorly. Again we return to this point in the next section.

3. A Simulation Study

A simulation was conducted to compare the use of the OLS predictor with the predictors discussed in the previous section and with predictors arising from other estimators of $\beta$ which have been discussed in the literature. For convenience we set $\sigma^2 = 1$ and take $X^TX = I$, i.e. $\hat{\beta}_{\text{OLS}} \sim N(\beta, I)$. Without loss of generality we set $\beta^* = 0$ and $X_0^TX_0 = 1$. Under this setup ridge estimators become $(1+\hat{a})^{-1}\hat{\beta}_{\text{OLS}}$ and Stein estimators become $(1 - c/\hat{\beta}_{\text{OLS}}^T\hat{\beta}_{\text{OLS}})\hat{\beta}_{\text{OLS}}$. In addition to $\hat{\beta}_{\text{OLS}}$ we consider the following six estimators of $\beta$ (4 ridge type, 2 Stein type).

(i) $\hat{\beta}_{\text{HK}}$ - arising from $\hat{a} = p/\hat{\beta}_{\text{OLS}}^T\hat{\beta}_{\text{OLS}}$. The ridge estimators discussed by Hoerl and Kennard (1970), Hoerl, Kennard and Baldwin (1975) and, in fact, Lawless and Wang (1976) reduce to $\hat{\beta}_{\text{HK}}$ in our setup.

(ii) $\hat{\beta}_{\text{RM}}$ - arising from $\hat{a} = p/(\hat{\beta}_{\text{OLS}}^T\hat{\beta}_{\text{OLS}} - p)$ with $(1+\hat{a})^{-1} = 0$ if $\hat{\beta}_{\text{OLS}}^T\hat{\beta}_{\text{OLS}} \leq p$. The RIDDM and STEINM estimators discussed by Dempster, Shatzoff, and Wermuth (1977) reduce to $\hat{\beta}_{\text{RM}}$.

(iii) $\hat{\beta}_{\text{MG}}$ - arising from $\hat{a} = (1 - p/\hat{\beta}_{\text{OLS}}^T\hat{\beta}_{\text{OLS}})^{-1/2}(1 - (1 - p/\hat{\beta}_{\text{OLS}}^T\hat{\beta}_{\text{OLS}})$ with $(1+\hat{a})^{-1} = 0$ if $\hat{\beta}_{\text{OLS}}^T\hat{\beta}_{\text{OLS}} \leq p$. The ridge estimator of McDonald and Galarneau (1975) reduces to $\hat{\beta}_{\text{MG}}$. 
(iv) $\hat{\beta}_a$ - arising from $\hat{\alpha}_0$ given in (9).

(v) $\hat{\beta}_{p-2}$ - arising from $c = p-2$, i.e. the "usual" Stein estimator.

(vi) $\hat{\beta}_c$ - arising from $\hat{c}_0$ given in (20), truncated to $[0, \infty)$.

"Positive part" restrictions were applied to all "shrinkages" in (v) and (vi).

We note that under the above assumptions the risks in (3) and (4) and, in fact, of the predictors arising from (i) - (vi) depend on $X_0$ and $\beta$ only through $(X_0^T \beta)^2$ and $\beta^T \beta$. Since $(X_0^T \beta)^2 = r \beta^T \beta$, $0 \leq r \leq 1$, we may summarize the results in terms of $\beta^T \beta$ and $r$. We consider $p = 3, 6, 10$. For a given $p$, we generated sets of $2p$ independent uniform random variables on the interval $[-1, 1]$. In each case we considered the first $p$ observations as a $p$ vector, standardized to length 1 and designated it as an $X_0$. Similarly, the second $p$ observations are considered as a $\beta$ vector with scaling by $1, 1, 10$. Hence we have large $\beta^T \beta$, i.e. $\beta^T \beta = 100$, moderate $\beta^T \beta$, i.e. $\beta^T \beta = 1$ and small $\beta^T \beta$, i.e. $\beta^T \beta = .01$. For each $X_0, \beta$ pair 1000 $\hat{\beta}_{OLS}$'s were generated from $N(\beta, I)$ and using $X_0$, each of the seven predictors were calculated for each of the 1000 replications. Bias, variance and mean square error (MSE) were estimated. A large number of $X_0, \beta$ pairs (approximately 400) were investigated enabling a wide range of $r$'s. Table 1 provides a brief summary indicating the best $\hat{\beta}$ for prediction over ranges.
for $r$ along with the typical percentage reduction in risk (using the best predictor over that range), $100(MSE \hat{X}_{0}^{T} \hat{\beta} - MSE \hat{X}_{0}^{T} \hat{\beta})/MSE \hat{X}_{0}^{T} \hat{\beta}$. Several comments are appropriate:

(i) The cross-over points in Table 1 are approximate, but in the vicinity of the cross-over competing predictors are indistinguishable with respect to MSE.

(ii) It is not surprising that regardless of $\beta$, if $\beta^T \beta$ large and $r$ large, the OLS predictor is best. In fact, if $\beta^T \beta$ large and $.01 < r < .5$, the percent improvement of the best predictor over OLS is never greater than 5%.

(iii) As expected, $\hat{\beta}_{a_0}^c$, $\hat{\beta}_{a_0}^c$ performed very badly, always sixth or seventh, doing well only when $\beta^T \beta$ large and $r$ very small (regardless of $p$). However, in such cases, improvements will be substantial, increasing as $r$ decreases, while the other five predictors are indistinguishable. Near $r = .01$ $\hat{\beta}_{a_0}^c$ is best; much below .01, $\hat{\beta}_{a_0}^c$ is best.

(iv) $\hat{\beta}_{p-2}$ is likely the best overall choice always amongst the two or three best apart from cases in (iii) above.

(v) When $\beta^T \beta$ is small or moderate, $\hat{\beta}_{RM}$, $\hat{\beta}_{HK}$, $\hat{\beta}_{p-2}$ and $\hat{\beta}_{MG}$ were always the best four. When $\beta^T \beta$ is small and $p = 3$, $\hat{\beta}_{MG}$ is close in performance to $\hat{\beta}_{p-2}$. When $\beta^T \beta$ is small and $p = 6$, $\hat{\beta}_{RM}$ and $\hat{\beta}_{MG}$ split for second best. When $\beta^T \beta$ is small and $p = 10$, $\hat{\beta}_{p-2}$ is second with $\hat{\beta}_{MG}$ third.
(vi) When \( r \) is large \( \hat{c}_0 \) is almost always \(<0 \) whence \( \hat{\beta}_0 = \hat{\beta}_{OLS} \).

Table 1 reveals that in many cases substantial reduction in squared error loss over the OLS predictor can be achieved. It further suggests the possibility of selecting the predictor according to \( \beta^T \beta \) and \( r \). However, finely detailed selection, e.g. according to \( X_0 \), will be unsuccessful as the performance of \( \hat{\beta}_\alpha \) and \( \hat{\beta}_0 \) reveals. In practice we will have \( \Delta/\sigma^2 \) instead of \( \beta_\alpha^T \hat{c}_0 \hat{a}_0 \) and \( r = (\Delta X_0^T(X^TX)^{-1}X_0)^{-1} \gamma^2 \), and we might define estimators \( \hat{\Delta}, \hat{r} \) with \( \beta \) replaced by \( \hat{\beta}_{OLS} \), \( \sigma^2 \) replaced by \( \hat{\sigma}^2 \). We may calculate
\[
E(\hat{\Delta}) = \frac{n-1}{n-3} (p+\Delta) \text{ whence } \hat{\Delta}' = \frac{n-3}{n-1} \hat{\Delta} - p \text{ is UMVU for } \Delta. 
\]
By an argument similar to that contained in Theorem 2, we may show
\[
E(\hat{r}) = E(p+2(L+M))^{-1}(2L+1) = r+(1-rp)\{(p+\Delta)^{-1}+(p+\Delta)^{-3}\Delta\}
\]
where \( L,M \) are distributed as in (13). For individual predictions, preliminary calculation of \( \hat{\Delta}' \) and \( \hat{r} \) should enable a judicious choice of predictor.

As Thisted and Morris (1980, p. 19) observe, the poorest estimation case for ridge procedures occurs when \((\text{with } X^T X = I, \beta^* = 0) \beta^T = (\beta_1,0,0,...,0) \) with \( \beta_1 \) large. This is also the poorest estimation case for Stein type procedures in the sense that the first coordinate will account for about half of the total risk and all coordinate risks would decrease if \( \beta_1 \) was excluded (see Baranchik (1964)). For prediction this implies \( \Delta \) large and \( r = \frac{X_{01}^2}{X_0^T X_0} \). Hence this is the poorest case for prediction as well, i.e. depending upon \( X_{01} \), improvement will be small or, in fact, the OLS predictor will be better.
How will multicollinearity in $X^TX$ affect prediction?  Let $X^TX = D$, a diagonal matrix with diagonal elements $d_1$ and assume $d_1 < d_2 < \ldots < d_p$. Then the extent of multicollinearity is usually measured in terms of how close $d_1$ is to 0. Although ridge methods have been advocated for improved estimation when $d_1$ is quite small, particularly relative to the other $d_i$, Thisted and Morris (p. 21) and others have established that in this case optimal ridge as well as Stein-type estimators will produce inconsequential improvement over $\hat{\beta}_{OLS}$. For prediction (with $\beta^* = 0$), $\Delta = \sum d_i \beta_i^2$ and $r = (X_0^T\beta)^2/\sum_{01}/d_1 \cdot \sum d_i \beta_i^2$. Bingham and Larntz (1977, p. 102) observe that (in this notation) the worst case for ridge estimation occurs when large $\beta_i$ are associated with small $d_1$. This tells us little about the magnitude of $\Delta$. However for fixed $X_0$ and $\beta$ as $X^TX$ becomes more severely multicollinear $\text{var}(X_0^T\hat{\beta}_{OLS}) = \sum_{01}/d_1$ will grow larger and $r$ will become smaller. As the simulation suggests when $r$ is small, using an appropriate predictor, we can expect significant improvement over the OLS predictor.

4. Multiple Prediction

In concluding we offer several comments regarding multiple or simultaneous prediction. Suppose we wish to make $r$ predictions defined by $X_1, X_2, \ldots, X_r$ and we set $X^* = (X_1, \ldots, X_r)$. For convenience we assume the $X_i$ are a linearly independent set whence rank $(X^*) = r \leq p$. What is an appropriate loss to employ? One
choice is unweighted sum of squared error loss, i.e.
\[ \Lambda(X_1^T(\hat{\beta}-\beta))^2 = (\hat{\beta}-\beta)^T G_1(\hat{\beta}-\beta) \] with \( G_1 = X^T X \). A second choice arises from the joint distribution of the \( X_1^T \hat{\beta}_{\text{OLS}} \), i.e.
\[ X^T \hat{\beta}_{\text{OLS}} \sim N(X^T \beta, \sigma^2 \gamma^2(X^T (X^T X)^{-1} X^*)^{-1}) \] suggests \( (\hat{\beta}-\beta)^T G_2(\hat{\beta}-\beta) \) where \( G_2 = X^T (X^T (X^T X)^{-1} X^*)^{-1} X^T \). Others may be envisioned as well. If \( r < p \), \( G_1, G_2 \) are positive semi-definite whence, as noted earlier for individual prediction, we may have loss equal to 0 but \( \hat{\beta} \) not close to \( \beta \). Nonetheless, it is well-known that if \( P \) is any positive definite matrix \( X^T \hat{\beta}_{\text{OLS}} \) is admissible for \( X^T \hat{\beta} \) under loss \( (\hat{\beta}-\beta)^T X^T P X^* T(\hat{\beta}-\beta) \) if \( p \leq 2 \), inadmissible if \( p \geq 3 \). In fact, work done by Berger (1976), Bhattacharyya (1966), Bock (1975), Casella (1977), Efron and Morris (1976) and Strawderman (1978) leads to explicit minimax predictors which improve upon \( X^T \hat{\beta}_{\text{OLS}} \). These predictors will be generalized adaptive ridge of the form
\[ (I + A(X^T \hat{\beta}_{\text{OLS}}, \sigma^2, X^T (X^T X)^{-1} X^*, P))^{-1} X^T \hat{\beta}_{\text{OLS}}, \] (see e.g. Strawderman, Theorem 6, p. 626, for a family of such predictors), paralleling, in a sense, the individual predictors \( X^T \hat{\beta}_{\alpha} \) and \( X^T \hat{\beta}_{\alpha} \).

As Strawderman notes (p. 626) there is no one predictor which will dominate the OLS predictor for all \( P \). For \( P = I \) (i.e. \( G_1 \)) a very simple procedure is to use estimates of \( A \) and \( r \) to select a good predictor for each \( X_1 \). If we are not prepared to specify \( P \) the simulation study suggests that using \( \hat{\beta}_{p-2} \) (i.e. \( c = \sigma^2 (p-2) \) in (4)) regardless of \( X_1 \) is a simple but perhaps adequate choice.
References


Footnotes

1. A possible refinement to using the predictor defined by $\hat{\beta}_S$ with $c = \hat{\sigma}^2(p-2)$ would employ the "limited translation" approach as discussed in Efron and Morris (1972, p. 136). Limiting the amount of shift for each coordinate of $\hat{\beta}_{OLS}$ toward the corresponding coordinate of $\hat{\beta}_S$ using a relevance function, $\rho$, leads to an estimator $\hat{\beta}_\rho$ and resulting predictor $X_0^T\hat{\beta}_\rho$. We also note that the estimator, $\hat{\beta}_S$, resulting from James and Stein (1961, p. 366) sets $c = \hat{\sigma}^2(p-2)(n-p+2)^{-1}(n-p)$. With this $c$ (19) becomes 

$$\sigma^2 x_0^T (x^T x)^{-1} x_0 (1+(n-p+2)^{-1}(n-p)(p^2-4)\Gamma_{1-2(n-p+2)^{-1}(n-2)(p-2)} \Gamma_2).$$

2. We recognize that these simplifying assumptions diminish the utility of the simulation study. In particular, certain estimators which differ in a more general model become equivalent. The study is only intended to be illustrative and suggestive. Certainly a more elaborate one might be undertaken. We also recognize that with these simplification expressions for the exact risks of some of the predictors below (ignoring restrictions) may be obtained using (17).
Figure 1: Risk of $(\gamma^2 + 1)^{-1} \gamma^3$
<table>
<thead>
<tr>
<th>$\beta_p$</th>
<th>Small</th>
<th>Moderate</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 $\hat{\beta}_{p-2}$ (95-99)*</td>
<td>$r &gt; .45, \hat{\beta}_{HK}$</td>
<td>$r &gt; .45, \hat{\beta}_{RM}$</td>
<td>$r &gt; .5, \hat{\beta}_{OLS}$</td>
</tr>
<tr>
<td></td>
<td>(40-60, + in r)</td>
<td>(60-80, + in r)</td>
<td>.35 &lt; $r$ &lt; .5, $\hat{\beta}_{p-2}$</td>
</tr>
<tr>
<td></td>
<td>$r &lt; .45, \hat{\beta}_{RM}$</td>
<td></td>
<td>.25 &lt; $r$ &lt; .35, $\hat{\beta}<em>{p-2}$, $\hat{\beta}</em>{MG}$, $\hat{\beta}_{HK}$</td>
</tr>
<tr>
<td></td>
<td>(60-80, + in r)</td>
<td></td>
<td>$\hat{\beta}_{HK}$</td>
</tr>
<tr>
<td>6 $\hat{\beta}_{p-2}$ (95-99)*</td>
<td>$r &gt; .45, \hat{\beta}_{HK}$</td>
<td>$r &gt; .4, \hat{\beta}_{OLS}$</td>
<td>.25 &lt; $r$ &lt; .4, $\hat{\beta}<em>{p-2}$, $\hat{\beta}</em>{MG}$, $\hat{\beta}_{HK}$</td>
</tr>
<tr>
<td></td>
<td>(40-60, + in r)</td>
<td></td>
<td>$\hat{\beta}_{HK}$</td>
</tr>
<tr>
<td></td>
<td>$r &gt; .45, \hat{\beta}_{RM}$</td>
<td></td>
<td>.1 &lt; $r$ &lt; .25, $\hat{\beta}_{RM}$</td>
</tr>
<tr>
<td></td>
<td>(60-80, + in r)</td>
<td></td>
<td>.01 &lt; $r$ &lt; .1, $\hat{\beta}_{RM}$</td>
</tr>
<tr>
<td></td>
<td>$r &lt; .01, \beta_a, \beta_a$</td>
<td></td>
<td>$\gamma_0, \alpha_0$</td>
</tr>
<tr>
<td>10 $\hat{\beta}_{RM}$ (90-95)*</td>
<td>$r &gt; .45, \hat{\beta}_{HK}$</td>
<td>$r &gt; .4, \hat{\beta}_{OLS}$</td>
<td>.25 &lt; $r$ &lt; .4, $\hat{\beta}<em>{p-2}$, $\hat{\beta}</em>{MG}$, $\hat{\beta}_{HK}$</td>
</tr>
<tr>
<td></td>
<td>(30-60, + in r)</td>
<td></td>
<td>$\hat{\beta}_{HK}$</td>
</tr>
<tr>
<td></td>
<td>$r &lt; .45, \hat{\beta}_{RM}$</td>
<td></td>
<td>.15 &lt; $r$ &lt; .25, $\hat{\beta}_{RM}$</td>
</tr>
<tr>
<td></td>
<td>(60-90, + in r)</td>
<td></td>
<td>.01 &lt; $r$ &lt; .15, $\hat{\beta}_{RM}$</td>
</tr>
<tr>
<td></td>
<td>$r &lt; .01, \beta_a, \beta_a$</td>
<td></td>
<td>$\gamma_0, \alpha_0$</td>
</tr>
</tbody>
</table>

*Figures in parentheses indicate range of expected percent improvement over OLS predictor using "best" predictor.

**Are indistinguishable.
**REPORT DOCUMENTATION PAGE**

1. **REPORT NUMBER**
   - 374

2. **GOVT ACCESSION NO.**
   - (Blank)

3. **RECIPIENT'S CATALOG NUMBER**
   - (Blank)

4. **TITLE (and Subtitle)**
   - On The Use Of Ridge And Stein-Type Estimators In Prediction

5. **TYPE OF REPORT & PERIOD COVERED**
   - TECHNICAL REPORT

6. **PERFORMING ORG. REPORT NUMBER**
   - (Blank)

7. **AUTHOR(s)**
   - Alan E. Gelfand

8. **CONTRACT OR GRANT NUMBER(s)**
   - N00014–86–K–0156

9. **PERFORMING ORGANIZATION NAME AND ADDRESS**
   - Department of Statistics
   - Stanford University
   - Stanford, CA 94305

10. **PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS**
    - NR–042–267

11. **CONTRACT OR GRANT NUMBER(s)**
    - (Blank)

12. **REPORT DATE**
    - May 21, 1986

13. **NUMBER OF PAGES**
    - 24

14. **MONITORING AGENCY NAME & ADDRESS (IF DIFFERENT FROM Controlling Office)**
    - Office of Naval Research
    - Statistics & Probability Program Code 1111

15. **SECURITY CLASS. (OF THIS REPORT)**
    - UNCLASSIFIED

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-type estimators, Stein-type estimators, Regression.

20. **ABSTRACT (Continue on reverse side if necessary and identify by block number)**
    - PLEASE SEE FOLLOWING PAGE.
20. ABSTRACT

For the usual regression model with fixed regressors, there is a considerable literature devoted to alternatives to ordinary least squares estimators of the regression parameters. These alternatives are biased with "small" variances resulting in reduced mean square error over some (perhaps all) of the parameter space. Two prominent classes of such estimators are ridge-type and Stein-type estimators.

Consider the simplest prediction problem in this context, i.e. prediction at a single new vector of prediction values. We calculate the risk (squared error) for predictors based on estimators in the above families. While the ordinary least squares predictor is admissible, a simulation study reveals that over regions of the parameter space substantial reduction in risk is possible using estimators in these families. A simple preliminary procedure based upon the vector of prediction values is given to select a "good" estimator from these families. It is apparent that in multiple prediction a single choice of estimator need not be best.