The Annals of Statistics
1980, Vol. 8, No. 1 64–74

ADAPTIVE MULTIVARIATE RIDGE REGRESSION

BY P. J. BROWN AND J. V. ZIDEK

Imperial College, London and University of British Columbia

A multivariate version of the Hoerl-Kennard ridge regression rule is introduced. The choice from among a large class of possible generalizations is guided by Bayesian considerations; the result is implicitly in the work of Lindley and Smith although not actually derived there. The proposed rule, in a variety of equivalent forms is discussed and the choice of its ridge matrix considered. As well, adaptive multivariate ridge rules and closely related empirical Bayes procedures are presented, these being for the most part formal extensions of certain univariate rules. Included is the Efron-Morris multivariate version of the James-Stein estimator. By means of an appropriate generalization of a result of Morris (see Thisted) the mean square error of these adaptive and empirical Bayes rules are compared.

1. Introduction. Consider a multivariate problem with q responses and n observations, $\mathbf{Y}$, assumed to satisfy the standard multivariate linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1.1}$$

with $\mathbf{X}$ a $(n \times p)$ matrix whose elements are treated as fixed known constants (see Brown and Zidek, 1978, page 5.5 for discussion) and $\boldsymbol{\beta}$ a $(p \times q)$ matrix of unknown coefficients. With $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}^1, \cdots, \boldsymbol{\varepsilon}^q)$, the usual assumptions on the error are

$$E(\boldsymbol{\varepsilon}^j) = 0, \operatorname{Cov}(\boldsymbol{\varepsilon}^j, \boldsymbol{\varepsilon}^l) = \gamma_{jl}\mathbf{I}_n \tag{1.2}$$

$j,l = 1, \cdots, q$ when the least squares estimator of $\boldsymbol{\beta}$ (Rao, 1965, section 8c.2) is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{1.3}$$

This has the additional property of being maximum likelihood when normality of the error distribution is assumed. Writing $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}^1, \cdots, \hat{\boldsymbol{\beta}}^q), \mathbf{Y} = (\mathbf{Y}^1, \cdots, \mathbf{Y}^q)$ so that $\mathbf{Y}^j (n \times 1)$ and $\hat{\boldsymbol{\beta}}^j$ pertain to the jth of the q responses, (1.3) asserts equivalently that

$$\hat{\boldsymbol{\beta}}^j = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}^j. \tag{1.4}$$

This familiar univariate result for the multiple regression of $\mathbf{Y}^j$ on $\mathbf{X}$ is perhaps unsatisfactory, as Sclove (1971) argues, in that (1.4) takes no account of $\boldsymbol{\Gamma} = (\gamma_{jl})$,

Received March 1978; revised September 1978.
AMS 1970 *subject classifications.* Primary 62J05, 62C99; secondary 62H99, 62C10, 62C15.
Key words and phrases. Empirical Bayes, minimax, unbiased risk estimation, Bayesian regression.

the between regressions covariance matrix. We will investigate estimators which utilize information across all q equations in order to estimate $\boldsymbol{\beta}^j$.

The focus for the method of estimation adopted lies in the record of $\hat{\boldsymbol{\beta}}$'s shortcomings; if X^TX is nearly singular $\hat{\boldsymbol{\beta}}$ is unstable, i.e. for even relatively small changes in $k > 0$

$$\hat{\boldsymbol{\beta}}^j(k) = \left(\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p\right)^{-1}\mathbf{X}^T\mathbf{Y}^j \tag{1.5}$$

may vary dramatically. This fact provided the major stimulus for the rapid development of the theory of univariate ridge regression following the papers of Hoerl and Kennard (1970) whose progress is traced and well documented in Thisted (1976). Although conceived in the context of an ill-conditioned design matrix, the ridge regression estimator's Bayesian roots (Lindley and Smith, 1972) suggest that it is useful even when $\mathbf{X}$ is well conditioned (columns of $\mathbf{X}$ near orthogonal).

The estimator given in equation (1.5) with a suitably chosen value of k is adopted by Brown and Payne (1975) in the multivariate case $q > 1$ where it proves effective in "election night forecasting". However, whilst this estimator is free of the second of $\hat{\boldsymbol{\beta}}$'s shortcomings, it retains the first in that it does not take any account of the covariance structure $\boldsymbol{\Gamma}$ between responses. We propose as our candidate for the role of multivariate ridge rule the estimator $\hat{\boldsymbol{\beta}}^*(K)$ where

$$\hat{\boldsymbol{\beta}}^*(K) = \left(\mathbf{X}^T\mathbf{X} \otimes \mathbf{I}_q + \mathbf{I}_p \otimes \mathbf{K}\right)^{-1}\left(\mathbf{X}^T\mathbf{X} \otimes \mathbf{I}_q\right)\hat{\boldsymbol{\beta}} \tag{1.6}$$

where $\mathbf{K}(q \times q) > 0$ is the ridge matrix. Here $\otimes$ denotes the usual Kronecker product and $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}^*(\mathbf{K})$ are $(pq \times 1)$ vectors of estimators of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_p)^T$ where $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_p$ are each $(1 \times q)$ row vectors of $\boldsymbol{\beta}$. Furthermore $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$ corresponding to $\mathbf{K} = 0$.

This choice is justified in Section 2. There a Bayesian model is shown to yield (1.6) as its Bayes estimator. These Bayesian roots are of primary importance for they establish within the broad sphere of model formulation when the rule proposed here is applicable. Furthermore they determine $\mathbf{K}$ and suggest how it might be chosen in practical applications.

By letting $\mathbf{K}$ depend on the data, we obtain the adaptive rules referred to in the title. In Section 4, the mean-square error performances of two such rules are assessed, along with those of two of their empirical Bayes counterparts. This assessment is carried out by means of Condition 3.4 derived in Section 3 as an extension of a result of Thisted (1976) and gives a sufficient condition for a class of estimators to be minimax with respect to various weighted or unweighted quadratic loss functions and a fortiori to dominate the maximum likelihood estimator. These examples are included to show that the domain of applicability of our results encompass special cases of particular interest, namely, natural generalisations of some distinguished univariate rules.

The results in this article are given for $\boldsymbol{\Gamma}$ known. Usually we retain $\boldsymbol{\Gamma}$ in the sequel even though we could, without loss of generality, let $\boldsymbol{\Gamma} = \boldsymbol{I}_q$. This is done to

clarify its role and facilitate its estimation if unknown. Sometimes we will assume $\Gamma = I_q$. This obtains from transforming $\mathbf{Y}$ into $\mathbf{Y}\Gamma^{-\frac{1}{2}}$, noting $\boldsymbol{\beta}$ transforms similarly and therefore a different quadratic loss function applies. Berger et al. (1977) investigate minimax properties of a class of estimators with Γ unknown but their class of estimators does not include our ridge class (taken in conjunction with the loss we shall use). Whilst a Bayes rule ignores the weighting in a quadratic loss, we shall see that dominance of the maximum likelihood estimator depends critically on the particular quadratic loss involved (see also Bunke, 1975 and L.D. Brown, 1975), maximum likelihood being the only rule that retains minimaxity with respect to all quadratic losses.

2. Deriving a class of multivariate rules. As with univariate ridge regression, the canonical form of model (1.1) allows properties to be readily perceived. Accordingly let

$$\mathbf{X} = \mathbf{Q}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{P} \quad , \quad \boldsymbol{\Lambda} = \operatorname{diag}\{\lambda_1, \cdots, \lambda_p\} \quad , \quad \lambda_1 \geqslant \cdots \geqslant \lambda_p \geqslant 0$$

where the $(p \times p)$ orthogonal matrix $\mathbf{P}$ is such that $\mathbf{P}\mathbf{X}^T\mathbf{X}\mathbf{P}^T = \boldsymbol{\Lambda}$ and the $n \times p$ matrix $\mathbf{Q}$ equals $\mathbf{X}\mathbf{P}^T\boldsymbol{\Lambda}^{-\frac{1}{2}}$ so that $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}_p$. Now the model (1.1) may be expressed as

$$\mathbf{Z} = \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}^* \tag{2.1}$$

with $\mathbf{Z} = \mathbf{Q}^T\mathbf{Y}$, $\boldsymbol{\alpha} = \mathbf{P}\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}^* = \mathbf{Q}^T\boldsymbol{\varepsilon}$. Note that $\mathbf{Q}$ just provides a linear reduction from n to p observations within each of the q responses. With Γ known this reduction retains the sufficient statistics for the $(p \times q)$ unknown matrix $\boldsymbol{\beta}$. It results in the loss of a Wishart variable with $(n - p)$ degrees of freedom. Naturally, the Wishart variable's distribution involves Γ and this variable would be used to estimate Γ were it unknown. Writing $\boldsymbol{\varepsilon}^* = (\boldsymbol{\varepsilon}^{*1}, \cdots, \boldsymbol{\varepsilon}^{*q})$ we have that (1.2) transforms to

$$E(\boldsymbol{\varepsilon}^{*j}) = 0 \quad , \quad \operatorname{Cov}(\boldsymbol{\varepsilon}^{*j}, \boldsymbol{\varepsilon}^{*l}) = \gamma_{jl}\mathbf{I}_p; \qquad j,l = 1, \cdots, q \tag{2.2}$$

so that the covariance structure is unchanged. In addition normality is assumed so that amongst other things zero covariances imply independence.

Formally transforming $\mathbf{X}$, $\boldsymbol{\beta}$, $\mathbf{Y}$ as above, the analogous estimator to (1.6) satisfies

$$\hat{\boldsymbol{\alpha}}^*(\mathbf{K}) = (\boldsymbol{\Lambda} \otimes \mathbf{I}_q + \mathbf{I}_p \otimes \mathbf{K})^{-1}(\boldsymbol{\Lambda} \otimes \mathbf{I}_q)\hat{\boldsymbol{\alpha}} \tag{2.3}$$

where $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\alpha}}^*(\mathbf{K})$ are $(pq \times 1)$ vectors of estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_p)^T$ is the $(pq \times 1)$ vector formed by stringing out $\boldsymbol{\alpha}$ row-by-row as a column vector. Generally, we will find it convenient to deal with row vectors (indexed by a suffix rather than a superfix) in what is to follow.

Equation (2.3) is equivalent to

$$\hat{\boldsymbol{\alpha}}_i^*(\mathbf{K}) = \hat{\boldsymbol{\alpha}}_i\lambda_i(\lambda_i\mathbf{I}_q + \mathbf{K})^{-1}, \qquad i = 1, \cdots, p \tag{2.4}$$

or

$$\hat{\boldsymbol{\alpha}}_i^*(\mathbf{K}) = \hat{\boldsymbol{\alpha}}_i\big[I_q - \mathbf{B}_i(\mathbf{K})\big], \qquad i = 1, \cdots, p$$

where $\mathbf{B}_i(\mathbf{K}) = \mathbf{K}(\lambda_i \mathbf{I}_q + \mathbf{K})^{-1}$ and is a matrix shrinkage factor. These shrinkage matrices satisfy the inequalities

$$\mathbf{B}_1 \leqslant \cdots \leqslant \mathbf{B}_p \tag{2.5}$$

where $\mathbf{A} \leqslant \mathbf{B}$ means $\mathbf{B} - \mathbf{A}$ is a nonegative definite matrix.

Many estimators other than $\hat{\boldsymbol{\alpha}}^*(\mathbf{K})$ of (2.3) satisfy (2.5). In selecting (2.3) from amongst many contenders we were guided by two further requirements:

(i) $\hat{\boldsymbol{\alpha}}^*(\mathbf{K})$ should be a Bayes rule for fixed known $\mathbf{K}$.

(ii) For a suitably chosen estimator $\hat{\mathbf{K}}$ of $\mathbf{K}$, in the special case of equal information ($\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$), $\hat{\boldsymbol{\alpha}}^*(\hat{\mathbf{K}})$ should correspond to the Efron and Morris (1972) multivariate extension of the James and Stein (1961) estimator.

Whilst requirements (i) and (ii) in conjunction with condition (2.5) by no means uniquely specify our class of estimators, they do summarize important features of them. Indeed appropriate prior assumptions yielding (2.3) satisfying (i), (ii) and (2.5) can be perceived from Efron and Morris (1972), and Lindley and Smith (1972) as $\boldsymbol{\alpha}_i(l \times q)$ independently distributed such that

$$\mathcal{L}(\boldsymbol{\alpha}_i|\mathbf{m}_\alpha^i, \boldsymbol{\Gamma}_\alpha^i) = N_q(\mathbf{m}_\alpha^i, \boldsymbol{\Gamma}_\alpha^i)$$

and if $\mathbf{Z}(p \times q) = (\mathbf{Z}_1^T, \cdots, \mathbf{Z}_p^T)^T$ so that $\mathbf{Z}_i$ is the ith row of $\mathbf{Z}$, then

$$\mathcal{L}(\mathbf{Z}_i|\boldsymbol{\alpha}) = \mathcal{L}(\mathbf{Z}_i|\boldsymbol{\alpha}_i) = N_{1\times q}(\lambda_i^{\frac{1}{2}}\boldsymbol{\alpha}_i, \boldsymbol{\Gamma}). \tag{2.6}$$

Here $\mathcal{L}(\mathbf{Z}_i|\boldsymbol{\alpha})$ denotes the conditional distribution of $\mathbf{Z}_i$ given $\boldsymbol{\alpha}$. A straightforward calculation shows

$$\mathcal{L}(\mathbf{Z}_i, \boldsymbol{\alpha}_i) = N_{1\times 2q}\left(\left[\lambda_i^{\frac{1}{2}}\mathbf{m}_\alpha^i, \mathbf{m}_\alpha^i\right], \Sigma_{Z,\alpha}^i\right) \tag{2.7}$$

where

$$\Sigma_{Z,\alpha}^i = \begin{bmatrix} \boldsymbol{\Gamma} + \lambda_i\boldsymbol{\Gamma}_\alpha^i & \lambda_i^{\frac{1}{2}}\boldsymbol{\Gamma}_\alpha^i \\ \lambda_i^{\frac{1}{2}}\boldsymbol{\Gamma}_\alpha^i & \boldsymbol{\Gamma}_\alpha^i \end{bmatrix}.$$

Furthermore

$$\mathcal{L}(\mathbf{Z}_i) = N_{1\times q}(\lambda_i^{\frac{1}{2}}\mathbf{m}_\alpha^i, \boldsymbol{\Gamma} + \lambda_i\boldsymbol{\Gamma}_\alpha^i). \tag{2.8}$$

Combining equations (2.7) and (2.8) gives

$$\mathcal{L}(\boldsymbol{\alpha}_i|\mathbf{Z}_i) = N_{1\times q}\big(\mathbf{m}_\alpha^i + (\mathbf{Z}_i - \lambda_i^{\frac{1}{2}}\mathbf{m}_\alpha^i)(\boldsymbol{\Gamma} + \lambda_i\boldsymbol{\Gamma}_\alpha^i)^{-1}\boldsymbol{\Gamma}_\alpha^i\lambda^{\frac{1}{2}}, \boldsymbol{\Gamma}_\alpha^i - \lambda_i\boldsymbol{\Gamma}_\alpha^i(\boldsymbol{\Gamma} + \lambda_i\boldsymbol{\Gamma}_\alpha^i)^{-1}\boldsymbol{\Gamma}_\alpha^i\big). \tag{2.9}$$

In conclusion

$$E(\boldsymbol{\alpha}_i|\mathbf{Z}_i) = \mathbf{m}_\alpha^i + (\hat{\boldsymbol{\alpha}}_i - \mathbf{m}^i)(\mathbf{K}_i + \lambda_i\mathbf{I}_q)^{-1}\lambda_i \tag{2.10}$$

where $\mathbf{K}_i = (\boldsymbol{\Gamma}_\alpha^i)^{-1}\boldsymbol{\Gamma}$ and $\hat{\boldsymbol{\alpha}}_i = \lambda_i^{-\frac{1}{2}}\mathbf{Z}_i$ denotes the least squares estimator of $\boldsymbol{\alpha}_i$.

To obtain equation (1.6) from (2.10) we specialize the above model, taking $\mathbf{K}_i = \mathbf{K}$, $\mathbf{m}_\alpha^i = 0$ for all i and setting

$$\hat{\boldsymbol{\alpha}}_i^*(K) = E(\boldsymbol{\alpha}_i|\mathbf{Z}_i) = \hat{\boldsymbol{\alpha}}_i(\mathbf{K} + \lambda_i\mathbf{I}_q)^{-1}\lambda_i. \tag{2.11}$$

Then with $\hat{\alpha}$ represented by $(\hat{\alpha}_1, \cdots, \hat{\alpha}_p)^T$, and $\hat{\alpha}^*(\mathbf{K})$ correspondingly "stretched out" as a $pq \times 1$ vector,

$$\hat{\alpha}^*(\mathbf{K}) = \operatorname{diag}\left\{\left(\mathbf{I}_q + \lambda_1^{-1}\mathbf{K}\right)^{-1}, \cdots, \left(\mathbf{I}_q + \lambda_p^{-1}\mathbf{K}\right)^{-1}\right\}\hat{\alpha}$$

$$= \operatorname{diag}^{-1}\left\{\mathbf{I}_q + \lambda_1^{-1}\mathbf{K}, \cdots, \mathbf{I}_q + \lambda_p^{-1}\mathbf{K}\right\}\hat{\alpha}.$$

Thus, since in general $(\mathbf{A} \otimes \mathbf{B}) \otimes (\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$

$$\hat{\alpha}^*(\mathbf{K}) = \left(\mathbf{I}_p \otimes \mathbf{I}_q + \Lambda^{-1} \otimes \mathbf{K}\right)^{-1}\hat{\alpha}$$

$$= (\Lambda \otimes \mathbf{I}_q + \mathbf{I}_p \otimes \mathbf{K})^{-1}(\Lambda \otimes \mathbf{I}_q)\hat{\alpha}.$$

Since $(\Lambda \otimes \mathbf{I}_q)\hat{\alpha} = \Lambda^{\frac{1}{2}}\mathbf{Q}^T\mathbf{Y} = \mathbf{P}\mathbf{X}^T\mathbf{Y}, \hat{\boldsymbol{\beta}}^*(\mathbf{K}) = \mathbf{P}^T\hat{\alpha}^*(\mathbf{K})$ and $\mathbf{P}(\Lambda \otimes \mathbf{I}_q + \mathbf{I}_p \otimes \mathbf{K})^{-1}\mathbf{P}^T = (\mathbf{X}^T\mathbf{X} \otimes \mathbf{I}_q + \mathbf{I}_p \otimes \mathbf{K})^{-1}$, equation (1.6) obtains.

REMARK 2.1. *The Brown-Payne ridge rule.* For $\Gamma_\alpha^i = k^{-1}\Gamma$, it follows that $\mathbf{K}_i = \mathbf{K} = k\mathbf{I}_q$ so $(\mathbf{K}_i + \lambda_i\mathbf{I}_q)^{-1}\lambda_i = \lambda_i(\lambda_i + k)^{-1}\mathbf{I}_q$. Under these conditions then the ridge rule of Brown and Payne (1975) is obtained.

REMARK 2.2. *Efron-Morris estimator.* With $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$, i.e., $\lambda_i = 1$ all i, $\Gamma = \mathbf{I}_q$, $\Gamma_\alpha^i = \Gamma_\alpha$, $m_\alpha^i = 0$ all i, equation (2.8) implies

$$\mathcal{L}(\hat{\alpha}_i) = N_{1\times q}(\mathbf{0}, \mathbf{I}_q + \Gamma_\alpha). \tag{2.12}$$

Equation (2.11) is, equivalently

$$\hat{\alpha}_i^* = \hat{\alpha}_i\left[\mathbf{I}_q - \mathbf{B}\right], \tag{2.13}$$

where $\mathbf{B} = (\mathbf{I}_q + \Gamma_\alpha)^{-1}$. The Efron-Morris(1972) estimator is obtained by replacing $\mathbf{B}$ in (2.13) by its natural estimator $\hat{\mathbf{B}}$ given by

$$\hat{\mathbf{B}} = (p - q - 1)(\Sigma_{i=1}^p \hat{\alpha}_i^T\hat{\alpha}_i)^{-1}. \tag{2.14}$$

REMARK 2.3. *Implied prior for* $\boldsymbol{\beta}$. The prior assumptions $\mathbf{m}_\alpha^i = \mathbf{m}_\alpha$, $\Gamma_\alpha^i = \Gamma_\alpha$, $i = 1, \cdots, p$ which lead to estimator (2.10) translate to the same prior assumptions on the original parameter $\boldsymbol{\beta}$. More generally, if $\mathbf{X}^T\mathbf{X}$ has a block diagonal structure, then it is easy to see that provided the priors on coefficients $\boldsymbol{\beta}_i$ for variables i in the same block are identical, then α will have the same prior structure as $\boldsymbol{\beta}$.

REMARK 2.4. *Known dependent variable singularities.* For example, suppose that for each of the n observations $\Sigma_{i=1}^q \mathbf{Y}_i$ were constant, that is, in the notation of Section 1, $\mathbf{Y}\mathbf{e}_1^T = c\mathbf{e}_2^T$ where $\mathbf{Y}$ is $n \times q$, $\mathbf{e}_1$, $\mathbf{e}_2$ are $(1 \times q)$, $(1 \times n)$ vectors of ones and c is the scalar constant. This would be the case if the q dependent variables were proportions. We leave aside here the added question as to whether the linear model is appropriate, assuming it to be, at least over the range of the data. Now, of course, of the q dependent variables one may be removed but for reasons of symmetry, this may not be desirable. If this constraint applies, then it is sufficient that $\mathbf{e}_1\boldsymbol{\beta}^T = (c, \mathbf{0}_{p-1})$ where $\mathbf{0}_{p-1}$ is a $1 \times (p-1)$ vector of zeros and $\operatorname{Var}(\boldsymbol{\varepsilon}\mathbf{e}_1^T) = 0$. The constraint on $\boldsymbol{\beta}$ translates to the same constraint on α (provided each of the

($p - 1$) explanatory variables have been centered). Thus appropriate prior assumptions might be

$$\mathbf{m}_\alpha^i \mathbf{e}_1^T = 0, i = 2, \cdots, p; \ \mathbf{m}_\alpha^1 \mathbf{e}_1^T = c$$

$$\mathbf{e}_1 \boldsymbol{\Gamma}_\alpha^i \mathbf{e}_1^T = \mathbf{e}_1 \boldsymbol{\Gamma} \mathbf{e}_1^T = 0, \qquad i = 1, \cdots, p.$$

These conditions ensure that the estimator (2.10) with generalised inverses, gives rise to predicted q vector responses which satisfy the constraint of this remark.

Finally, if the constraint was thought to hold only approximately then the prior assumptions could be modified accordingly with $\mathbf{e}_1 \boldsymbol{\Gamma}_\alpha^i \mathbf{e}_1^T, \mathbf{e}_1 \boldsymbol{\Gamma} \mathbf{e}_1^T$ small rather than zero.

3. LS estimator dominators. The natural, multivariate extensions in Section 3 of univariate empirical Bayes and adaptive ridge rules have a common form, namely,

$$\hat{\boldsymbol{\alpha}}_i^* = \hat{\boldsymbol{\alpha}}_i \left[\mathbf{I}_q - \hat{\mathbf{B}}_i \right], i = 1, \cdots, p \tag{3.1}$$

where $\hat{\boldsymbol{\alpha}}$ denotes the LS estimators in the canonical case, $\hat{\boldsymbol{\alpha}}^{\mathrm{T}} = (\hat{\boldsymbol{\alpha}}_1^T, \cdots, \hat{\boldsymbol{\alpha}}_p^T)$,

$$\hat{\mathbf{B}}_i = \tau_i v_i w_i^i \left(c_i \mathbf{I}_q + \Sigma_{j=1}^p w_j^i \hat{\boldsymbol{\alpha}}_j^T \hat{\boldsymbol{\alpha}}_j \right)^{-1}, \tag{3.2}$$

$\tau_i > 0, c_i > 0, w_j^i > 0$ are arbitrary scalars and $v_i = \lambda_i^{-1}$ for all i. When $q = 1$, the form proposed by Morris (see Thisted, 1976) results. Conditions are derived below generalizing those of Morris and of Thisted (see Thisted, 1976) to the case $q > 1$, which, if satisfied by $\hat{\boldsymbol{\alpha}}_i^*$, imply for prespecified scalars $L_i \geqslant 0$, all i,

$$E \Sigma L_i (\hat{\boldsymbol{\alpha}}_i^* - \boldsymbol{\alpha}_i)(\hat{\boldsymbol{\alpha}}_i^* - \boldsymbol{\alpha}_i)^T < E \Sigma L_i (\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i)(\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i)^T. \tag{3.3}$$

Note that if $L_i = 1$ all i, the right hand side of inequality (3.3) is $E\mathrm{tr}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})^T = E\mathrm{tr}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T$ since $\hat{\boldsymbol{\alpha}} = \mathbf{P}\hat{\boldsymbol{\beta}}, \boldsymbol{\alpha} = \mathbf{P}\boldsymbol{\beta}$. Another case of particular interest is that in which $L_i = \lambda_i$ for all i as Dempster (1973) points out, this case arises when the mean of the sum of squares of prediction errors at the n design points $E\mathrm{tr}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T$, is used to measure the performance of $\hat{\boldsymbol{\beta}}$. Quadratic prediction loss at m future points has been adopted by Goldstein and Brown (1978) following Brown (1974). Different m point designs lead to different $L_i, i = 1, \cdots, p$, where L_i may be zero (inevitably so when $m < p$). Prediction at the design points, $L_i = \lambda_i$, favours estimators which only slightly shrink the poorly estimated coefficients (small λ_i).

Our derivation of the required sufficient conditions follows that of Thisted (1976) which in turn depends on the method of unbiased risk estimation proposed by Stein (1973). Consider firstly, the case $\boldsymbol{\Gamma} = \mathbf{1}_q$ and the estimator in (3.1) for arbitrary differentiable functions $\{\mathbf{B}_i\}$. Let

$$\mathbf{R}_i(\hat{\boldsymbol{\alpha}}_i^*) = E(\hat{\boldsymbol{\alpha}}_i^* - \boldsymbol{\alpha}_i)(\hat{\boldsymbol{\alpha}}_i^* - \boldsymbol{\alpha}_i)^T, \qquad i = 1, \cdots, p.$$

Then, as is easily shown

$$R_i(\hat{\boldsymbol{\alpha}}_i^*) - R_i(\hat{\boldsymbol{\alpha}}_i) = \lambda_i^{-1}E\left[\mathbf{Z}_i\hat{\mathbf{B}}_i\hat{\mathbf{B}}_i^T\mathbf{Z}_i^T - 2\mathbf{Z}_i\hat{\mathbf{B}}_i\left\{\mathbf{Z}_i^T - \lambda_i^{\frac{1}{2}}\boldsymbol{\alpha}_i^T\right\}\right]. \tag{3.4}$$

Write

$$\mathbf{B}_i = (\hat{\mathbf{B}}_i^1, \cdots, \hat{\mathbf{B}}_i^q),\ \mathbf{Z}_i = (\mathbf{Z}_i^1, \cdots, \mathbf{Z}_i^q), \qquad i = 1, \cdots, p.$$

Then integration by parts yields

$$\Sigma L_i\left[R_i(\hat{\boldsymbol{\alpha}}_i^*) - R_i(\hat{\boldsymbol{\alpha}}_i)\right] = E\Sigma_{i=1}^p\lambda_i^{-1}L_i\left[\mathbf{Z}_i\hat{\boldsymbol{\beta}}_i\hat{\boldsymbol{\beta}}_i^T\mathbf{Z}_i^T - 2\Sigma_{j=1}^q\partial(\mathbf{Z}_i\hat{\mathbf{B}}_i^j)/\partial\mathbf{Z}_i^j\right].$$

Thus if

CONDITION 3.1. $\Sigma\lambda_i^{-1}L_i[\mathbf{Z}_i\hat{\boldsymbol{\beta}}_i\hat{\boldsymbol{\beta}}_i^T\mathbf{Z}_i^T - 2\Sigma_{j=1}^q\partial(\mathbf{Z}_i\hat{\mathbf{B}}_i^j)/\partial\mathbf{Z}_i^j] < 0$ holds with probability one, $\hat{\boldsymbol{\alpha}}^*$ is a superior alternative to $\hat{\boldsymbol{\alpha}}$.

Condition 3.1 is not very useful for picking a specific alternative to $\hat{\boldsymbol{\alpha}}$. Successive and more useful refinements are given below in Conditions 3.2-3.4.

Set $\hat{\mathbf{B}}_i = [\mathbf{C}_i(\mathbf{Z})]^{-1}$; $\partial(\mathbf{Z}_i\hat{\mathbf{B}}_i^j)/\partial\mathbf{Z}_i^j$ may now be computed. The result is

$$\partial(\mathbf{Z}_i\hat{\mathbf{B}}_i)/\partial\mathbf{Z}_i^j = \mathbf{e}_j\mathbf{C}_i^{-1}(\mathbf{Z}) - \mathbf{Z}_i\mathbf{C}_i^{-1}(\partial\mathbf{C}_i/\partial\mathbf{Z}_i^j)\mathbf{C}_i^{-1}$$

where $\partial\mathbf{A}(\mathbf{x})/\partial\mathbf{x}$ in general denotes the elementwise matrix of derivatives w.r.t. the real variable $\mathbf{x}$ and $\mathbf{e}_j$: $1 \times q$ is the vector whose elements are 0 except the jth, it being 1. Thus

$$\begin{aligned}\Sigma_j\partial(\mathbf{Z}_i\hat{\mathbf{B}}_i^j)/\partial\mathbf{Z}_i^j &= \mathrm{tr}\left\{\left[\partial(\mathbf{Z}_i\hat{\mathbf{B}}_i)/\partial\mathbf{Z}_i^1\right]^T, \cdots, \left[\partial(\mathbf{Z}_i\hat{\mathbf{B}}_i)/\partial\mathbf{Z}_i^q\right]^T\right\}^T \\ &= \mathrm{tr}\mathbf{C}_i^{-1} - \mathrm{tr}\mathbf{D}_i\end{aligned}$$

where

$$\mathbf{D}_i = \mathbf{D}_i(\mathbf{Z}) = \left\{\left[\mathbf{Z}_i\mathbf{C}_i^{-1}(\partial\mathbf{C}_i/\partial\mathbf{Z}_i^1)\mathbf{C}_i^{-1}\right]^T, \cdots, \left[\mathbf{Z}_i\mathbf{C}_i^{-1}(\partial\mathbf{C}_i/\partial\mathbf{Z}_i^q)\mathbf{C}_i^{-1}\right]^T\right\}. \tag{3.5}$$

These calculations may be summarized as follows. If

$$\hat{\boldsymbol{\alpha}}_i^* = \hat{\boldsymbol{\alpha}}_i\left[\mathbf{I}_q - \mathbf{C}_i^{-1}(\mathbf{Z})\right],\ \mathbf{C}_i > 0, \tag{3.6}$$

and

CONDITION 3.2. $\Sigma_{i=1}^p\lambda_i^{-1}L_i[\mathbf{Z}_i\mathbf{C}_i^{-2}\mathbf{Z}_i^T - 2\mathrm{tr}\mathbf{C}_i^{-1} + 2\mathrm{tr}\mathbf{D}_i] < 0$ holds with probability one when D_i is given in equation (3.5), then $\hat{\boldsymbol{\alpha}}^*$ dominates $\hat{\boldsymbol{\alpha}}$.

Finally, suppose $\mathbf{C}_i(\mathbf{Z}) = (\tau_i w_i^i v_i)^{-1}[c_i\mathbf{I}_q + \Sigma_{j=1}^p w_j^i\hat{\boldsymbol{\alpha}}_j^T\boldsymbol{\alpha}_j]$, τ_i, c_i, $w_j^i > 0$ being arbitrary constants and $v_i = \lambda_i^{-1}$. Then $\partial\mathbf{C}_i/\partial\mathbf{Z}_i^j = (\tau_i w_i^i v_i)^{-1}\Sigma_k w_k^i v_k\partial(\mathbf{Z}_k^T\mathbf{Z}_k)/\partial\mathbf{Z}_i^j = \tau_i^{-1}[\mathbf{e}_j^T\mathbf{Z}_i + \mathbf{Z}_i^T\mathbf{e}_j]$. It follows that $\mathrm{tr}\mathbf{D}_i = \tau_i^{-1}[\mathrm{tr}\mathbf{A}_i + \mathrm{tr}\mathbf{B}_i]$ where the jth rows of $\mathbf{A}_i$ and $\mathbf{B}_i$ are, respectively, $\mathbf{Z}_i\mathbf{C}_i^{-1}\mathbf{e}_j^T\mathbf{Z}_i\mathbf{C}_i^{-1}$ and $\mathbf{Z}_i\mathbf{C}_i^{-1}\mathbf{Z}_i^T\mathbf{e}_j\mathbf{C}_i^{-1}$. Since $\mathbf{Z}_i\mathbf{C}_i^{-1}\mathbf{e}_j^T$ and $\mathbf{Z}_i\mathbf{C}_i^{-1}\mathbf{Z}_i^T$ are scalars, $\mathrm{tr}\mathbf{A}_i = (\mathbf{Z}_i\mathbf{C}_i^{-1}\mathbf{e}_1^T, \cdots, \mathbf{Z}_i\mathbf{C}_i^{-1}\mathbf{e}_q^T)[\mathbf{Z}_i\mathbf{C}_i^{-1}] = \mathbf{Z}_i\mathbf{C}_i^{-2}\mathbf{Z}_i^T$ and $\mathrm{tr}\mathbf{B}_i = \mathbf{Z}_i\mathbf{C}_i^{-1}\mathbf{Z}_i^{-1}\mathrm{tr}\mathbf{C}_i^{-1}$. Thus

$$\mathrm{tr}\mathbf{D}_i = \tau_i^{-1}\left[\mathbf{Z}_i\mathbf{C}_i^{-2}\mathbf{Z}_i^T + (\mathrm{tr}\mathbf{C}_i^{-1})\mathbf{Z}_i\mathbf{C}_i^{-1}\mathbf{Z}_i^T\right]. \tag{3.7}$$

From Condition 3.2 we deduce for this choice of $\mathbf{C}_i$ that $\hat{\boldsymbol{\alpha}}^*$dominates $\hat{\boldsymbol{\alpha}}$ if

CONDITION 3.3. $\Sigma_{i=1}^p L_i v_i[\mathbf{Z}_i\mathbf{C}_i^{-2}\mathbf{Z}_i^T\{1 + 2\tau_i^{-1}\} - 2\mathrm{tr}\mathbf{C}_i^{-1}\{1 - \tau_i^{-1}\mathbf{Z}_i\mathbf{C}_i^{-1}\mathbf{Z}_i^T\}] < 0$, holds.

To complete the analysis we give a generalization for $q > 1$ of an important special case ($r \equiv 1$) of Thisted's Corollary 5.9 (1976, page 55). Set

$$w_j^i = w_j,\ Q_i = L_i v_i^2 w_i,$$

and let $\mathbf{F} = \operatorname{diag}\{F_1, \cdots, F_q\}$ where $F_1 > F_2 > \cdots > F_q$ denote the eigenvalues of $\Sigma w_j \hat{\boldsymbol{\alpha}}_j^T \hat{\boldsymbol{\alpha}}_j$. The orthogonal matrix, $O(q \times q)$ is chosen to satisfy $OFO' = \Sigma_{j=1}^p w_j \hat{\boldsymbol{\alpha}}_j^T \hat{\boldsymbol{\alpha}}_j$ so that $\mathbf{F} = \Sigma_{j=1}^p w_j \mathbf{U}_j^T \mathbf{U}_j$ where $\mathbf{U}_j = \hat{\boldsymbol{\alpha}}_j \mathbf{O}$. Condition 3.3 may then alternatively be written

(3.8)

$$O > \Sigma_{i=1}^p Q_i \tau_i \Big[w_i \big\{ \mathbf{U}_i (c_i \mathbf{I}_q + \mathbf{F})^{-2} \mathbf{U}_i^T (\tau_i + 2) + 2\operatorname{tr}(c_i \mathbf{I}_q + \mathbf{F})^{-1} \cdot \mathbf{U}_i (c_i \mathbf{I}_q + \mathbf{F})^{-1} \mathbf{U}_i^T \big\} - 2\operatorname{tr}\big[c_i \mathbf{I}_q + \mathbf{F} \big]^{-1} \Big].$$

The right hand side of this last inequality is, with H_i^{jj} the jth diagonal element of $\mathbf{H}_i := (c_i \mathbf{I}_q + \mathbf{F})^{-1} \cdot w_i \mathbf{U}_i^T \mathbf{U}_i$,

$$\begin{aligned}
&\Big[\operatorname{tr} \Sigma_i Q_i \tau_i \big[(\tau_i + 2)(c_i \mathbf{I}_q + \mathbf{F})^{-1} + 2\Sigma_m (c_i + F_m)^{-1} \mathbf{I}_q \big] \mathbf{H}_i \Big] - 2\Sigma_j \Sigma_i Q_i \tau_i (c_i + F_j)^{-1} \\
&\quad = \Sigma_j \Big\{ \Sigma_i Q_i \tau_i \big[(c_i + F_j)^{-1} (\tau_i + 4) + 2\Sigma_{m \neq j} (c_i + F_m)^{-1} \big] H_i^{jj} \Big\} \\
&\qquad - 2\Sigma_j \Sigma_i Q_i \tau_i (c_i + F_j)^{-1} \\
&\quad = \Sigma_j \Big\{ \Sigma_i (c_i + F_j)^{-1} \big[Q_i \tau_i (\tau_i + 4) H_i^{jj} + 2 Q_i \tau_i \Sigma_{m \neq j} H_i^{mm} \big] \\
&\qquad - 2\Sigma_i Q_i \tau_i (c_i + F_j)^{-1} \Big\} \\
&\quad < \Sigma \Big\{ \max_i Q_i \tau_i (\tau_i + 4)(c_i + F_j)^{-1} + 2(q - 1) \max_i Q_i \tau_i (c_i + F_j)^{-1} \\
&\qquad - 2\Sigma_i Q_i \tau_i (c_i + F_j)^{-1} \Big\}
\end{aligned}$$

since $\mathbf{I}_q - \Sigma_i \mathbf{H}_i > 0$. It follows that $\hat{\boldsymbol{\alpha}}^*$ dominates $\hat{\boldsymbol{\alpha}}$ if

CONDITION 3.4. $\max_i Q_i \tau_i (\tau_i + 4)(c_i + t)^{-1} + 2(q - 1)\max_i Q_i \tau_i (c_i + t)^{-1} < 2\Sigma Q_i \tau_i (c_i + t)^{-1}$, $t \geqslant 0$ holds.

This last condition reduces to Thisted's (1976, page 55) when $q = 1$ and his function r, satisfies $r(t) \equiv t$.

REMARK. When $\boldsymbol{\Gamma} \neq \mathbf{I}_q$, $\hat{\boldsymbol{\alpha}}$, etc. are computed with $\mathbf{Y}$ replaced by $\mathbf{Y}\boldsymbol{\Gamma}^{-\frac{1}{2}}$.

4. Adaptive multivariate rules. A number of existing empirical Bayes [cf. Subsections 4.1, 4.2] and adaptive ridge [cf. Subsections 4.3, 4.4] rules for $q = 1$ have formal extensions to the case $q > 1$. In this section these rules are presented along with sufficient conditions derived from Condition 3.4 for their dominance of the *LS* estimator. As the works of Stein (1960), Sclove (1971), Efron and Morris (1972), Baranchik (1973), and Zidek (1976) would suggest, dominance over the *LS* estimator is increasingly harder to achieve as $q \to p$. This is reflected in Condition 3.4 by the increasing size of the second term, $2(q - 1)\max_{1 < i < p} Q_i \tau_i (c_i + t)^{-1}$ so

that as $q \to p$ the inequality of that condition is increasingly difficult to satisfy. Finding estimators which dominate the *LS* estimators is also made increasingly more difficult as the breadth of the spectrum of X^TX increases (Thisted, 1976). Work relevant to these considerations is found in Berger (1978).

Many regression procedures have been proposed; we will confine our considerations to the following notable examples, estimators suggested by the works of (i) Berger (1975) and Hudson (1974) (see Thisted, 1976, page 75), (ii) Efron and Morris (1972) and James and Stein (1961), (iii) Hoerl, Kennard and Baldwin (1975) and (iv) Sclove (1973). For some other methods of estimating K see Brown and Zidek (1978).

In the following examples $L_i = 1, i = 1, \cdots, p$ and $p - q - 1 > 0$ are assumed.

4.1. *Hudson-Berger rule.* Here

$$\hat{\boldsymbol{\alpha}}_i^{HB} = \hat{\boldsymbol{\alpha}}_i\left[\mathbf{I}_q - v_i^{-1}(p - q - 1)\left(\Sigma v_j^{-2}\hat{\boldsymbol{\alpha}}_j^T\hat{\boldsymbol{\alpha}}_j\right)^{-1}\right]$$

is suggested so, comparing this with equation (3.6) and the work immediately following,

$$c_i = 0,\ w_i = v_i^{-2},\ \tau_i = (p - q - 1),\ Q_i = 1 \text{ all } i.$$

Condition 3.4 becomes in this case,

$$(p - q - 1)(p - q + 3) + 2(q - 1)(p - q - 1) < 2p(p - q - 1)$$
$$\text{i.e. } p - q - 1 > 0.$$

CONCLUSION. $\hat{\alpha}^{HB}$ dominates $\hat{\alpha}$ provided $p - q - 1 > 0$.

4.2. *Modified James-Stein, Efron-Morris rule.* In this case

$$\hat{\boldsymbol{\alpha}}_i^M = \hat{\boldsymbol{\alpha}}_i\left[I_q - \left(p\bar{v} - (q + 1)v_p\right)^+\left(\Sigma\hat{\boldsymbol{\alpha}}_j^T\hat{\boldsymbol{\alpha}}_j\right)^{-1}\right] \tag{4.2.1}$$

so

$$c_i = 0,\ w_i = 1,\ \tau_i = \left(p\bar{v} - (q + 1)v_p\right)^+ v_i^{-1},\ Q = v_i^2 \quad \text{all} \quad i,$$

where $(x)^+ = \max(0,x)$ all x. This obtains from Section 2.2 by replacing $p - q - 1$ by $p\bar{v} - (q + 1)v_p$ when the latter is positive. It is straightforward to show that this results in an improved estimator and that Condition 3.4 is automatically satisfied. This estimator, $\hat{\boldsymbol{\alpha}}^M$, dominates $\hat{\boldsymbol{\alpha}}$ if $v_i \equiv 1$ and $p > q + 1$ thus proving anew the main results of James and Stein (1961) and Efron and Morris (1972). The positivity of $p\bar{v} - (q + 1)v_p$ implies this dominance of $\hat{\boldsymbol{\alpha}}$ will persist until the spectrum of $\mathbf{X}^T\mathbf{X}$ becomes too widely dispersed.

4.3. *Modified Hoerl-Kennard-Baldwin rule.*

$$\hat{\boldsymbol{\alpha}}_i^{MHKB} = \hat{\boldsymbol{\alpha}}_i\left[\mathbf{I}_q - (p - q - 1)v_i\left((p - q - 1)v_i\mathbf{I}_q + \Sigma_j\hat{\boldsymbol{\alpha}}_j^T\hat{\boldsymbol{\alpha}}_j\right)^{-1}\right]. \tag{4.3.1}$$

Thus

$$c_i = (p - q - 1)v_i,\ w_i = 1,\ \tau_i = (p - q - 1),\ Q_i = v_i^2 \quad \text{for all} \quad i.$$

Now Condition 3.4 is

$$v_p^2(p+q+1)\big([p-q-1]v_p+t\big)^{-1} < 2\Sigma v_j^2\big([p-q-1]v_j+t\big)^{-1}$$

which is implied by

$$v_p^2(p+q+1) < 2p\,\overline{v^2} \tag{4.3.2}$$

where $\overline{v^2} = \Sigma v_j^2/p$. Thus if (4.3.2) holds $\hat{\boldsymbol{\alpha}}^{MHKB}$ dominates $\hat{\boldsymbol{\alpha}}$; in particular if $v_i = 1$ for all i, this dominance is assured by $p > q+1$.

4.4. *Sclove's rule.*

$$\hat{\boldsymbol{\alpha}}_i^s = \hat{\boldsymbol{\alpha}}_i\Big[\mathbf{I}_q - v_i(p-q-1)\bar{v}^{-1}\big\{v_i(p-q-1)\bar{v}^{-1}\mathbf{I}_q + \Sigma_{j=1}^{p} v_j^{-1}\hat{\boldsymbol{\alpha}}_j^T\hat{\boldsymbol{\alpha}}_j\big\}^{-1}\Big],$$

where $\bar{v}^{-1} = \Sigma v^{-1}/p$.

Thus

$$c_i = (p-q-1)\bar{v}^{-1}v_i,\ w_i = v_j^{-1},\ \tau_i = (p-q-1)\,\overline{v^{-1}}\,v_i,\ Q_i = v_i \quad \text{for all} \quad i.$$

Now Condition 3.4 is implied by

$$(p-q-1)\,\overline{v^{-1}}\,v_p^3 - 2\Big[p\,\overline{v^2} - (q+1)v_p^2\Big] < 0 \tag{4.4.1}$$

where $\overline{v^2} = \Sigma v_j^2/p$. Thus if (4.4.1) holds $\hat{\boldsymbol{\alpha}}^s$ dominates $\hat{\boldsymbol{\alpha}}$. Indeed (4.4.1) indicates that this will be so provided the eigenvalue spectrum of $\mathbf{X}^T\mathbf{X}$ is not too broad.

REFERENCES

[1] BARANCHIK, A. J. (1973). Inadmissibility of maximum likelihood estimators in some multiple regression problems with three or more independent variables. *Ann. Statist.* **1** 312-321.

[2] BERGER, J. (1975). Minimax estimation of location vectors for a wide class of densities. *Ann. Statist.* **3** 1318-1328.

[3] BERGER, J. (1977). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. Tech. Rep. #480, Statistics Department, Purdue University.

[4] BERGER, J., BOCK, M. E., BROWN, L. D., CASELLA, G. and GLESER, L. (1977). Minimax estimation of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix. *Ann. Statist.* **4** 763-771.

[5] BROWN, L. D. (1975). Estimation with incompletely specified loss functions (the case of several location parameters). *J. Amer. Statist. Assoc.* **70** 417-427.

[6] BROWN, P. J. (1974). Predicting by ridge regression. Unpublished report, Imperial College London.

[7] BROWN, P. J. and PAYNE, C. (1975). Election night forecasting (with discussion). *J. Roy. Statist. Soc. Ser. A* **138** 463-498.

[8] BROWN, P. J. and ZIDEK, J. V. (1978). Multivariate ridge regression. Inst. of Applied Math. and Statistics, Tech. Rep. 78-8, Univ. of British Columbia.

[9] BUNKE, O. (1975). Improved inference in linear models with additional information. *Math. Operationsforsch. Statist.* **6** 817-829.

[10] DEMPSTER, A. P. (1973). Alternatives to least squares in multiple regression. In *Multivariate Statistical Analysis.* (D. Kabe and R. P. Gupta, eds.) 25-40. Amsterdam: North-Holland Publishing Co.

[11] DEMPSTER, A. P., SCHATZOFF, M. and WERMUTH, M. (1977). A simulation study of alternatives to ordinary least squares. *J. Amer. Statist. Soc.* **72** 77-106.

[12] EFRON, B. and MORRIS, C. (1972). Empirical Bayes on vector observations: an extension of Stein's method. *Biometrika* **59** 335-347.

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

[14] GOLDSTEIN, M. and BROWN, P. J. (1978). Prediction with shrinkage estimators. *Math. Operationsforsch. Statist.* **9** (1) 3-7.

[15] HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12** 55-67.

[16] HOERL, A. E., KENNARD, R. W., and BALDWIN, K. F. (1975). Ridge regression: some simulations. *Comm. Statist.* **4** 105-123.

[17] HUDSON, M. (1974). Empirical Bayes estimation. Ph.D. thesis, Dept. of Statistics, Stanford Univ., Tech. Rep. #58.

[18] JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berk Symp. Math. Statist. Probability* **1** 361-379. Univ. of California Press.

[19] LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc., Ser. B.* **34** 1-41.

[20] RAO, C. R. (1975). *Linear Statistical Inference and its Applications.* New York: John Wiley and Sons.

[21] SCLOVE, S. L. (1971). Improved estimation of parameters in multivariate regression. *Sankhya, Ser. A* **33** 61-66.

[22] SCLOVE, S. L. (1973). Least squares problems with random coefficients. Tech. Rep. 72, Dept. of Statistics, Stanford University.

[23] STEIN, C. (1960). Multiple regression. In *Contributions to Probability and Statistics: Essay in Honor of Harold Hotelling.* (Ingram Olkon, Sudhish G. Ghurye, Wassily Hoeffding, William G. Madow and Henry B. Mann, eds) 424-443. Stanford Univ. Press.

[24] STEIN, C. (1973). Estimation of the mean of a multivariate normal distribution. *Proc. Prague Symp. Asymptotic Statist.* **2** 345-381.

[25] THISTED, R. A. (1976). Ridge regression, minimax estimation, and empirical Bayes methods. Tech. Rep. 28, Statistics Department, Stanford University.

[26] ZIDEK, J. V. (1978). Deriving unbiased risk estimators of multinormal mean and regression coefficient estimators using zonal polynomials. *Ann. Statist.* **6** 769-782.

DEPARTMENT OF MATHEMATICS
IMPERIAL COLLEGE OF SCIENCE AND TECHNOLOGY
HUXLEY BUILDING,
QUEENS GATE,
LONDON, U.K. SW72BZ

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF BRITISH COLUMBIA
2075 WESBROOK MALL
VANCOUVER, B.C.
CANADA V6T 1W5