CIRJE-F-170

Minimax Empirical Bayes Ridge-Principal Component Regression Estimators

Tatsuya Kubokawa
University of Tokyo

M. S. Srivastava
University of Toronto

September 2002

Minimax Empirical Bayes Ridge-Principal Component Regression Estimators

T. Kubokawa and M.S. Srivastava
University of Tokyo and *University of Toronto*

In this paper, we consider the problem of estimating the regression parameters in a multiple linear regression model with design matrix $\boldsymbol{A}$ when the multicollinearity is present. Minimax empirical Bayes estimators are proposed under the assumption of normality and loss function $(\boldsymbol{\delta}-\boldsymbol{\beta})^t(\boldsymbol{A}^t\boldsymbol{A})^2(\boldsymbol{\delta}-\boldsymbol{\beta})/\sigma^2$, where $\boldsymbol{\delta}$ is an estimator of the vector $\boldsymbol{\beta}$ of p regression parameters, and σ^2 is the unknown variance of the model. The minimax estimators are also obtained under linear constraints on $\boldsymbol{\beta}$ such as $\boldsymbol{\beta} = \boldsymbol{C}\boldsymbol{\alpha}$ for some $p \times q$ known matrix $\boldsymbol{C}$, $q \leq p$. For a particular $\boldsymbol{C}$, this combines the principal component regression and ridge regression. These results are also applicable for estimating the p means θ_i when the p observations x_i are independently distributed as $\mathcal{N}(\theta_i, d_i\sigma^2)$, d_i's are known but σ^2 is unknown.

Key words and phrases: Multiple regression, multicolliearity, ridge regression, empirical Bayes method, principal component method, hednic regression, minimaxity.

AMS subject classifications: Primary 62J05, 62J07, Secondary 62F10, 62C12, 62C20.

1 Introduction

The primary purpose of regression models is prediction with the help of many independent variables called predictors. However, when there are many independent variables, it is very likely that some of them may be highly correlated among themselves leading to the phenomenon of near multicollinearity. To avoid multicollinearity, fewer independent variables are selected by various methods avalable in the literature. As an alternative, Hoerl and Kennard (1970) proposed the so-called ridge regression method which is unaffected by the multicollinearity among the many independent variables. Although ridge regression has been studied extensively in the literature, certain aspects remained unresolved. To focus on this aspect, we consider the regression model

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\boldsymbol{\epsilon}$ has normal distribution $\mathcal{N}_N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_N)$ with unknown disturbance σ^2, $\boldsymbol{\beta}$ is a p-vector of unknown parameters and $\boldsymbol{A}$ is an $N \times p$ design matrix of rank p. When the design matrix $\boldsymbol{A}$ is a matrix of observations on p independent variables, some of these variables may be highly correlated. Thus, the matrix $\boldsymbol{A}^t\boldsymbol{A}$ may have some very small eigenvalues. Consequently, the least squares estimator $\widehat{\boldsymbol{\beta}} = (\boldsymbol{A}^t\boldsymbol{A})^{-1}\boldsymbol{A}^t\boldsymbol{y}$ whose covariance matrix is given by $\mathbf{Cov}\,(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{A}^t\boldsymbol{A})^{-1}$ is not a suitable estimator since some components of $\widehat{\boldsymbol{\beta}}$ or some linear combinations of $\widehat{\boldsymbol{\beta}}$ may have a very large variance; in particular $E[\widehat{\boldsymbol{\beta}}^t\widehat{\boldsymbol{\beta}}]$ would be very large. This led Hoerl and Kennard (1970) to propose that the residual sum of square should be minimized (with respect to $\boldsymbol{\beta}$) subject to the constraint that $\boldsymbol{\beta}^t\boldsymbol{\beta} \leq M$ for positive constant M. Thus, using Langrange's multiplier, it amounts to minimizing $(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\beta})^t(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\beta}) + k(\boldsymbol{\beta}^t\boldsymbol{\beta} - M)$. This is minimized by

$$\widehat{\boldsymbol{\beta}}^R(\lambda) = [\boldsymbol{A}^t\boldsymbol{A} + k\boldsymbol{I}]^{-1}\boldsymbol{A}^t\boldsymbol{y} = \widehat{\boldsymbol{\beta}} - [\boldsymbol{I} + \lambda\boldsymbol{A}^t\boldsymbol{A}]^{-1}\widehat{\boldsymbol{\beta}} \tag{1.2}$$

for $\lambda = 1/k$, $k > 0$, and is called a *ridge regression estimator* of $\boldsymbol{\beta}$.

The ridge regression estimator can also be shown to be a Bayes estimator corresponding to the prior distribution of $\boldsymbol{\beta}$ as $\mathcal{N}_p(\mathbf{0}, \lambda\sigma^2\boldsymbol{I}_p)$. However, no matter which interpretation is taken, a suitable choice for the value of λ has been the subject of many studies over the last three decades. Corresponding to the above prior distribution, the marginal distribution of $\widehat{\boldsymbol{\beta}}$ can be shown to be $\mathcal{N}_p(\mathbf{0}, \sigma^2\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda\boldsymbol{I}_p\})$, and since σ^2 can be estimated by aS for some a and $S = (\boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{\beta}})^t(\boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{\beta}})$, Efron and Morris (1975) proposed an estimator of λ by solving the likelihood equation of the marginal distribution of $\widehat{\boldsymbol{\beta}}$. The ridge regression estimator with this estimated value of λ is called an empirical Bayes or adaptive ridge regression estimator of $\boldsymbol{\beta}$. However, the question remained as to the best or most appropriate choice of an estimator of λ no matter how it is obtained. Strawderman (1978) and Casella (1980) gave a class of minimax estimators under a very general quadratic loss function

$$L(\omega, \boldsymbol{\delta}, \boldsymbol{Q}) = (\boldsymbol{\delta} - \boldsymbol{\beta})^t\boldsymbol{Q}(\boldsymbol{\delta} - \boldsymbol{\beta})/\sigma^2 \tag{1.3}$$

where $\boldsymbol{\delta}$ is an estimator of $\boldsymbol{\beta}$, $\boldsymbol{Q}$ is a known $p \times p$ positive definite matrix and $\omega = (\boldsymbol{\beta}, \sigma^2)$. These minimax estimators are, however, not applicable to the multicollinearity case as the conditions imposed for minimaxity are not satisfied here except in the case when $\boldsymbol{Q} = (\boldsymbol{A}^t\boldsymbol{A})^2$, considered by Strawderman (1978). When $\boldsymbol{Q} = (\boldsymbol{A}^t\boldsymbol{A})^2$ in (1.3), we shall call it Strawderman's loss function. A miniamx estimator of λ under Strawderman's loss function is given by

$$\hat{\lambda}_{AD} = (n+2)d_1\widehat{\boldsymbol{\beta}}^t\boldsymbol{A}^t\boldsymbol{A}\widehat{\boldsymbol{\beta}} + \lambda_0$$

where $d_1 \geq \cdots \geq d_p$ are the ordered eigenvalues of $(\boldsymbol{A}^t\boldsymbol{A})^{-1}$ and λ_0 is the solution of

$$\sum_{i=1}^{p}(d_i - d_p)/(d_i + \lambda_0) = (p-2)/2.$$

Our numerical study shows that $\hat{\lambda}_{AD}$ or a truncated version of it considered in this paper are not good choices. Thus, we consider a modified version of the choice made by Fay and Herriot (1979) and Shinozaki and Chang (1993) who obtained an estimator of λ by solving the equation

$$\widehat{\boldsymbol{\beta}}\left[(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda\boldsymbol{I}\right]^{-1}\widehat{\boldsymbol{\beta}} = (p-2)S/(n+2). \tag{1.4}$$

Using the implicit function theorem, Shinozaki and Chang (1993) showed that such an adaptive ridge regression estimator is minimax under the loss function

$$L(\omega, \boldsymbol{\delta}, \boldsymbol{I}) = (\boldsymbol{\delta} - \boldsymbol{\beta})^t(\boldsymbol{\delta} - \boldsymbol{\beta}) \tag{1.5}$$

provided

$$\sum_{i=1}^{p} d_i^2/d_1^2 - 2 \geq (p-2)/2. \tag{1.6}$$

However, in the case of multicollinearity d_1 would be very large and the condition (1.6) would rarely be satisfied. These results were later extended by Shinozaki and Chang (1996) to the situation when it is suspected that the hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{C\alpha}$ may be true. They proposed minimax adaptive ridge regression estimators shrunken toward the hypothesis under the quadratic loss function (1.5). Again the minimaxity conditions are not satisfied in the multicollinearity case.

The above idea of Shinozaki and Chang (1996) is an interesting one and has been used in the past in Stein estimators, see Lindley (1962). In the multicollinearity case it makes sense to consider the case of suspected hypothesis. For if $\boldsymbol{H}$ is an orthogonal matrix such that $\boldsymbol{H}(\boldsymbol{A}^t\boldsymbol{A})^{-1}\boldsymbol{H}^t = \boldsymbol{D}$ and $\boldsymbol{H}\boldsymbol{H}^t = \boldsymbol{I}$, where $\boldsymbol{D} = \mathrm{diag}\,(d_1, \ldots, d_p)$ and $d_1 \geq \cdots \geq d_p$, we may write with $\boldsymbol{H}^t = (\boldsymbol{H}_1^t, \boldsymbol{H}_2^t)$,

$$\begin{aligned} \boldsymbol{\beta} &= \boldsymbol{H}^t\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{H}_1^t\boldsymbol{H}_1\boldsymbol{\beta} + \boldsymbol{H}_2^t\boldsymbol{H}_2\boldsymbol{\beta} \\ &= \boldsymbol{H}_1^t\boldsymbol{\gamma}_1 + \boldsymbol{H}_2^t\boldsymbol{\gamma}_2 \end{aligned} \tag{1.7}$$

where $\boldsymbol{\gamma}_1$ corresponds to the smaller eigenvalues of $\boldsymbol{A}^t\boldsymbol{A}$ and should not be included in the model. Thus, it would be desirable to include the constraint that $\boldsymbol{\beta} = \boldsymbol{H}_2^t\boldsymbol{\gamma}_2$ or more generally that $\boldsymbol{\beta} = \boldsymbol{C\alpha}$ for some known $p \times q$ matrix $\boldsymbol{C}$. It may be noted that from (1.7), we get $\boldsymbol{\beta} = \boldsymbol{H}^t\boldsymbol{\gamma}$ for $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^t, \boldsymbol{\gamma}_2^t)^t$, and hence the model (1.1) becomes

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{H}^t\boldsymbol{\gamma} + \boldsymbol{\epsilon} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

where $\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{H}^t$ and $\boldsymbol{Z}^t\boldsymbol{Z} = \boldsymbol{H}\boldsymbol{A}^t\boldsymbol{A}\boldsymbol{H}^t = \boldsymbol{D}^{-1}$. Thus, dropping $\boldsymbol{\gamma}_1$ from this model is equivalent to doing principal component regression, see Sen and Srivastava (1990, p 253-255). Clearly, then the approach here is a combination of principal component regression and ridge regression.

Although the adaptive ridge regression estimators that estimates λ from (1.4) cannot be theoretically justified under the squared loss function $L(\omega, \boldsymbol{\delta}, \boldsymbol{I})$ or predictor error loss function $L(\omega, \boldsymbol{\delta}, \boldsymbol{A}^t\boldsymbol{A})$, it can be shown to be minimax under the Strawderman's loss function

$$L(\omega, \boldsymbol{\delta}, (\boldsymbol{A}^t\boldsymbol{A})^2) = (\boldsymbol{\delta} - \boldsymbol{\beta})^t(\boldsymbol{A}^t\boldsymbol{A})^2(\boldsymbol{\delta} - \boldsymbol{\beta})/\sigma^2. \tag{1.8}$$

On the other hand, the Monte Carlo simulations given in Section 4 show that these estimators perform much better for $L(\omega, \boldsymbol{\delta}, \boldsymbol{I})$ and $L(\omega, \boldsymbol{\delta}, \boldsymbol{A}^t\boldsymbol{A})$ loss functions than for $L(\omega, \boldsymbol{\delta}, (\boldsymbol{A}^t\boldsymbol{A})^2)$ for which it has been shown to be minimax.

In this paper, we also treat the empirical Bayes estimator shrunken toward the subspace $\{\boldsymbol{C\alpha} \,|\, \boldsymbol{\alpha} \in \boldsymbol{R}^q\}$, given by

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^B(\hat{\lambda}, \widehat{\boldsymbol{\alpha}}) &= \left(\boldsymbol{A}^t\boldsymbol{A} + \hat{\lambda}^{-1}\boldsymbol{I}\right)^{-1}\boldsymbol{A}^t\boldsymbol{A}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\widehat{\boldsymbol{\alpha}}\right) + \boldsymbol{C}\widehat{\boldsymbol{\alpha}} \\ &= \widehat{\boldsymbol{\beta}} - \left(\boldsymbol{I} + \hat{\lambda}(\boldsymbol{A}^t\boldsymbol{A})^{-1}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\widehat{\boldsymbol{\alpha}}\right) \end{aligned} \tag{1.9}$$

for

$$\widehat{\boldsymbol{\alpha}} = (\boldsymbol{C}^t\boldsymbol{A}^t\boldsymbol{A}\boldsymbol{C})^{-1}\boldsymbol{C}^t\boldsymbol{A}^t\boldsymbol{A}\widehat{\boldsymbol{\beta}}.$$

In Section 2, we propose an empirical Bayes estimator of $\boldsymbol{\beta}$. In Section 3, we show that this estimator is minimax under Strawderman's loss function. We also propose in Section 2 a hierarchical empirical Bayes estimator of $\boldsymbol{\beta}$. However, the minimaxity result is not available at this time. In Section 4, a comparison between several estimators under the loss function $L_j(\omega, \boldsymbol{\delta}, (\boldsymbol{A}^t\boldsymbol{A})^j) = (\boldsymbol{\delta}-\boldsymbol{\beta})^t(\boldsymbol{A}^t\boldsymbol{A})^j(\boldsymbol{\delta}-\boldsymbol{\beta})$, $j = 0, 1, 2$, are carried out by Monte Carlo simulation along with two examples. In Section 5, the results are applied in estimating the means θ_i when the variances of the associated random variables are $d_i\sigma^2$ where d_i's are known numbers.

2 Proposed Empirical Bayes Ridge-Principal component Regression Estimators

For the multiple regressin model (1.1) under the assumption of normality, $\widehat{\boldsymbol{\beta}}$ and S are independently distributed, where

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{A}^t\boldsymbol{A})^{-1}) \quad \text{and} \quad S/\sigma^2 \sim \chi^2_n, \quad n = N - p.$$

Consider the situation in which the following hypothesis may be suspected:

$$H_0 : \ \boldsymbol{\beta} = \boldsymbol{C}\boldsymbol{\alpha}$$

where $\boldsymbol{C}$ is a $p \times q$ matrix with rank q and $\boldsymbol{\alpha} \in \boldsymbol{R}^q$ is an unknown vector. It may be reasonable to consider adaptive ridge regression estimators shrunken toward the hypothesis. To derive such a shrinkage procedure, we employ two types of empirical Bayes methods, which are here called an empirical Bayes ridge regression estimator (EB) and a hierarchical empirical Bayes ridge regression estimator (HB).

2.1 Empirical Bayes Ridge Regression Estimator (EB)

Suppose that $\boldsymbol{\beta}$ has prior distribution $\mathcal{N}_p(\boldsymbol{C}\boldsymbol{\alpha}, \sigma^2\lambda\boldsymbol{I}_p)$ for unknown $\lambda > 0$. Then the posterior distribution of $\boldsymbol{\beta}$ given $\widehat{\boldsymbol{\beta}}$ and the marginal distribution of $\widehat{\boldsymbol{\beta}}$ are, respectively, given by

$$\begin{aligned} \boldsymbol{\beta} \,|\, \widehat{\boldsymbol{\beta}} &\sim \mathcal{N}_p\left(\widehat{\boldsymbol{\beta}}^B(\lambda, \boldsymbol{\alpha}), \sigma^2(\boldsymbol{A}^t\boldsymbol{A} + \lambda^{-1}\boldsymbol{I})^{-1}\right), \\ \widehat{\boldsymbol{\beta}} &\sim \mathcal{N}_p\left(\boldsymbol{C}\boldsymbol{\alpha}, \sigma^2\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda\boldsymbol{I}\}\right), \end{aligned}$$

where $\widehat{\boldsymbol{\beta}}^B(\lambda, \boldsymbol{\alpha})$ is the Bayes estimator of $\boldsymbol{\beta}$ given by

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^B(\lambda, \boldsymbol{\alpha}) &= (\boldsymbol{A}^t\boldsymbol{A} + \lambda^{-1}\boldsymbol{I})^{-1}\boldsymbol{A}^t\boldsymbol{A}(\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\boldsymbol{\alpha}) + \boldsymbol{C}\boldsymbol{\alpha} \\ &= \widehat{\boldsymbol{\beta}} - \left(\boldsymbol{I} + \lambda\boldsymbol{A}^t\boldsymbol{A}\right)^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\boldsymbol{\alpha}). \end{aligned}$$

Since $\boldsymbol{\alpha}$ and λ are unknown, they need to be estimated. First, $\boldsymbol{\alpha}$ may be estimated by the weighted least squares estimator

$$\widehat{\boldsymbol{\alpha}} = (\boldsymbol{C}^t\boldsymbol{A}^t\boldsymbol{A}\boldsymbol{C})^{-1}\boldsymbol{C}^t\boldsymbol{A}^t\boldsymbol{A}\widehat{\boldsymbol{\beta}},$$

which can be obtained by minimizing the weighted squared loss $(\widehat{\boldsymbol{\beta}}-\boldsymbol{C}\boldsymbol{\alpha})^t\boldsymbol{A}^t\boldsymbol{A}(\widehat{\boldsymbol{\beta}}-\boldsymbol{C}\boldsymbol{\alpha})$. Substituting $\widehat{\boldsymbol{\alpha}}$ into $\widehat{\boldsymbol{\beta}}^B(\lambda,\boldsymbol{\alpha})$, we get the estimator

$$\widehat{\boldsymbol{\beta}}^B(\lambda,\widehat{\boldsymbol{\alpha}})=\widehat{\boldsymbol{\beta}}-(\boldsymbol{I}+\lambda\boldsymbol{A}^t\boldsymbol{A})^{-1}(\widehat{\boldsymbol{\beta}}-\boldsymbol{C}\widehat{\boldsymbol{\alpha}}). \tag{2.1}$$

A reasonable method to estimate λ is from the marginal distribution of $\widehat{\boldsymbol{\beta}}$. Using the sample moments, we propose an estimator which we call an empirical Bayes estimator.

Let λ^* be a root of the equation

$$(\widehat{\boldsymbol{\beta}}-\boldsymbol{C}\widehat{\boldsymbol{\alpha}})^t\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1}+\lambda^*\boldsymbol{I}\right\}^{-1}(\widehat{\boldsymbol{\beta}}-\boldsymbol{C}\widehat{\boldsymbol{\alpha}})=\frac{p-q-2}{n+2}S, \tag{2.2}$$

and λ_0 is the root of the equation

$$\sum_{i=1}^{p}(1-b_{ii})\frac{d_i-d_p}{d_i+\lambda_0}=(p-q-2)/2, \tag{2.3}$$

where $\boldsymbol{H}(\boldsymbol{A}^t\boldsymbol{A})^{-1}\boldsymbol{H}^t=\boldsymbol{D}=\mathrm{diag}\,(d_1,\ldots,d_p)$, $d_1\geq\cdots\geq d_p$ and b_{ii} is the (i,i)-th element of $\boldsymbol{H}\boldsymbol{C}(\boldsymbol{C}^t\boldsymbol{H}^t\boldsymbol{D}^{-1}\boldsymbol{H}\boldsymbol{C})^{-1}\boldsymbol{C}^t\boldsymbol{H}^t\boldsymbol{D}^{-1}$. Then we show in Section 3 that the empirical Bayes ridge regression estimator (EB) $\widehat{\boldsymbol{\beta}}^{EB}=\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{EB},\widehat{\alpha})$ defined in (2.1) for

$$\hat{\lambda}_{EB}=\max(\lambda^*,\lambda_0) \tag{2.4}$$

is minimax under the loss function (1.8). Monte Carlo simulations in Section 4 show that this estimator performs well under the loss functions

$$L(\omega,\boldsymbol{\delta},(\boldsymbol{A}^t\boldsymbol{A})^j)=(\boldsymbol{\delta}-\boldsymbol{\beta})^t(\boldsymbol{A}^t\boldsymbol{A})^j(\boldsymbol{\delta}-\boldsymbol{\beta})/\sigma^2,\quad j=0,1,2$$

for any estimator $\boldsymbol{\delta}$ of $\boldsymbol{\beta}$.

Using the root of the equation (2.2) as an estimator of λ was suggested by Fay and Herriot (1979) and Shinozaki and Chang (1996) when $\boldsymbol{\alpha}$ is present but they used another weighted least squares estimator of $\boldsymbol{\alpha}$.

2.2 Hierarchical Empirical Bayes Ridge Regression Estimator (HB)

When the rank q of the matirix $\boldsymbol{C}$ is large, it may be reasonable to shrink an esitmator of $\boldsymbol{\alpha}$. For the purpose, consider the hierarchical structure of the prior distributions:

$$\begin{aligned}\boldsymbol{\beta}\,|\,\boldsymbol{\alpha} &\sim \mathcal{N}_p(\boldsymbol{C}\boldsymbol{\alpha},\sigma^2\lambda\boldsymbol{I}_p),\\ \boldsymbol{\alpha} &\sim \mathcal{N}_q(\boldsymbol{\alpha}_0,\sigma^2\tau\boldsymbol{I}_q),\end{aligned}$$

where λ and τ are unknown and $\boldsymbol{\alpha}_0$ is a known value. Such hierarchical prior distributions have been proposed in the literature (for example, see Lindley and Smith (1972) and DuMouchel and Harris (1983)).

Integrating out the joint prior distribution with respect to $\boldsymbol{\alpha}$, we can see that the mariginal prior distribution of $\boldsymbol{\beta}$ has

$$\boldsymbol{\beta}\sim\mathcal{N}_p(\boldsymbol{C}\boldsymbol{\alpha}_0,\sigma^2\boldsymbol{\Psi}),\quad \boldsymbol{\Psi}=\lambda\boldsymbol{I}_p+\tau\boldsymbol{C}\boldsymbol{C}^t.$$

Since $\widehat{\boldsymbol{\beta}}|\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{A}^t\boldsymbol{A})^{-1})$, given $\widehat{\boldsymbol{\beta}}$ the posterior distribution of $\boldsymbol{\beta}$ has

$$\boldsymbol{\beta} \,|\, \widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\widehat{\boldsymbol{\beta}}^{HB}(\lambda,\tau), (\boldsymbol{A}^t\boldsymbol{A} + \boldsymbol{\Psi}^{-1})^{-1}),$$

and the marginal distribution of $\widehat{\boldsymbol{\beta}}$ has

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\boldsymbol{C}\boldsymbol{\alpha}_0, \boldsymbol{\Psi} + (\boldsymbol{A}^t\boldsymbol{A})^{-1}),$$

where $\widehat{\boldsymbol{\beta}}^{HB}(\lambda,\tau)$ is the Bayes estimator of $\boldsymbol{\beta}$, given by

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^{HB}(\lambda,\tau) &= (\boldsymbol{A}^t\boldsymbol{A} + \boldsymbol{\Psi}^{-1})^{-1}(\boldsymbol{A}^t\boldsymbol{A}\widehat{\boldsymbol{\beta}} + \boldsymbol{\Psi}^{-1}\boldsymbol{C}\boldsymbol{\alpha}_0) \\ &= \widehat{\boldsymbol{\beta}} - (\boldsymbol{A}^t\boldsymbol{A} + \boldsymbol{\Psi}^{-1})^{-1}\boldsymbol{\Psi}^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\boldsymbol{\alpha}_0) \\ &= \widehat{\boldsymbol{\beta}} - (\boldsymbol{A}^t\boldsymbol{A})^{-1}\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda\boldsymbol{I}_p\right\}^{-1}\left\{\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\widehat{\boldsymbol{\alpha}}^S(\lambda,\tau)\right\}, \end{aligned} \tag{2.5}$$

where

$$\widehat{\boldsymbol{\alpha}}^S(\lambda,\tau) = \widehat{\boldsymbol{\alpha}}(\lambda) - \left[\boldsymbol{I}_q + \tau\boldsymbol{C}^t\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda\boldsymbol{I}_p\right\}^{-1}\boldsymbol{C}\right]^{-1}(\widehat{\boldsymbol{\alpha}}(\lambda) - \boldsymbol{\alpha}_0), \tag{2.6}$$

as shown in the Appendix. It is interesting to note that $\widehat{\boldsymbol{\alpha}}^S(\lambda,\tau)$ shrinks the weighted LSE $\widehat{\boldsymbol{\alpha}}(\lambda)$ towards the prior value $\boldsymbol{\alpha}_0$. Hence $\widehat{\boldsymbol{\beta}}^{HB}(\lambda,\tau)$ is interpreted as a double shrinkage procedure that shrinks the LSE $\widehat{\boldsymbol{\beta}}$ towards the shrunken value $\boldsymbol{C}\widehat{\boldsymbol{\alpha}}^S(\lambda,\tau)$.

By considering the marginal distribution of $\widehat{\boldsymbol{\beta}}$, we can obtain the maximum likelihood estimates of λ and τ. For σ^2, however, as before we shall use the usual estimator S/n. Let λ^{**} and τ^{**} be the solution of the following two equations:

$$\begin{aligned} &n(\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\boldsymbol{\alpha}_0)^t(\boldsymbol{G} + \tau\boldsymbol{C}\boldsymbol{C}^t)^{-2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\boldsymbol{\alpha}_0) \\ &\qquad = S\cdot\mathrm{tr}\,(\boldsymbol{G} + \tau\boldsymbol{C}\boldsymbol{C}^t)^{-1}, \\ &n(\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\boldsymbol{\alpha}_0)^t(\boldsymbol{G} + \tau\boldsymbol{C}\boldsymbol{C}^t)^{-1}\boldsymbol{C}\boldsymbol{C}^t(\boldsymbol{G} + \tau\boldsymbol{C}\boldsymbol{C}^t)^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\boldsymbol{\alpha}_0) \\ &\qquad = S\cdot\mathrm{tr}\,\boldsymbol{C}\boldsymbol{C}^t(\boldsymbol{G} + \tau\boldsymbol{C}\boldsymbol{C}^t)^{-1}, \end{aligned}$$

for $\boldsymbol{G} = \boldsymbol{G}(\lambda) = (\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda\boldsymbol{I}_p$. Define

$$\hat{\lambda}_{HB} = \max(\lambda^{**}, 0), \quad \hat{\tau}_{HB} = \max(\tau^{**}, 0).$$

Then, the hierarchical empirical Bayes ridge regression estimator (HB) is given by

$$\widehat{\boldsymbol{\beta}}^{HB} = \widehat{\boldsymbol{\beta}} - (\boldsymbol{A}^t\boldsymbol{A})^{-1}\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \hat{\lambda}_{HB} + \hat{\tau}_{HB}\boldsymbol{C}\boldsymbol{C}^t\right\}^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\boldsymbol{\alpha}_0). \tag{2.7}$$

The prior mean $\boldsymbol{\alpha}_0$ is given from a prior information. If there are no prior information available, $\boldsymbol{\alpha}_0$ may be chosen to be a zero vector.

When $\tau = 0$ and $\boldsymbol{C} = \boldsymbol{0}$, the estimator $\widehat{\boldsymbol{\beta}}^{HB}(\lambda, 0)$ yields

$$\widehat{\boldsymbol{\beta}}^{HB}(\hat{\lambda}_{ML}, 0) = \widehat{\boldsymbol{\beta}} - (\boldsymbol{I}_p + \hat{\lambda}_{ML}\boldsymbol{A}^t\boldsymbol{A})^{-1}\widehat{\boldsymbol{\beta}}, \tag{2.8}$$

where $\hat{\lambda}_{ML} = \max(\lambda^{**}, 0)$ and λ^{**} is a solution of the equation

$$n\widehat{\boldsymbol{\beta}}^t\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda^{**}\boldsymbol{I}_p\}^{-2}\widehat{\boldsymbol{\beta}} = S\cdot\mathrm{tr}\,\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda^{**}\boldsymbol{I}_p\}^{-1}. \tag{2.9}$$

The estimator is rewitten as the empirical Bayes estimator $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{ML}, \boldsymbol{0})$ in (2.1) and $\hat{\lambda}_{ML}$ is the maximum likelihood estimator of λ in the marginal distribution of $\widehat{\boldsymbol{\beta}}$.

3 Minimaxity of the Empirical Bayes Estimator under Strawderman's Loss Function

To handle the estimators more conveniently, we treat them in a canonical form. Let $\boldsymbol{H}$ be a $p \times p$ orthogonal matrix, $\boldsymbol{H}\boldsymbol{H}^t = \boldsymbol{I}_p$ such that

$$\boldsymbol{H}(\boldsymbol{A}^t\boldsymbol{A})^{-1}\boldsymbol{H}^t = \boldsymbol{D} = \mathrm{diag}\,(d_1, \ldots, d_p)$$

where $d_1 \geq \ldots \geq d_p > 0$. Define $\boldsymbol{x} = \boldsymbol{H}\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\theta} = \boldsymbol{H}\boldsymbol{\beta}$. Then

$$\boldsymbol{x} \sim \mathcal{N}_p(\boldsymbol{\theta}, \sigma^2\boldsymbol{D}). \tag{3.1}$$

That is x_i's are independently normally distributed $\mathcal{N}(\theta_i, \sigma^2 d_i)$ where x_i and θ_i are the respective ith component of the vectors $\boldsymbol{x}$ and $\boldsymbol{\theta}$. The estimator $\hat{\lambda}$ of λ can be represented as a function of $x_1, \ldots, x_p$ and S since it is a function of $\widehat{\boldsymbol{\beta}}$ and S. Letting $\widehat{\boldsymbol{\theta}}^B(\lambda, \boldsymbol{\alpha}) = \boldsymbol{H}\widehat{\boldsymbol{\beta}}^B(\lambda, \boldsymbol{\alpha})$ and $\boldsymbol{Z} = \boldsymbol{H}\boldsymbol{C}$, we see that

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}^B(\lambda, \boldsymbol{\alpha}) &= \boldsymbol{x} - (\boldsymbol{D} + \lambda\boldsymbol{I})^{-1}\boldsymbol{D}(\boldsymbol{x} - \boldsymbol{Z}\boldsymbol{\alpha}) \\
\widehat{\boldsymbol{\theta}}^B(\lambda, \widehat{\boldsymbol{\alpha}}) &= \boldsymbol{x} - (\boldsymbol{D} + \lambda\boldsymbol{I})^{-1}\boldsymbol{D}\,(\boldsymbol{x} - \boldsymbol{B}\boldsymbol{x})\,,
\end{aligned}$$

where $\boldsymbol{B} = \boldsymbol{Z}(\boldsymbol{Z}^t\boldsymbol{D}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}^t\boldsymbol{D}^{-1} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p)^t$. Then the estimator $\widehat{\boldsymbol{\theta}}^B(\lambda, \widehat{\boldsymbol{\alpha}})$ is represented componentwise by

$$\hat{\theta}_i^B(\lambda, \widehat{\boldsymbol{\alpha}}) = x_i - \frac{d_i}{d_i + \lambda}(x_i - \boldsymbol{b}_i^t\boldsymbol{x}).$$

Now we are ready to give general conditions on the estimator $\hat{\lambda}$ of λ under which the minimaxity is guaranteed for the adaptive ridge regression estimator $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}, \widehat{\boldsymbol{\alpha}})$ under the loss function (1.8).

Theorem 1. *The empirical Bayes or adaptive ridge regression estimator $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}, \widehat{\boldsymbol{\alpha}})$ is minimax, that is, improves on the least squares estimator $\widehat{\boldsymbol{\beta}}$ relative to the loss $L(\omega, \boldsymbol{\delta}, (\boldsymbol{A}^t\boldsymbol{A})^2)$ given by (1.8) if the following conditions are satisfied for $p \geq q + 3$:*

(a) $\hat{\lambda} \geq \lambda_m$ for nonnegative constant λ_m, and $\hat{\lambda}$ is absolutely continuous with respect to $x_1, \ldots, x_p$ and S.

(b) $(x_i - \boldsymbol{b}_i^t\boldsymbol{x})\partial\hat{\lambda}/\partial x_i \geq 0$ for $i = 1, \ldots, p$, and

$$\sum_{i=1}^{p} \frac{(x_i - \boldsymbol{b}_i^t\boldsymbol{x})}{d_i + \hat{\lambda}} \frac{\partial\hat{\lambda}}{\partial x_i} \leq 2. \tag{3.2}$$

(c) $\partial\hat{\lambda}/\partial S \leq 0$ and for positive constants α and β,

$$\sum_{i=1}^{p} \frac{(x_i - \boldsymbol{b}_i^t\boldsymbol{x})^2/S}{d_i + \hat{\lambda}} \leq \alpha \quad \text{and} \quad -\sum_{i=1}^{p} \frac{(x_i - \boldsymbol{b}_i^t\boldsymbol{x})^2}{(d_i + \hat{\lambda})^2} \frac{\partial\hat{\lambda}}{\partial S} \leq \beta. \tag{3.3}$$

(d) The constants λ_m, α and β satisfy the inequality:

$$\sum_{i=1}^{p} (1 - b_{ii}) \frac{d_i - d_p}{d_i + \lambda_m} + \frac{(n-2)\alpha}{2} + 2\beta \le p - q - 2, \tag{3.4}$$

where b_{ii} is the (i,i)-th element of the matrix $\boldsymbol{B} = \boldsymbol{Z}(\boldsymbol{Z}^t\boldsymbol{D}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}^t\boldsymbol{D}^{-1}$.

Theorem 1, proved in the Appendix, can be applied to get the sufficient conditions for several adaptive or empirical Bayes ridge regression estimators to be minimax.

3.1 Minimaxity of the Empirical Bayes Ridge Regression Estimator

We first show the minimaxity of the empirical Bayes ridge regression estimator $\boldsymbol{\beta}^B(\hat{\lambda}_{EB}, \widehat{\boldsymbol{\alpha}})$ proposed in Section 2. The λ^* defined as a root of the equation (2.2) is expressed in the notation of the model (3.1) as

$$\sum_{i=1}^{p} \frac{(x_i - \boldsymbol{b}_i^t\boldsymbol{x})^2/S}{d_i + \lambda^*} = \frac{p-q-2}{n+2}. \tag{3.5}$$

To check the conditions of Theorem 1, we need to calculate the derivatives $\partial\lambda^*/\partial x_i$ and $\partial\lambda^*/\partial S$. The theorem of the implicit function can be applied to get these quantities. Letting

$$F(x_1, \ldots, x_p, S, \lambda) = \sum_{i=1}^{p} \frac{(x_i - \boldsymbol{b}_i^t\boldsymbol{x})^2}{d_i + \lambda} - \frac{p-q-2}{n+2} S,$$

we see that $F(x_1, \ldots, x_p, S, \lambda^*) = 0$. Then we observe that

$$\begin{aligned}
\frac{\partial\lambda^*}{\partial x_i} &= -\frac{\partial F}{\partial x_i}\left(\frac{\partial F}{\partial\lambda^*}\right)^{-1} = 2(1-b_{ii})\frac{x_i - \boldsymbol{b}_i^t\boldsymbol{x}}{d_i+\lambda^*}\left(\sum_{j=1}^{p}\frac{(x_j - \boldsymbol{b}_j^t\boldsymbol{x})^2}{(d_j+\lambda^*)^2}\right)^{-1} \\
\frac{\partial\lambda^*}{\partial S} &= -\frac{\partial F}{\partial S}\left(\frac{\partial F}{\partial\lambda^*}\right)^{-1} = -\frac{p-q-2}{n+2}\left(\sum_{j=1}^{p}\frac{(x_j - \boldsymbol{b}_j^t\boldsymbol{x})^2}{(d_j+\lambda^*)^2}\right)^{-1}.
\end{aligned}$$

By using these quantities and the equation (3.5), since $0 \le b_{ii} \le 1$, it can be seen that

$$\begin{aligned}
\sum_{i=1}^{p}\frac{x_i - \boldsymbol{b}_i^t\boldsymbol{x}}{d_i+\hat{\lambda}_{EB}}\frac{\partial\hat{\lambda}_{EB}}{\partial x_i} &\le 2I(\lambda^* > \lambda_0) \le 2, \\
-\sum_{i=1}^{p}\frac{(x_i - \boldsymbol{b}_i^t\boldsymbol{x})^2}{(d_i+\hat{\lambda}_{EB})^2}\frac{\partial\hat{\lambda}_{EB}}{\partial S} &= \frac{p-q-2}{n+2}I(\lambda^* > \lambda_0) \le \frac{p-q-2}{n+2}, \\
\sum_{i=1}^{p}\frac{(x_i - \boldsymbol{b}_i^t\boldsymbol{x})^2/S}{d_i+\hat{\lambda}_{EB}} &\le \sum_{i=1}^{p}\frac{(x_i - \boldsymbol{b}_i^t\boldsymbol{x})^2/S}{d_i+\lambda^*} = \frac{p-q-2}{n+2}.
\end{aligned}$$

Hence, the conditions (b) and (c) are satisfied by putting $\alpha = \beta = (p-q-2)/(n+2)$. The constant λ_m in the condition (d) is required to satisfy the inequality

$$\sum_{i=1}^{p}(1-b_{ii})\frac{d_i - d_p}{d_i+\lambda_0} \le \frac{p-q-2}{2}, \tag{3.6}$$

which is guaranteed by the equation (2.3) by putting $\lambda_m = \lambda_0$. Hence all the conditions in Theorem 1 are satisfied, and we get the following proposition:

Proposition 1. *Assume that $p \geq q+3$ and that λ_0 satifies the equation (2.3). Let $\hat{\lambda}_{EB} = \max(\lambda^*, \lambda_0)$ for the root λ^* of the equation (3.5). Then the EB estimator $\widehat{\boldsymbol{\beta}}^{EB} = \widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{EB}, \widehat{\boldsymbol{\alpha}})$ is minimax under the loss (1.8).*

Next, we shall apply Theorem 1 to show that Strawderman's type of estimators are also minimax.

3.2 Minimaxity of an Adaptive Ridge Regression Estimator

Consider the adaptive estimator $\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{AD}(a, \lambda_a), \widehat{\boldsymbol{\alpha}})$ discussed in the introduction where

$$\hat{\lambda}_{AD}(a, \lambda_a) = (\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\widehat{\boldsymbol{\alpha}})^t \boldsymbol{A}^t \boldsymbol{A} (\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\widehat{\boldsymbol{\alpha}})/(aS) + \lambda_a;$$

this can be expressed in the notation of the model (3.1) as

$$\hat{\lambda}_{AD}(a, \lambda_a) = \sum_{i=1}^{p} (x_i - \boldsymbol{b}_i^t \boldsymbol{x})^2/(a d_i S) + \lambda_a. \tag{3.7}$$

We shall now show that conditions (a)-(d) are satisfied by the estimator given in (3.7) for a suitable choice of a and λ_a. The condition (a) is satisfied by putting $\lambda_m = \lambda_a$. The condition (b) is verified as

$$\begin{aligned} \sum_{i=1}^{p} \frac{x_i - \boldsymbol{b}_i^t \boldsymbol{x}}{d_i + \hat{\lambda}_{AD}(a, \lambda_a)} \frac{\partial \hat{\lambda}_{AD}(a, \lambda_a)}{\partial x_i} &\leq 2 \sum_{i=1}^{p} (1 - b_{ii}) \frac{(x_i - \boldsymbol{b}_i^t \boldsymbol{x})^2/(a d_i S)}{d_i + \hat{\lambda}_{AD}(a, \lambda_a)} \\ &\leq \frac{2\hat{\lambda}_{AD}(a, \lambda_a)}{d_p + \hat{\lambda}_{AD}(a, \lambda_a)} \leq 2, \end{aligned}$$

since $0 \leq b_{ii} \leq 1$. For the condition (c), we observe that

$$\begin{aligned} \sum_{i=1}^{p} \frac{(x_i - \boldsymbol{b}_i^t \boldsymbol{x})^2/S}{d_i + \hat{\lambda}_{AD}(a, \lambda_a)} \leq a d_1 \frac{\sum_{i=1}^{p} (x_i - \boldsymbol{b}_i^t \boldsymbol{x})^2/(a d_i S)}{d_p + \hat{\lambda}_{AD}(a, \lambda_a)} &\leq a d_1, \\ -\sum_{i=1}^{p} \frac{(x_i - \boldsymbol{b}_i^t \boldsymbol{x})^2}{(d_i + \hat{\lambda}_{AD}(a, \lambda_a))^2} \frac{\partial \hat{\lambda}_{AD}(a, \lambda_a)}{\partial S} \leq \frac{a d_1 \{\hat{\lambda}_{AD}(a, \lambda_a)\}^2}{(d_p + \hat{\lambda}_{AD}(a, \lambda_a))^2} &\leq a d_1, \end{aligned}$$

which imply that the condition is satisfied by putting $\alpha = \beta = a d_1$. Hence the condition (d) is given by

$$\sum_{i=1}^{p} (1 - b_{ii}) \frac{d_i - d_p}{d_i + \lambda_a} + \frac{(n+2) d_1}{2} a \leq p - q - 2. \tag{3.8}$$

A reasonable choice of a is $a = (p-q-2)/[(n+2)d_1]$, and then λ_a should be chosen as a root such that the equality holds in the inequality (3.8). This root is equal to the solution λ_0 of the equation (2.3).

Proposition 2. *Assume that $p \geq q+3$ and let λ_0 be a solution of the equation given by (2.3). Then the adaptive ridge regression estimator $\widehat{\boldsymbol{\beta}}^{AD} = \widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{AD}, \widehat{\boldsymbol{\alpha}})$ with*

$$\begin{aligned} \hat{\lambda}_{AD} &= \frac{n+2}{p-q-2} \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\widehat{\boldsymbol{\alpha}})^t \boldsymbol{A}^t \boldsymbol{A} (\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\widehat{\boldsymbol{\alpha}})}{\mathrm{ch}_{min}(\boldsymbol{A}^t \boldsymbol{A}) S} + \lambda_0 \\ &= \frac{(n+2)d_1}{p-q-2} \sum_{i=1}^{p} (x_i - \boldsymbol{b}_i^t \boldsymbol{x})^2 / (d_i S) + \lambda_0 \end{aligned} \tag{3.9}$$

is minimax under the loss (1.8), where $\mathrm{ch}_{min}(\boldsymbol{M})$ denotes the minimum eigen value of the matrix $\boldsymbol{M}$. When there is no restriction on $\boldsymbol{\beta}$ belonging to the subspace, a similar estimator has been considered by Strawderman (1978) under the same loss function as we do.

3.3 Minimaxity of a Modified Adaptive Ridge Regression Estimator

It is noted that the estimator (3.9) has a shortcoming for smaller d_p. In fact, as d_p tends to zero, $\hat{\lambda}_{AD}$ goes to infinity, so that the adaptive ridge regression estimator $\widehat{\boldsymbol{\beta}}^{AD}$ approaches the unstable estimator $\widehat{\boldsymbol{\beta}}$ in the case of large d_1. To eliminate this shortcoming, we modify $\hat{\lambda}_{AD}$ as

$$\begin{aligned} \hat{\lambda}_{TR} &= \max\left\{ \frac{(n+2)(d_1+1)}{(p-q-2)S} (\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\widehat{\boldsymbol{\alpha}})^t [(\boldsymbol{A}^t \boldsymbol{A})^{-1} + \boldsymbol{I}_p]^{-1} (\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\widehat{\boldsymbol{\alpha}}), \lambda_0 \right\} \\ &= \max\left\{ \frac{(n+2)(d_1+1)}{(p-q-2)S} \sum_{i=1}^{p} \frac{(x_i - \boldsymbol{b}_i^t \boldsymbol{x})^2}{d_i + 1}, \lambda_0 \right\}, \end{aligned} \tag{3.10}$$

where $d_1 = \{\mathrm{ch}_{min}(\boldsymbol{A}^t \boldsymbol{A})\}^{-1}$. It is easy to see that $\hat{\lambda}_{TR}$ is bounded for d_p going to zero as well as $\hat{\lambda}_{TR} \leq \hat{\lambda}_{AD}$. This means that the modified estimator $\widehat{\boldsymbol{\beta}}^{TR} = \widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{TR}, \widehat{\boldsymbol{\alpha}})$ is shrunken more than $\widehat{\boldsymbol{\beta}}^{AD}$. The minimaxity of $\widehat{\boldsymbol{\beta}}^{TR}$ can be verified by the same argument as in the above proposition.

Proposition 3. *The modified adaptive ridge regression estimator $\widehat{\boldsymbol{\beta}}^{TR} = \widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{TR}, \widehat{\boldsymbol{\alpha}})$ is minimax under the loss (1.8) for λ_0 defined by the equation (2.3) if $p \geq q+3$.*

4 Multicollinearity Cases and Simulation Studies

In the multicollinearity case, usual ridge regression estimators shrink the least squares estimator toward zero, that is, $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$. In this case, the adaptive or empirical Bayes ridge regression estimators are written by

$$\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}, \boldsymbol{0}) = \left[\boldsymbol{A}^t \boldsymbol{A} + \hat{\lambda}^{-1} \boldsymbol{I}\right]^{-1} \boldsymbol{A}^t \boldsymbol{y} = \widehat{\boldsymbol{\beta}} - [\boldsymbol{I} + \hat{\lambda} \boldsymbol{A}^t \boldsymbol{A}]^{-1} \widehat{\boldsymbol{\beta}}. \tag{4.1}$$

Three estimators $\hat{\lambda}_{AD}$, $\hat{\lambda}_{TR}$ and $\hat{\lambda}_{EB}$ of λ are given by (3.9), (3.10) and (2.4) with $\widehat{\boldsymbol{\alpha}} = \boldsymbol{0}$, $\boldsymbol{B} = \boldsymbol{0}$ and $q = 0$, and these estimators of λ yield the estimators

$$\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{AD}, \boldsymbol{0}),\ \widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{TR}, \boldsymbol{0}),\ \widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{EB}, \boldsymbol{0}),\ \text{denoted by } AD,\ TR,\ EB,$$

respectively, whose minimaxities were shown by Propositions 1, 2 and 3 with $b_{ii} = 0$ and $q = 0$. Other estimators of λ treated in the literature are the F-ratio

$$\hat{\lambda}_{RR} = \frac{\widehat{\boldsymbol{\beta}}^t \boldsymbol{A}^t \boldsymbol{A} \widehat{\boldsymbol{\beta}}/p}{S/n} \tag{4.2}$$

and the maximum likelihood estimator $\hat{\lambda}_{ML}$ given by (2.9), which yield the estimators

$$\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{RR}, \mathbf{0}),\ \widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{ML}, \mathbf{0}),\ \text{denoted by } RR,\ ML,$$

respectively. Although the minimaxity of their estimators is not discussed, we include them for numerical comparison of the risk behaviors.

In the multicollieарity case, we can construct more reasonable ridge-type regression estimators by using the information about which eigenvalues are smaller. This gives a good motivation about empirical Bayes ridge regression estimators shrunken toward a subspace of $\boldsymbol{\beta}$. Let $\boldsymbol{H}$ be an orthogonal matrix such that

$$\boldsymbol{H}\boldsymbol{A}^t\boldsymbol{A}\boldsymbol{H}^t = \boldsymbol{D}^{-1} = \begin{pmatrix} \boldsymbol{D}_1^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{D}_2^{-1} \end{pmatrix}$$

where $\boldsymbol{D}_1^{-1} = \text{diag}\,(d_1^{-1}, \dots, d_{p-q}^{-1})$, $(p-q)\times(p-q)$ diagonal matrix with smaller eigenvalues. Corresponding to this decomposition, the orthogonal matrix $\boldsymbol{H}$ is decomposed as $\boldsymbol{H}^t = (\boldsymbol{H}_1^t; \boldsymbol{H}_2^t)$ for $q \times p$ matrix $\boldsymbol{H}_2$. Then, as in (1.7),

$$\boldsymbol{\beta} = \boldsymbol{H}_1^t \boldsymbol{\gamma}_1 + \boldsymbol{H}_2^t \boldsymbol{\gamma}_2.$$

Since $\boldsymbol{\gamma}_1$ corresponds to the smaller eigenvalues of $\boldsymbol{A}^t\boldsymbol{A}$, it should not be included in the model. Thus, it may be reasonable to shrink $\widehat{\boldsymbol{\beta}}$ towards the linear constraint:

$$H_0 \ :\ \boldsymbol{\beta} = \boldsymbol{H}_2^t \boldsymbol{\gamma}_2, \quad \boldsymbol{\gamma}_2 \in \boldsymbol{R}^q. \tag{4.3}$$

The estimator (1.9) is applied to this situation, resulting in the empirical Bayes ridge regression estimator

$$\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}, \widehat{\boldsymbol{\gamma}}_2) = \widehat{\boldsymbol{\beta}} - \left(\boldsymbol{I} + \hat{\lambda}\boldsymbol{A}^t\boldsymbol{A}\right)^{-1}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{H}_2\widehat{\boldsymbol{\gamma}}_2\right) \tag{4.4}$$

for $\widehat{\boldsymbol{\gamma}}_2 = (\boldsymbol{H}_2\boldsymbol{A}^t\boldsymbol{A}\boldsymbol{H}_2^t)^{-1}\boldsymbol{H}_2\boldsymbol{A}^t\boldsymbol{y} = \boldsymbol{H}_2\widehat{\boldsymbol{\beta}}$. Since the principal component estimator of $\boldsymbol{\beta}$ is given by $\widehat{\boldsymbol{\beta}}^{PC} = \boldsymbol{H}_2\widehat{\boldsymbol{\gamma}}_2 = \boldsymbol{H}_2^t\boldsymbol{H}_2\widehat{\boldsymbol{\beta}}$, the empirical Bayes estimator is rewritten by

$$\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}, \widehat{\boldsymbol{\gamma}}_2) = \widehat{\boldsymbol{\beta}} - \left(\boldsymbol{I} + \hat{\lambda}\boldsymbol{A}^t\boldsymbol{A}\right)^{-1}\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{PC}\right).$$

It is known that the principal component estimator and the ridge regression estimator are useful in predicting a response variable in the presence of multicollieаrity. It is anticipated that the empirical Bayes ridge regression estimators $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}, \widehat{\boldsymbol{\gamma}}_2)$ elliminate the shortcomings of both the least squares estimator $\widehat{\boldsymbol{\beta}}$ and the principal component estimator $\widehat{\boldsymbol{\beta}}^{PC}$. Three estimators $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{AD}, \widehat{\boldsymbol{\gamma}}_2)$, $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{TR}, \widehat{\boldsymbol{\gamma}}_2)$ and $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{EB}, \widehat{\boldsymbol{\gamma}}_2)$ with minimaxity properties are given by Propositions 1, 2 and 3 with $\boldsymbol{C} = \boldsymbol{H}_2^t$ and $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\gamma}}_2$. In this case, the Fay-Herriot type estimator $\hat{\lambda}_{EB}$ of λ is of the form $\hat{\lambda}_{EB} = \max(\lambda^*, \lambda_0)$ where λ^* and λ_0 are solutions of the equations

$$\sum_{i=1}^{p-q} \frac{x_i^2/S}{d_i + \lambda^*} = \frac{p-q-2}{n+2} \quad \text{and} \quad \sum_{i=1}^{p-q} \frac{d_i - d_p}{d_i + \lambda_0} = \frac{p-q-2}{2}, \tag{4.5}$$

for $\boldsymbol{x} = (x_1, \ldots, x_p)^t = \boldsymbol{H}\widehat{\boldsymbol{\beta}}$. The hierarchical empirical Bayes ridge regression estimator $\widehat{\boldsymbol{\beta}}^{HB}(\hat{\lambda}_{HB}, \hat{\tau}_{HB})$ given by (2.7) is expressed as

$$\widehat{\boldsymbol{\beta}}^{HB}(\hat{\lambda}_{HB}, \hat{\tau}_{HB}) = \widehat{\boldsymbol{\beta}} - (\boldsymbol{A}^t\boldsymbol{A})^{-1}\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \hat{\lambda}_{HB}\boldsymbol{I}_p\right\}^{-1}\left\{\widehat{\boldsymbol{\beta}} - \boldsymbol{H}_2^t\widehat{\boldsymbol{\alpha}}^S(\hat{\lambda}_{HB}, \hat{\tau}_{HB})\right\}, \tag{4.6}$$

and $\widehat{\boldsymbol{\alpha}}^S(\hat{\lambda}_{HB}, \hat{\tau}_{HB})$ can be rewritten by

$$\widehat{\boldsymbol{\alpha}}^S(\hat{\lambda}_{HB}, \hat{\tau}_{HB}) = \hat{\tau}_{HB}\left(\boldsymbol{D}_2 + (\hat{\lambda}_{HB} + \hat{\tau}_{HB})\boldsymbol{I}_q\right)^{-1}\boldsymbol{H}_2\widehat{\boldsymbol{\beta}}^{PC},$$

where $\hat{\lambda}_{HB} = \max(\lambda^{**}, 0)$ and $\hat{\tau}_{HB} = \max(\tau^{**}, 0)$ and λ^{**}, $\hat{\tau}^{**}$ are solutions of the equations

$$\sum_{i=1}^{p-q} \frac{x_i^2}{(d_i + \lambda^{**})^2} = \frac{S}{n}\sum_{i=1}^{p-q}\frac{1}{d_i + \lambda^{**}}, \tag{4.7}$$

$$\sum_{i=p-q+1}^{p} \frac{x_i^2}{(d_i + \lambda^{**} + \tau^{**})^2} = \frac{S}{n}\sum_{i=p-q+1}^{p}\frac{1}{d_i + \lambda^{**} + \tau^{**}}. \tag{4.8}$$

In the case that $q = 0$, the maximum likelihood estimator $\hat{\lambda}_{ML}$ is given by $\hat{\lambda}_{HB}$ with $q = 0$ and $\hat{\tau} = 0$.

In the simulation experiments given below, we treat the case $\boldsymbol{\gamma}_1 \in \boldsymbol{R}^5$ in (1.7). The principal component estimator $\widehat{\boldsymbol{\beta}}_5^{PC} = \boldsymbol{H}_2^t\widehat{\boldsymbol{\gamma}}_2 = \boldsymbol{H}_2^t\boldsymbol{H}_2\widehat{\boldsymbol{\beta}}$, denoted by PC_5, is obtained by deleting the eigenvectors corresponding to the five largest eigenvalues of $(\boldsymbol{A}^t\boldsymbol{A})^{-1}$. The corresponding adaptive ridge regression estimators (4.4) with $\hat{\lambda}_{AD}$ and $\hat{\lambda}_{TR}$ given by (3.9) and (3.10), the empirical Bayes estimator (4.4) with $\hat{\lambda}_{EB}$ given by (4.5) and the hierarchical empirical Bayes estimator (4.6) are, respectivey, given by

$$\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{AD}, \widehat{\boldsymbol{\gamma}}_2),\ \widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{TR}, \widehat{\boldsymbol{\gamma}}_2),\ \widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{EB}, \widehat{\boldsymbol{\gamma}}_2),\ \widehat{\boldsymbol{\beta}}^{HB}(\hat{\lambda}_{HB}, \hat{\tau}_{HB}),$$

denoted by AD_5, TR_5, EB_5, HB_5.

The principal component estimator deleting the eigenvector corresponding to the largest eigenvalue of $(\boldsymbol{A}^t\boldsymbol{A})^{-1}$ is also treated and denoted by PC_1.

Now we are ready to investigate the risk-performances of estimators of $\boldsymbol{\beta}$ numerically. The estimators we want to compare are AD, TR, EB, ML and RR for $q = 0$; AD_5, TR_5, EB_5, HB_5 and PC_5 for $\boldsymbol{\gamma}_1 \in \boldsymbol{R}^5$; PC_1 for $\boldsymbol{\gamma}_1 \in \boldsymbol{R}^1$. Every estimator $\boldsymbol{\delta}$ is evaluated by three types of risk functions $R_j(\omega, \boldsymbol{\delta})$ under the loss functions $L_j(\omega, \boldsymbol{\delta}, (\boldsymbol{A}^t\boldsymbol{A})^j) = (\boldsymbol{\delta} - \boldsymbol{\beta})^t(\boldsymbol{A}^t\boldsymbol{A})^j(\boldsymbol{\delta} - \boldsymbol{\beta})/\sigma^2$, called the L_j-loss, for $j = 0, 1, 2$. The risk functions of the above estimators and the least squares estimator $\widehat{\boldsymbol{\beta}}$ are obtained from 50,000 replications

Table 1. Relative Efficiencies of the Estimators under L_0, L_1, L_2 Losses for $n = 30$, $\theta_i = \eta \times i$, $i = 1, \ldots, 10$, and $\boldsymbol{D} = \text{diag}(500, 50, 30, 10, 5, 1, 1, 1, 1, 1)$

	η	AD	TR	EB	ML	RR	AD_5	TR_5	EB_5	HB_5	PC_5	PC_1
L_0	0.0	0.460	0.397	0.024	0.001	0.003	0.574	0.540	0.207	0.008	0.008	0.167
	0.5	0.879	0.796	0.063	0.037	0.029	0.622	0.589	0.210	0.046	0.031	0.167
	1.0	0.964	0.933	0.124	0.088	0.068	0.721	0.692	0.222	0.101	0.100	0.169
	1.5	0.983	0.968	0.190	0.141	0.113	0.810	0.787	0.242	0.168	0.215	0.171
	2.5	0.994	0.988	0.332	0.264	0.215	0.908	0.895	0.307	0.327	0.582	0.177
L_1	0.0	0.910	0.894	0.362	0.071	0.154	0.933	0.926	0.801	0.096	0.498	0.900
	0.5	0.982	0.970	0.662	0.633	0.629	0.942	0.936	0.804	0.650	0.673	0.900
	1.0	0.994	0.990	0.767	0.744	0.722	0.959	0.954	0.812	0.756	1.196	0.900
	1.5	0.997	0.995	0.823	0.804	0.782	0.973	0.969	0.827	0.819	2.068	0.900
	2.5	0.999	0.998	0.882	0.867	0.851	0.987	0.985	0.870	0.884	4.858	0.901
L_2	0.0	0.994	0.992	0.592	0.129	0.278	0.998	0.998	0.989	0.131	0.933	0.999
	0.5	0.999	0.999	0.959	0.962	0.998	0.998	0.998	0.989	0.962	0.987	0.999
	1.0	0.999	0.999	0.984	0.985	0.999	0.999	0.999	0.990	0.983	1.151	0.999
	1.5	0.999	0.999	0.991	0.992	0.998	0.999	0.999	0.992	0.991	1.425	0.999
	2.5	0.999	0.999	0.996	0.996	0.998	0.999	0.999	0.997	0.996	2.301	0.999

Table 2. Relative Efficiencies of the Estimators under L_0, L_1, L_2 Losses for $n = 30$, $\theta_i = \eta \times (11 - i)$, $i = 1, \ldots, 10$, and $\boldsymbol{D} = \text{diag}(500, 50, 30, 10, 5, 1, 0.5, 0.01, 0.005, 0.001)$

	η	AD	TR	EB	ML	RR	AD_5	TR_5	EB_5	HB_5	PC_5	PC_1
L_0	0.0	0.458	0.346	0.024	0.000	0.001	0.572	0.537	0.205	0.001	0.002	0.162
	0.5	0.982	0.600	0.119	0.119	0.133	0.674	0.646	0.250	0.123	0.141	0.204
	1.0	0.995	0.819	0.296	0.320	0.316	0.817	0.799	0.373	0.295	0.557	0.330
	1.5	0.997	0.906	0.463	0.507	0.489	0.897	0.886	0.520	0.457	1.251	0.540
	2.5	0.999	0.963	0.739	0.816	0.719	0.957	0.953	0.740	0.716	3.470	1.213
L_1	0.0	0.911	0.882	0.530	0.024	0.379	0.933	0.926	0.805	0.029	0.498	0.900
	0.5	0.997	0.939	0.707	0.797	0.766	0.952	0.947	0.823	0.804	0.899	0.905
	1.0	0.999	0.974	0.883	0.953	0.871	0.974	0.972	0.874	0.915	2.102	0.920
	1.5	0.999	0.986	0.937	0.996	0.916	0.986	0.984	0.924	0.947	4.107	0.945
	2.5	0.999	0.994	0.977	1.015	0.958	0.994	0.993	0.965	0.975	10.521	1.025
L_2	0.0	0.999	0.999	0.998	0.083	0.993	0.999	0.999	0.999	0.084	0.999	0.999
	0.5	1.000	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	1.000	0.999
	1.0	1.000	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	1.001	0.999
	1.5	1.000	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	1.003	0.999
	2.5	1.000	0.999	0.999	0.999	0.999	1.000	1.000	0.999	0.999	1.009	1.000

Table 3. Relative Efficiencies of the Estimators under L_0, L_1, L_2 Losses for $n = 30$, $\theta_i = \eta \times i$, $i = 1, \ldots, 10$, and $\boldsymbol{D} = \text{diag}\,(300, 250, 200, 150, 100, 4, 3, 2, 1, 1)$

	η	AD	TR	EB	ML	RR	AD_5	TR_5	EB_5	HB_5	PC_5	PC_1
L_0	0.0	0.501	0.410	0.091	0.000	0.001	0.704	0.593	0.499	0.001	0.009	0.704
	0.5	0.999	0.826	0.111	0.031	0.846	0.706	0.596	0.501	0.137	0.023	0.704
	1.0	0.999	0.944	0.185	0.115	0.957	0.713	0.607	0.507	0.187	0.064	0.705
	1.5	0.999	0.974	0.292	0.226	0.980	0.724	0.625	0.516	0.272	0.132	0.706
	2.5	0.999	0.990	0.492	0.431	0.992	0.755	0.674	0.549	0.480	0.349	0.710
L_1	0.0	0.761	0.713	0.515	0.007	0.159	0.863	0.810	0.764	0.007	0.498	0.900
	0.5	0.999	0.920	0.533	0.477	0.929	0.865	0.812	0.765	0.536	0.509	0.900
	1.0	0.999	0.974	0.591	0.556	0.980	0.868	0.818	0.768	0.595	0.540	0.900
	1.5	0.999	0.988	0.658	0.625	0.991	0.874	0.827	0.774	0.649	0.593	0.901
	2.5	0.999	0.995	0.764	0.735	0.996	0.890	0.853	0.792	0.758	0.761	0.902
L_2	0.0	0.999	0.999	0.999	0.071	0.996	0.999	0.999	0.999	0.071	0.999	0.999
	0.5	1.000	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
	1.0	1.000	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
	1.5	1.000	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
	2.5	1.000	1.000	0.999	0.999	1.000	0.999	0.999	0.999	0.999	0.999	0.999

Table 4. Relative Efficiencies of the Estimators under L_0, L_1, L_2 Losses for $n = 30$, $\theta_i = \eta \times (11 - i)$, $i = 1, \ldots, 10$, and $\boldsymbol{D} = \text{diag}\,(300, 250, 200, 150, 100, 4, 3, 2, 1, 1)$

	η	AD	TR	EB	ML	RR	AD_5	TR_5	EB_5	HB_5	PC_5	PC_1
L_0	0.0	0.499	0.425	0.087	0.001	0.001	0.703	0.593	0.498	0.058	0.010	0.704
	0.5	0.600	0.528	0.144	0.091	0.087	0.716	0.612	0.514	0.176	0.092	0.729
	1.0	0.755	0.696	0.306	0.312	0.331	0.748	0.666	0.562	0.373	0.336	0.803
	1.5	0.853	0.811	0.524	0.599	0.708	0.791	0.736	0.639	0.582	0.744	0.927
	2.5	0.936	0.915	0.862	1.061	1.770	0.873	0.860	0.817	0.889	2.048	1.322
L_1	0.0	0.759	0.720	0.506	0.065	0.089	0.863	0.810	0.763	0.097	0.498	0.900
	0.5	0.810	0.774	0.537	0.437	0.363	0.869	0.819	0.770	0.485	0.540	0.908
	1.0	0.886	0.857	0.624	0.693	0.803	0.884	0.844	0.792	0.730	0.665	0.933
	1.5	0.932	0.912	0.742	0.814	1.120	0.904	0.877	0.827	0.816	0.872	0.975
	2.5	0.970	0.961	0.918	1.014	1.617	0.941	0.934	0.910	0.948	1.537	1.108
L_2	0.0	0.989	0.985	0.935	0.161	0.228	0.997	0.996	0.995	0.160	0.990	0.998
	0.5	0.992	0.990	0.938	0.709	0.575	0.997	0.996	0.996	0.708	0.991	0.999
	1.0	0.996	0.995	0.949	0.978	1.033	0.998	0.997	0.996	0.977	0.9938	0.999
	1.5	0.997	0.997	0.966	0.987	1.220	0.998	0.997	0.996	0.987	0.997	0.999
	2.5	0.999	0.998	0.989	0.993	1.229	0.999	0.998	0.998	0.995	1.010	1.001

Table 5. Relative Efficiencies of the Estimators under L_0, L_1, L_2 Losses in the case that $n = 30$, $\theta_i = 0$ for $i = 1, \cdots, 5$, $\theta_i = \eta \times i$, $i = 6, \ldots, 10$, and $\boldsymbol{D} = \text{diag}\,(300, 250, 200, 150, 100, 4, 3, 2, 1, 0.001)$

	η	AD	TR	EB	ML	RR	AD_5	TR_5	EB_5	HB_5	PC_5	PC_1
L_0	0.0	0.478	0.383	0.073	0.000	0.001	0.704	0.593	0.499	0.001	0.009	0.704
	0.2	0.997	0.547	0.074	0.005	0.432	0.704	0.593	0.499	0.120	0.009	0.704
	0.8	0.999	0.912	0.102	0.045	0.934	0.704	0.593	0.499	0.138	0.009	0.704
	1.6	0.999	0.975	0.238	0.192	0.982	0.704	0.593	0.499	0.249	0.009	0.704
	3.0	0.999	0.992	0.490	0.448	0.995	0.704	0.593	0.499	0.549	0.009	0.704
L_1	0.0	0.749	0.699	0.505	0.005	0.155	0.863	0.810	0.764	0.007	0.498	0.900
	0.2	0.998	0.785	0.506	0.335	0.727	0.863	0.810	0.764	0.406	0.498	0.900
	0.8	0.999	0.959	0.534	0.506	0.970	0.863	0.810	0.764	0.559	0.498	0.900
	1.6	0.999	0.988	0.625	0.602	0.992	0.863	0.810	0.764	0.632	0.498	0.900
	3.0	0.999	0.996	0.758	0.739	0.997	0.863	0.810	0.764	0.788	0.498	0.900
L_2	0.0	0.999	0.999	0.999	0.050	0.996	0.999	0.999	0.999	0.071	0.999	0.999
	0.2	1.000	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
	0.8	1.000	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
	1.6	1.000	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
	3.0	1.000	1.000	0.999	0.999	1.000	0.999	0.999	0.999	0.999	0.999	0.999

through simulation experiments, and the relative efficiencies $R_j(\omega, \boldsymbol{\delta})/R_j(\omega, \widehat{\boldsymbol{\beta}})$, $j = 0, 1, 2$, of estimator $\boldsymbol{\delta}$ over $\widehat{\boldsymbol{\beta}}$ are reported. The simulation experiments are done in the following five cases for $p = 10$, $n = 30$:

Case 1: $\boldsymbol{D} = \text{diag}\,(500, 50, 30, 10, 5, 1, 1, 1, 1, 1)$, $\theta_i = \eta \times i$, $i = 1, \ldots, 10$ and $\eta = 0.0$, 0.5, 1.0, 1.5 and 2.5.

Case 2: $\boldsymbol{D} = \text{diag}\,(500, 50, 30, 10, 5, 1, 0.5, 0.01, 0.005, 0.001)$, $\theta_i = \eta \times (11 - i)$, $i = 1, \ldots, 10$ and $\eta = 0.0$, 0.5, 1.0, 1.5 and 2.5.

Case 3: $\boldsymbol{D} = \text{diag}\,(300, 250, 200, 150, 100, 4, 3, 2, 1, 1)$, $\theta_i = \eta \times i$, $i = 1, \ldots, 10$ and $\eta = 0.0$, 0.5, 1.0, 1.5 and 2.5.

Case 4: $\boldsymbol{D} = \text{diag}\,(300, 250, 200, 150, 100, 4, 3, 2, 1, 1)$, $\theta_i = \eta \times (11 - i)$, $i = 1, \ldots, 10$ and $\eta = 0.0$, 0.5, 1.0, 1.5 and 2.5.

Case 5: $\boldsymbol{D} = \text{diag}\,(300, 250, 200, 150, 100, 4, 3, 2, 1, 0.001)$, $\theta_i = 0$ for $i = 1, \cdots, 5$, $\theta_i = \eta \times i$, $i = 6, \ldots, 10$ and $\eta = 0.0$, 0.2, 0.8, 1.6 and 3.0.

The relative efficiencies of the above estimators for the five cases are given in Tables 1, 2, 3, 4 and 5, respectively. Form these tables, the following conclusions can be drawn.

(1) The empirical Bayes estimators EB and ML for $q = 0$ have very nice risk behaviors; they are highly recommended in the case of multicollinearity. Since EB is minimax, its risk is well behaved.

(2) The adaptive ridge regression estimator AD is always improved on by the trancated estimator TR. However, their risk performances are much worse than EB and ML. The ordinary adaptive ridge regression estimator RR exceeds the minimax risk for large η as seen in Table 4 since it is not minimax.

(3) When $\boldsymbol{\gamma}_1 \in \boldsymbol{R}^5$, the risk gains of AD_5 and TR_5 are not so much. The shrinkage of $\boldsymbol{\alpha}$ in the hierarchical empirical Bayes estimator HB_5 is effective, and so it behaves much better than AD_5 and TR_5. In the situation, HB_5 is recomended.

(4) The estimators ML and HB_5 have significant risk gains near $\eta = 0.0$ under the L_1 and L_2 loss functions as well as L_0 loss.

(5) Table 5 treats the risk performances under the null hypothsis $H_0 : \theta_1 = \cdots = \theta_5 = 0$, so that the principal coponent estimator PC_5 is an appropriate procedure. In this case, the estimators EB, ML and HB also have nice risk behaviours with their risks much smaller than those of AD, TR and RR.

(6) Through the first four tables, we see that the principal component estimator PC_5 has the smallest risks for $\boldsymbol{\theta}$ near $\mathbf{0}$, while the risk of it gets larger as $\|\boldsymbol{\theta}\|$ increases. This means that the use of the principal component estimator PC_5 is risky.

We shall provide empirical studies for two sets of data.

Example 1. (*Response Surface*) We first consider the acetylene data analyzed by Marquardt and Snee (1975). The data consisted of 16 observations on the response variable y (conversion of n-heptane to acetylene), three predictor variables a_1 (reactor temperature), a_2 (ratio of H_2 to n-heptane) and a_3 (contact time). It is anticipated that the response y is on a quadratic response surface, that is, y is expressed by the model

$$y = \beta_0 + \sum_{i=1}^{3} \beta_i a_i + \sum_{i=1}^{3} \beta_{ii} a_i^2 + \sum_{i=1}^{3} \sum_{j=i+1}^{3} \beta_{ij} a_i a_j + \varepsilon.$$

Such an analysis includes multicollearity and the above data have been repeatedly analyzed by Beisley (1984), Casella (1985) and Wetherill (1986). Before any computation were done, the means were removed from the variables y, a_1, a_2 and a_3. Then the squares and cross products of the predictor variables were computed and standardized.

The eigenvalues of the matrix $\boldsymbol{A}^t\boldsymbol{A}$ are 4.205, 2.162, 1.138, 1.040, 0.385, 0.0495, 0.0136, 0.00512 and 0.0000969, and so the eigenvalues of $(\boldsymbol{A}^t\boldsymbol{A})^{-1}$ are given by

$$\boldsymbol{D} = \text{diag}\,(10316., 195.015, 73.393, 20.186, 2.595, 0.961, 0.878, 0.462, 0.237),$$

which means that the problem is highly ill-conditioned. The ridge curves of the ridge regression estimate $\widehat{\boldsymbol{\beta}}^R(\lambda)$ given by (1.2) are drawn for $k = 1/\lambda \in [0, 0.005]$ in Figure 1 where the horizontal axis denotes the value of $k = 1/\lambda$. This figure demonstrates that each ridge regression estimator is instable for smaller k, or larger λ because of the multicollinearity.

We shall investigate how the proposed ridge-type regression estimators of the coefficients $\boldsymbol{\beta}$ behave for the ill-conditioned data. The estimators we treat are the least squares $\widehat{\boldsymbol{\beta}}$ (denoted by LS), the adaptive ridge regression estimators shrunken towards zero: $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{TR}, \mathbf{0})$ (TR) and $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{RR}, \mathbf{0})$ (RR) and the empirical Bayes ridge regression estimators shrunken towards zero: $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{EB}, \mathbf{0})$ (EB) and $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{ML}, \mathbf{0})$ (ML). Since the first four eigenvalues d_1, d_2, d_3, d_4 are not small, we may consider the linear subspace (4.3) constructed by eigenvectors of $(\boldsymbol{A}^t\boldsymbol{A})^{-1}$ with deleting the eigenvectors corresponding to the

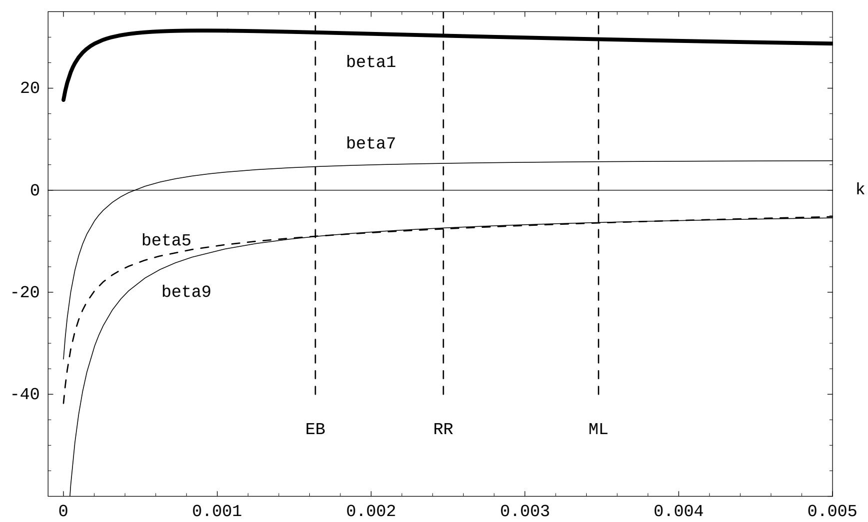

Figure 1. Curves of the Ridge Estimates of β_1, β_5, β_7 and β_9 (The horizontal axis denotes the values of $k = 1/\lambda$. The dotted lines EB, RR, ML show the values of $1/\hat{\lambda}_{EB}$, $1/\hat{\lambda}_{RR}$, $1/\hat{\lambda}_{ML}$, respectively.)

Table 6. Estimates of λ (or k) and $\boldsymbol{\beta}$ for the Eight Estimators LS, TR, EB, ML, RR, EB_4, HB_4 and PC_4

	LS	TR	EB	ML	RR	EB_4	HB_4	PC_4
$\hat{\lambda}$		$1.7 \cdot 10^6$	610	287	404	2280	398	
$\hat{k}$		$5.6 \cdot 10^{-7}$	0.0016	0.0034	0.0024	0.0004	0.0025	
$\widehat{\beta}_1$	17.2	17.9	30.9	29.6	30.3	30.7	30.3	23.2
$\widehat{\beta}_2$	10.7	10.7	10.3	10.2	10.2	10.4	10.2	10.1
$\widehat{\beta}_3$	-28.0	-28.0	-10,0	-11.8	-10.9	-10.4	-10.9	-18.8
$\widehat{\beta}_4$	-22.0	-22.0	-19.9	-18.3	-19.1	-21.2	-19.1	-5.1
$\widehat{\beta}_5$	-83.2	-82.8	-9.1	-6.3	-7.4	-19.3	-7.3	-2.7
$\widehat{\beta}_6$	-12.0	-12.0	-9.1	-7.4	-8.3	-10.7	-8.2	6.3
$\widehat{\beta}_7$	-33.1	-32.9	4.6	5.59	5.3	-0.3	5.3	8.0
$\widehat{\beta}_8$	-4.1	-4.1	-3.2	-2.9	-3.0	-3.5	-3.0	-2.4
$\widehat{\beta}_9$	-41.8	-41.6	-9.0	-6.4	-7.6	-14.7	-7.5	-2.1
CV	299	276	106	103	100	219	110	100

Table 7. Estimates of λ (or k) and $\boldsymbol{\theta}$ for the Eight Estimators LS, TR, EB, ML, RR, EB_4, HB_4 and PC_4

	d_i	LS	TR	EB	ML	RR	EB_4	HB_4	PC_4
$\hat{\lambda}$			$1.7 \cdot 10^6$	610	287	404	2280	398	
$\hat{k}$			$5.6 \cdot 10^{-7}$	0.0016	0.0034	0.0024	0.0004	0.0025	
$\hat{\theta}_1$	10316.04	97.5	97.0	5.4	2.6	3.7	17.7	3.6	0.0
$\hat{\theta}_2$	195.02	23.9	23.9	18.1	14.2	16.1	22.0	16.0	0.0
$\hat{\theta}_3$	73.39	-17.5	-17.5	-15.6	-14.0	-14.8	-17.0	-14.8	0.0
$\hat{\theta}_4$	20.19	-9.9	-9.9	-9.6	-9.3	-9.5	-9.8	-9.5	0.0
$\hat{\theta}_5$	2.60	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4
$\hat{\theta}_6$	0.96	-10.9	-10.8	-10.8	-10.8	-10.8	-10.8	-10.8	-10.8
$\hat{\theta}_7$	0.88	27.6	27.6	27.6	27.6	27.6	27.6	27.6	27.6
$\hat{\theta}_8$	0.46	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
$\hat{\theta}_9$	0.24	-16.2	-16.2	-16.2	-16.2	-16.2	-16.2	-16.2	-16.2

four largest eigenvalues. We thus deal with the principal component estimator $\widehat{\boldsymbol{\beta}}^{PC}$ (PC_4) under the subspace and empirical Bayes and hierarchical empirical Bayes ridge regression estimators shrunken towards the subspace : $\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{EB_4}, \widehat{\boldsymbol{\gamma}}_2)$ (EB_4) and $\widehat{\boldsymbol{\beta}}^{HB}(\hat{\lambda}_{HB}, \hat{\tau}_{HB})$ (HB_4).

The estimates of λ (or k) and $\boldsymbol{\beta}$ for the above procedures are given in Table 6. Since $\hat{\lambda}_{TR}$ is very large, the minimax adaptive ridge regression estimate $\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{TR}, \mathbf{0})$ is so close to the LS estimate $\widehat{\boldsymbol{\beta}}$, which implies that $\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{TR}, \mathbf{0})$ is not useful in the multicollearity case. From Figure 1 and Table 6, on the other hand, it is seen that $\hat{\lambda}_{EB}$, $\hat{\lambda}_{ML}$, $\hat{\lambda}_{RR}$ are estimated appropriately and that the resulting estimators $\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{EB}, \mathbf{0})$, $\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{ML}, \mathbf{0})$ and $\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{RR}, \mathbf{0})$ are well stabilized. For the hierarchical empirical Bayes estimate $\widehat{\boldsymbol{\beta}}^{HB}(\hat{\lambda}_{HB}, \hat{\tau}_{HB})$, we observe that $\hat{\lambda}_{HB} = 398$ and $\hat{\tau}_{HB} = 0$, so that the estimate is close to that of $\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{RR}, \mathbf{0})$ in this case. The principal component estimator $\widehat{\boldsymbol{\beta}}^{PC}$ gives estimates different from the ridge type estimators. Table 7 gives similar estimates in the canonical model with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_9)^t = \boldsymbol{H}\boldsymbol{\beta}$ and it reveals that the estimates by EB, ML, RR and HB get more shrunken for larger d_i.

The primary purpose of regression models may be prediction with the help of many independent variables, and the predictors constructed by the ridge-type estimators proposed in this paper are anticipated to have good performances. The prediction error may be estimated by the cross-varidation method as described in Srivastava (2002, p322). The estimates of the prediction errors for the above considered estimators are given at the last row as CV in Table 6. It reveals that the use of the estimators EB, ML, RR, HB_4 and PC_4 provides much smaller prediction errors than the least squares estimator LS. It is interesting to note that the ridge-type estimators EB, ML, RR and HB_4 give estimates different from the principal component estimator PC_4, but the estimates of the prediction errors for both procedures are quite similar. The estimate of the prediction error of PC_1 by the cross-varidation method is given by 270, which is much larger than that of PC_4. In use of the principal component procedure, statisticians need to determine how many eigenvalues should be dropped, which means that there is a room for arbitrariness by analysts. On the other hand, the ridge-type procedures can be computed automatically without any transformation or any decision by analysts, and the resulting predictiors have as small prediction errors as the principal component procedure as well as their estimates are quite stable as seen in Figure 1.

Through this example, it is seen that the empirical Bayes ridge regression estimators $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{EB}, \mathbf{0})$, $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{ML}, \mathbf{0})$ and the hierarchical type estimator $\widehat{\boldsymbol{\beta}}^{HB}(\hat{\lambda}_{HB}, \hat{\tau}_{HB})$ are quite useful from practical sense as well as $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{RR}, \mathbf{0})$ gives stable estimates. ■

Example 2. (*Hednic Regression*) Hednic regressions regress the prices of goods on the characteristics that describe the goods, and they are used in many applications in the fields of Business and Econometrics. It is well known that the hednic regression is often faced with the multicollinearity among the characteristics. For the hednic regression and the related problems, see Rosen (1974), Feenstra (1995), Gilley and Pace (1995) and their references. As an illustlative example, we here treat the data given by Exhibit 2.2 of Sen and Srivastava (1990), who regressed, based on 26 observations, selling price of house on 13 explanatory variables: a_1 (number of bedrooms), a_2 (floor space), a_3 (number of fireplaces), a_4 (number of rooms), a_5 (storm windows), a_6 (front footage of lot), a_7 (annual taxes), a_8 (number of bathrooms), a_9 (construction), a_{10} (garage size), a_{11} (condition), a_{12} (location), a_{13} (location). It is anticipated that correlations exist among the characteristics such as number of bedrooms, number of rooms, floor space, front footage of lot and number of bathrooms.

When we analyse the original data by a linear regression model with an intercept term, the eigenvalues of the matrix $\boldsymbol{A}^t\boldsymbol{A}$ are 6.06×10^7, 1.92×10^6, 4.06×10^4, 42.113, 17.654, 11.306, 7.272, 5.240, 3.694, 2.960, 2.027, 1.580, 1.134, 0.325, and $\boldsymbol{D}$ is given by

$$\boldsymbol{D} = \mathrm{diag}\,(3.074, 0.881, 0.632, 0.493, 0.337, 0.270, 0.190, 0.137, \\ 0.088, 0.056, 0.023, 2.45 \times 10^{-4}, 5.18 \times 10^{-7}, 1.64 \times 10^{-8}).$$

Although $d_1 = 3.074$ is not large, the condition number $d_1/d_{13} = 1.8 \times 10^8$ is very large and the problem is ill-conditioned. Although the least squares estimate of the regression coefficient vector is

$$\widehat{\boldsymbol{\beta}} = (17.3, -5.7, 0.0, 4.3, 2.2, 9.7, 0.3, 0.0, 1.3, 2.6, 3.8, -0.7, 1.7, 6.8),$$

the ridge curves for β_0, β_3 and β_{11} in Figure 2 imply that the LS estimate is not stable. On the other hand, we have that $\hat{\lambda}_{RR} = 311.27$, $\hat{\lambda}_{TR} = 4.25$, $\hat{\lambda}_{EB} = 1.56$ and $\hat{\lambda}_{ML} = 0.89$, which yield the adaptive and empirical Bayes ridge regression estimates of $\boldsymbol{\beta}$. Their estimates of β_0, β_3 and β_{11} are described as RR, TR, EB and ML in Figure 2, which demonstrates that RR is very close to the LS estimate, so that it is not useful. On the other hand, it seems that EB and ML present more stable estimates, and the empirical Bayes ridge regression estimates of $\boldsymbol{\beta}$ are give by

$$\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{EB}, \mathbf{0}) = (6.9, -5.0, 0.0, 1.2, 3.4, 8.1, 0.3, 0.0, 2.1, 4.4, 4.0, 0.1, 0.8, 6.1),$$

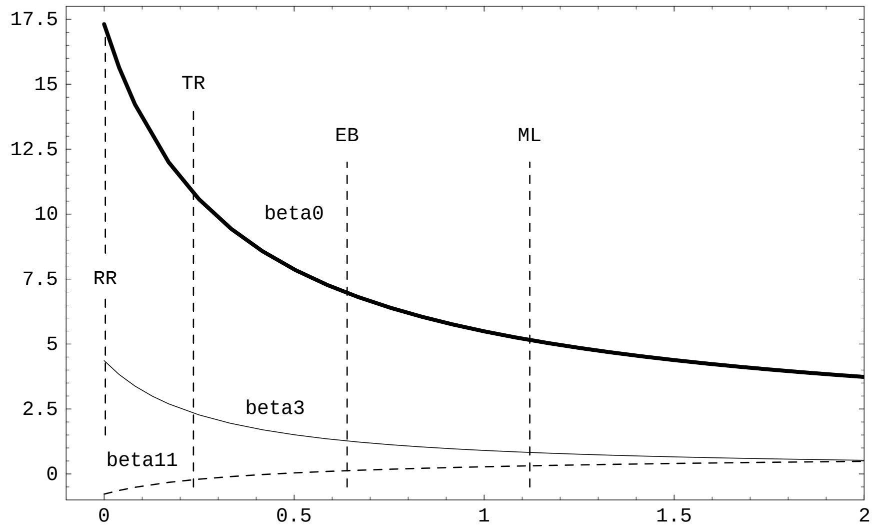

Figure 2. Curves of the Ridge Estimates of β_0, β_3 and β_{11} (The dotted lines RR, TR, EB, ML show the values of $1/\hat{\lambda}_{RR}$, $1/\hat{\lambda}_{TR}$, $1/\hat{\lambda}_{EB}$, $1/\hat{\lambda}_{ML}$, respectively.)

$$\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{ML}, \mathbf{0}) = (5.1, -4.7, 0.0, 0.8, 3.6, 7.4, 0.3, 0.0, 2.3, 4.6, 3.9, 0.3, 0.6, 5.6).$$

Since $\hat{\lambda}_{EB}$ and $\hat{\lambda}_{ML}$ are automatically computed without any transformation or any consideration, we can obtain stable estimates of the regression coefficients directly from the original data without any careful analysis by statisticians.

We next analyse the data after standardizing some of the explanatory variables. Then the eigenvalues of the matrix $\boldsymbol{A}^t\boldsymbol{A}$ are 131.141, 16.928, 8.321, 6.777, 4.765, 3.962, 3.183, 2.315, 1.448, 0.752, 0.572, 0.478, 0.053, 0.048, and $\boldsymbol{D}$ is given by

$$\boldsymbol{D} = \mathrm{diag}\,(20.556, 18.554, 2.088, 1.745, 1.329, 0.690, \\ 0.431, 0.314, 0.252, 0.209, 0.147, 0.120, 0.059, 0.008).$$

The least squares estimate of $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}} = (43.6, -37.6, 45.4, 4.3, 19.4, 9.7, 13.8, -5.7, 1.3, 2.6, 3.8, -0.7, 1.7, 6.8)$$

and the ridge curves for β_1, β_2, β_7 and β_{11} are drawn in Figure 3. Since $\hat{\lambda}_{RR} = 342.08$ and $\hat{\lambda}_{TR} = 186.59$, the resulting estimates $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{RR}, \mathbf{0})$ and $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{TR}, \mathbf{0})$ are close to the LS estimate. For the empirical Bayes estimates of λ, we observe that $\hat{\lambda}_{EB} = 21.22$ and $\hat{\lambda}_{ML} = 11.04$, which result in the empirical Bayes ridge regression estimates

$$\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{EB}, \mathbf{0}) = (38.6, -18.4, 29.1, 1.9, 10.2, 8.5, 13.1, 4.9, 3.2, 4.5, 4.5, 1.3, 2.0, 8.3),$$

$$\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{ML}, \mathbf{0}) = (36.3, -12.3, 23.1, 1.2, 7.2, 8.0, 12.3, 8.3, 4.2, 5.0, 4.8, 2.0, 2.4, 8.9).$$

Figure 3 illustrates that both procedures give more stable estimates.

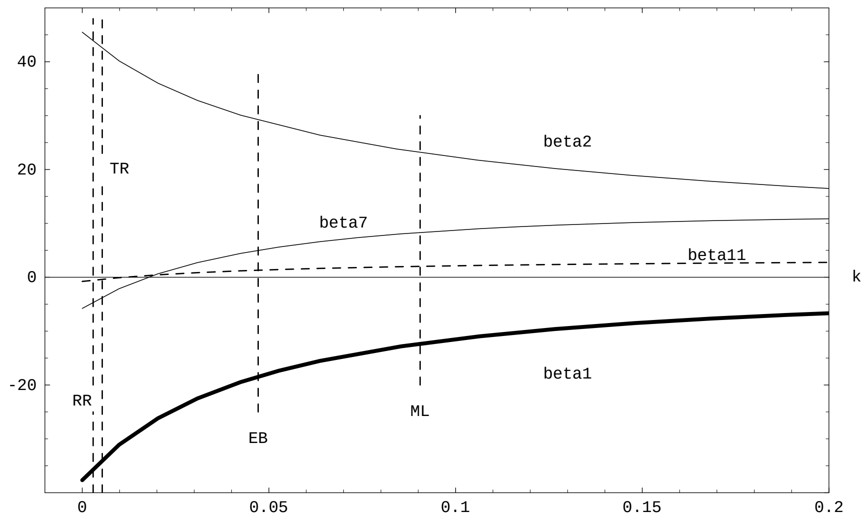

Figure 3. Curves of the Ridge Estimates of β_1, β_2, β_7 and β_{11} (The dotted lines RR, TR, EB, ML show the values of $1/\hat{\lambda}_{RR}$, $1/\hat{\lambda}_{TR}$, $1/\hat{\lambda}_{EB}$, $1/\hat{\lambda}_{ML}$, respectively.)

When the eigenvectors corresponding to the five largest eigenvalues of $(\boldsymbol{A}^t\boldsymbol{A})^{-1}$ are deleted, the principal component estimate is

$$\widehat{\boldsymbol{\beta}}^{PC} = (12.4, -7.4, -2.1, -2.4, 22.6, 1.8, -5.1, 4.4, 1.6, 11.1, 7.1, -0.9, 8.5, 14.6),$$

which is different from the ridge type estimates. For the empirical Bayes estimate shrunken toward the $\widehat{\boldsymbol{\beta}}^{PC}$, we have $\hat{\lambda}_{EB} = 82.77$, and then the empirical Bayes ridge regression estimate shrunken toward the $\widehat{\boldsymbol{\beta}}^{PC}$ is

$$\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{EB}, \widehat{\boldsymbol{\gamma}}_2) = (38.4, -20.7, 21.8, 1.2, 19.3, 8.0, 11.7, 9.0, 3.2, 5.1, 4.6, 1.1, 2.3, 8.3).$$

Since $\hat{\lambda}_{HB} = 31.09$ and $\hat{\tau}_{HB} = 0.0$, the hierarchical empirical Bayes ridge regression estimate is given by

$$\widehat{\boldsymbol{\beta}}^{HB}(\hat{\lambda}_{HB}, \hat{\tau}_{HB}) = (39.7, -22.1, 32.4, 2.3, 12.0, 8.8, 13.4, 2.9, 2.8, 4.1, 4.4, 0.8, 1.9, 8.0).$$

It seems that the empirical Bayes ridge regression estimators $\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{EB}, \mathbf{0})$, $\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{ML}, \mathbf{0})$, the empirical Bayes estimator shrunken toward the subspace $\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{EB}, \widehat{\boldsymbol{\gamma}}_2)$ and the hierarchical empirical Bayes ridge regression estimator $\widehat{\boldsymbol{\beta}}^{HB}(\hat{\lambda}_{HB}, \hat{\tau}_{HB})$ give stable estimates. The adaptive ridge regression estimators $\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{TR}, \mathbf{0})$ and $\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{RR}, \mathbf{0})$ are not good in this example while $\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{RR}, \mathbf{0})$ gives the stable estimate in Example 1.

The estimates of the prediction errors by the cross-varidation method are given by $\{LS, 60\}$, $\{TR, 57\}$, $\{EB, 51\}$, $\{ML, 55\}$, $\{RR, 57\}$, $\{EB_5, 56\}$, $\{HB, 54\}$, $\{PC_1, 77\}$, $\{PC_5, 136\}$. In this example, the risk gains by the ridge-type estimators are small. Of them, the empirical Bayes estimator EB gives the smallest prediction-error estimate while the principal component estimators are not good.

■

5 Application to k Sample Problem with Unequal Variances

The results given in Sections 2 and 3 can be applied to the problem of estimating the means of k populations simultaneously based on samples with unequal sizes.

For $i = 1, \ldots, k$, let $X_{i1}, \ldots, X_{ir_i}$ be a random sample from a population following normal distribution $\mathcal{N}(\mu_i, \sigma^2)$ with the unknown mean μ_i and unknown common variance σ^2. Let $\overline{X}_{i\cdot} = \sum_{j=1}^{r_i} X_{ij}/r_i$, $i = 1, \ldots, k$, and $S = \sum_{i=1}^{k}\sum_{j=1}^{r_i}(X_{ij} - \overline{X}_{i\cdot})^2$. Then $\overline{X}_{i\cdot}, \ldots, \overline{X}_{k\cdot}$ and S are mutually independent and have distributions

$$\begin{aligned} \overline{X}_{i\cdot} &\sim \mathcal{N}(\mu_i, \sigma^2/r_i), \quad i = 1, \ldots, k \\ S &\sim \sigma^2\chi_n^2, \quad n = \sum_{i=1}^{k}(r_i - 1). \end{aligned}$$

Let $\boldsymbol{X} = (\overline{X}_{1\cdot}, \ldots, \overline{X}_{k\cdot})^t$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^t$ and $\boldsymbol{R} = \mathrm{diag}\,(r_1, \ldots, r_k)$. We consider the problem of estimating $\boldsymbol{\mu}$ by $\widehat{\boldsymbol{\mu}}$ under the loss function

$$L(\omega, \widehat{\boldsymbol{\mu}}, \boldsymbol{R}^2) = (\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})^t \boldsymbol{R}^2 (\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})/\sigma^2, \tag{5.1}$$

for $\omega = (\boldsymbol{\mu}, \sigma^2)$, the unknown parameters. It is seen that this problem is in the framework described in (3.1). Since the sample sizes $r_1, \ldots, r_k$ are unequal, the sample means with smaller sizes should be shrunken more. It may also be reasonable that $\boldsymbol{X}$ is shrunken toward the subspace $V(\boldsymbol{j}) = \{\mu\boldsymbol{j} \,|\, \mu \in \boldsymbol{R}, \boldsymbol{j} = (1, \ldots, 1)^t \in \boldsymbol{R}^k\}$ with the same mean μ. Then the empirical Bayes estimator given by $\widehat{\boldsymbol{\theta}}^B(\lambda, \widehat{\boldsymbol{\alpha}})$ in Section 2 can be written in the k sample model as

$$\widehat{\boldsymbol{\mu}}^B(\lambda, \overline{X}_{\cdot\cdot}) = \boldsymbol{X} - (\boldsymbol{I} + \lambda\boldsymbol{R})^{-1}(\boldsymbol{X} - \overline{X}_{\cdot\cdot}\boldsymbol{j}), \tag{5.2}$$

where $\overline{X}_{\cdot\cdot}$ is the total mean $\sum_{i=1}^{k} r_i\overline{X}_{i\cdot}/\sum_{i=1}^{k} r_i = \sum_{i=1}^{k}\sum_{j=1}^{k} X_{ij}/\sum_{i=1}^{k} r_i$. It is also written componentwise as

$$\widehat{\mu}_i^B(\lambda, \overline{X}_{\cdot\cdot}) = \overline{X}_{i\cdot} - \frac{1}{1 + \lambda r_i}(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot}).$$

The estimator $\hat{\lambda}$ of λ is based on $\boldsymbol{X} - \overline{X}_{\cdot\cdot}\boldsymbol{j}$ and S, and the conditions for the minimaxity of the empical Bayes estimator $\widehat{\boldsymbol{\mu}}^B(\hat{\lambda}, \overline{X}_{\cdot\cdot})$ are obtained from Theorem 1.

As candidates of estimators of λ, we employ the estimators (3.9), (3.10) and (2.4), which are, respectively, given by

$$\begin{aligned} \hat{\lambda}_{AD} &= \frac{n+2}{k-3}\sum_{i=1}^{k} r_i(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2/\{S \min_i r_i\} + \lambda_0, \\ \hat{\lambda}_{TR} &= \max\left\{\frac{(n+2)\{\max_i(r_i^{-1}) + 1\}}{(k-3)S}\sum_{i=1}^{k}\frac{r_i(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2}{1 + r_i},\ \lambda_0\right\}, \\ \hat{\lambda}_{EB} &= \max\left(\lambda^*, \lambda_0\right), \end{aligned} \tag{5.3}$$

where λ^* is a root of the equation

$$\sum_{i=1}^{k}\frac{r_i(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2/S}{1 + r_i\lambda^*} = \frac{k-3}{n+2}, \tag{5.4}$$

and λ_0 is a solution of the equation

$$\sum_{i=1}^k \left(1 - \frac{r_i^{-1}}{\sum_{j=1}^k r_j^{-1}}\right) \frac{1 - r_i \min_j(r_j^{-1})}{1 + r_i\lambda_0} = \frac{k-3}{2}. \tag{5.5}$$

Then from Propositions 1, 2 and 3, it follows that $\widehat{\boldsymbol{\mu}}^B(\hat{\lambda}_{AD}, \overline{X}..)$, $\widehat{\boldsymbol{\mu}}^B(\hat{\lambda}_{TR}, \overline{X}..)$ and $\widehat{\boldsymbol{\mu}}^B(\hat{\lambda}_{EB}, \overline{X}..)$, denoted by AD_9, TR_9 and EB_9, improve on $\boldsymbol{X}$ under the loss function (5.1). We can also consider a simple estimator of λ given by

$$\hat{\lambda}_{RR} = \frac{\sum_{i=1}^k r_i(\overline{X}_{i\cdot} - \overline{X}..)^2/(k-1)}{S/n},$$

yielding the simple estimator $\widehat{\boldsymbol{\mu}}^B(\hat{\lambda}_{RR}, \overline{X}..)$, denoted by RR_9.

The hierarchical empirical Bayes estimator (2.7) is written in this case as

$$\widehat{\mu}_i^{HB} = \overline{X}_{i\cdot} - \frac{1}{1 + r_i\hat{\lambda}_{HB}} \left(\overline{X}_{i\cdot} - \frac{\hat{\tau}_{HB}\sum_{j=1}^k r_j\overline{X}_{j\cdot}/(1 + r_j\hat{\lambda}_{HB})}{1 + \hat{\tau}_{HB}\sum_{j=1}^k r_j/(1 + r_j\hat{\lambda}_{HB})}\right),$$

where $\hat{\lambda}_{HB} = \max(\lambda^{**}, 0)$ and $\hat{\tau}_{HB} = \max(\tau^{**}, 0)$, and $(\lambda^{**}, \tau^{**})$ is a solution of the equations

$$\begin{aligned}
&\sum_{i=1}^k \frac{r_i^2}{(1 + r_i\lambda^{**})^2} \left(\overline{X}_{i\cdot} - \frac{\tau^{**}\sum_{j=1}^k r_j\overline{X}_{j\cdot}/(1 + r_j\lambda^{**})}{1 + \tau^{**}\sum_{j=1}^k r_j/(1 + r_j\lambda^{**})}\right)^2 \\
&\qquad = \frac{S}{n}\left(\sum_{j=1}^k \frac{r_j}{1 + r_j\lambda^{**}} - \frac{\tau^{**}\sum_{j=1}^k r_j^2/(1 + r_j\lambda^{**})^2}{1 + \tau^{**}\sum_{j=1}^k r_j/(1 + r_j\lambda^{**})}\right), \\
&\left(\sum_{j=1}^k \frac{r_j\overline{X}_{j\cdot}}{1 + r_j\lambda^{**}}\right)^2 = \frac{S}{n}\sum_{j=1}^k \frac{r_j}{1 + r_j\lambda^{**}}\left(1 + \tau^{**}\sum_{j=1}^k \frac{r_j}{1 + r_j\lambda^{**}}\right).
\end{aligned}$$

We shall compare the risk-performances of the empirical Bayes estimators shrunken towards the subspace $V(\boldsymbol{j})$ and the hierarchical emprirical Bayes estimator, given by

$$\widehat{\boldsymbol{\mu}}^B(\hat{\lambda}_{AD}, \overline{X}..),\ \widehat{\boldsymbol{\mu}}^B(\hat{\lambda}_{TR}, \overline{X}..),\ \widehat{\boldsymbol{\mu}}^B(\hat{\lambda}_{EB}, \overline{X}..),\ \widehat{\boldsymbol{\mu}}^{HB},$$

denoted by AD_9, TR_9, EB_9, HB_9;

and the simple empirical Bayes estimators shrunken towards 0 as

$$\widehat{\mu}_i^B(\hat{\lambda}, 0) = \overline{X}_{i\cdot} - (1 + r_i\hat{\lambda})^{-1}\overline{X}_{i\cdot},$$

given by

$$\widehat{\boldsymbol{\mu}}^B(\hat{\lambda}_{AD}, 0),\ \widehat{\boldsymbol{\mu}}^B(\hat{\lambda}_{TR}, 0),\ \widehat{\boldsymbol{\mu}}^B(\hat{\lambda}_{EB}, 0),\ \widehat{\boldsymbol{\mu}}^B(\hat{\lambda}_{ML}, 0), \widehat{\boldsymbol{\mu}}^B(\hat{\lambda}_{RR}, 0),$$

denoted by AD, TR, EB, ML, RR,

where $\hat{\lambda}_{ML}$ is the maximum likelihood estimator of λ in the marginal distribution, correspoding to (2.9), and

$$\hat{\lambda}_{RR} = \frac{(\sum_{i=1}^k r_i\overline{X}_{i\cdot}{}^2)/p}{S/n}.$$

Table 8. Relative Efficiencies of the Estimators under L_0, L_1, L_2 Losses for $\theta_i = \eta \times (10 - i)$, $i = 1, \ldots, 10$, and $\boldsymbol{R} = \boldsymbol{D}^{-1} = \text{diag}\,(2, 5, 5, 10, 10, 10, 20, 20, 20, 30)$

	η	AD	TR	EB	ML	RR	AD_9	TR_9	EB_9	HB_9	RR_9
L_0	0.00	0.592	0.333	0.118	0.029	0.677	0.614	0.378	0.187	0.265	0.689
	0.03	0.651	0.434	0.277	0.251	0.728	0.634	0.411	0.233	0.674	0.707
	0.05	0.726	0.567	0.504	0.562	0.790	0.667	0.464	0.312	0.935	0.735
	0.07	0.795	0.695	0.717	0.843	0.846	0.707	0.533	0.422	0.999	0.769
	0.10	0.869	0.835	0.905	1.049	0.902	0.770	0.645	0.611	1.000	0.821
L_1	0.00	0.742	0.477	0.210	0.062	0.804	0.764	0.535	0.320	0.281	0.817
	0.03	0.783	0.563	0.349	0.275	0.838	0.778	0.560	0.354	0.667	0.829
	0.05	0.833	0.671	0.549	0.562	0.878	0.798	0.601	0.413	0.921	0.846
	0.07	0.878	0.769	0.733	0.804	0.912	0.824	0.652	0.497	0.993	0.867
	0.10	0.924	0.868	0.883	0.959	0.946	0.863	0.732	0.642	0.999	0.898
L_2	0.00	0.829	0.579	0.288	0.093	0.875	0.851	0.645	0.427	0.294	0.889
	0.03	0.857	0.649	0.398	0.272	0.898	0.860	0.665	0.454	0.645	0.896
	0.05	0.891	0.734	0.564	0.514	0.924	0.874	0.697	0.501	0.896	0.907
	0.07	0.921	0.810	0.722	0.722	0.945	0.891	0.737	0.567	0.984	0.920
	0.10	0.951	0.886	0.858	0.869	0.966	0.916	0.796	0.683	0.999	0.940

Table 9. Relative Efficiencies of the Estimators under L_0, L_1, L_2 Losses for $\theta_i = 2 \times \eta$, $i = 1, \ldots, 10$, and $\boldsymbol{R} = \boldsymbol{D}^{-1} = \text{diag}\,(2, 10, 10, 50, 50, 80, 100, 100, 100, 100)$

	η	AD	TR	EB	ML	RR	AD_9	TR_9	EB_9	HB_9	RR_9
L_0	0.00	0.518	0.230	0.037	0.009	0.621	0.529	0.257	0.067	0.197	0.626
	0.03	0.557	0.243	0.074	0.050	0.658	0.529	0.257	0.067	0.564	0.626
	0.05	0.613	0.265	0.122	0.104	0.709	0.529	0.257	0.067	0.866	0.626
	0.07	0.676	0.294	0.172	0.156	0.764	0.529	0.257	0.067	0.979	0.626
	0.10	0.761	0.345	0.243	0.224	0.833	0.529	0.257	0.067	0.999	0.626
L_1	0.00	0.856	0.592	0.151	0.056	0.895	0.864	0.635	0.270	0.228	0.899
	0.03	0.873	0.615	0.297	0.241	0.908	0.864	0.635	0.270	0.631	0.899
	0.05	0.894	0.648	0.458	0.448	0.925	0.864	0.635	0.270	0.907	0.899
	0.07	0.916	0.686	0.575	0.583	0.941	0.864	0.635	0.270	0.989	0.899
	0.10	0.941	0.739	0.677	0.677	0.960	0.864	0.635	0.270	0.999	0.899
L_2	0.00	0.953	0.738	0.206	0.083	0.969	0.959	0.786	0.370	0.241	0.973
	0.03	0.961	0.762	0.406	0.342	0.975	0.959	0.786	0.370	0.666	0.973
	0.05	0.971	0.796	0.616	0.620	0.981	0.959	0.786	0.370	0.930	0.973
	0.07	0.979	0.832	0.754	0.783	0.986	0.959	0.786	0.370	0.995	0.973
	0.10	0.986	0.878	0.849	0.865	0.991	0.959	0.786	0.370	1.000	0.973

Every estimator $\boldsymbol{\delta}$ is evaluated by three types of risk functions $R_j(\omega, \boldsymbol{\delta})$ under the loss functions $L_j(\omega, \boldsymbol{\delta}, \boldsymbol{R}^j) = (\boldsymbol{\delta} - \boldsymbol{\mu})^t \boldsymbol{R}^j (\boldsymbol{\delta} - \boldsymbol{\mu})/\sigma^2$, called the L_j-loss, for $j = 0, 1, 2$.

The relative efficiencies $R_j(\omega, \boldsymbol{\delta})/R_j(\omega, \boldsymbol{X})$, $j = 0, 1, 2$, of estimator $\boldsymbol{\delta}$ over $\boldsymbol{X}$ are reported by simulation experiments with 50,000 replications, where the simulation experiments are done in the following two cases for $k = 10$ and $\eta =$ 0.0, 0.03, 0.05, 0.07 and 0.1:

Case 1: $\boldsymbol{R} = \text{diag}\,(2, 5, 5, 10, 10, 10, 20, 20, 20, 30)$ and $\theta_i = \eta \times (10 - i)$, $i = 1, \ldots, 10$.

Case 2: $\boldsymbol{R} = \text{diag}\,(2, 10, 10, 50, 50, 80, 100, 100, 100, 100)$ and $\theta_i = 2 \times \eta$, $i = 1, \ldots, 10$.

Since n is very large in the above two cases, S/n is supposed to be equal to σ^2 in the simulation experiment. The relative efficiensies of the above estimators for the two cases are given by Tables 8 and 9, respectively. These tables reveal that the estimators EB, ML and EB_9 have superior risk performances. Especially from Table 9, the empirical Bayes estimator EB_9 shrunken towards the subspace $V(\boldsymbol{j})$ has significantly small risks when the parameters are in $V(\boldsymbol{j})$. The risk performances of the simple estimators RR and RR_9 are poor.

6 Concluding Remarks

In a linear regression model with the multicollearity, the ridge regression estimator is known to be useful for providing stable estimates. Thus, it is important to have a suitable estimate of the parameter $\lambda = 1/k$ because it can adjust the stability of the ridge estimates as illustrated by Figures 1, 2 and 3. A reasonable method is to use an estimtator $\hat{\lambda}$ such that the resulting adaptive ridge regression estimator $\boldsymbol{\beta}^R(\hat{\lambda}) = \widehat{\boldsymbol{\beta}}^B(\hat{\lambda}, \mathbf{0})$ is better than the least squares estimator, namely, minimax in terms of risk. As demonstrated in Casella (1980, 85) and Shinozaki and Chang (1993), when the problem is ill-conditioned, usual adaptive ridge regression estimators and empirical Bayes estimators do not satisfy the conditions for the minimaxity under the squared error loss function. We thus, in this paper, have employed the Strawderman's loss function $L(\omega, \boldsymbol{\delta}, (\boldsymbol{A}^t\boldsymbol{A})^2)$, given by (1.8), and have obtained the general conditions on $\hat{\lambda}$ under which the resulting ridge-type estimators are minimax. The empirical Bayes ridge regression estimator $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{EB}, \mathbf{0})$ we propose satisfies the minimaxity conditions, and through the simulation experiments and the empirical studies given in Section 4, it has revealed that the estimator $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{EB}, \mathbf{0})$ performs reasonablly well in the multicollearity cases under the squared error loss $L(\omega, \boldsymbol{\delta}, \boldsymbol{I}_p)$ and the prediction error loss $L(\omega, \boldsymbol{\delta}, \boldsymbol{A}^t\boldsymbol{A})$ as well as under the Strawderman's loss. It is seen that the other minimax adaptive ridge regression estimators $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{AD}, \mathbf{0})$ and $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{TR}, \mathbf{0})$ are not useful in the multicollearity cases.

From a practical point of view, we have looked into the usual adaptive ridge regression estimator $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{RR}, \mathbf{0})$ and the empirical Bayes ridge regression estimator $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{ML}, \mathbf{0})$ for the maximum likelihood estimator $\hat{\lambda}_{ML}$ of λ although their minimaxity are not guaranteed. The numerical studies in Section 4 have shown that the estimator $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{ML}, \mathbf{0})$ behaves well, but the performance of the estimator $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{RR}, \mathbf{0})$ depends on cases, and in fact it is not good in the cases given in Table 4 and Example 2.

In this paper, we also have considered the ridge-type estimators shrunken towards a suspected subspace. In the ill-conditioned cases, this subspace is constructed by eigenvectors of $(\boldsymbol{A}^t\boldsymbol{A})^{-1}$ with deleting the eigenvectors corresponding to some largest eigenvalues. Then the resulting ridge-type estimators shrink the least squares estimator toward the principal component estimator. The estimators we propose in this situation are the empirical Bayes ridge-principal component estimator $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{EB}, \widehat{\boldsymbol{\gamma}}_2)$ and the hierarchical empirical Bayes ridge regression estimator $\widehat{\boldsymbol{\beta}}^{HB}(\hat{\lambda}_{HB}, \hat{\tau}_{HB})$. Although the minimaxity of $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{EB}, \widehat{\boldsymbol{\gamma}}_2)$ has been shown in Section 3, we could not gurantee the minimaxity of $\widehat{\boldsymbol{\beta}}^{HB}(\hat{\lambda}_{HB}, \hat{\tau}_{HB})$ at this time. However the numerical studies in Section 4 have revealed that $\widehat{\boldsymbol{\beta}}^{HB}(\hat{\lambda}_{HB}, \hat{\tau}_{HB})$ performs reasonably well and gives stable estimates while $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{EB}, \widehat{\boldsymbol{\gamma}}_2)$ is not bad.

Through the simulation and empirical studies, we can conclude that the empirical Bayes ridge regression estimators $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{EB}, \boldsymbol{0})$, $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{ML}, \boldsymbol{0})$ and $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{EB}, \widehat{\boldsymbol{\gamma}}_2)$ and the hierarchical empirical Bayes ridge regression estimator $\widehat{\boldsymbol{\beta}}^{HB}(\hat{\lambda}_{HB}, \hat{\tau}_{HB})$ are highly recommended when the linear model is ill-conditioned. Especially, the estimators $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{EB}, \boldsymbol{0})$ and $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{ML}, \boldsymbol{0})$ can provide stable estimates automatically from the original data without any transformation or any decision by statistician. Thus, the main goal of regression analysis, that is prediction can easily be carried out without the necessity of carrying out ‘ selection of variables’ or ‘testing the significance’ of regression coefficients. An estimate of the prediction error can be obtained by the cross-varidation method, and it has been shown that the estimates of the prediction errors for the proposed ridge-type estimators are much smaller than that for the least squares estimator.

The results in Sections 2 and 3 have been applied to estimation of small area means in a one-way random effects model, and we have obtained the minimax empirical Bayes ridge-type estimator $\widehat{\boldsymbol{\mu}}^B(\hat{\lambda}_{EB}, \overline{X}..)$ shrunken toward the total sample mean. Through the simulation studies, it has revealed that the estimator has superior risk performance.

7 Appendix

7.1 Derivation of the expression of the hierarchical Bayes estimator

We here show the following equation in the expression (2.5) of the hierarchical Bayes estimator $\widehat{\boldsymbol{\beta}}^{HB}(\lambda, \tau)$:

$$\begin{aligned} &\widehat{\boldsymbol{\beta}} - (\boldsymbol{A}^t\boldsymbol{A} + \boldsymbol{\Psi}^{-1})^{-1}\boldsymbol{\Psi}^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\boldsymbol{\alpha}_0) \\ &\quad = \ \widehat{\boldsymbol{\beta}} - (\boldsymbol{A}^t\boldsymbol{A})^{-1}\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda \boldsymbol{I}_p\right\}^{-1}\left\{\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\widehat{\boldsymbol{\alpha}}^S(\lambda, \tau)\right\}. \end{aligned} \tag{7.1}$$

Since $\boldsymbol{\Psi} = \lambda \boldsymbol{I}_p + \tau \boldsymbol{C}\boldsymbol{C}^t$, the l.h.s. of (7.1) is expressed by

$$\begin{aligned} &\widehat{\boldsymbol{\beta}} - (\boldsymbol{A}^t\boldsymbol{A})^{-1}\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda \boldsymbol{I}_p + \tau \boldsymbol{C}\boldsymbol{C}^t\right\}^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\boldsymbol{\alpha}_0) \qquad (7.2) \\ &\quad = \ \widehat{\boldsymbol{\beta}} - (\boldsymbol{A}^t\boldsymbol{A})^{-1}\boldsymbol{G}^{-1}(\boldsymbol{I}_p + \tau \boldsymbol{C}\boldsymbol{C}^t\boldsymbol{G}^{-1})^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\boldsymbol{\alpha}_0), \end{aligned}$$

for $\boldsymbol{G} = \boldsymbol{G}(\lambda) = (\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda\boldsymbol{I}_p$. Noting that

$$\begin{aligned}(\boldsymbol{I}_p + \tau\boldsymbol{C}\boldsymbol{C}^t\boldsymbol{G}^{-1})^{-1} &= \boldsymbol{I}_p - \tau\boldsymbol{C}(\boldsymbol{I}_q + \tau\boldsymbol{C}^t\boldsymbol{G}^{-1}\boldsymbol{C})^{-1}\boldsymbol{C}^t\boldsymbol{G}^{-1} \\ &= \boldsymbol{I}_p - \tau\boldsymbol{C}\left\{(\boldsymbol{C}^t\boldsymbol{G}^{-1}\boldsymbol{C})^{-1} + \tau\boldsymbol{I}_q\right\}^{-1}(\boldsymbol{C}^t\boldsymbol{G}^{-1}\boldsymbol{C})^{-1}\boldsymbol{C}^t\boldsymbol{G}^{-1},\end{aligned}$$

we see that

$$\begin{aligned}&(\boldsymbol{I}_p + \tau\boldsymbol{C}\boldsymbol{C}^t\boldsymbol{G}^{-1})^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\boldsymbol{\alpha}_0) \\ &= \widehat{\boldsymbol{\beta}} - \boldsymbol{C}\boldsymbol{\alpha}_0 - \tau\boldsymbol{C}\left\{(\boldsymbol{C}^t\boldsymbol{G}^{-1}\boldsymbol{C})^{-1} + \tau\boldsymbol{I}_q\right\}^{-1}(\widehat{\boldsymbol{\alpha}}(\lambda) - \boldsymbol{\alpha}_0) \\ &= \widehat{\boldsymbol{\beta}} - \boldsymbol{C}\left\{\widehat{\boldsymbol{\alpha}}(\lambda) - (\boldsymbol{I}_q + \tau\boldsymbol{C}^t\boldsymbol{G}^{-1}\boldsymbol{C})^{-1}(\widehat{\boldsymbol{\alpha}}(\lambda) - \boldsymbol{\alpha}_0)\right\},\end{aligned}$$

where $\widehat{\boldsymbol{\alpha}}(\lambda) = (\boldsymbol{C}^t\boldsymbol{G}^{-1}\boldsymbol{C})^{-1}\boldsymbol{C}^t\boldsymbol{G}^{-1}\widehat{\boldsymbol{\beta}}$, being the weighted least squares estimator with the weight $\boldsymbol{G}^{-1} = (\lambda\boldsymbol{I} + \boldsymbol{D})^{-1}$. Hence from (7.2), we get the expression in the r.h.s. of the equation (7.1). ■

7.2 Proof of Theorem 1

In this subsection, we shall provide the proof of the theorem. The derivation of an unbiased estimator of risk of the adaptive ridge regression estimator

$$\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}, \widehat{\boldsymbol{\alpha}}) = \widehat{\boldsymbol{\beta}} - \left(\boldsymbol{I} + \hat{\lambda}(\boldsymbol{A}^t\boldsymbol{A})^{-1}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\widehat{\boldsymbol{\alpha}}\right) \tag{7.3}$$

is essential for the proof. The canonical representation is here treated, that is,

$$\widehat{\boldsymbol{\theta}}^B(\hat{\lambda}, \widehat{\boldsymbol{\alpha}}) = \boldsymbol{x} - (\boldsymbol{D} + \hat{\lambda}\boldsymbol{I})^{-1}\boldsymbol{D}(\boldsymbol{x} - \boldsymbol{B}\boldsymbol{x}),$$

where $\boldsymbol{B} = (b_{ij}) = \boldsymbol{Z}(\boldsymbol{Z}^t\boldsymbol{D}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}^t\boldsymbol{D}^{-1} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p)^t$, and $\hat{\lambda}$ is a function of $x_1, \ldots, x_p$ and S.

Lemma A. *The risk function of the estimator $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}, \widehat{\boldsymbol{\alpha}})$ under the loss (1.8) is given by $R(\omega, \widehat{\boldsymbol{\beta}}^B(\hat{\lambda}, \widehat{\boldsymbol{\alpha}})) = R(\omega, \widehat{\boldsymbol{\beta}}) + E[\widetilde{\Delta}(\hat{\lambda})]$, where*

$$\begin{aligned}\widetilde{\Delta}(\hat{\lambda}) &= -2\sum_{i=1}^p \frac{1 - b_{ii}}{d_i + \hat{\lambda}} + 2\sum_{i=1}^p \frac{x_i - \boldsymbol{b}_i^t\boldsymbol{x}}{(d_i + \hat{\lambda})^2}\frac{\partial\hat{\lambda}}{\partial x_i} \\ &\quad + (n-2)\sum_{i=1}^p \frac{(x_i - \boldsymbol{b}_i^t\boldsymbol{x})^2/S}{(d_i + \hat{\lambda})^2} - 4\sum_{i=1}^p \frac{(x_i - \boldsymbol{b}_i^t\boldsymbol{x})^2}{(d_i + \hat{\lambda})^3}\frac{\partial\hat{\lambda}}{\partial S}.\end{aligned} \tag{7.4}$$

Proof. The risk function of $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}, \widehat{\boldsymbol{\alpha}})$ is written by

$$\begin{aligned}&R(\omega, \widehat{\boldsymbol{\beta}}^B(\hat{\lambda}, \widehat{\boldsymbol{\alpha}})) = R(\omega, \boldsymbol{X}) \\ &\quad -2\sum_{i=1}^p E\left[(x_i - \theta_i)\frac{(x_i - \boldsymbol{b}_i^t\boldsymbol{x})/d_i}{d_i + \hat{\lambda}}\right]/\sigma^2 + \sum_{i=1}^p E\left[\frac{(x_i - \boldsymbol{b}_i^t\boldsymbol{x})^2}{(d_i + \hat{\lambda})^2}\right]/\sigma^2.\end{aligned} \tag{7.5}$$

Using the Stein identity given by Stein (1973, 81), we observe that

$$E\left[(x_i-\theta_i)\frac{(x_i-\boldsymbol{b}_i^t\boldsymbol{x})/(d_i\sigma^2)}{d_i+\hat{\lambda}}\right]=E\left[\frac{1-b_{ii}}{d_i+\hat{\lambda}}-\frac{x_i-\boldsymbol{b}_i^t\boldsymbol{x}}{(d_i+\hat{\lambda})^2}\frac{\partial\hat{\lambda}}{\partial x_i}\right]. \tag{7.6}$$

Using the chi-square identity given by Efron and Morris (1976) gives that

$$\begin{aligned}
&E\left[\sum_{i=1}^p\frac{(x_i-\boldsymbol{b}_i^t\boldsymbol{x})^2/S}{(d_i+\hat{\lambda})^2}\frac{S}{\sigma^2}\right]\\
&\quad= E\left[(n-2)\sum_{i=1}^p\frac{(x_i-\boldsymbol{b}_i^t\boldsymbol{x})^2/S}{(d_i+\hat{\lambda})^2}-4\sum_{i=1}^p\frac{(x_i-\boldsymbol{b}_i^t\boldsymbol{x})^2}{(d_i+\hat{\lambda})^3}\frac{\partial\hat{\lambda}}{\partial S}\right].
\end{aligned} \tag{7.7}$$

Combining (7.5), (7.6) and (7.7) proves Lemma A. ■

Proof of the theorem. From the condition (b) of Theorem 1, it is seen that

$$\sum_{i=1}^p\frac{x_i-\boldsymbol{b}_i^t\boldsymbol{x}}{(d_i+\hat{\lambda})^2}\frac{\partial\hat{\lambda}}{\partial x_i}\le\frac{2}{d_p+\hat{\lambda}}. \tag{7.8}$$

From the condition (c) of Theorem 1, it follows that

$$(n-2)\sum_{i=1}^p\frac{(x_i-\boldsymbol{b}_i^t\boldsymbol{x})^2/S}{(d_i+\hat{\lambda})^2}-4\sum_{i=1}^p\frac{(x_i-\boldsymbol{b}_i^t\boldsymbol{x})^2}{(d_i+\hat{\lambda})^3}\frac{\partial\hat{\lambda}}{\partial S}\le\frac{n-2}{d_p+\hat{\lambda}}\alpha+\frac{4\beta}{d_p+\hat{\lambda}}. \tag{7.9}$$

Combining (7.4), (7.8) and (7.9) gives that

$$\widetilde{\Delta}(\lambda)\le-\sum_{i=1}^p\frac{2(1-b_{ii})}{d_i+\hat{\lambda}}+\frac{(n-2)\alpha+4(\beta+1)}{d_p+\hat{\lambda}} \tag{7.10}$$

which is not positive if

$$-2\sum_{i=1}^p(1-b_{ii})\frac{d_p+\hat{\lambda}}{d_i+\hat{\lambda}}+(n-2)\alpha+4(\beta+1)\le 0. \tag{7.11}$$

From the condition (a), it is noted that

$$\begin{aligned}
\sum_{i=1}^p(1-b_{ii})\frac{d_p+\hat{\lambda}}{d_i+\hat{\lambda}} &\ge \sum_{i=1}^p(1-b_{ii})\frac{d_p+\lambda_m}{d_i+\lambda_m}\\
&= p-\sum_{i=1}^p b_{ii}-\sum_{i=1}^p(1-b_{ii})\frac{d_i-d_p}{d_i+\lambda_m},
\end{aligned}$$

which is used to get the following condition from (7.11):

$$2\sum_{i=1}^p(1-b_{ii})\frac{d_i-d_p}{d_i+\lambda_m}+(n-2)\alpha+4\beta\le 2(p-q-2),$$

since $\sum_{i=1}^p b_{ii}=q$. This inequality is just given by the condition (d) of Theorem 1, which has therefore been proved. ■

Acknowledgements. The research of the first author was supported in part by Natural Sciences and Engineering Research Council of Canada. The research of the second author was supported in part by a grant from the Ministry of Education, Japan, No. 13680371. This work was done during the visit of the first author to the University of Toronto, 2001 and 2002, summer. The authors thank Mr. K. Tanaka for his help in the simulation experiments.

REFERENCES

Beisley, D.A. (1984). Demeaning conditioning diagnostics through centering (with discussion). *Amer. Statist.*, **38**, 73-93.

Casella, G. (1980). Minimax ridge regression estimation. it Ann. Statist., **8**, 1036-1056.

Casella, G. (1985). Condition numbers and minimax ridge regression estimators. *J. Amer. Statist. Assoc.*, **80**, 753-758.

DuMouchel, W.M. and Harris, J.E. (1983). Bayes methods for combining the results of cancer studies in humans and other species (with discussion). *J. Amer. Statist. Assoc.*, **78**, 293-315.

Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amr. Statist. Assoc.*, **70**, 311-319.

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

Fay, R.E. and Herriot, R. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.*, **74**, 269-277.

Feenstra, R.C. (1995). Exact hednic price indexes. *Review of Economics and Statistics*, **77**, 634-653.

Gilley, O.W. and Pace, R.K. (1995). Improving hednic estimation with an inequality restricted estimator. *Review of Economics and Statistics*, **77**, 609-621.

Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthoganal problems. *Technometrics*, **12**, 55-67.

James, W., and Stein, C. (1961). Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* , **1**, 361-379. University of California Press, Berkeley.

Lawless, J.F. and Wang, P. (1976). A simulation study of ridge and other regression estimators. *Commun. Statist. - Theory Method*, **A5**, 307-323.

Lindley, D.V. (1962). Discussion of "Confidence sets for the mean of a multivariate normal distribution", by C. Stein. *J. Roy. Statist. Soc.*, **B 24**, 285-287.

Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc.*, **B 34**, 1-41.

Marquardt, D.W., and Snee, R.E. (1975). Ridge regression in practice. *Amer. Statist.*, **12**, 3-19.

Rosen, S. (1974). Hednic prices and implicit markets: Product differentiation in pure competition. *J. Political Economy*, ***i*82**, 34-55.

Sen, A. and Srivastava, M. (1990). *Regression Analysis: Theory, Methods, and Applications.* Springer, New York.

Shinozaki, N. and Chang, Y.-T. (1993). Minimaxity of empirical Bayes estimators of the means of independent normal variables with unequal variances. *Commun. Statist. - Theory Method*, **22**, 2147-2169.

Shinozaki, N. and Chang, Y.-T. (1996). Minimaxity of empirical Bayes estimators shrinking toward the grand mean when variances are unequal. *Commun. Statist. - Theory Method*, **25**, 183-199.

Srivastava, M.S. (2002). *Methods of Multivariate Statistics*. Wiley, New York.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. Third Berkeley Symp. Math. Statist. Probab.*, **1**, 197-206. University of California University, Berkeley.

Stein, C. (1973). Estimation of the mean of a multivariate normal distribution. In *Proc. Prague Symp. Asymptotic Statist.*, 345-381.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, **9**, 1135-1151.

Strawderman, W.E. (1978). Minimax adaptive generalized ridge regression estimators. *J.Amer. Statist. Assoc.*, **73**, 623-627.

Wetherill, G.B. (1986). *Regression Analysis with Applications*. Chapman and Hall.