CIRJE-F-190

Improved Empirical Bayes Ridge Regression Estimators under Multicollinearity

Tatsuya Kubokawa
University of Tokyo

M. S. Srivastava
University of Toronto

January 2003

Improved Empirical Bayes Ridge Regression Estimators under Multicollinearity

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

January 29, 2003

In this paper we consider the problem of estimating the regression parameters in a multiple linear regression model when the multicollinearity is present. Under the assumption of normality, we present three empirical Bayes estimators. One of them shrinks the least squares (LS) estimator towards the principal component. The second one is a hierarchical empirical Bayes estimator shrinking the LS estimator twice. The third one is obtained by choosing different priors for the two sets of regression parameters that arise in the case of multicollinearity; this estimator is termed decomposed empirical Bayes estimator. These proposed estimators are not only proved to be uniformly better than the LS estimator, that is, minimax in terms of risk under the Strawderman's loss function, but also shown to be useful in the multicollinearity cases through simulation and empirical studies.

Key words and phrases: Multiple regression, multicollinearity, ridge regression, empirical Bayes method, principal component method, 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 available 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. A more general method called 'continuum regression' has been proposed by Stone and Brooks (1990). This procedure depends on a parameter, say, 'γ' which is recommended to be determined by cross validation. However, except for two special values of γ, (0 and 1), Sundberg (1993) and Björkström and Sundberg (1996) have shown that it is equivalent to ridge

[*] Faculty of Economics, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN; E-Mail: tatsuya@e.u-tokyo.ac.jp ; Fax: +81-3-5841-5521

[†] Department of Statistics, University of Toronto, 100 St George Street, Toronto, Ontario, CANADA M5S 3G3; E-Mail: srivasta@utstat.utstat.toronto.edu ; Fax: +1-416-978-5133

regression. Although the corresponding parameter of ridge regression can also be found by the method of cross validation, it is not guaranteed to be optimum in any sense, at least better than the least squares estimator. The problem of finding an optimum value of the parameter in the ridge regression has been elusive.

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 (LS) 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. This led Hoerl and Kennard (1970) to propose the estimator

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

for $\lambda = 1/k$, $k > 0$, and is called a *ridge regression estimator* of $\boldsymbol{\beta}$. The estimator (1.2), however, depends on λ as well as it is not always better than the LS estimator in terms of risk under any quadratic loss. When an estimator is uniformly better than the LS estimator, we say in this paper that it is *minimax*.

Strawderman (1978) and Casella (1980) gave a class of estimators of λ which result in minimax estimators of $\boldsymbol{\beta}$ 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 minimax estimator of λ under Strawderman's loss function is given by

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

for $S = (\boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{\beta}})^t(\boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{\beta}})$, 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 Shinozaki and Chang (1993) who obtained an estimator of λ by solving the equation

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

and 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 a linear hypothesis on $\boldsymbol{\beta}$ is suspected. 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 $\boldsymbol{H}^t = (\boldsymbol{H}_1^t, \boldsymbol{H}_2^t)$,

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

where $\boldsymbol{\gamma}$ 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 $\boldsymbol{\beta} = \boldsymbol{H}_2^t\boldsymbol{\alpha}$. Under the Strawderman's loss function

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

we propose in Section 2 of this paper, the following three empirical Bayes estimators when the linear hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{H}_2^t\boldsymbol{\alpha}$ for $\boldsymbol{\alpha} \in \boldsymbol{R}^q$ is suspected.

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{EB} &= \widehat{\boldsymbol{\beta}} - (\boldsymbol{A}^t\boldsymbol{A})^{-1}\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \hat{\lambda}_{EB}\boldsymbol{I}_p\right\}^{-1}\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{PC}\right),\\ \widehat{\boldsymbol{\beta}}^{HB} &= \widehat{\boldsymbol{\beta}} - (\boldsymbol{A}^t\boldsymbol{A})^{-1}\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \hat{\lambda}_{EB}\boldsymbol{I}_p + \hat{\tau}_{HB}\boldsymbol{H}_2^t\boldsymbol{H}_2\right\}^{-1}\widehat{\boldsymbol{\beta}},\\ \widehat{\boldsymbol{\beta}}^{DB} &= \widehat{\boldsymbol{\beta}} - (\boldsymbol{A}^t\boldsymbol{A})^{-1}\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \hat{\lambda}_{EB}\boldsymbol{H}_1^t\boldsymbol{H}_1 + \widehat{\psi}_{DB}\boldsymbol{H}_2^t\boldsymbol{H}_2\right\}^{-1}\widehat{\boldsymbol{\beta}},\end{aligned}$$

where $\hat{\lambda}_{EB}$, $\hat{\tau}_{HB}$ and $\widehat{\psi}_{DB}$ are defined by (2.8), (2.15) and (2.17), respectively, and $\widehat{\boldsymbol{\beta}}^{PC} = \boldsymbol{H}_2^t\boldsymbol{H}_2\widehat{\boldsymbol{\beta}}$ is the principal component regression (PC) estimator. It is shown in Sections 3 and 4 that these estimators are minimax, that is, uniformly better than the LS estimator in terms of risk under the Strawderman's loss function. In Section 5, 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 an example. These simulations show that the proposed estimators perform well for all the three loss functions.

2 Proposed Empirical Bayes Ridge Regression Estimators

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

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

In the multicollinearity case, we can construct reasonable ridge-type regression estimators by using the information about which eigenvalues are smaller. Let $\boldsymbol{H}$ be an orthogonal matrix such that

$$\boldsymbol{H}\boldsymbol{A}^t\boldsymbol{A}\boldsymbol{H}^t = \boldsymbol{D}^{-1} = \mathrm{diag}\,(d_1, \ldots, d_p) = \begin{pmatrix} \boldsymbol{D}_1^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_2^{-1} \end{pmatrix} \tag{2.1}$$

where $\boldsymbol{D}_1^{-1} = \mathrm{diag}\,(d_1^{-1}, \ldots, d_{p-q}^{-1})$, $(p-q) \times (p-q)$ diagonal matrix with smaller eigenvalues. Corresponding to this decomposition, the orthogonal matrix $\boldsymbol{H}$ is decomposed as

$$\boldsymbol{H}^t = (\boldsymbol{H}_1^t; \boldsymbol{H}_2^t) \tag{2.2}$$

for $q \times p$ matrix $\boldsymbol{H}_2$. Then, as in (1.7),

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

Since $\boldsymbol{\gamma}$ 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:

$$H_0 \,:\, \boldsymbol{\beta} = \boldsymbol{H}_2^t\boldsymbol{\alpha}, \quad \boldsymbol{\alpha} \in \boldsymbol{R}^q. \tag{2.3}$$

It may be reasonable to consider adaptive ridge regression estimators shrunken toward the hypothesis. To derive such a shrinkage procedure, we employ three types of empirical Bayes methods, which are here called an empirical Bayes ridge regression estimator (EB), a hierarchical empirical Bayes ridge regression estimator (HB) and a decomposed empirical Bayes estimator (DB).

2.1 Empirical Bayes ridge regression estimator (EB)

Suppose that $\boldsymbol{\beta}$ has prior distribution $\mathcal{N}_p(\boldsymbol{H}_2^t\boldsymbol{\alpha}, \sigma^2\lambda\boldsymbol{I}_p)$ 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

$$\begin{aligned} \boldsymbol{\beta}\,|\,\widehat{\boldsymbol{\beta}} &\sim \mathcal{N}_p\left(\widehat{\boldsymbol{\beta}}^B(\lambda, \boldsymbol{\alpha}), \sigma^2(\boldsymbol{A}^t\boldsymbol{A} + \lambda^{-1}\boldsymbol{I})^{-1}\right), \\ \widehat{\boldsymbol{\beta}} &\sim \mathcal{N}_p\left(\boldsymbol{H}_2^t\boldsymbol{\alpha}, \sigma^2\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda\boldsymbol{I}\}\right), \end{aligned}$$

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

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^B(\lambda, \boldsymbol{\alpha}) &= (\boldsymbol{A}^t\boldsymbol{A} + \lambda^{-1}\boldsymbol{I})^{-1}\boldsymbol{A}^t\boldsymbol{A}(\widehat{\boldsymbol{\beta}} - \boldsymbol{H}_2^t\boldsymbol{\alpha}) + \boldsymbol{H}_2^t\boldsymbol{\alpha} \\ &= \widehat{\boldsymbol{\beta}} - \left(\boldsymbol{I} + \lambda\boldsymbol{A}^t\boldsymbol{A}\right)^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{H}_2^t\boldsymbol{\alpha}). \end{aligned} \tag{2.4}$$

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

$$\widehat{\boldsymbol{\alpha}} = (\boldsymbol{H}_2\boldsymbol{A}^t\boldsymbol{A}\boldsymbol{H}_2^t)^{-1}\boldsymbol{H}_2\boldsymbol{A}^t\boldsymbol{A}\widehat{\boldsymbol{\beta}},$$

which can be obtained by minimizing the weighted squared loss $(\widehat{\boldsymbol{\beta}} - \boldsymbol{H}_2^t\boldsymbol{\alpha})^t\boldsymbol{A}^t\boldsymbol{A}(\widehat{\boldsymbol{\beta}} - \boldsymbol{H}_2^t\boldsymbol{\alpha})$. Using the decomposition given by (2.1) and (2.2), we see that $\widehat{\boldsymbol{\alpha}} = \boldsymbol{H}_2\widehat{\boldsymbol{\beta}}$. Since the principal component (PC) regression estimator $\widehat{\boldsymbol{\beta}}^{PC}$ of $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}}^{PC} = \boldsymbol{H}_2^t\boldsymbol{H}_2\widehat{\boldsymbol{\beta}},$$

we observe that $\boldsymbol{H}_2^t\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\beta}}^{PC}$. Substituting $\boldsymbol{H}_2^t\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\beta}}^{PC}$ into $\widehat{\boldsymbol{\beta}}^{B}(\lambda, \boldsymbol{\alpha})$, we get the estimator

$$\widehat{\boldsymbol{\beta}}^{B}(\lambda, \widehat{\boldsymbol{\alpha}}) = \widehat{\boldsymbol{\beta}} - (\boldsymbol{I} + \lambda \boldsymbol{A}^t\boldsymbol{A})^{-1}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{PC}). \tag{2.5}$$

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

$$(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{PC})^t \left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda^*\boldsymbol{I}\right\}^{-1} (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{PC}) = \frac{p-q-2}{n+2}S, \tag{2.6}$$

and λ_0 is the root of the equation

$$\sum_{i=1}^{p-q} \frac{d_i - d_{p-q}}{d_i + \lambda_0} = (p-q-2)/2. \tag{2.7}$$

Then we propose the estimator $\hat{\lambda}_{EB}$ of λ, given by the maximum of λ^* and λ_0, that is,

$$\hat{\lambda}_{EB} = \max(\lambda^*, \lambda_0). \tag{2.8}$$

Substituting $\widehat{\boldsymbol{\alpha}}$ and $\hat{\lambda}_{EB}$ into (2.4), we get *the empirical Bayes ridge regression estimator* (EB)

$$\widehat{\boldsymbol{\beta}}^{EB} = \widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{EB}, \widehat{\boldsymbol{\alpha}}) = \widehat{\boldsymbol{\beta}} - \left(\boldsymbol{I} + \hat{\lambda}_{EB}\boldsymbol{A}^t\boldsymbol{A}\right)^{-1} \left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{PC}\right), \tag{2.9}$$

which shrinks the LS estimator $\widehat{\boldsymbol{\beta}}$ towards the PC estimator $\widehat{\boldsymbol{\beta}}^{PC}$. It is known that the principal component estimator and the ridge regression estimator are useful in predicting a response variable in the presence of multicollinearity. It is interesting to note that both methods of ridge regression and principal components are incorporated in the proposed estimator $\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{EB}, \widehat{\boldsymbol{\alpha}})$.

The single empirical Bayes ridge regression estimator given by (2.9) can be shown in Section 3 to be minimax under the loss function (1.8), namely, $\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{EB}, \widehat{\boldsymbol{\alpha}})$ has uniformly smaller risk than $\widehat{\boldsymbol{\beta}}$.

2.2 Hierarchical empirical Bayes ridge regression estimator (HB)

When the dimension q of the vector $\boldsymbol{\alpha}$ is large, it may be reasonable to shrink the estimator of $\boldsymbol{\alpha}$. Thus we consider the hierarchical type of the prior distributions:

$$\begin{aligned} \boldsymbol{\beta} \,|\, \boldsymbol{\alpha} &\sim \mathcal{N}_p(\boldsymbol{H}_2^t\boldsymbol{\alpha}, \sigma^2\lambda\boldsymbol{I}_p), \\ \boldsymbol{\alpha} &\sim \mathcal{N}_q(\boldsymbol{\alpha}_0, \sigma^2\tau\boldsymbol{I}_q), \end{aligned}$$

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)).

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

$$\boldsymbol{\beta} \sim \mathcal{N}_p\left(\boldsymbol{H}_2^t\boldsymbol{\alpha}_0, \sigma^2\left(\lambda\boldsymbol{I}_p + \tau\boldsymbol{H}_2^t\boldsymbol{H}_2\right)\right).$$

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 $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} \,|\, \widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\widehat{\boldsymbol{\beta}}^{HB}(\lambda,\tau), \left\{\boldsymbol{A}^t\boldsymbol{A} + (\lambda \boldsymbol{I}_p + \tau \boldsymbol{H}_2^t\boldsymbol{H}_2)^{-1}\right\}^{-1}),$$

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

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\boldsymbol{H}_2^t\boldsymbol{\alpha}_0, \lambda\boldsymbol{I}_p + \tau\boldsymbol{H}_2^t\boldsymbol{H}_2 + (\boldsymbol{A}^t\boldsymbol{A})^{-1}),$$

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

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{HB}(\lambda,\tau) &= \left\{\boldsymbol{A}^t\boldsymbol{A} + (\lambda\boldsymbol{I}_p + \tau\boldsymbol{H}_2^t\boldsymbol{H}_2)^{-1}\right\}^{-1}\left\{\boldsymbol{A}^t\boldsymbol{A}\widehat{\boldsymbol{\beta}} + (\lambda\boldsymbol{I}_p + \tau\boldsymbol{H}_2^t\boldsymbol{H}_2)^{-1}\boldsymbol{H}_2^t\boldsymbol{\alpha}_0\right\}\\ &= \widehat{\boldsymbol{\beta}} - (\boldsymbol{A}^t\boldsymbol{A})^{-1}\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda\boldsymbol{I}_p + \tau\boldsymbol{H}_2^t\boldsymbol{H}_2\right\}^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{H}_2^t\boldsymbol{\alpha}_0).\end{aligned}$$

As shown in the Appendix, $\widehat{\boldsymbol{\beta}}^{HB}(\lambda,\tau)$ can be rewritten as

$$\widehat{\boldsymbol{\beta}}^{HB}(\lambda,\tau) = \widehat{\boldsymbol{\beta}} - (\boldsymbol{A}^t\boldsymbol{A})^{-1}\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda\boldsymbol{I}_p\right\}^{-1}\left\{\widehat{\boldsymbol{\beta}} - \boldsymbol{H}_2^t\widehat{\boldsymbol{\alpha}}^S(\lambda,\tau)\right\}, \tag{2.10}$$

where

$$\widehat{\boldsymbol{\alpha}}^S(\lambda,\tau) = \widehat{\boldsymbol{\alpha}}(\lambda) - \left[\boldsymbol{I}_q + \tau\boldsymbol{H}_2\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda\boldsymbol{I}_p\right\}^{-1}\boldsymbol{H}_2^t\right]^{-1}(\widehat{\boldsymbol{\alpha}}(\lambda) - \boldsymbol{\alpha}_0), \tag{2.11}$$

for $\widehat{\boldsymbol{\alpha}}(\lambda) = [\boldsymbol{H}_2\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda\boldsymbol{I}_p\}^{-1}\boldsymbol{H}_2^t]^{-1}\boldsymbol{H}_2\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda\boldsymbol{I}_p\}^{-1}\widehat{\boldsymbol{\beta}}$. It is interesting to note that $\widehat{\boldsymbol{\alpha}}^S(\lambda,\tau)$ shrinks the weighted LS estimator $\widehat{\boldsymbol{\alpha}}(\lambda)$ towards the prior mean $\boldsymbol{\alpha}_0$. Hence $\widehat{\boldsymbol{\beta}}^{HB}(\lambda,\tau)$ is interpreted as a double shrinkage procedure that shrinks the LS estimator $\widehat{\boldsymbol{\beta}}$ towards the shrunken value $\boldsymbol{H}_2^t\widehat{\boldsymbol{\alpha}}^S(\lambda,\tau)$.

The hyper-parameters λ and τ are estimated from the marginal distribution of $\widehat{\boldsymbol{\beta}}$. We here employ the estimator $\hat{\lambda}_{EB}$ given by (2.8) for λ. To estimate τ, let ψ^* be the solution of the equation:

$$(\boldsymbol{H}_2\widehat{\boldsymbol{\beta}} - \boldsymbol{\alpha}_0)^t\left\{\boldsymbol{H}_2(\boldsymbol{A}^t\boldsymbol{A})^{-1}\boldsymbol{H}_2^t + \psi^*\boldsymbol{I}_q\right\}^{-1}(\boldsymbol{H}_2\widehat{\boldsymbol{\beta}} - \boldsymbol{\alpha}_0) = (q-2)S/(n+2). \tag{2.12}$$

Also let ψ_0 and ψ_1 be the solution of the following two equations:

$$\sum_{i=p-q+1}^{p}\frac{d_i - d_p}{d_i + \psi_0} = \frac{q-2}{2}, \tag{2.13}$$

$$\sum_{i=p-q+1}^{p}\frac{d_i - d_p}{d_i + \psi_1} + 2\frac{q-2}{n+2}\frac{d_1 - d_p}{d_p + \psi_1} = \frac{q+2}{2}. \tag{2.14}$$

Define $\widehat{\psi}_{HB}$ and $\hat{\tau}_{HB}$ by

$$\widehat{\psi}_{HB} = \max(\psi^*, \psi_0, \psi_1), \quad \hat{\tau}_{HB} = \max(\widehat{\psi}_{HB} - \hat{\lambda}_{EB}, 0). \tag{2.15}$$

Then, we get the estimator

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{HB} &= \widehat{\boldsymbol{\beta}}^{HB}(\hat{\lambda}_{EB}, \hat{\tau}_{HB})\\ &= \widehat{\boldsymbol{\beta}} - (\boldsymbol{A}^t\boldsymbol{A})^{-1}\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \hat{\lambda}_{EB}\boldsymbol{I}_p + \hat{\tau}_{HB}\boldsymbol{H}_2^t\boldsymbol{H}_2\right\}^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{H}_2^t\boldsymbol{\alpha}_0),\end{aligned} \tag{2.16}$$

which we shall call the hierarchical empirical Bayes ridge regression estimator (HB). 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.

The hierarchical empirical Bayes ridge regression estimator given by (2.16) is shown in Section 4 to be minimax under the loss function (1.8).

2.3 Decomposed empirical Bayes ridge regression estimator

The regression coefficients vector $\boldsymbol{\beta}$ is expressed as $\boldsymbol{\beta} = \boldsymbol{H}_1^t\boldsymbol{\gamma} + \boldsymbol{H}_2^t\boldsymbol{\alpha}$ by the decomposition (2.1) and (2.2). Since $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ correspond to the smaller and larger eigenvalues of $\boldsymbol{A}^t\boldsymbol{A}$, respectively, it may be reasonable to suppose that the decomposed parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ have different prior distributions. We thus suppose that $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ are independently distributed as

$$\begin{aligned}\boldsymbol{\gamma} &\sim \mathcal{N}(\boldsymbol{0}, \lambda\sigma^2\boldsymbol{I}_{p-q}),\\ \boldsymbol{\alpha} &\sim \mathcal{N}(\boldsymbol{\alpha}_0, \psi\sigma^2\boldsymbol{I}_q),\end{aligned}$$

where λ, ψ are unknown parameters and $\boldsymbol{\alpha}_0$ is a known prior mean. Then $\boldsymbol{\beta}$ has a prior distribution $\mathcal{N}(\boldsymbol{H}_2^t\boldsymbol{\alpha}_0, \sigma^2(\lambda\boldsymbol{H}_1^t\boldsymbol{H}_1 + \psi\boldsymbol{H}_2^t\boldsymbol{H}_2))$. It can be seen that the posterior distribution of $\boldsymbol{\beta}$ given $\widehat{\boldsymbol{\beta}}$ and the marginal distribution of $\widehat{\boldsymbol{\beta}}$ are given by

$$\begin{aligned}\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}} &\sim \mathcal{N}\left(\widehat{\boldsymbol{\beta}}^{DB}(\lambda,\psi), \sigma^2\left\{\boldsymbol{A}^t\boldsymbol{A} + \frac{1}{\lambda}\boldsymbol{H}_1^t\boldsymbol{H}_1 + \frac{1}{\psi}\boldsymbol{H}_2^t\boldsymbol{H}_2\right\}^{-1}\right),\\ \widehat{\boldsymbol{\beta}} &\sim \mathcal{N}\left(\boldsymbol{H}_2^t\boldsymbol{\alpha}_0, \sigma^2\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda\boldsymbol{H}_1^t\boldsymbol{H}_1 + \psi\boldsymbol{H}_2^t\boldsymbol{H}_2\right\}\right),\end{aligned}$$

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

$$\widehat{\boldsymbol{\beta}}^{DB}(\lambda,\psi) = \widehat{\boldsymbol{\beta}} - (\boldsymbol{A}^t\boldsymbol{A})^{-1}\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda\boldsymbol{H}_1^t\boldsymbol{H}_1 + \psi\boldsymbol{H}_2^t\boldsymbol{H}_2\right\}^{-1}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{H}_2^t\boldsymbol{\alpha}_0\right).$$

The unknown hyper-parameters λ and ψ are estimated from the marginal distribution of $\widehat{\boldsymbol{\beta}}$. We here employ the estimator $\hat{\lambda}_{EB}$ given by (2.8) for λ and define the estimator $\widehat{\psi}_{DB}$ by

$$\widehat{\psi}_{DB} = \max(\psi^*, \psi_0), \tag{2.17}$$

for ψ^* and ψ_0 given in (2.12) and (2.13), respectively. Then, we get the estimator

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{DB} &= \widehat{\boldsymbol{\beta}}^{DB}(\hat{\lambda}_{EB}, \widehat{\psi}_{DB})\\ &= \widehat{\boldsymbol{\beta}} - (\boldsymbol{A}^t\boldsymbol{A})^{-1}\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \hat{\lambda}_{EB}\boldsymbol{H}_1^t\boldsymbol{H}_1 + \widehat{\psi}_{DB}\boldsymbol{H}_2^t\boldsymbol{H}_2\right\}^{-1}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{H}_2^t\boldsymbol{\alpha}_0\right),\end{aligned} \tag{2.18}$$

which is here called *the decomposed empirical Bayes ridge regression estimator* (DB).

The decomposed empirical Bayes ridge regression estimator given by (2.18) will be shown in Section 4 to be minimax under the loss function (1.8).

3 Minimaxity of the Empirical Bayes Estimators

In this section, we not only show the minimaxity of the single empirical Bayes ridge regression estimator $\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{EB}, \widehat{\boldsymbol{\alpha}})$ given by (2.9), but also derive other minimax adaptive ridge regression estimators. For the purpose, in the next subsection, we shall obtain the general conditions on an estimator $\hat{\lambda}$ of λ under which the resulting adaptive ridge regression estimator is minimax under the Strawderman's loss function (1.8).

3.1 General conditions for the minimaxity

Let $\hat\lambda$ be a nonnegative function of $\widehat{\boldsymbol\beta} - \widehat{\boldsymbol\beta}^{PC}$ and S and consider the empirical Bayes or adaptive ridge regression estimator

$$\widehat{\boldsymbol\beta}^{B}(\hat\lambda, \widehat{\boldsymbol\alpha}) = \widehat{\boldsymbol\beta} - (\boldsymbol{I} + \hat\lambda \boldsymbol{A}^t \boldsymbol{A})^{-1}(\widehat{\boldsymbol\beta} - \widehat{\boldsymbol\beta}^{PC}). \tag{3.1}$$

To handle the estimators more conveniently, we treat them in a canonical form. For the orthogonal matrix $\boldsymbol{H}$ given by (2.1), define $\boldsymbol{x} = \boldsymbol{H}\widehat{\boldsymbol\beta}$ and $\boldsymbol\theta = \boldsymbol{H}\boldsymbol\beta$. Then

$$\begin{aligned} &\boldsymbol{x} \sim \mathcal{N}_p(\boldsymbol\theta, \sigma^2 \boldsymbol{D}), \\ &\boldsymbol{D} = \mathrm{diag}\,(d_1, \ldots, d_p), \quad d_1 \geq \ldots \geq d_p > 0. \end{aligned} \tag{3.2}$$

That is x_i's are independently normally distributed as $x_i \sim \mathcal{N}(\theta_i, d_i\sigma^2)$ where x_i and θ_i are the respective ith component of the vectors $\boldsymbol{x}$ and $\boldsymbol\theta$. Letting $\widehat{\boldsymbol\theta}^{B}(\hat\lambda, \widehat{\boldsymbol\alpha}) = \boldsymbol{H}\widehat{\boldsymbol\beta}^{B}(\hat\lambda, \widehat{\boldsymbol\alpha})$ and noting that $\boldsymbol{H}\widehat{\boldsymbol\beta}^{PC} = \boldsymbol{H}\boldsymbol{H}_2^t\boldsymbol{H}_2\widehat{\boldsymbol\beta} = (0, \ldots, 0, x_{p-q+1}, \ldots, x_p)^t$, we see that

$$\begin{aligned} \widehat{\boldsymbol\theta}^{B}(\hat\lambda, \widehat{\boldsymbol\alpha}) &= \boldsymbol{x} - (\boldsymbol{D} + \hat\lambda \boldsymbol{I})^{-1}\boldsymbol{D}(\boldsymbol{x} - \boldsymbol{H}\widehat{\boldsymbol\beta}^{PC}) \\ &= \begin{pmatrix} \boldsymbol{x}_{(1)} - (\boldsymbol{D}_1 + \hat\lambda \boldsymbol{I}_{p-q})^{-1}\boldsymbol{D}_1\boldsymbol{x}_{(1)} \\ \boldsymbol{x}_{(2)} \end{pmatrix}, \end{aligned} \tag{3.3}$$

where $\boldsymbol{x}_{(1)} = (x_1, \ldots, x_{p-q})^t$ and $\boldsymbol{x}_{(2)} = (x_{p-q+1}, \ldots, x_p)^t$ and the estimator $\hat\lambda$ of λ can be represented as a function of $\boldsymbol{x}_{(1)}$ and S.

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 on the $\hat\lambda$ are satisfied for $p \geq q+3$:*

(a) $\hat\lambda \geq \lambda_m$ for a nonnegative constant λ_m, and $\hat\lambda$ is an absolutely continuous function of $x_1, \ldots, x_{p-q}$ and S.

(b) $x_i\partial\hat\lambda/\partial x_i \geq 0$ for $i = 1, \ldots, p-q$, and

$$\sum_{i=1}^{p-q} \frac{x_i}{d_i + \hat\lambda}\frac{\partial\hat\lambda}{\partial x_i} \leq 2. \tag{3.4}$$

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

$$\sum_{i=1}^{p-q} \frac{x_i^2/S}{d_i + \hat\lambda} \leq \alpha \quad \text{and} \quad -\sum_{i=1}^{p-q} \frac{x_i^2}{(d_i + \hat\lambda)^2}\frac{\partial\hat\lambda}{\partial S} \leq \beta. \tag{3.5}$$

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

$$\sum_{i=1}^{p-q} \frac{d_i - d_{p-q}}{d_i + \lambda_m} + \frac{(n-2)\alpha}{2} + 2\beta \leq p - q - 2. \tag{3.6}$$

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

$$\begin{aligned} R(\omega, \widehat{\boldsymbol\beta}^{B}(\hat\lambda, \widehat{\boldsymbol\alpha})) &= R(\omega, \boldsymbol{X}) \\ &-2\sum_{i=1}^{p-q} E\left[(x_i - \theta_i)\frac{x_i/d_i}{d_i + \hat\lambda}\right]/\sigma^2 + \sum_{i=1}^{p-q} E\left[\frac{x_i^2}{(d_i + \hat\lambda)^2}\right]/\sigma^2. \end{aligned} \tag{3.7}$$

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

$$E\left[(x_i-\theta_i)\frac{x_i/(d_i\sigma^2)}{d_i+\hat{\lambda}}\right]=E\left[\frac{1}{d_i+\hat{\lambda}}-\frac{x_i}{(d_i+\hat{\lambda})^2}\frac{\partial\hat{\lambda}}{\partial x_i}\right]. \tag{3.8}$$

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

$$E\left[\sum_{i=1}^{p-q}\frac{x_i^2/S}{(d_i+\hat{\lambda})^2}\frac{S}{\sigma^2}\right]=E\left[(n-2)\sum_{i=1}^{p-q}\frac{x_i^2/S}{(d_i+\hat{\lambda})^2}-4\sum_{i=1}^{p-q}\frac{x_i^2}{(d_i+\hat{\lambda})^3}\frac{\partial\hat{\lambda}}{\partial S}\right]. \tag{3.9}$$

Combining (3.7), (3.8) and (3.9) gives the expression as $R(\omega,\widehat{\boldsymbol{\beta}}^B(\hat{\lambda},\widehat{\boldsymbol{\alpha}}))=R(\omega,\widehat{\boldsymbol{\beta}})+E[\widetilde{\Delta}(\hat{\lambda})]$, where

$$\begin{aligned}\widetilde{\Delta}(\hat{\lambda})=&-2\sum_{i=1}^{p-q}\frac{1}{d_i+\hat{\lambda}}+2\sum_{i=1}^{p-q}\frac{x_i}{(d_i+\hat{\lambda})^2}\frac{\partial\hat{\lambda}}{\partial x_i}\\&+(n-2)\sum_{i=1}^{p-q}\frac{x_i^2/S}{(d_i+\hat{\lambda})^2}-4\sum_{i=1}^{p-q}\frac{x_i^2}{(d_i+\hat{\lambda})^3}\frac{\partial\hat{\lambda}}{\partial S}.\end{aligned} \tag{3.10}$$

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

$$\sum_{i=1}^{p-q}\frac{x_i}{(d_i+\hat{\lambda})^2}\frac{\partial\hat{\lambda}}{\partial x_i}\le\frac{2}{d_{p-q}+\hat{\lambda}}. \tag{3.11}$$

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

$$(n-2)\sum_{i=1}^{p-q}\frac{x_i^2/S}{(d_i+\hat{\lambda})^2}-4\sum_{i=1}^{p-q}\frac{x_i^2}{(d_i+\hat{\lambda})^3}\frac{\partial\hat{\lambda}}{\partial S}\le\frac{n-2}{d_{p-q}+\hat{\lambda}}\alpha+\frac{4\beta}{d_{p-q}+\hat{\lambda}}. \tag{3.12}$$

Combining (3.10), (3.11) and (3.12) gives that

$$\widetilde{\Delta}(\lambda)\le-\sum_{i=1}^{p-q}\frac{2}{d_i+\hat{\lambda}}+\frac{(n-2)\alpha+4(\beta+1)}{d_{p-q}+\hat{\lambda}} \tag{3.13}$$

which is not positive if

$$-2\sum_{i=1}^{p-q}\frac{d_{p-q}+\hat{\lambda}}{d_i+\hat{\lambda}}+(n-2)\alpha+4(\beta+1)\le 0. \tag{3.14}$$

From the condition (a), it is noted that

$$\sum_{i=1}^{p-q}\frac{d_{p-q}+\hat{\lambda}}{d_i+\hat{\lambda}}\ge\sum_{i=1}^{p-q}\frac{d_{p-q}+\lambda_m}{d_i+\lambda_m}=p-q-\sum_{i=1}^{p-q}\frac{d_i-d_{p-q}}{d_i+\lambda_m},$$

which is used to get the following condition from (3.14):

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

This inequality is just given by the condition (d) of Theorem 1, which has therefore been proved. ■

3.2 Minimaxity of the empirical Bayes ridge regression estimator

Theorem 1 can be applied to get the sufficient conditions for several adaptive or empirical Bayes ridge regression estimators to be minimax. We first show the minimaxity of the empirical Bayes ridge regression estimator $\boldsymbol{\beta}^{EB} = \widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{EB}, \widehat{\boldsymbol{\alpha}})$ proposed by (2.9). The λ^* defined as a root of the equation (2.6) is expressed in the notation of the model (3.2) as

$$\sum_{i=1}^{p-q} \frac{x_i^2/S}{d_i + \lambda^*} = \frac{p-q-2}{n+2}. \tag{3.15}$$

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

$$F(x_1, \dots, x_{p-q}, S, \lambda) = \sum_{i=1}^{p-q} \frac{x_i^2}{d_i + \lambda} - \frac{p-q-2}{n+2} S,$$

we see that $F(x_1, \dots, x_{p-q}, S, \lambda^*) = 0$. Then we observe that

$$\frac{\partial\lambda^*}{\partial x_i} = -\frac{\partial F}{\partial x_i}\left(\frac{\partial F}{\partial\lambda^*}\right)^{-1} = 2\frac{x_i}{d_i + \lambda^*}\left(\sum_{j=1}^{p-q} \frac{x_j^2}{(d_j + \lambda^*)^2}\right)^{-1} \tag{3.16}$$

$$\frac{\partial\lambda^*}{\partial S} = -\frac{\partial F}{\partial S}\left(\frac{\partial F}{\partial\lambda^*}\right)^{-1} = -\frac{p-q-2}{n+2}\left(\sum_{j=1}^{p-q} \frac{x_j^2}{(d_j + \lambda^*)^2}\right)^{-1}. \tag{3.17}$$

By using these quantities and the equation (3.15), it can be seen that

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

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

$$\sum_{i=1}^{p-q} \frac{d_i - d_{p-q}}{d_i + \lambda_m} \le \frac{p-q-2}{2}, \tag{3.18}$$

which is guaranteed by the equation (2.7) 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 λ_0 satisfies the equation (2.7). Let $\hat{\lambda}_{EB} = \max(\lambda^*, \lambda_0)$ for the root λ^* of the equation (2.6) or (3.15). Then the EB estimator (2.9) $\widehat{\boldsymbol{\beta}}^{EB} = \widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{EB}, \widehat{\boldsymbol{\alpha}})$ is minimax under the loss (1.8) for $p \ge q + 3$.*

3.3 Another minimax adaptive ridge regression estimator

Next, we shall apply Theorem 1 to derive another minimax 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

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

this can be expressed in the notation of the model (3.2) as

$$\hat{\lambda}_{AD}(a, \lambda_a) = \sum_{i=1}^{p-q} x_i^2/(ad_i S) + \lambda_a. \tag{3.19}$$

We shall now show that the conditions (a)-(d) are satisfied by the estimator given in (3.19) 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

$$\begin{aligned}\sum_{i=1}^{p-q} \frac{x_i}{d_i + \hat{\lambda}_{AD}(a, \lambda_a)} \frac{\partial \hat{\lambda}_{AD}(a, \lambda_a)}{\partial x_i} &\leq 2 \sum_{i=1}^{p-q} \frac{x_i^2/(ad_i S)}{d_i + \hat{\lambda}_{AD}(a, \lambda_a)} \\ &\leq \frac{2\hat{\lambda}_{AD}(a, \lambda_a)}{d_{p-q} + \hat{\lambda}_{AD}(a, \lambda_a)} \leq 2.\end{aligned}$$

For the condition (c), we observe that

$$\sum_{i=1}^{p-q} \frac{x_i^2/S}{d_i + \hat{\lambda}_{AD}(a, \lambda_a)} \leq ad_1 \frac{\sum_{i=1}^{p-q} x_i^2/(ad_i S)}{d_{p-q} + \hat{\lambda}_{AD}(a, \lambda_a)} \leq ad_1,$$

$$-\sum_{i=1}^{p-q} \frac{x_i^2}{(d_i + \hat{\lambda}_{AD}(a, \lambda_a))^2} \frac{\partial \hat{\lambda}_{AD}(a, \lambda_a)}{\partial S} \leq \frac{ad_1 \{\hat{\lambda}_{AD}(a, \lambda_a)\}^2}{(d_{p-q} + \hat{\lambda}_{AD}(a, \lambda_a))^2} \leq ad_1,$$

which imply that the condition is satisfied by putting $\alpha = \beta = ad_1$. Hence the condition (d) is given by

$$\sum_{i=1}^{p-q} \frac{d_i - d_{p-q}}{d_i + \lambda_a} + \frac{(n+2)d_1}{2} a \leq p - q - 2. \tag{3.20}$$

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 (3.20). This root is equal to the solution λ_0 of the equation (2.7).

Proposition 2. *Let λ_0 be a solution of the equation (2.7). Then the adaptive ridge regression estimator $\widehat{\boldsymbol{\beta}}^{AD} = \widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{AD}, \widehat{\boldsymbol{\alpha}})$ with*

$$\begin{aligned}\hat{\lambda}_{AD} &= \frac{n+2}{p-q-2} \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{PC})^t \boldsymbol{A}^t \boldsymbol{A}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{PC})}{\mathrm{ch}_{min}(\boldsymbol{A}^t \boldsymbol{A}) S} + \lambda_0 \\ &= \frac{(n+2)d_1}{p-q-2} \sum_{i=1}^{p-q} x_i^2/(d_i S) + \lambda_0\end{aligned} \tag{3.21}$$

is minimax under the loss (1.8) for $p \geq q+3$, 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.

3.4 A modified adaptive ridge regression estimator

It is noted that the estimator (3.21) has a shortcoming for smaller d_{p-q}. In fact, when $q=0$ and 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

$$\begin{aligned}\hat{\lambda}_{TR} &= \max\left\{\frac{(n+2)(d_1+1)}{(p-q-2)S}(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}^{PC})^t[(\boldsymbol{A}^t\boldsymbol{A})^{-1}+\boldsymbol{I}_p]^{-1}(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}^{PC}), \lambda_0\right\} \qquad (3.22)\\ &= \max\left\{\frac{(n+2)(d_1+1)}{(p-q-2)S}\sum_{i=1}^{p-q}\frac{x_i^2}{d_i+1}, \lambda_0\right\}.\end{aligned}$$

It is easy to see that $\hat{\lambda}_{TR}$ is bounded for d_p going to zero as well as $\hat{\lambda}_{TR} \le \hat{\lambda}_{AD}$. 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 (2.7) if $p \ge q+3$.*

4 Minimaxity of the Hierarchical and Decomposed Empirical Bayes Estimators

We here show that the hierarchical and decomposed empirical Bayes ridge regression estimators $\widehat{\boldsymbol{\beta}}^{HB}$ given by (2.16) and $\widehat{\boldsymbol{\beta}}^{DB}$ given by (2.18) have uniformly smaller risks than the LS estimator $\widehat{\boldsymbol{\beta}}$ relative to the Strawderman's loss.

To handle the estimator $\widehat{\boldsymbol{\beta}}^{HB}$ more easily, we use the canonical model (3.2) to get the expression

$$\begin{aligned}\widehat{\boldsymbol{\theta}}^{HB} &= \boldsymbol{H}\widehat{\boldsymbol{\beta}}^{HB}\\ &= \begin{pmatrix}\boldsymbol{x}_{(1)} - \boldsymbol{D}_1(\boldsymbol{D}_1+\hat{\lambda}_{EB}\boldsymbol{I}_{p-q})^{-1}\boldsymbol{x}_{(1)}\\ \boldsymbol{x}_{(2)} - \boldsymbol{D}_2(\boldsymbol{D}_2+(\hat{\lambda}_{EB}+\hat{\tau}_{HB})\boldsymbol{I}_q)^{-1}(\boldsymbol{x}_{(2)}-\boldsymbol{\alpha}_0)\end{pmatrix}.\end{aligned} \qquad (4.1)$$

Also the equation (2.13) is rewritten by

$$(\boldsymbol{x}_{(2)}-\boldsymbol{\alpha}_0)^t(\boldsymbol{D}_2+\psi^*\boldsymbol{I}_q)^{-1}(\boldsymbol{x}_{(2)}-\boldsymbol{\alpha}_0) = (q-2)S/(n+2). \qquad (4.2)$$

Under the same notations, the estimator $\widehat{\boldsymbol{\beta}}^{DB}$ is expressed by

$$\begin{aligned}\widehat{\boldsymbol{\theta}}^{DB} &= \boldsymbol{H}\widehat{\boldsymbol{\beta}}^{DB}\\ &= \begin{pmatrix}\boldsymbol{x}_{(1)} - \boldsymbol{D}_1(\boldsymbol{D}_1+\hat{\lambda}_{EB}\boldsymbol{I}_{p-q})^{-1}\boldsymbol{x}_{(1)}\\ \boldsymbol{x}_{(2)} - \boldsymbol{D}_2(\boldsymbol{D}_2+\widehat{\psi}_{DB}\boldsymbol{I}_q)^{-1}(\boldsymbol{x}_{(2)}-\boldsymbol{\alpha}_0)\end{pmatrix}.\end{aligned} \qquad (4.3)$$

The minimaxities of the estimators $\widehat{\boldsymbol{\beta}}^{HB}$ and $\widehat{\boldsymbol{\beta}}^{DB}$ are established by the following theorems.

Theorem 2. *Assume that $p-q \geq 3$ and $q \geq 3$. Then the hierarchical empirical Bayes estimator $\widehat{\boldsymbol{\beta}}^{HB}$ dominates the LS estimator $\widehat{\boldsymbol{\beta}}$ under the Strawderman's loss (1.8).*

Theorem 3. *Assume that $p-q \geq 3$ and $q \geq 3$. Then the decomposed empirical Bayes estimator $\widehat{\boldsymbol{\beta}}^{DB}$ dominates the LS estimator $\widehat{\boldsymbol{\beta}}$ under the Strawderman's loss (1.8).*

Proof of Theorem 2. For simplicity, in this proof, we use the notations $\hat{\lambda}$, $\hat{\tau}$ and $\widehat{\psi}$ instead of $\hat{\lambda}_{EB}$, $\hat{\tau}_{HB}$ and $\widehat{\psi}_{HB}$ respectively. From the expression (4.1), the risk function of $\widehat{\boldsymbol{\beta}}^{HB}$ is written by

$$\begin{aligned}R(\omega, \widehat{\boldsymbol{\beta}}^{HB}) =& \sum_{i=1}^{p-q} E\left[\left(x_i - \theta_i - \frac{d_i}{d_i + \hat{\lambda}} x_i\right)^2 / (d_i^2 \sigma^2)\right] \\ &+ \sum_{i=p-q+1}^{p} E\left[\left(x_i - \theta_i - \frac{d_i}{d_i + \hat{\lambda} + \hat{\tau}}(x_i - \alpha_i)\right)^2 / (d_i^2 \sigma^2)\right]\end{aligned}$$

for $\boldsymbol{\alpha}_0 = (\alpha_{01}, \dots, \alpha_{0q})^t$. By the same arguments as in (3.10), the risk $R(\omega, \widehat{\boldsymbol{\beta}}^{HB})$ can be rewritten as $R(\omega, \widehat{\boldsymbol{\beta}}^{HB}) = R(\omega, \widehat{\boldsymbol{\beta}}) + E[\widetilde{\Delta}_1 + \widetilde{\Delta}_2]$, where

$$\begin{aligned}\widetilde{\Delta}_1 =& -2 \sum_{i \leq p-q} \frac{1}{d_i + \hat{\lambda}} + 2 \sum_{i \leq p-q} \frac{x_i}{(d_i + \hat{\lambda})^2} \frac{\partial \hat{\lambda}}{\partial x_i} \\ &+ (n-2) \sum_{i \leq p-q} \frac{x_i^2 / S}{(d_i + \hat{\lambda})^2} - 4 \sum_{i \leq p-q} \frac{x_i^2}{(d_i + \hat{\lambda})^3} \frac{\partial \hat{\lambda}}{\partial S}\end{aligned} \tag{4.4}$$

and

$$\begin{aligned}\widetilde{\Delta}_2 =& -2 \sum_{i > p-q} \frac{1}{d_i + \hat{\lambda} + \hat{\tau}} + 2 \sum_{i > p-q} \frac{x_i - \alpha_{0i}}{(d_i + \hat{\lambda} + \hat{\tau})^2} \frac{\partial (\hat{\lambda} + \hat{\tau})}{\partial x_i} \\ &+ (n-2) \sum_{i > p-q} \frac{(x_i - \alpha_{0i})^2 / S}{(d_i + \hat{\lambda} + \hat{\tau})^2} - 4 \sum_{i > p-q} \frac{(x_i - \alpha_{0i})^2}{(d_i + \hat{\lambda} + \hat{\tau})^3} \frac{\partial (\hat{\lambda} + \hat{\tau})}{\partial S}.\end{aligned} \tag{4.5}$$

Since $\widetilde{\Delta}_1$ is equal to $\widetilde{\Delta}(\hat{\lambda})$ given by (3.10) and $\hat{\lambda}_{EB} \geq \lambda_0$, combining the proof of Theorem 1 and the arguments in Subsection 3.2 gives that

$$\widetilde{\Delta}_1 \leq \frac{1}{d_{p-q} + \hat{\lambda}} \left(2 \sum_{i=1}^{p-q} \frac{d_i - d_{p-q}}{d_i + \lambda_0} - (p-q-2)\right), \tag{4.6}$$

which is equal to zero from the definition of λ_0.

To prove that $\widetilde{\Delta}_2 \leq 0$, note that $\hat{\lambda}$ is a function of $x_1, \dots, x_{p-q}$ and that $\hat{\lambda} + \hat{\tau} = \widehat{\psi} I_A + \hat{\lambda} I_{A^c}$ for $A = \{\boldsymbol{x} | \widehat{\psi} > \hat{\lambda}\}$ and $A^c = \{\boldsymbol{x} | \widehat{\psi} \leq \hat{\lambda}\}$. Then, the same arguments as in

(3.16) and (3.17) give that for $p-q<i\le p$,

$$\begin{aligned}\frac{\partial(\hat{\lambda}+\hat{\tau})}{\partial x_i} =&\frac{\partial\psi^*}{\partial x_i}I(\psi^*>\psi_m)I_A\\ =&2\frac{x_i-\alpha_{0i}}{d_i+\psi^*}\left(\sum_{j>p-q}\frac{(x_j-\alpha_{0j})^2}{(d_j+\psi^*)^2}\right)^{-1}I(\psi^*>\psi_m)I_A \qquad (4.7)\\ \frac{\partial(\hat{\lambda}+\hat{\tau})}{\partial S} =&\frac{\partial\psi^*}{\partial S}I(\psi^*>\psi_m)I_A+\frac{\partial\lambda^*}{\partial S}I(\lambda^*>\lambda_0)I_{A^c}\\ =&-\frac{q-2}{n+2}\left(\sum_{j>p-q}\frac{(x_j-\alpha_{0j})^2}{(d_j+\psi^*)^2}\right)^{-1}I(\psi^*>\psi_m)I_A\\ &-\frac{p-q-2}{n+2}\left(\sum_{j\le p-q}\frac{x_j^2}{(d_j+\lambda^*)^2}\right)^{-1}I(\lambda^*>\lambda_0)I_{A^c}, \qquad (4.8)\end{aligned}$$

for $\psi_m=\max(\psi_0,\psi_1)$. From (4.7), the second term in the r.h.s. of (4.5) is evaluated as

$$\begin{aligned}2\sum_{i>p-q}\frac{x_i-\alpha_{0i}}{(d_i+\hat{\lambda}+\hat{\tau})^2}\frac{\partial(\hat{\lambda}+\hat{\tau})}{\partial x_i} =&4\frac{\sum_{i>p-q}(x_i-\alpha_{0i})^2/(d_i+\psi^*)^3}{\sum_{i>p-q}(x_i-\alpha_{0i})^2/(d_i+\psi^*)^2}I(\psi^*>\psi_m)I_A\\ \le&\frac{4}{d_p+\psi^*}I(\psi^*>\psi_m)I_A. \qquad (4.9)\end{aligned}$$

From (4.8), the fourth term in the r.h.s. of (4.5) is also evaluated by

$$\begin{aligned}-4&\sum_{i>p-q}\frac{(x_i-\alpha_{0i})^2}{(d_i+\hat{\lambda}+\hat{\tau})^3}\frac{\partial(\hat{\lambda}+\hat{\tau})}{\partial S}\\ =&4\frac{q-2}{n+2}\frac{\sum_{i>p-q}(x_i-\alpha_{0i})^2/(d_i+\psi^*)^3}{\sum_{i>p-q}(x_i-\alpha_{0i})^2/(d_i+\psi^*)^2}I(\psi^*>\psi_m)I_A\\ &+4\frac{p-q-2}{n+2}\frac{\sum_{i>p-q}(x_i-\alpha_{0i})^2/(d_i+\lambda^*)^3}{\sum_{i\le p-q}x_i^2/(d_i+\lambda^*)^2}I(\lambda^*>\lambda_0)I_{A^c}\\ \le&4\frac{q-2}{n+2}\frac{1}{d_p+\psi^*}I(\psi^*>\psi_m)I_A\\ &+4\frac{p-q-2}{n+2}\frac{d_1+\lambda^*}{(d_p+\lambda^*)^2}\frac{\sum_{i>p-q}(x_i-\alpha_{0i})^2/(d_i+\lambda^*)}{\sum_{i\le p-q}x_i^2/(d_i+\lambda^*)}I(\lambda^*>\lambda_0)I_{A^c}. \qquad (4.10)\end{aligned}$$

Note that the λ^* is the solution of the equation (3.15). Since $(d_1+x)/(d_p+x)$ is decreasing in x, we can evaluate the second term in the r.h.s. of the inequality (4.10) as

$$\begin{aligned}4&\frac{p-q-2}{n+2}\frac{d_1+\lambda^*}{(d_p+\lambda^*)^2}\frac{\sum_{i>p-q}(x_i-\alpha_{0i})^2/(d_i+\lambda^*)}{\sum_{i\le p-q}x_i^2/(d_i+\lambda^*)}I(\lambda^*>\lambda_0)I_{A^c}\\ &\le4\frac{d_1+\lambda^*}{(d_p+\lambda^*)^2}\sum_{i>p-q}\frac{(x_i-\alpha_{0i})^2}{d_i+\lambda^*}I(\lambda^*>\lambda_0)I_{A^c}. \qquad (4.11)\end{aligned}$$

Since $\hat{\lambda}\ge\widehat{\psi}\ge\psi^*$ on A^c and $\lambda^*=\hat{\lambda}$ on $\{\lambda^*>\lambda_0\}$, we see that the r.h.s. of the inequality

(4.11) is evaluated as

$$\begin{aligned}
4\frac{d_1+\hat{\lambda}}{(d_p+\hat{\lambda})^2}\sum_{i>p-q}\frac{(x_i-\alpha_{0i})^2}{d_i+\hat{\lambda}}I(\lambda^*>\lambda_0)I_{A^c} \le & 4\frac{d_1+\hat{\lambda}}{(d_p+\hat{\lambda})^2}\sum_{i>p-q}\frac{(x_i-\alpha_{0i})^2}{d_i+\psi^*}I(\lambda^*>\lambda_0)I_{A^c} \\
= & 4\frac{q-2}{n+2}\frac{d_1+\hat{\lambda}}{(d_p+\hat{\lambda})^2}I(\lambda^*>\lambda_0)I_{A^c},
\end{aligned} \tag{4.12}$$

where the last equality follows from the fact that ψ^* is the solution of the equation (4.2).

Noting again that $\hat{\lambda}+\hat{\tau}=\widehat{\psi}I_A+\hat{\lambda}I_{A^c}$ and combining (4.5), (4.9), (4.10) and (4.12), we see that

$$\begin{aligned}
\widetilde{\Delta}_2 = & I_A\left\{-\sum_{i>p-q}\frac{2}{d_i+\widehat{\psi}}+\frac{4}{d_p+\widehat{\psi}}I(\psi^*>\psi_m)+\frac{q-2}{n+2}\frac{n-2}{d_p+\widehat{\psi}}+\frac{4}{n+2}\frac{q-2}{d_p+\widehat{\psi}}I(\psi^*>\psi_m)\right\} \\
& + I_{A^c}\left\{-\sum_{i>p-q}\frac{2}{d_i+\hat{\lambda}}+\frac{q-2}{n+2}\frac{n-2}{d_p+\hat{\lambda}}+4\frac{q-2}{n+2}\frac{d_1+\hat{\lambda}}{(d_p+\hat{\lambda})^2}I(\lambda^*>\lambda_0)\right\} \qquad (4.13) \\
= & \widetilde{\Delta}_{21}+\widetilde{\Delta}_{22}. \quad (say)
\end{aligned}$$

Since $\widehat{\psi}\ge\psi_m=\max(\psi_0,\psi_1)\ge\psi_0$, it is seen that

$$\begin{aligned}
\widetilde{\Delta}_{21} \le & I_A\frac{1}{d_p+\widehat{\psi}}\left\{-2\sum_{i>p-q}\frac{d_p+\widehat{\psi}}{d_i+\widehat{\psi}}+4+(n-2)\frac{q-2}{n+2}+4\frac{q-2}{n+2}\right\} \\
\le & I_A\frac{1}{d_p+\widehat{\psi}}\left\{-2\sum_{i>p-q}\frac{d_p+\psi_0}{d_i+\psi_0}+q+2\right\} \\
= & I_A\frac{1}{d_p+\widehat{\psi}}\left\{2\sum_{i>p-q}\frac{d_i-d_p}{d_i+\psi_0}-(q-2)\right\},
\end{aligned} \tag{4.14}$$

which is equal to zero from the definition (2.13) of ψ_0. Since $\hat{\lambda}\ge\widehat{\psi}\ge\psi_1$ on the set A^c, we have that

$$\begin{aligned}
\widetilde{\Delta}_{22} \le & I_{A^c}\frac{1}{d_p+\hat{\lambda}}\left\{-2\sum_{i>p-q}\frac{d_p+\hat{\lambda}}{d_i+\hat{\lambda}}+(n-2)\frac{q-2}{n+2}+4\frac{q-2}{n+2}\frac{d_1+\hat{\lambda}}{d_p+\hat{\lambda}}\right\} \\
\le & I_{A^c}\frac{1}{d_p+\hat{\lambda}}\left\{-2\sum_{i>p-q}\frac{d_p+\psi_1}{d_i+\psi_1}+(n-2)\frac{q-2}{n+2}+4\frac{q-2}{n+2}\frac{d_1+\psi_1}{d_p+\psi_1}\right\} \\
= & I_{A^c}\frac{1}{d_p+\hat{\lambda}}\left\{2\sum_{i>p-q}\frac{d_i-d_p}{d_i+\psi_1}+4\frac{q-2}{n+2}\frac{d_1-d_p}{d_p+\psi_1}-(q+2)\right\},
\end{aligned}$$

which is equal to zero from the definition of ψ_1. Therefore the proof of Theorem 2 is complete. ■

Proof of Theorem 3. From the expression (4.3), the risk function of $\widehat{\boldsymbol{\beta}}^{DB}$ is written by

$$R(\omega, \widehat{\boldsymbol{\beta}}^{DB}) = \sum_{i=1}^{p-q} E\left[\left(x_i - \theta_i - \frac{d_i}{d_i + \hat{\lambda}_{EB}} x_i\right)^2 / (d_i^2 \sigma^2)\right]$$
$$+ \sum_{i=p-q+1}^{p} E\left[\left(x_i - \theta_i - \frac{d_i}{d_i + \widehat{\psi}_{DB}} (x_i - \alpha_i)\right)^2 / (d_i^2 \sigma^2)\right].$$

By the same arguments as in the proof of Theorem 2, the risk $R(\omega, \widehat{\boldsymbol{\beta}}^{DB})$ can be rewritten as $R(\omega, \widehat{\boldsymbol{\beta}}^{DB}) = R(\omega, \widehat{\boldsymbol{\beta}}) + E[\widetilde{\Delta}_1 + \widetilde{\Delta}_2^*]$, where $\widetilde{\Delta}_1$ is given by (4.4) and

$$\widetilde{\Delta}_2^* = -2 \sum_{i>p-q} \frac{1}{d_i + \widehat{\psi}_{DB}} + 2 \sum_{i>p-q} \frac{x_i - \alpha_{0i}}{(d_i + \widehat{\psi}_{DB})^2} \frac{\partial \widehat{\psi}_{DB}}{\partial x_i}$$
$$+ (n-2) \sum_{i>p-q} \frac{(x_i - \alpha_{0i})^2 / S}{(d_i + \widehat{\psi}_{DB})^2} - 4 \sum_{i>p-q} \frac{(x_i - \alpha_{0i})^2}{(d_i + \widehat{\psi}_{DB})^3} \frac{\partial \widehat{\psi}_{DB}}{\partial S}. \tag{4.15}$$

From (4.6), it follows that $\widetilde{\Delta}_1 \leq 0$. $\widetilde{\Delta}_2^*$ corresponds to the case of $I_A = 1$ in $\widetilde{\Delta}_2$ in the proof of Theorem 2. Hence from (4.14), it is seen that

$$\widetilde{\Delta}_2^* \leq \frac{1}{d_p + \widehat{\psi}} \left\{ 2 \sum_{i>p-q} \frac{d_i - d_p}{d_i + \psi_0} - (q-2) \right\},$$

which is zero from the definition of ψ_0, proving Theorem 3. ■

5 Simulation and Empirical Studies

Now we investigate the risk-performances of estimators of $\boldsymbol{\beta}$ numerically. The estimators we want to investigate are described below: The usual ridge regression estimators in the multicollinearity case shrink the LS estimator toward zero, that is, $H_0 : \boldsymbol{\beta} = \mathbf{0}$. In this case, the adaptive or empirical Bayes ridge regression estimators are written by

$$\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}, \mathbf{0}) = \left[\boldsymbol{A}^t \boldsymbol{A} + \hat{\lambda}^{-1} \boldsymbol{I}\right]^{-1} \boldsymbol{A}^t \boldsymbol{y} = \widehat{\boldsymbol{\beta}} - [\boldsymbol{I} + \hat{\lambda} \boldsymbol{A}^t \boldsymbol{A}]^{-1} \widehat{\boldsymbol{\beta}}. \tag{5.1}$$

Three estimators $\hat{\lambda}_{AD}$, $\hat{\lambda}_{TR}$ and $\hat{\lambda}_{EB}$ of λ are given by (3.21), (3.22) and (2.8) with $q = 0$, and these estimators of λ yield the estimators

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

respectively, whose minimaxities were shown by Propositions 1, 2 and 3 with $q = 0$ for $p \geq 3$. Adaptive or empirical Bayes ridge regression estimators shrunken towards the linear hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{H}_2^t \boldsymbol{\alpha}$ for $\boldsymbol{\alpha} \in \boldsymbol{R}^q$, are given by

$$\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{AD}, \widehat{\boldsymbol{\alpha}}), \widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{TR}, \widehat{\boldsymbol{\alpha}}), \widehat{\boldsymbol{\beta}}^{EB} = \widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{EB}, \widehat{\boldsymbol{\alpha}}),\ \text{denoted by } AD_{p-q}, TR_{p-q}, EB_{p-q},$$

respectively. Their minimaxities are guaranteed by Propositions 1, 2 and 3. We also treat the hierarchical and decomposed empirical Bayes ridge regression estimators

$$\widehat{\boldsymbol{\beta}}^{HB} \text{ and } \widehat{\boldsymbol{\beta}}^{DB},\ \text{denoted by } HB_{p-q} \text{ and } DB_{p-q},$$

Table 1: Relative Efficiencies of the Estimators under L_0, L_1, L_2 Losses for $\boldsymbol{D} = \mathrm{diag}\,(10316., 195.0, 73.4, 20.2, 2.6, 1.0, 0.9, 0.5, 0.2)$, $p = 9$, $q = 5$, $n = 6$, $\theta_i = i \times \eta$, $i = 1, \ldots, 9$

	η	AD	TR	EB	AD_4	TR_4	EB_4	HB_4	DB_4	PC_4	PC_1
L_0	0	0.494	0.417	0.059	0.639	0.625	0.309	0.309	0.309	0.000	0.026
	1	0.985	0.945	0.066	0.670	0.656	0.310	0.310	0.310	0.005	0.026
	2	0.996	0.985	0.079	0.739	0.728	0.312	0.312	0.312	0.021	0.026
	3	0.998	0.993	0.095	0.811	0.803	0.315	0.315	0.315	0.047	0.027
	4	0.999	0.996	0.114	0.866	0.860	0.320	0.320	0.320	0.083	0.027
L_1	0	0.938	0.927	0.607	0.956	0.954	0.894	0.894	0.637	0.445	0.887
	1	0.998	0.993	0.730	0.960	0.958	0.894	0.894	0.890	1.635	0.887
	2	0.999	0.998	0.800	0.968	0.967	0.895	0.895	0.894	5.204	0.887
	3	0.999	0.999	0.837	0.977	0.976	0.896	0.896	0.896	11.15	0.887
	4	0.999	0.999	0.858	0.984	0.983	0.898	0.898	0.898	19.48	0.887
L_2	0	0.999	0.999	0.952	0.999	0.999	0.999	0.999	0.670	0.952	0.999
	1	1.000	1.000	0.995	0.999	0.999	0.999	0.999	0.996	1.346	0.999
	2	1.000	1.000	0.998	0.999	0.999	0.999	0.999	0.998	2.526	0.999
	3	1.000	1.000	0.999	0.999	0.999	0.999	0.999	0.999	4.494	0.999
	4	1.000	1.000	0.999	1.000	1.000	0.999	0.999	0.999	7.249	0.999

which are minimax from Theorems 2 and 3. As alternative estimators, we deal with the principal component regression estimators $\widehat{\boldsymbol{\beta}}^{PC}_{p-q} = \boldsymbol{H}_2^t \widehat{\boldsymbol{\alpha}} = \boldsymbol{H}_2^t \boldsymbol{H}_2 \widehat{\boldsymbol{\beta}}$, denoted by PC_{p-q}, and $\widehat{\boldsymbol{\beta}}^{PC}_1$, denoted by PC_1, where PC_{p-q} is obtained by deleting the eigenvectors corresponding to the $p-q$ largest eigenvalues of $(\boldsymbol{A}^t\boldsymbol{A})^{-1}$ and PC_1 corresponds to the one obtained by deleting the largest eigenvalue.

We thus compare the estimators AD, TR and EB for $q = 0$; AD_{p-q}, TR_{p-q}, EB_{p-q}, HB_{p-q}, DB_{p-q} and PC_{p-q} for $\boldsymbol{\alpha} \in \boldsymbol{R}^q$; PC_1 for $\boldsymbol{\alpha} \in \boldsymbol{R}^{p-1}$. 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 LS estimator $\widehat{\boldsymbol{\beta}}$ are obtained from 1,000 replications 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 three cases:

Case 1: $\boldsymbol{D} = \mathrm{diag}\,(10316., 195.0, 73.4, 20.2, 2.6, 1.0, 0.9, 0.5, 0.2)$, $p = 9$, $q = 5$, $n = 6$ and $\theta_i = i \times \eta$, $i = 1, \ldots, 9$

Case 2: $\boldsymbol{D} = \mathrm{diag}\,(300, 250, 200, 150, 100, 100, 100, 80, 80, 80, 1, 1, 1, 1, 1)$, $p = 15$, $q = 5$, $n = 50$ and $\theta_i = (3i + 1) \times \eta$, $i = 1, \ldots, 15$

Case 3: $\boldsymbol{D} = \mathrm{diag}\,(300, 250, 200, 150, 100, 10, 10, 10, 5, 5, 5, 1, 1, 1, 1)$, $p = 15$, $q = 10$, $n = 50$, $\theta_i = (p - i + 2)\sqrt{\eta}$, $i = 1, \ldots, 10$

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

(1) The empirical Bayes estimator EB for $q = 0$, namely $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{EB}, \boldsymbol{0})$ has a very nice risk behavior for L_0- and L_1- losses; it is highly recommended in the case of multicollinear-

Table 2: Relative Efficiencies of the Estimators under L_0, L_1, L_2 Losses for $\boldsymbol{D} = \mathrm{diag}\,(300, 250, 200, 150, 100, 100, 100, 80, 80, 80, 1, 1, 1, 1, 1)$, $p = 15$, $q = 5$, $n = 50$, $\theta_i = (3i + 1) \times \eta$, $i = 1, \ldots, 15$

	η	AD	TR	EB	AD_{10}	TR_{10}	EB_{10}	HB_{10}	DB_{10}	PC_{10}	PC_1
L_0	0	0.541	0.456	0.115	0.721	0.693	0.525	0.348	0.240	0.309	0.803
	1	0.998	0.996	0.763	0.812	0.800	0.689	0.662	0.662	0.716	0.814
	2	0.999	0.999	0.923	0.909	0.906	0.850	0.843	0.843	1.935	0.847
	3	0.999	0.999	0.963	0.951	0.951	0.916	0.913	0.913	3.967	0.902
	4	0.999	0.999	0.978	0.970	0.970	0.948	0.946	0.946	6.813	0.979
L_1	0	0.732	0.675	0.415	0.875	0.862	0.777	0.578	0.437	0.663	0.936
	1	0.999	0.998	0.877	0.921	0.916	0.870	0.842	0.842	0.958	0.940
	2	0.999	0.999	0.961	0.964	0.964	0.946	0.939	0.939	1.843	0.950
	3	0.999	0.999	0.981	0.981	0.981	0.970	0.967	0.967	3.319	0.968
	4	0.999	0.999	0.989	0.989	0.989	0.981	0.979	0.979	5.384	0.993
L_2	0	0.989	0.985	0.956	0.998	0.997	0.996	0.975	0.901	0.994	0.999
	1	0.999	0.999	0.997	0.998	0.998	0.998	0.997	0.997	1.001	0.999
	2	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	1.023	0.999
	3	1.000	0.999	0.999	0.999	0.999	0.999	0.999	0.999	1.059	0.999
	4	1.000	1.000	0.999	0.999	0.999	0.999	0.999	0.999	1.110	0.999

Table 3: Relative Efficiencies of the Estimators under L_0, L_1, L_2 Losses for $\boldsymbol{D} = \mathrm{diag}\,(300, 250, 200, 150, 100, 10, 10, 10, 5, 5, 5, 1, 1, 1, 1)$, $p = 15$, $q = 10$, $n = 50$, $\theta_i = (p - i + 2)\sqrt{\eta}$, $i = 1, \ldots, 15$

	η	AD	TR	EB	AD_5	TR_5	EB_5	HB_5	DB_5	PC_5	PC_1
L_0	0	0.460	0.405	0.033	0.514	0.469	0.112	0.111	0.105	0.046	0.728
	1	0.886	0.851	0.614	0.869	0.855	0.596	0.595	0.596	0.988	0.971
	2	0.936	0.915	0.823	0.924	0.916	0.767	0.767	0.767	1.930	1.215
	3	0.955	0.940	0.917	0.947	0.941	0.845	0.845	0.845	2.872	1.459
	4	0.966	0.954	0.964	0.959	0.954	0.887	0.887	0.887	3.814	1.702
L_1	0	0.806	0.782	0.411	0.830	0.811	0.618	0.599	0.365	0.663	0.936
	1	0.962	0.950	0.851	0.957	0.952	0.851	0.844	0.847	1.016	0.993
	2	0.979	0.972	0.928	0.975	0.973	0.914	0.911	0.913	1.368	1.050
	3	0.985	0.980	0.961	0.983	0.981	0.942	0.940	0.942	1.721	1.107
	4	0.988	0.985	0.977	0.986	0.985	0.957	0.955	0.957	2.073	1.163
L_2	0	0.986	0.983	0.789	0.993	0.991	0.964	0.930	0.322	0.994	0.999
	1	0.998	0.997	0.972	0.998	0.998	0.992	0.978	0.971	1.000	0.999
	2	0.999	0.998	0.986	0.999	0.999	0.995	0.989	0.985	1.007	1.000
	3	0.999	0.999	0.991	0.999	0.999	0.997	0.992	0.989	1.014	1.001
	4	0.999	0.999	0.993	0.999	0.999	0.997	0.994	0.992	1.021	1.001

ity. As seen from Table 1, the risk performance is quite well when d_1 is extremely large.

(2) As seen from Tables 2 and 3, the empirical Bayes estimators EB_{p-q}, HB_{p-q} and DB_{p-q} are much better than the LS estimator for L_0- and L_1- losses. The estimators HB_{p-q} and DB_{p-q} have slightly smaller risks than EB except for small values of η. In the case where several eigenvalues d_i's are large, the estimators HB_{p-q} and DB_{p-q} are also recommended.

(3) Although the minimaxity of the proposed estimators are guaranteed under the L_2-loss, their risk performances are much better than the LS estimator under L_0- and L_1-loss functions.

(4) Through the tables, we see that the principal component regression estimator PC_{p-q} has the smallest risks for smaller values of $\boldsymbol{\theta}$ and gets larger as $\|\boldsymbol{\theta}\|$ increases.

We shall provide an empirical study for a set of data.

Example 1. (*Response Surface*) We 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 multicollinearity and the above data have been repeatedly analyzed by Beisley (1984) 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

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

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 $k = 1/\lambda \in [0, 0.07]$ 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 estimator shrunken towards zero $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{TR}, \mathbf{0})$ (TR) and the empirical Bayes ridge regression estimator shrunken towards zero $\widehat{\boldsymbol{\beta}}^B(\hat{\lambda}_{EB}, \mathbf{0})$ (EB). Since the first four eigenvalues d_1, d_2, d_3, d_4 are not small, we may consider the linear subspace (2.3) constructed by eigenvectors of $(\boldsymbol{A}^t\boldsymbol{A})^{-1}$ with deleting the eigenvectors corresponding to the four largest eigenvalues. We thus deal with the principal component (PC) regression estimator $\widehat{\boldsymbol{\beta}}^{PC}$ (PC_4) under the subspace and

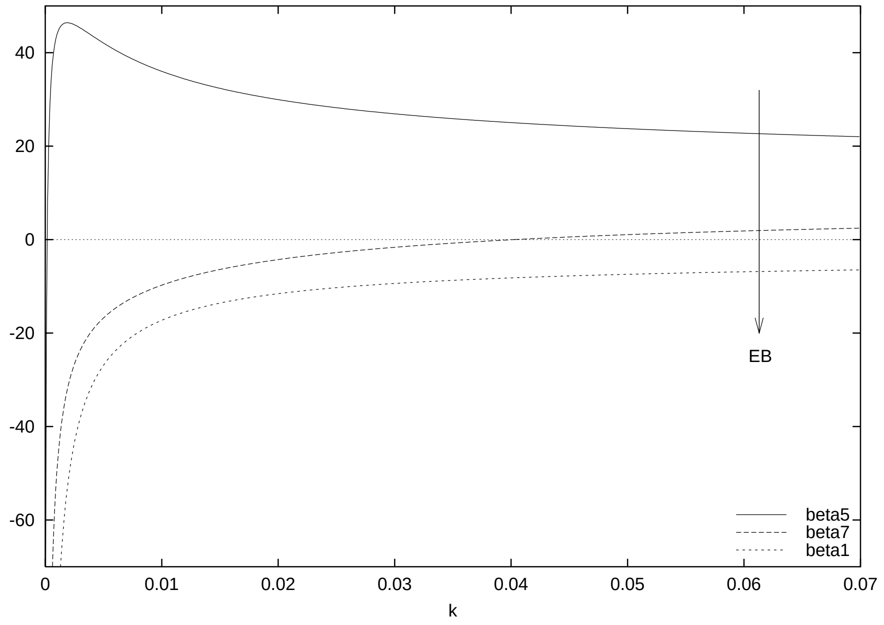

Figure 1: Curves of the Ridge Estimates of β_1, β_5 and β_7 (The horizontal axis denotes the values of $k = 1/\lambda$. The line EB shows the values of $1/\hat{\lambda}_{EB}$ for $q = 0$.)

Table 4: Estimates of $\boldsymbol{\beta}$ and prediction-error estimates for the Eight Estimators LS, TR, EB, EB_4, HB_4, DB_4, PC_4 and PC_1

	LS	TR	EB	EB_4	HB_4	DB_4	PC_4	PC_1
$\widehat{\beta_1}$	-108.5	-45.2	22.6	38.5	38.4	36.5	17.8	65.7
$\widehat{\beta_2}$	21.2	20.0	14.5	18.1	18.1	14.4	15.7	17.9
$\widehat{\beta_3}$	-197.5	-111.4	-9.5	2.7	2.7	4.4	-14.6	39.3
$\widehat{\beta_4}$	7.2	8.1	-2.6	9.1	9.1	9.1	-4.9	9.7
$\widehat{\beta_5}$	-814.7	-522.9	-6.8	-120.5	-120.5	-119.6	-4.4	-17.3
$\widehat{\beta_6}$	11.3	14.3	8.1	18.4	18.4	18.4	5.4	19.6
$\widehat{\beta_7}$	-426.5	-275.7	1.9	-67.6	-67.6	-69.0	8.6	-14.8
$\widehat{\beta_8}$	-20.5	-18.6	-10.2	-15.6	-15.6	-12.5	-11.5	-15.5
$\widehat{\beta_9}$	-331.5	-210.9	4.2	-43.6	-43.6	-44.4	1.3	-2.6
PE	299	276	114	267	267	252	100	270

Table 5: Estimates of $\boldsymbol{\theta}$ for the Eight Estimators LS, TR, EB, EB_4, HB_4, DB_4, PC_4 and PC_1

	d_i	LS	TR	EB	EB_4	HB_4	DB_4	PC_4	PC_1
$\hat{\theta}_1$	10316.04	-1000.0	-634.1	-1.5	-130.9	-130.9	-130.9	0.0	0.0
$\hat{\theta}_2$	195.01	-65.4	-64.6	-5.0	-58.1	-58.1	-58.1	0.0	-65.4
$\hat{\theta}_3$	73.39	-44.3	-44.1	-8.0	-42.3	-42.3	-42.3	0.0	-44.3
$\hat{\theta}_4$	20.18	11.3	11.3	5.0	11.2	11.2	11.2	0.0	11.3
$\hat{\theta}_5$	2.59	-11.3	-11.3	-9.7	-11.3	-11.3	-7.4	-11.3	-11.3
$\hat{\theta}_6$	0.96	-12.7	-12.7	-12.0	-12.7	-12.7	-10.7	-12.7	-12.7
$\hat{\theta}_7$	0.87	-25.5	-25.5	-24.2	-25.5	-25.4	-21.7	-25.5	-25.5
$\hat{\theta}_8$	0.46	4.1	4.1	3.9	4.1	4.1	3.7	4.1	4.1
$\hat{\theta}_9$	0.23	-10.2	-10.2	-10.0	-10.2	-10.2	-9.7	-10.2	-10.2

usual, hierarchical and decomposed empirical Bayes ridge regression estimators shrunken towards the subspace : $\widehat{\boldsymbol{\beta}}^{EB} = \widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{EB_4}, \widehat{\boldsymbol{\alpha}})$ (EB_4), $\widehat{\boldsymbol{\beta}}^{HB}$ (HB_4) and $\widehat{\boldsymbol{\beta}}^{DB}$ (DB_4).

The estimates of λ (or k), τ and ψ are given by $\hat{\lambda}_{TR} = 17,878.4$, $\hat{\lambda}_{EB} = 16.3$, $\hat{\lambda}_{EB_4} = 1,554.2$, $\hat{\tau}_{HB} = 1,965.7$ and $\widehat{\psi}_{DB} = 5.0$. The estimates of $\boldsymbol{\beta}$ for the above procedures are given in Table 4. Since $\hat{\lambda}_{TR}$ is very large, the minimax adaptive ridge regression estimate $\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{TR}, \mathbf{0})$ is very 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 multicollinearity case. From Figure 1 and Table 4, on the other hand, it is seen that $\hat{\lambda}_{EB}$ is estimated appropriately and that the resulting estimator $\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{EB}, \mathbf{0})$ is well stabilized. The hierarchical empirical Bayes estimate $\widehat{\boldsymbol{\beta}}^{HB}$ and the decomposed empirical Bayes estimate $\widehat{\boldsymbol{\beta}}^{DB}$ are almost identical to the empirical Bayes estimate $\widehat{\boldsymbol{\beta}}^{B}(\hat{\lambda}_{EB_4}, \widehat{\boldsymbol{\alpha}})$ shrunken towards the PC estimate. The PC estimator $\widehat{\boldsymbol{\beta}}^{PC}$ gives estimates different from the ridge type estimators. Table 5 gives similar estimates in the canonical model with $\boldsymbol{\theta} = (\theta_1, \dots, \theta_9)^t = \boldsymbol{H}\boldsymbol{\beta}$ and it explains how the proposed procedures work in the presence of the large eigenvalues of $(\boldsymbol{A}^t\boldsymbol{A})^{-1}$. The tabel reveals that the estimates by EB, EB_4, HB_4 and DB_4 gets 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 of the methods considered may be estimated via the leave-one-out cross-validation as described in Srivastava (2002, p322). That is, 16 predictive errors are obtained by leaving out one observation each time. The estimates of the prediction errors for the above considered estimators are given at the last row as PE in Table 4. It reveals that the use of the estimators EB, EB_4, HB_4, DB_4 and PC_4 provides smaller prediction errors than the least squares estimator (LS). Of these, EB and PC_4 give much smaller prediction error estimates. It is interesting to note that the ridge-type estimator EB gives estimates different from the PC estimator PC_4, but the estimates of the prediction errors for both procedures are similar. The estimate of the prediction error of PC_1 by the cross-validation method is 270, which is much larger than that of PC_4 and EB. ■

6 Concluding Remarks

We have proposed and compared several empirical Bayes estimators which are minimax under the Strawderman's loss function. Although the idea of shrinking the estimators towards the subhypothesis is an interesting one, it does introduce some arbitrariness as to the selection of the subspace, same as in the principal component regression estimator. On the other hand, the empirical Bayes estimator obtained under the hypothesis that $\boldsymbol{\beta} = \boldsymbol{0}$, given in (5.1), performs reasonably well and requires no special attention. The final choice, however, rests with the analyst.

7 Appendix

We here show the following equation in the expression (2.10) of the hierarchical Bayes estimator $\widehat{\boldsymbol{\beta}}^{HB}(\lambda, \tau)$:

$$\begin{aligned} &\widehat{\boldsymbol{\beta}} - \left\{\boldsymbol{A}^t\boldsymbol{A} + (\lambda\boldsymbol{I}_p + \tau\boldsymbol{H}_2^t\boldsymbol{H}_2)^{-1}\right\}^{-1}(\lambda\boldsymbol{I}_p + \tau\boldsymbol{H}_2^t\boldsymbol{H}_2)^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{H}_2^t\boldsymbol{\alpha}_0) \\ &\quad = \widehat{\boldsymbol{\beta}} - (\boldsymbol{A}^t\boldsymbol{A})^{-1}\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda\boldsymbol{I}_p\right\}^{-1}\left\{\widehat{\boldsymbol{\beta}} - \boldsymbol{H}_2^t\widehat{\boldsymbol{\alpha}}^S(\lambda, \tau)\right\}. \end{aligned} \tag{7.1}$$

For $\boldsymbol{G} = \boldsymbol{G}(\lambda) = (\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda\boldsymbol{I}_p$, the l.h.s. of (7.1) is expressed by

$$\begin{aligned} &\widehat{\boldsymbol{\beta}} - (\boldsymbol{A}^t\boldsymbol{A})^{-1}\left\{(\boldsymbol{A}^t\boldsymbol{A})^{-1} + \lambda\boldsymbol{I}_p + \tau\boldsymbol{H}_2^t\boldsymbol{H}_2\right\}^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{H}_2^t\boldsymbol{\alpha}_0) \\ &\quad = \widehat{\boldsymbol{\beta}} - (\boldsymbol{A}^t\boldsymbol{A})^{-1}\boldsymbol{G}^{-1}(\boldsymbol{I}_p + \tau\boldsymbol{H}_2^t\boldsymbol{H}_2\boldsymbol{G}^{-1})^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{H}_2^t\boldsymbol{\alpha}_0). \end{aligned} \tag{7.2}$$

Noting that

$$\begin{aligned} (\boldsymbol{I}_p + \tau\boldsymbol{H}_2^t\boldsymbol{H}_2\boldsymbol{G}^{-1})^{-1} &= \boldsymbol{I}_p - \tau\boldsymbol{H}_2^t(\boldsymbol{I}_q + \tau\boldsymbol{H}_2\boldsymbol{G}^{-1}\boldsymbol{H}_2^t)^{-1}\boldsymbol{H}_2\boldsymbol{G}^{-1} \\ &= \boldsymbol{I}_p - \tau\boldsymbol{H}_2^t\left\{(\boldsymbol{H}_2\boldsymbol{G}^{-1}\boldsymbol{H}_2^t)^{-1} + \tau\boldsymbol{I}_q\right\}^{-1}(\boldsymbol{H}_2\boldsymbol{G}^{-1}\boldsymbol{H}_2^t)^{-1}\boldsymbol{H}_2\boldsymbol{G}^{-1}, \end{aligned}$$

we see that

$$\begin{aligned} &(\boldsymbol{I}_p + \tau\boldsymbol{H}_2^t\boldsymbol{H}_2\boldsymbol{G}^{-1})^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{H}_2^t\boldsymbol{\alpha}_0) \\ &\quad = \widehat{\boldsymbol{\beta}} - \boldsymbol{H}_2^t\boldsymbol{\alpha}_0 - \tau\boldsymbol{H}_2^t\left\{(\boldsymbol{H}_2\boldsymbol{G}^{-1}\boldsymbol{H}_2^t)^{-1} + \tau\boldsymbol{I}_q\right\}^{-1}(\widehat{\boldsymbol{\alpha}}(\lambda) - \boldsymbol{\alpha}_0) \\ &\quad = \widehat{\boldsymbol{\beta}} - \boldsymbol{H}_2^t\left\{\widehat{\boldsymbol{\alpha}}(\lambda) - (\boldsymbol{I}_q + \tau\boldsymbol{H}_2\boldsymbol{G}^{-1}\boldsymbol{H}_2^t)^{-1}(\widehat{\boldsymbol{\alpha}}(\lambda) - \boldsymbol{\alpha}_0)\right\}, \end{aligned}$$

where $\widehat{\boldsymbol{\alpha}}(\lambda) = (\boldsymbol{H}_2\boldsymbol{G}^{-1}\boldsymbol{H}_2^t)^{-1}\boldsymbol{H}_2\boldsymbol{G}^{-1}\widehat{\boldsymbol{\beta}}$, being the weighted least squares estimator. Hence from (7.2), we get the expression in the r.h.s. of the equation (7.1). ■

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

REFERENCES

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

Björkström, A. and Sundberg, R. (1996). Continuum regression is not always continuous. *J. R. Statist. Soc.*, **B, 58**, 703-710.

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

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

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

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.

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. (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.

Stone, M. and Brooks, R.J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *J. R. Statist. Soc.*, **B, 52**, 237-269.

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

Sundberg, R. (1993). Continuum regression and ridge regression. *J. R. Statist. Soc.*, **B, 55**, 653-659.

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