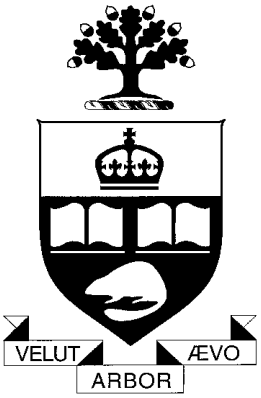

Empirical Bayes Regression Analysis with Many Regressors but Fewer Observations

by

Muni S. Srivastava
Department of Statistics
University of Toronto

and

Tatsuya Kubokawa
Faculty of Economics
University of Tokyo

Technical Report No. 0410 September 15, 2004

TECHNICAL REPORT SERIES

University of Toronto
Department of Statistics

Empirical Bayes Regression Analysis with Many Regressors but Fewer Observations

Muni S. Srivastava*and Tatsuya Kubokawa†
University of Toronto and University of Tokyo

December 20, 2005

In this paper, we consider the prediction problem in multiple linear regression model in which the number of predictor variables, p, is extremely large compared to the number of available observations, n. The least squares predictor based on a generalized inverse is not efficient. We propose six empirical Bayes estimators of the regression parameters. Three of them are shown to have uniformly lower prediction error than the least squares predictors when the vector of regressor variables are assumed to be random with mean vector zero and the covariance matrix $(1/n)\boldsymbol{X}^t\boldsymbol{X}$ where $\boldsymbol{X}^t = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is the $p \times n$ matrix of observations on the regressor vector centered from their sample means. For other estimators, we use simulation to show its superiority over the least squares predictor.

Key words and phrases: Empirical Bayes method, James-Stein estimator, linear regression model, many regressors, prediction, ridge regression, shrinkage, stability.

1 Introduction

In this paper, we consider the prediction problem in multiple linear regression model in which the number of predictor variables is extremely large compared to the number of observations available. More specifically, we consider the problem in which the n response variables $y_1, \ldots, y_n$ are linearly related to the p regressor variables as

$$\boldsymbol{y} = \beta_0 \mathbf{1_n} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1.1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^t$, $\boldsymbol{X}$ is the $n \times p$ observation matrix on the p predictor variables $x_1, \ldots, x_p$, centered from their sample means, $\mathbf{1}_n = (1, \ldots, 1)^t$, an n-vector of ones, β_0 is an unknown constant, $\boldsymbol{\beta}$ is the p-vector of the regression parameters $\beta_1, \ldots, \beta_p$,

*Department of Statistics, University of Toronto, 100 St George Street, Toronto, Ontario, CANADA M5S 3G3, E-Mail: srivasta@utstat.utstat.toronto.edu

†Faculty of Economics, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN, E-Mail: tatsuya@e.u-tokyo.ac.jpFaculty

$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$, and $\boldsymbol{\epsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^t$ is the vector of errors. We shall assume that given $\boldsymbol{\beta}$, $\boldsymbol{y}$ has a multivariate normal distribution with mean vector $\boldsymbol{\beta}_0 \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}$, and covariance matrix $\sigma^2 \boldsymbol{I}_n$. We shall write this as

$$\boldsymbol{y}|\boldsymbol{\beta} \sim \mathcal{N}_n(\boldsymbol{\beta}_0 \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}_n), \tag{1.2}$$

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. When there is no confusion, we may not always mention that the distribution of $\boldsymbol{y}$ or $\boldsymbol{\epsilon}$ is conditional given $\boldsymbol{\beta}$. Often, $n > p$, and this problem has been extensively discussed in the literature.

In this paper, we consider the case when n is substantially smaller than p, namely $n << p$. From the singular value decomposition of $\boldsymbol{X}$, we can write

$$\boldsymbol{X}^t = \boldsymbol{A}\boldsymbol{D}\boldsymbol{F}, \boldsymbol{A}^t\boldsymbol{A} = \boldsymbol{I}_r, \boldsymbol{F}\boldsymbol{F}^t = \boldsymbol{I}_r, \tag{1.3}$$

where $\boldsymbol{A}$ is the $p \times r$ matrix of r eigenvectors corresponding to the r nonzero eigenvalues of $\boldsymbol{X}^t\boldsymbol{X}$, $\boldsymbol{D} = \text{diag}\,(d_1, \ldots, d_r)$, $d_1 \geq \cdots \geq d_r$, is the diagonal matrix of the positive square roots of the r nonzero ordered eigenvalues of $\boldsymbol{X}^t\boldsymbol{X}$, and $\boldsymbol{F}$ is the $n \times r$ matrix of the r eigenvectors corresponding to the r nonzero eigenvalues of the $n \times n$ matrix $\boldsymbol{X}\boldsymbol{X}^t$ which we assume to be of rank $r \leq n-1$. The nonzero eigenvalues of $\boldsymbol{X}\boldsymbol{X}^t$ are the same as the nonzero eigenvalues of $\boldsymbol{X}^t\boldsymbol{X}$. The model (1.1) can be thus be rewritten as

$$\begin{aligned} \boldsymbol{y} =& \beta_0 \mathbf{1}_n + \boldsymbol{F}^t\boldsymbol{D}\boldsymbol{A}^t\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ =& \beta_0 \mathbf{1}_n + \boldsymbol{F}^t\boldsymbol{D}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \end{aligned} \tag{1.4}$$

where

$$\boldsymbol{\gamma} = \boldsymbol{A}^t\boldsymbol{\beta}.$$

The least squares estimators of β_0 and $\boldsymbol{\gamma}$ are given by

$$\begin{aligned} \widehat{\beta}_0 =& \overline{y} = n^{-1}\mathbf{1}_n'\boldsymbol{y}, \\ \widehat{\boldsymbol{\gamma}} =& (\boldsymbol{D}\boldsymbol{F}\boldsymbol{F}^t\boldsymbol{D})^{-1}\boldsymbol{D}\boldsymbol{F}\boldsymbol{y} = \boldsymbol{D}^{-1}\boldsymbol{F}\boldsymbol{y}. \end{aligned}$$

Using the Moore-Penrose inverse (see the Appendix for the definition), we find that the least squares estimator of $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{A}\boldsymbol{D}^2\boldsymbol{A}^t)^+\boldsymbol{A}\boldsymbol{D}\boldsymbol{F}\boldsymbol{y} = \boldsymbol{A}\boldsymbol{D}^{-1}\boldsymbol{F}\boldsymbol{y} = \boldsymbol{A}\widehat{\boldsymbol{\gamma}}. \tag{1.5}$$

Let $(\boldsymbol{A}, \boldsymbol{B})$ be a $p \times p$ orthogonal matrix and $\boldsymbol{\delta} = \boldsymbol{B}^t\boldsymbol{\beta}$, where $\boldsymbol{A}$: $p \times r$ and $\boldsymbol{B}$: $p \times (p-r)$. Then, since $\boldsymbol{A}\boldsymbol{A}^t + \boldsymbol{B}\boldsymbol{B}^t = \boldsymbol{I}_p$, it follows that

$$\boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\gamma} + \boldsymbol{B}\boldsymbol{\delta}.$$

This is a one-to-one and onto transformation. See Srivastava and Khatri (1979, p.139) for more discussion. Thus, $\widehat{\boldsymbol{\beta}}$ is not an unbiased estimator of $\boldsymbol{\beta}$. The mean of $\widehat{\boldsymbol{\beta}}$ is only $\boldsymbol{A}\boldsymbol{\gamma}$.

The reader is reminded that since the matrix $\boldsymbol{X}^t$ is centered from their means, $\boldsymbol{X}^t\mathbf{1}_n = \mathbf{0}$, and hence $\boldsymbol{F}\mathbf{1}_n = \mathbf{0}$. It may be noted that

$$\begin{aligned}\widehat{\boldsymbol{\gamma}} &\sim \mathcal{N}_r(\boldsymbol{\gamma}, \sigma^2\boldsymbol{D}^{-2}),\\ \widehat{\boldsymbol{\beta}} &\sim \mathcal{N}_p(\boldsymbol{A}\boldsymbol{\gamma}, \sigma^2\boldsymbol{A}\boldsymbol{D}^{-2}\boldsymbol{A}^t),\end{aligned}$$

which is a singular multivariate normal distribution.

The above discussion suggests to consider other biased estimators of $\boldsymbol{\beta}$. In particular, the ones obtained by empirical Bayes methods which includes Stein type estimators. The need of considering biased estimators also arises from the fact that the number of nonzero eigenvaluess of $\boldsymbol{X}^t\boldsymbol{X}$ assumed r above, will usually be not known. Thus, the dimension of the subspace in which the mean of $\widehat{\boldsymbol{\beta}}$ lies will not be known. However, we can begin with a judicious choice of r and further reduce it by considering the prior distribution of $\boldsymbol{\beta}$ which has a mean lying in a reduced subspace as considered below. The initial value of r can be obtained by deleting the eigenvaluess of $\boldsymbol{X}^t\boldsymbol{X}$ which are zeros or near zeros. For example, in the data analyzed by West (2003), and Brown *et al*. (1999) in which there are 300 regressors and only 78 observations, it was found that only sixteen d_i^2's account for 99.995% of the variation. That is,

$$\sum_{i=1}^{16} d_i^2 / \sum_{i=1}^{77} d_i^2 = 99.995\%.$$

Thus, it may be reasonable to initially assume that at most r, $r < n$, d_i's are different from zero and then search for a most appropriate number of d_i that may be taken to be different from zero through the analysis of the data. This is achieved by assuming that $\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{A}_1\boldsymbol{\gamma}_1, \sigma^2\boldsymbol{A}\boldsymbol{T}\boldsymbol{A}^t)$, where $\boldsymbol{T}$ is positive definite,

$$\begin{aligned}&\boldsymbol{A} = (\boldsymbol{A}_1, \boldsymbol{A}_2), \quad \boldsymbol{\gamma}^t = (\boldsymbol{\gamma}_1^t, \boldsymbol{\gamma}_2^t),\\ &\boldsymbol{A}_1 : p \times m,\ \boldsymbol{A}_2 : p \times m_1, \boldsymbol{\gamma}_1 : m \times 1,\ \boldsymbol{\gamma}_2 : m_1 \times 1,\\ &\boldsymbol{A}_1^t\boldsymbol{A}_1 = \boldsymbol{I}_m,\ \boldsymbol{A}_2^t\boldsymbol{A}_2 = \boldsymbol{I}_{m_1},\ \text{and } m_1 = r - m.\end{aligned} \tag{1.6}$$

Although $\boldsymbol{T}$ can be any $r \times r$ positive definite matrix, we will only consider the case where

$$\boldsymbol{T} = \lambda\boldsymbol{D}^{c-2}, \quad \lambda > 0,\ c = 0,\ \text{or } c = 2. \tag{1.7}$$

The Bayes estimator or the posterior mean of $\boldsymbol{\beta}$ will involve $\boldsymbol{\gamma}_1$ and λ. These quantities are estimated by using the marginal distribution of $\widehat{\boldsymbol{\beta}}$ and an estimate of σ^2. Such an estimator, known as empirical Bayes estimator is given by

$$\widehat{\boldsymbol{\beta}}_c^{EB}(\hat{\lambda}, \widehat{\boldsymbol{\gamma}}_1) = \boldsymbol{A}\widehat{\boldsymbol{\gamma}} - \left[\boldsymbol{I}_p - \boldsymbol{A}(\boldsymbol{I}_r + \hat{\lambda}^{-1}\boldsymbol{D}^{-c})^{-1}\boldsymbol{A}^t\right](\boldsymbol{A}\widehat{\boldsymbol{\gamma}} - \boldsymbol{A}_1\widehat{\boldsymbol{\gamma}}_1), \tag{1.8}$$

where $\widehat{\boldsymbol{\gamma}} = \boldsymbol{A}^t\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\gamma}}_1 = \boldsymbol{D}_1^{-1}\boldsymbol{F}_1\boldsymbol{y}$, $\boldsymbol{F}^t = (\boldsymbol{F}_1^t, \boldsymbol{F}_2^t)$ and $\boldsymbol{D} = \text{diag}\,(\boldsymbol{D}_1, \boldsymbol{D}_2)$. Here, $\boldsymbol{F}_1^t : n \times m$, $\boldsymbol{F}_1\boldsymbol{F}_1^t = \boldsymbol{I}_m$, $\boldsymbol{F}_2^t : n \times m_1$, $\boldsymbol{F}_2\boldsymbol{F}_2^t = \boldsymbol{I}_{m_1}$, $m_1 = r - m$, $\boldsymbol{D}_1 = \text{diag}\,(d_1, \ldots, d_m)$, and

$\boldsymbol{D}_2 = \text{diag}\,(d_{m+1}, \ldots, d_r)$. By choosing $c = 0$ or $c = 2$ and using different estimators of λ, we obtain several empirical Bayes estimators which are described in Section 2. In Section 3, we prove optimality of three of these estimators. In Section 4, we reanalyze the data of Osborne, Feran, Miller and Douglas (1984). We also present some simulation results in this section. The paper concludes in Section 5.

2 Empirical Bayes Estimators

In this section, we present our analysis of the linear regression model (1.1) or equivalently (1.4) in terms of singular value decomposition of $\boldsymbol{X}$. The least squares estimate of $\boldsymbol{\beta}$ given in (1.5) can be written in terms of a singular value decomposition of the matrix $\boldsymbol{X}$ as

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{D}^{-1}\boldsymbol{F}\boldsymbol{y}$$

which is distributed as multivariate normal with the mean vector $\boldsymbol{A}\boldsymbol{A}^t\boldsymbol{\beta}$, and the covariance matrix $\sigma^2\boldsymbol{A}\boldsymbol{D}^{-2}\boldsymbol{A}^t$. We shall assume that the prior distribution of $\boldsymbol{\beta}$ is a singular multivariate normal distribution with the mean vector $\boldsymbol{A}_1\boldsymbol{\gamma}_1$ and covariance matrix $\sigma^2\boldsymbol{A}\boldsymbol{T}\boldsymbol{A}^t$, that is,

$$\pi(\boldsymbol{\beta}\,|\,\boldsymbol{T}, \boldsymbol{\gamma}_1) \sim \mathcal{N}_p(\boldsymbol{A}_1\boldsymbol{\gamma}_1, \sigma^2\boldsymbol{A}\boldsymbol{T}\boldsymbol{A}^t),$$

where $\boldsymbol{T}$ is an $r \times r$ positive definite matrix and $\boldsymbol{A}_1$ is the $p \times m$ matrix defined in (1.6). Noting that

$$\boldsymbol{A}\boldsymbol{A}^t\boldsymbol{A}_1 = (\boldsymbol{A}_1, \boldsymbol{A}_2)\begin{pmatrix} \boldsymbol{A}_1^t \\ \boldsymbol{A}_2^t \end{pmatrix}\boldsymbol{A}_1 = (\boldsymbol{A}_1, \boldsymbol{A}_2)\begin{pmatrix} \boldsymbol{I}_m \\ \boldsymbol{0} \end{pmatrix} = \boldsymbol{A}_1,$$

we find from the above prior distribution that the joint distribution of $(\widehat{\boldsymbol{\beta}}^t, \boldsymbol{\beta}^t)^t$ is given by

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \boldsymbol{\beta} \end{pmatrix} \sim \mathcal{N}_{2p}\left(\begin{pmatrix} \boldsymbol{A}_1\boldsymbol{\gamma}_1 \\ \boldsymbol{A}_1\boldsymbol{\gamma}_1 \end{pmatrix},\ \sigma^2\begin{pmatrix} \boldsymbol{A}(\boldsymbol{T} + \boldsymbol{D}^{-2})\boldsymbol{A}^t & \boldsymbol{A}\boldsymbol{T}\boldsymbol{A}^t \\ \boldsymbol{A}\boldsymbol{T}\boldsymbol{A}^t & \boldsymbol{A}\boldsymbol{T}\boldsymbol{A}^t \end{pmatrix}\right).$$

Hence, from Theorem 2.2.7 of Srivastava and Khatri (1979, page 47), the conditional distribution of $\boldsymbol{\beta}$ given $\widehat{\boldsymbol{\beta}}$ or the posterior distribution of $\boldsymbol{\beta}$ is $\mathcal{N}_p(\widehat{\boldsymbol{\beta}}^B(\boldsymbol{G}, \boldsymbol{\gamma}_1), \boldsymbol{\Lambda})$, where

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^B(\boldsymbol{G}, \boldsymbol{\gamma}_1) =& [\boldsymbol{I} - \boldsymbol{A}\boldsymbol{G}\boldsymbol{A}^t]\boldsymbol{A}_1\boldsymbol{\gamma}_1 + \boldsymbol{A}\boldsymbol{G}\widehat{\boldsymbol{\gamma}}, \\ \boldsymbol{G} =& \boldsymbol{T}(\boldsymbol{T} + \boldsymbol{D}^{-2})^{-1}, \\ \widehat{\boldsymbol{\gamma}} =& \boldsymbol{A}^t\widehat{\boldsymbol{\beta}} = \boldsymbol{D}^{-1}\boldsymbol{F}\boldsymbol{y}, \end{aligned} \tag{2.1}$$

and

$$\boldsymbol{\Lambda} = \sigma^2\boldsymbol{A}\boldsymbol{T}(\boldsymbol{T} + \boldsymbol{D}^{-2})^{-1}\boldsymbol{T}\boldsymbol{A}^t.$$

When $\boldsymbol{\beta} = \boldsymbol{A}_1\boldsymbol{\gamma}_1$, the model becomes

$$\begin{aligned}\boldsymbol{y} =& \boldsymbol{F}^t\boldsymbol{D}\boldsymbol{A}^t\boldsymbol{A}_1\boldsymbol{\gamma}_1 + \boldsymbol{\epsilon} \\ =& \boldsymbol{F}_1^t\boldsymbol{D}_1\boldsymbol{\gamma}_1 + \boldsymbol{\epsilon},\end{aligned}$$

and the least squares estimate of $\boldsymbol{\gamma}_1$ is given by

$$\widehat{\boldsymbol{\gamma}}_1 = \boldsymbol{D}_1^{-1}\boldsymbol{F}_1\boldsymbol{y}, \tag{2.2}$$

which can be interpreted as a principal component estimator. Hence, an empirical Bayes estimator of $\boldsymbol{\beta}$ is given by

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^B(\boldsymbol{G}, \widehat{\boldsymbol{\gamma}}_1) =& [\boldsymbol{I} - \boldsymbol{A}\boldsymbol{G}\boldsymbol{A}^t]\boldsymbol{A}_1\widehat{\boldsymbol{\gamma}}_1 + \boldsymbol{A}\boldsymbol{G}\widehat{\boldsymbol{\gamma}} \\ =& \boldsymbol{A}\widehat{\boldsymbol{\gamma}} - (\boldsymbol{I} - \boldsymbol{A}\boldsymbol{G}\boldsymbol{A}^t)(\boldsymbol{A}\widehat{\boldsymbol{\gamma}} - \boldsymbol{A}_1\widehat{\boldsymbol{\gamma}}_1)\end{aligned} \tag{2.3}$$

where $\widehat{\boldsymbol{\gamma}}_1$ is given by (2.2) and $\widehat{\boldsymbol{\gamma}}$ by (2.1). The estimator $\widehat{\boldsymbol{\beta}}^B(\boldsymbol{G}, \widehat{\boldsymbol{\gamma}}_1)$ in (2.3) shrinks the LS estimator $\boldsymbol{A}\widehat{\boldsymbol{\gamma}}$ towards the principal component regression (PCR) estimator $\boldsymbol{A}_1\widehat{\boldsymbol{\gamma}}_1$. However in (2.3), we need to specify $\boldsymbol{G}$ or equivalently $\boldsymbol{T}$.

We choose $\boldsymbol{T}$ as given in the equation (1.7). This gives

$$\boldsymbol{G} = \left(\boldsymbol{I}_r + \lambda^{-1}\boldsymbol{D}^{-c}\right)^{-1}.$$

Thus, our empirical Bayes estimator depends only on the estimators of $(\lambda, \boldsymbol{\gamma}_1)$, denoted by $(\hat{\lambda}, \widehat{\boldsymbol{\gamma}}_1)$, and is given by

$$\widehat{\boldsymbol{\beta}}_c^{EB}(\hat{\lambda}, \widehat{\boldsymbol{\gamma}}_1) = \boldsymbol{A}\widehat{\boldsymbol{\gamma}} - \left[\boldsymbol{I}_p - \boldsymbol{A}(\boldsymbol{I}_r + \hat{\lambda}^{-1}\boldsymbol{D}^{-c})^{-1}\boldsymbol{A}^t\right](\boldsymbol{A}\widehat{\boldsymbol{\gamma}} - \boldsymbol{A}_1\widehat{\boldsymbol{\gamma}}_1). \tag{2.4}$$

From the least squares theory, we choose an estimator of σ^2 given by

$$\hat{\sigma}^2 = \|\boldsymbol{y} - \overline{y}\boldsymbol{1} - \boldsymbol{F}^t\boldsymbol{D}\boldsymbol{A}^t\widehat{\boldsymbol{\beta}}\|^2/(n - r + 1), \tag{2.5}$$

provided $r < n$. To obtain an estimator of λ, we consider the marginal distribution of $\widehat{\boldsymbol{\beta}}$ which is a singular distribution. Thus, we shall consider the distribution of $\boldsymbol{A}^t\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\gamma}}$, which is nonsingular, given by

$$\widehat{\boldsymbol{\gamma}} \sim \mathcal{N}_r\left(\begin{pmatrix}\boldsymbol{\gamma}_1 \\ \boldsymbol{0}\end{pmatrix},\ \sigma^2\left(\lambda\boldsymbol{D}^{c-2} + \boldsymbol{D}^{-2}\right)\right).$$

Thus if $\boldsymbol{\gamma}_1$ and σ^2 are known, the moment estimator of λ is obtained by solving

$$\left(\widehat{\boldsymbol{\gamma}} - \begin{pmatrix}\boldsymbol{\gamma}_1 \\ \boldsymbol{0}\end{pmatrix}\right)^t\left(\lambda\boldsymbol{D}^{c-2} + \boldsymbol{D}^{-2}\right)^{-1}\left(\widehat{\boldsymbol{\gamma}} - \begin{pmatrix}\boldsymbol{\gamma}_1 \\ \boldsymbol{0}\end{pmatrix}\right)^t = a\sigma^2 \tag{2.6}$$

where a is chosen appropriately to get biased or unbiased estimator. For σ^2 and $\boldsymbol{\gamma}_1$, we use the estimators given by (2.5) and (2.2), respectively. Thus, to get an estimate of λ, we solve the equation

$$\widehat{\boldsymbol{\gamma}}_2^t \left[\lambda \boldsymbol{D}_2^{c-2} + \boldsymbol{D}_2^{-2}\right]^{-1} \widehat{\boldsymbol{\gamma}}_2 = a\hat{\sigma}^2, \tag{2.7}$$

where $\widehat{\boldsymbol{\gamma}}_2$ is an m_1-vector, $m_1 = r - m$, given by

$$\widehat{\boldsymbol{\gamma}}_2 = \boldsymbol{D}_2^{-1} \boldsymbol{F}_2 \boldsymbol{y}.$$

Let λ^* be a solution of (2.6). In order that it remains positive, we choose

$$\hat{\lambda} = \max(\lambda^*,\ 0)$$

as an estimator of λ.

Next, we describe several estimators by choosing the values of a, c etc. While we choose only two values of c, namely $c = 0$ or 2, we choose five values of a, given by

$$\begin{aligned} a_1 =& a_2 = r - m - 2, \\ a_3 =& 2 \sum_{i=m+1}^{r} (d_r^2/d_i^2) - 4, \\ a_4 =& r - 2 \quad \text{and} \quad a_5 = a_3. \end{aligned}$$

2.1 James-Stein type estimator (JS)

Let $c = 0$, or $\boldsymbol{T} = \lambda \boldsymbol{D}^{-2}$, and $a = a_1 = r - m - 2$. Then from (2.6), we get an estimator of λ or equivalently of α given by

$$\widehat{\alpha} \equiv \frac{1}{1 + \hat{\lambda}} = \min \left\{ \frac{(r - m - 2)\hat{\sigma}^2}{\widehat{\boldsymbol{\gamma}}_2^t \boldsymbol{D}_2^2 \widehat{\boldsymbol{\gamma}}_2},\ 1 \right\}.$$

The estimator of λ obtained from the above equation will be denoted by $\hat{\lambda}_1$. Thus,

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_1^{EB}(\hat{\lambda}_1, \widehat{\boldsymbol{\gamma}}_1) = \widehat{\boldsymbol{\beta}}^{JS} =& \boldsymbol{A}\widehat{\boldsymbol{\gamma}} - \widehat{\alpha}\left\{\boldsymbol{A}\widehat{\boldsymbol{\gamma}} - \boldsymbol{A}_1\widehat{\boldsymbol{\gamma}}_1\right\} \\ =& \boldsymbol{A}_1\widehat{\boldsymbol{\gamma}}_1 + (1 - \widehat{\alpha})\boldsymbol{A}_2\widehat{\boldsymbol{\gamma}}_2. \end{aligned} \tag{2.8}$$

Since it is identical to the so-called positive part Stein estimator proposed by James and Stein (1961) for $p = r$, we will call it the James-Stein type estimator (JS).

2.2 Ridge-principal component estimator (RP)

Let $c = 2$, $m_1 = r - m$, $a = a_2 = r - m - 2$ and $\hat{\lambda}_2 = \max(0, \hat{\lambda})$, where $\hat{\lambda}$ is the solution of

$$\widehat{\boldsymbol{\gamma}}_2^t \left(\hat{\lambda}\boldsymbol{I}_{m_1} + \boldsymbol{D}_2^{-2}\right)^{-1} \widehat{\boldsymbol{\gamma}}_2 = (r - m - 2)\hat{\sigma}^2. \tag{2.9}$$

Then, we obtain an estimator of $\widehat{\boldsymbol{\beta}}$, which we call the ridge-principal component regression estimator (RP), given by

$$\widehat{\boldsymbol{\beta}}_2^{EB}(\hat{\lambda}_2, \widehat{\boldsymbol{\gamma}}_1) = \widehat{\boldsymbol{\beta}}^{RP} = \boldsymbol{A}\widehat{\boldsymbol{\gamma}} - (\boldsymbol{I}_p - \boldsymbol{A}(\boldsymbol{I}_r + \hat{\lambda}_2^{-1}\boldsymbol{D}^{-2})^{-1}\boldsymbol{A}^t)(\boldsymbol{A}\widehat{\boldsymbol{\gamma}} - \boldsymbol{A}_1\widehat{\boldsymbol{\gamma}}_1). \tag{2.10}$$

This estimator is similar to the one proposed by Kubokawa and Srivastava (2004). However, it is difficult to verify analytically the superiority of the $\widehat{\boldsymbol{\beta}}^{RP}$ estimator over the least squares estimator. Thus, as dictated by the proof of the dominance result of Theorem 3.1 given in Section 3, we modify the equation (2.9) to obtain the solution of λ in the next subsection.

2.3 Modified ridge-regression principal component estimator (RPM)

Let $c = 2$, $m_1 = r - m$, and $a = a_3 = 2\sum_{i=m+1}^{r}(d_r^2/d_i^2) - 4$. Define

$$\hat{\lambda}_3 = \max(0, \hat{\lambda})$$

where $\hat{\lambda}$ is the solution of the equation

$$\widehat{\boldsymbol{\gamma}}_2^t \left(\hat{\lambda}\boldsymbol{I}_{m_1} + \boldsymbol{D}_2^{-2}\right)^{-1} \widehat{\boldsymbol{\gamma}}_2 = a_3\hat{\sigma}^2.$$

Then the modified ridge-regression principal component estimator is given by

$$\widehat{\boldsymbol{\beta}}^{RPM} = \widehat{\boldsymbol{\beta}}_2^{EB}(\hat{\lambda}_3, \widehat{\boldsymbol{\gamma}}_1), \tag{2.11}$$

whose expression can be obtained from (2.10) with $\hat{\lambda}_2$ replaced by $\hat{\lambda}_3$.

2.4 Ridge-regression type estimator (RR)

We now consider the case when $c = 2$, $m = 0$, and $a = a_4 = r - 2$. In this case $\boldsymbol{\gamma}_1 = \boldsymbol{0}$, and the estimate of λ is obtained by solving the equation

$$\widehat{\boldsymbol{\gamma}}^t \left(\hat{\lambda}\boldsymbol{I}_r + \boldsymbol{D}^{-2}\right)^{-1} \widehat{\boldsymbol{\gamma}} = (r-2)\hat{\sigma}^2. \tag{2.12}$$

We then estimate λ by

$$\hat{\lambda}_4 = \max(\hat{\lambda}, 0).$$

Using $\hat{\lambda}_4$ as an estimator of λ, we obtain an empirical Bayes estimator of Shinozaki and Chang (1993) type. We shall denote this estimator of $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}}^{RR}$ to reflect its similarity with ridge regression type of estimators. It is given by

$$\widehat{\boldsymbol{\beta}}^{RR} = \widehat{\boldsymbol{\beta}}_2^{EB}(\hat{\lambda}_4, \boldsymbol{0}) = \boldsymbol{A}\left(\boldsymbol{I}_r + \hat{\lambda}_4^{-1}\boldsymbol{D}^{-2}\right)^{-1} \widehat{\boldsymbol{\gamma}}. \tag{2.13}$$

However, the superiority of $\widehat{\boldsymbol{\beta}}^{RR}$ over the LS estimator $\widehat{\boldsymbol{\beta}}$ has not been established. We therefore modify (2.12) to obtain an estimator which we show to dominate the LS estimator $\widehat{\boldsymbol{\beta}}$.

2.5 Modified ridge-regression type estimator (RRM)

In this section, we assume as in Section 2.4 that $c = 2$, $m = 0$, but $a = a_5 = 2\sum_{i=m+1}^{r}(d_r^2/d_i^2) - 4$. Let

$$\hat{\lambda}_5 = \max(\hat{\lambda}, 0)$$

where $\hat{\lambda}$ is the solution of the equation

$$\widehat{\boldsymbol{\gamma}}^t(\hat{\lambda}\boldsymbol{I}_r + \boldsymbol{D}^{-2})^{-1}\widehat{\boldsymbol{\gamma}} = a_5\hat{\sigma}^2.$$

Then the resulting empirical Bayes estimator of $\boldsymbol{\beta}$ is given by (2.13) with $\hat{\lambda}_4$ replaced by $\hat{\lambda}_5$. Thus,

$$\widehat{\boldsymbol{\beta}}^{RRM} = \widehat{\boldsymbol{\beta}}_2^{EB}(\hat{\lambda}_5, 0). \tag{2.14}$$

In Section 3, and in the Appendix we show that the estimator $\widehat{\boldsymbol{\beta}}^{RRM}$ dominates the least squares estimator $\widehat{\boldsymbol{\beta}}$, in terms of prediction error when the vector of p predictor variables has a random distribution described in Section 3.

2.6 Crude empirical Bayes estimator (EB)

The above estimation procedures are available when the rank r of the matrix $\boldsymbol{X}$ is less than $n-1$. However, the following empirical Bayes estimator can be utilized even for $r = n-1$:

$$\widehat{\boldsymbol{\beta}}^{EB} = \left(\boldsymbol{X}^t\boldsymbol{X} + (1/\hat{\lambda}_{EB})\boldsymbol{I}_p\right)^{-1}\boldsymbol{X}^t\boldsymbol{y}, \tag{2.15}$$

where $\hat{\lambda}_{EB} = \max(0, \lambda_*)$ for the solution λ_* of the equation

$$\boldsymbol{y}^t\left(\hat{\lambda}_*\boldsymbol{X}\boldsymbol{X}^t + \boldsymbol{I}_n\right)^{-1}\boldsymbol{y}/\hat{\sigma}^2 = n-2.$$

If $\boldsymbol{X}$ has full rank, no variance estimator is available. However, we suggest the use of the estimator $\hat{\sigma}^2$ given by (2.5) for appropriately chosen r. When $\text{rank}\,(\boldsymbol{X}) = r < n-1$, this empirical Bayes estimator $\widehat{\boldsymbol{\beta}}^{EB}$ is close to the above one $\widehat{\boldsymbol{\beta}}^{RR}$.

3 Empirical Bayes predictors and their risk properties

Now we consider the problem of predicting the observation y when $\boldsymbol{x}$ is given as a predictor variable, where the model is described by

$$y = \beta_0 + \boldsymbol{x}^t\boldsymbol{\beta} + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The usual predictor based on the least squares estimators $\widehat{\beta}_0 = \overline{y}$ and $\widehat{\boldsymbol{\beta}}$ is given by

$$\hat{y} = \overline{y} + \boldsymbol{x}^t \widehat{\boldsymbol{\beta}}, \tag{3.1}$$

while the shrinkage predictor employing the empirical Bayes estimator of $\boldsymbol{\beta}$ is provided by

$$\hat{y}_c^{EB}(\hat{\lambda}, \widehat{\boldsymbol{\gamma}}_1) = \overline{y} + \boldsymbol{x}^t \widehat{\boldsymbol{\beta}}_c^{EB}(\hat{\lambda}, \widehat{\boldsymbol{\gamma}}_1), \tag{3.2}$$

where the empirical Bayes estimator $\widehat{\boldsymbol{\beta}}_c^{EB}(\hat{\lambda}, \widehat{\boldsymbol{\gamma}}_1)$ is given by (2.4).

To evaluate the predictors, we utilize the arguments as in Copas (1983), who assumes that $\boldsymbol{x}$ follows the empirical distribution $\widehat{F}_x$ based on the sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, which has the mean $E[\boldsymbol{x}|\widehat{F}_x] = 0$ and the covariance matrix $E[\boldsymbol{x}\boldsymbol{x}^t|\widehat{F}_x] = n^{-1}\boldsymbol{X}^t\boldsymbol{X}$, where the notation $E[\cdot|\widehat{F}_x]$ denotes the expectation with respect to the distribution $\widehat{F}_x$. When the expectation is taken with respect to ε and $\widehat{F}_x$, the prediction error of the usual predictor $\hat{y}$ under the squared loss is written by

$$\begin{aligned}
E^{\varepsilon}\left[E\left[(\hat{y}-y)^2|\widehat{F}_x\right]\right] =& E^{\varepsilon}\left[E\left[(\overline{y} + \boldsymbol{x}^t\widehat{\boldsymbol{\beta}} - \beta_0 - \boldsymbol{x}^t\boldsymbol{\beta} - \varepsilon)^2|\widehat{F}_x\right]\right] \\
=& (\overline{y}-\beta_0)^2 + (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^t E\left[\boldsymbol{x}\boldsymbol{x}^t|\widehat{F}_x\right](\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) + \sigma^2 \\
&+ 2(\overline{y}-\beta_0)E\left[\boldsymbol{x}^t|\widehat{F}_x\right](\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\
=& (\overline{y}-\beta_0)^2 + n^{-1}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^t\boldsymbol{X}^t\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) + \sigma^2.
\end{aligned} \tag{3.3}$$

Taking the expectation of (3.3) with respect to all the random variables turns out to be

$$E[(\hat{y}-y)^2] = n^{-1}E\left[(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^t\boldsymbol{X}^t\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right] + n^{-2}\mathbf{1}_n^t\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^t\boldsymbol{X}^t\mathbf{1}_n + (n+1)n^{-1}\sigma^2.$$

Since $\boldsymbol{X}^t\mathbf{1}_n = \boldsymbol{A}\boldsymbol{D}\boldsymbol{F}\mathbf{1}_n = \mathbf{0}$ as noted in Section 1, the prediction error is expressed by

$$E[(\hat{y}-y)^2] = n^{-1}E\left[(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^t\boldsymbol{X}^t\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right] + (n+1)n^{-1}\sigma^2, \tag{3.4}$$

which implies that the prediction problem is reduced to that of estimating $\boldsymbol{\beta}$ under the weighted loss $(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^t\boldsymbol{X}^t\boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = (\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})^t\boldsymbol{D}^2(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})$.

Next, we state a theorem whose proof is given in the Appendix.

Theorem 3.1 *Let $r - m \geq 3$. Then, under the assumptions that $\boldsymbol{x} \sim \widehat{F}_x$, the empirical Bayes predictor $\hat{y}_c^{EB}(\hat{\lambda}, \widehat{\boldsymbol{\gamma}}_1)$ given by* (3.2) *has a uniformly smaller prediction error than the least square predictor $\hat{y}$ given by* (3.1) *if the chosen values of $a > 0$, and c satisfy the inequality*

$$\sum_{i=m+1}^{r}(d_r/d_i)^c - 2 \geq a/2. \tag{3.5}$$

Corollary 3.1 *Assume that* $\boldsymbol{x} \sim \widehat{F}_x$.

(1) *For* $c = 0$ *and* $a = r - m - 2 > 0$, *the inequality* (3.5) *is satisfied. Hence, the empirical Bayes predictor* $\hat{y}^{JS} = \overline{y} + \boldsymbol{x}^t \widehat{\boldsymbol{\beta}}^{JS}$ *based on the James-Stein type estimator* (2.8) *improves on* $\hat{y}$.

(2) *For* $c = 2$ *and* $a = 2\left(\sum_{i=m+1}^{r} d_r^2/d_i^2 - 2\right) > 0$, *the inequality* (3.5) *is satisfied. Hence, the empirical Bayes predictor* $\hat{y}^{RPM} = \overline{y} + \boldsymbol{x}^t \widehat{\boldsymbol{\beta}}^{RPM}$ *based on the ridge-principal components regression estimator* (2.11) *improves on* $\hat{y}$.

(3) *For* $c = 2$, $m = 0$ *and* $a = 2\left(\sum_{i=1}^{r} d_r^2/d_i^2 - 2\right) > 0$, *the inequality* (3.5) *is satisfied. Hence, the empirical Bayes predictor* $\hat{y}^{RRM} = \overline{y} + \boldsymbol{x}^t \widehat{\boldsymbol{\beta}}^{RRM}$ *based on the ridge type estimator* (2.14) *improves on* $\hat{y}$.

4 Empirical and Simulation Studies

In this section, we first consider a dataset, initially analyzed by Osborne *et al*. (1984), and later by Brown *et al*. (1999, 2001) and West (2003). The data consist of measurements, on the constituents of the biscuit dough such as fat, sucrose, flour and water, which were obtained using near infrared (NIR) spectroscopy. In our analysis, we shall, however, focus on 'fat content' only which will be our dependent variable $\boldsymbol{y}$. The predictor variables consist of spectral readings at 700 points measured from 1100 to 2498 nanometers in steps of 2 nm. Thus, $p = 700$. The training or calibration set consist of 40 measurements on $\boldsymbol{y}$ and 700 predictor variables $x_1, \ldots, x_{700}$, which have been centered from their sample means.

The centered and scaled observations matrix $\boldsymbol{X}$ of the predictor variables is a huge 40×700 matrix of rank 39 with singularvalues given by

$$\begin{aligned}\boldsymbol{D}_0 =&(25.295, 6.604, 2.812, 1.953, 1.473, 1.287, 0.626, 0.465, 0.301, 0.252, 0.238, 0.196,\\ &0.193, 0.138, 0.127, 0.105, 0.098, 0.092, 0.085, 0.075, 0.070, 0.058, 0.057, 0.055,\\ &0.052, 0.046, 0.045, 0.039, 0.039, 0.034, 0.032, 0.029, 0.029, 0.026, 0.024, 0.022,\\ &0.019, 0.019, 0.017),\end{aligned}$$

which includes lots of small values. Taking the r largest eigenvalues of the 39, we can calculate their contribution rate

$$CR = CR(r) = 100 \times \sum_{i=1}^{r} d_i^2 / \sum_{i=1}^{39} d_i^2.$$

For instance, we have $CR(30) = 99.99\%$, $CR(17) = 99.99\%$, $CR(10) = 99.96\%$ and $CR(5) = 99.85\%$, which are given in the column CR in Table 1. The 17 largest eigen values account for 99.99% of the variation. Although $\boldsymbol{X}$ has many small eigenvalues, it is of rank $n - 1 = 39$, and $\boldsymbol{X}'$ is decomposed as $\boldsymbol{X}^t = \boldsymbol{A}_0 \boldsymbol{D}_0 \boldsymbol{F}_0$ where $\boldsymbol{A}_0$ is the $p \times (n-1)$ matrix of $n - 1$ eigenvectors corresponding to the $n - 1$ positive eigenvalues of $\boldsymbol{X}^t \boldsymbol{X}$,

Table 1: Prediction Error Estimates by the Cross-validation Method for $p = 700$ and $n = 40$ in the Two Cases of $\boldsymbol{X}$ Scaled and Non-scaled

Case of both $\boldsymbol{y}$ and $\boldsymbol{X}$ Scaled											
r	m	CR	LS	PC	JS	RPM	RRM	RP	RR	EB	$\overline{y}$
30	6	99.99	3.798	2.022	1.184	1.129	1.197	1.033	1.202	0.677	1.051
20	6	99.99	3.798	1.308	1.025	1.069	1.015	1.047	1.029	0.572	1.051
17	5	99.99	3.798	1.297	0.963	0.951	0.963	0.921	0.904	0.503	1.051
10	5	99.96	3.798	0.924	0.864	0.896	0.919	0.877	1.098	0.678	1.051
6	2	99.85	3.798	0.978	1.072	1.071	1.090	1.064	1.103	0.611	1.051
Case of $\boldsymbol{y}$ Scaled but $\boldsymbol{X}$ Non-scaled											
r	m	CR	LS	PC	JS	RPM	RRM	RP	RR	EB	$\overline{y}$
35	6	99.99	3.642	3.452	1.950	2.092	2.007	1.874	1.661	0.352	1.051
30	6	99.99	3.642	2.832	1.611	1.591	1.416	1.507	1.165	0.419	1.051
20	6	99.99	3.642	1.969	1.409	1.407	1.296	1.384	1.095	0.448	1.051
10	5	99.97	3.642	1.585	1.397	1.379	1.279	1.372	1.106	0.513	1.051
6	2	99.87	3.642	1.321	1.192	1.274	1.249	1.217	1.140	0.531	1.051

$\boldsymbol{D}_0 = \text{diag}\,(d_1, \ldots, d_{n-1})$, $d_1 \geq \ldots \geq d_{n-1}$, is the diagonal matrix of the positive square roots of the $n-1$ ordered eigenvalues of $\boldsymbol{X}^t\boldsymbol{X}$, known as the singularvalues of the matrix $\boldsymbol{X}$, and $\boldsymbol{F}_0$ is the $(n-1) \times n$ matrix of the $n-1$ eigenvectors corresponding to the $n-1$ nonzero eigenvalues of the $n \times n$ matrix $\boldsymbol{X}\boldsymbol{X}^t$. We may note that $\boldsymbol{A}_0^t\boldsymbol{A}_0 = \boldsymbol{I}_{n-1}$, $\boldsymbol{F}_0\boldsymbol{F}_0^t = \boldsymbol{I}_{n-1}$ and $\boldsymbol{F}_0\boldsymbol{1}_n = \boldsymbol{0}$. Then, we can calculate the least squares estimator as

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{A}_0\boldsymbol{D}_0^{-1}\boldsymbol{F}_0\boldsymbol{y}, \quad \text{denoted by LS}, \tag{4.1}$$

which is not stable. In the $\boldsymbol{D}_0$ described above, many eigenvalues d_i's are very close to zero. For example, $CR(17) = 99.99\%$, which means that the rank of $\boldsymbol{X}$ may be practically presumed to be $r = 17$. Since $\boldsymbol{X}^t$ is decomposed as $\boldsymbol{X}^t = \boldsymbol{A}\boldsymbol{D}\boldsymbol{F}$ for such a rank r, we can consider the principal component regression estimator

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{D}^{-1}\boldsymbol{F}\boldsymbol{y}, \quad \text{denoted by PC},$$

because the $r \times r$ diagonal matrix $\boldsymbol{D}$ deletes the $n-r-1$ smallest eigenvalues from the $(n-1)\times(n-1)$ matrix $\boldsymbol{D}_0$. Based on the estimator $\widehat{\boldsymbol{\beta}}$, we can compute the following estimators suggested in Section 2: $\widehat{\boldsymbol{\beta}}^{JS}$, $\widehat{\boldsymbol{\beta}}^{RP}$, $\widehat{\boldsymbol{\beta}}^{RPM}$, $\widehat{\boldsymbol{\beta}}^{RR}$, $\widehat{\boldsymbol{\beta}}^{RRM}$ and $\widehat{\boldsymbol{\beta}}^{EB}$, which are denoted by JS, RP, RPM, RR, RRM and EB, respectively. These estimators are compared below in terms of the prediction error estimates by the cross-validation method and estimation errors by simulation.

Corresponding to the estimators LS, PC, JS, RPM, RRM, RP, RR and EB of $\boldsymbol{\beta}$, we provide the predictors of the the form (3.2). In this analysis, we add the sample mean $\overline{y}$

as a simple predictor, which means that $\boldsymbol{\beta}$ is estimated by zeros. The prediction errors of these methods may be estimated via the leave-one-out cross-validation as described, for example, in Srivastava (2002, p322). That is, 40 predictive errors are obtained by leaving out one observation each time. When both $\boldsymbol{X}$ and $\boldsymbol{y}$ are scaled, the estimates of the prediction errors for the above considered predictors are given in the top part of Table 1 for the cases $(r, m) = (30, 6)$, $(20, 6)$, $(17, 5)$, $(10, 5)$ and (6.2). It reveals that the smallest prediction errors of EB and RR are 0.503 and 0.904, respectively, which are provided in the case of $r = 17$. Both minimum values are smaller than those of other predictors. Especially, the use of the estimator EB provides the smallest prediction errors among its competitors. The other emprical Bayes predictors give smaller prediction errors for $(r, m) = (10, 5)$ than the LS, PC and $\overline{y}$ predictiors.

When the observations matrix $\boldsymbol{X}$ is centered but not scaled, the $n - 1$ singularvalues of $\boldsymbol{X}$ are given by

$$\begin{aligned}\boldsymbol{D}_0 =&(10.487, 2.379, 0.950, 0.818, 0.544, 0.329, 0.261, 0.140, 0.107, 0.096, 0.088, 0.073,\\ &0.070, 0.054, 0.049, 0.043, 0.037, 0.034, 0.032, 0.029, 0.028, 0.022, 0.020, 0.020,\\ &0.019, 0.017, 0.016, 0.015, 0.014, 0.013, 0.012, 0.011, 0.010, 0.009, 0.009, 0.008,\\ &0.008, 0.007, 0.006).\end{aligned}$$

We carry out the same analysis for this non-scaled version as was done for the scaled version. The prediction error estimates are reported at the bottom in Table 1, which illustrates that the EB predictor provides the smallest prediction error 0.352 for $r = 35$, which are smaller than those of the other predictors. The other empirical predictors JS, RPM, RRM, RP and RR are better than LS and PS, but worse than $\overline{y}$.

In the case of $\boldsymbol{X}$ scaled but $\boldsymbol{y}$ non-scaled for $r = 6$ and $m = 2$, the values of $\hat{\lambda}$ in the empirical Bayes estimators JS, RPM, RRM, RP, RR and EB are given by 13.54, 7.80, 6.13, 7.72, 3.55 and 5432.27, and the resulting predicted values $\hat{y}_i^{EB} = \overline{y} + \boldsymbol{x}_i^t \widehat{\boldsymbol{\beta}}_c^{EB}(\hat{\lambda}, \widehat{\boldsymbol{\gamma}}_1)$ for some i's are provided in Table 2. Also predicted values by some predictors are drawn in Figures 1 and 2, where the x-axis indicates the observed number. From the table and the figures, it is found that the predicts based on EB are very close to the real data of the response variable y_i and that the line graphs of EB and the real data y_i overlap each other as seen in Figure 1, while the predicts by LS have a large dispersion. It also seems that the predictors based on the other empirical Bayes estimators take reasonable values.

We next treat the smaller data set of the predictor variables consisting of 300 points which are centered after deleting the first 200 and the last 200 points from the original data set. In this case, the maximum and minimum eigenvalues are, respectively, given by 17.087 and 0.00354 for $\boldsymbol{X}$ scaled, and 8.0108 and 0.00166 for $\boldsymbol{X}$ non-scaled. Through the same analysis as stated above, we provide the prediction errors in Table 3. The performance of each predictor is similar to the case of $p = 700$. The prediction error of EB for $r = 19$ is 0.117, and is the smallest in the case of both $\boldsymbol{y}$ and $\boldsymbol{X}$ scaled. Brown *et al.* (1999) gives a prediction error of 0.14, with West's (2003) method only slightly improving on it, while Osborne *et al.*'s (1984) method has prediction error 0.11. This

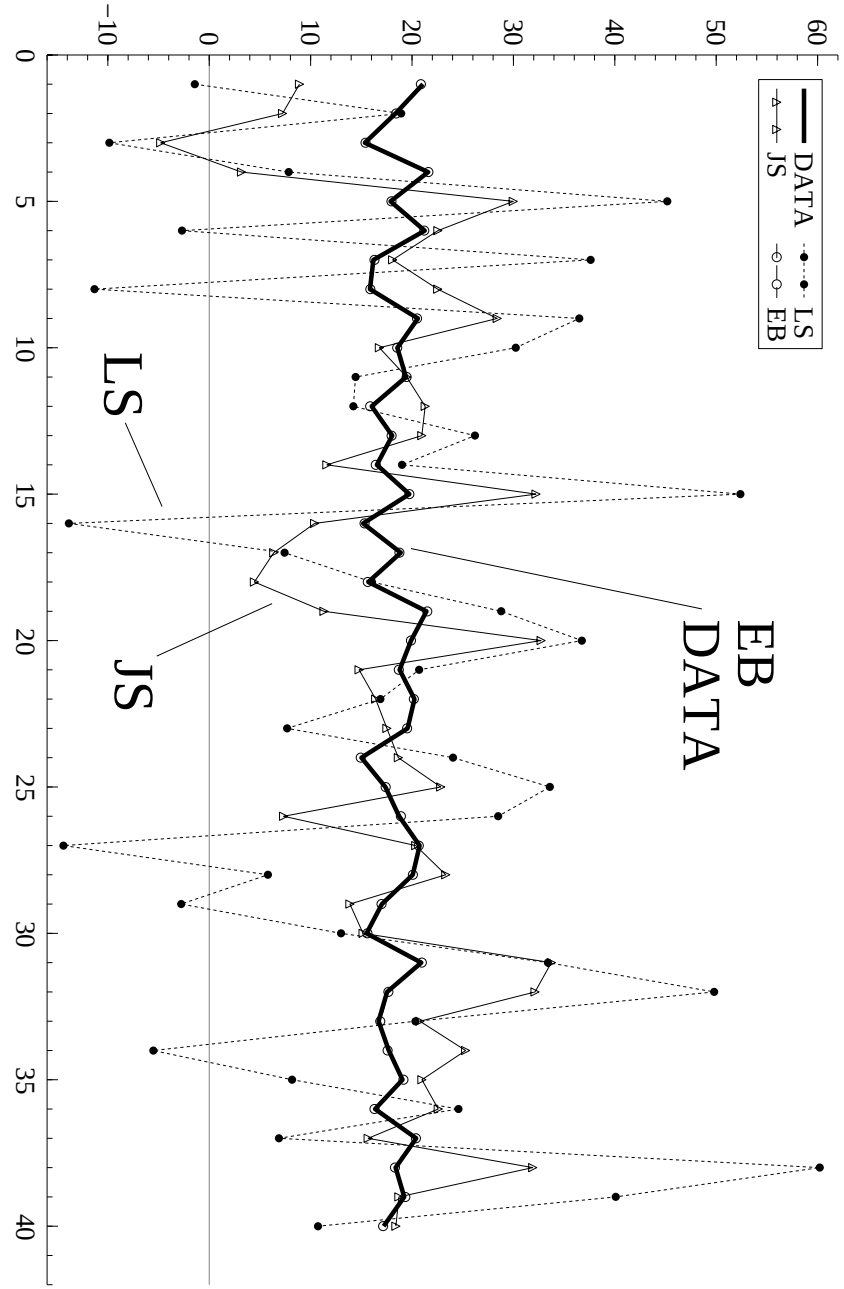

Figure 1: Line Graphs of the Data y_i and Predicts by LS, JS and EB for $p = 700$, $n = 40$, $r = 6$ and $m = 2$ in the Case of $\boldsymbol{X}$ Scaled but $\boldsymbol{y}$ Non-scaled (Line graphs of EB and y_i overlap each other.)

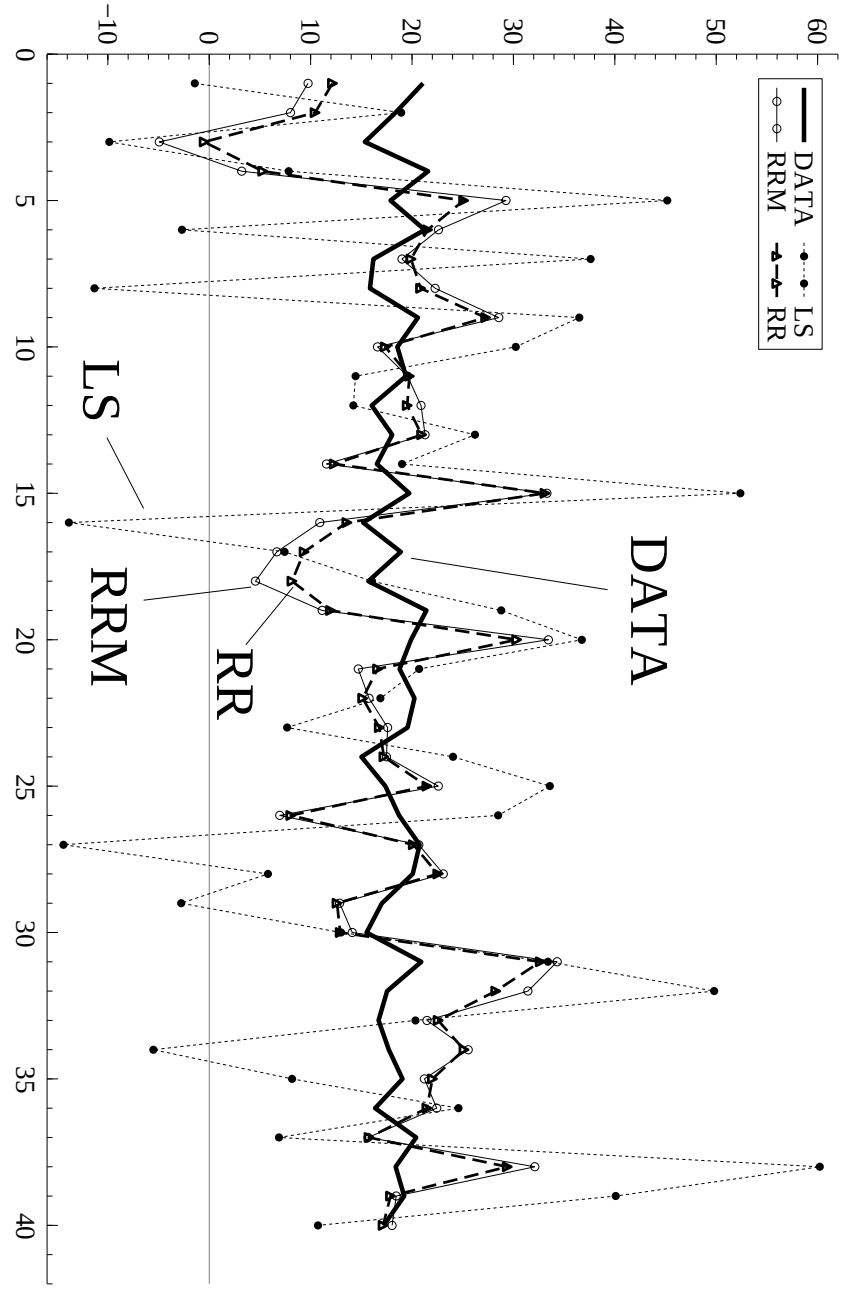

Figure 2: Line Graphs of the Data y_i and Predicts by LS, RRM and RR for $p = 700$, $n = 40$, $r = 6$ and $m = 2$ in the Case of $\boldsymbol{X}$ Scaled but $\boldsymbol{y}$ Non-scaled

Table 2: Predicted Values for $p = 700$, $n = 40$, $r = 6$ and $m = 2$ in the Case of $\boldsymbol{X}$ Scaled but $\boldsymbol{y}$ Non-scaled

i	y_i	LS	PC	JS	RPM	RRM	RR	EB	$\overline{y}$
3	15.350	-9.863	-7.225	-4.787	-5.612	-4.930	-0.499	15.413	18.351
7	16.190	37.614	18.558	18.082	18.833	19.004	19.898	16.307	18.351
11	19.400	14.422	19.403	19.473	19.523	19.550	19.739	19.466	18.351
15	19.730	52.377	33.096	32.230	33.286	33.307	33.116	19.737	18.351
19	21.420	28.793	10.639	11.311	11.014	11.131	11.957	21.490	18.351
23	19.560	7.675	17.903	17.520	17.688	17.592	16.771	19.520	18.351
27	20.750	-14.382	20.751	20.369	20.723	20.664	20.120	20.673	18.351
31	20.910	33.409	34.937	33.740	34.546	34.309	32.648	20.957	18.351
35	19.060	8.177	20.795	20.974	21.112	21.223	22.013	19.150	18.351
39	19.230	40.091	18.808	18.699	18.547	18.439	17.882	19.345	18.351

suggests that our empirical Bayes EB predictor gives prediction error similar to the ones proposed in the above papers.

Now we investigate the risk-performance of predictors through simulation experiments, which, from (3.4), is equivalent to comparing the risk-behaviors of the corresponding estimators of $\boldsymbol{\beta}$. Every estimator $\boldsymbol{\delta}$ is evaluated by the risk function $R(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta})$ under the loss function $L(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}) = (\boldsymbol{\delta} - \boldsymbol{\beta})^t \boldsymbol{X}^t \boldsymbol{X} (\boldsymbol{\delta} - \boldsymbol{\beta})/\sigma^2$. The risk functions of the above estimators are obtained from 1,000 replications through simulation experiments in the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

for the original data matrix $\boldsymbol{X}$ given above for $p = 300$ and $n = 40$. The relative efficiencies $R(\boldsymbol{\beta}, 1, \boldsymbol{\delta})/R(\boldsymbol{\beta}, 1, \widetilde{\boldsymbol{\beta}})$ for $\sigma^2 = 1$ are reported in Table 4 for three cases of $(r, m) = (30, 10)$, $(15, 3)$ and $(5, 2)$, where $\widetilde{\boldsymbol{\beta}}$ is the LS estimator given by (4.1), and

$$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t, \quad \beta_i = p^{-1}(k-1)[1 + (i-1)/100], \tag{4.2}$$

for $k = 1, \ldots, 7$. Table 5 reports the similar quantities for $(r, m) = (20, 8)$, $(10, 5)$ and $(5, 2)$, where

$$\boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\gamma}, \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_r), \gamma_i = r^{-1}(k-1)[1 + (i-1)/100], \tag{4.3}$$

for $k = 1, \ldots, 7$. Since the rank of $\boldsymbol{A}$ is r, the setup (4.3) corresponds to the case that the rank r is less than $n - 1$. These simulation results show that all the empirical Bayes estimators have nice risk performances, and especially EB gets significant risk gains.

Table 3: Prediction Error Estimates by the Cross-validation Method for $p = 300$ and $n = 40$ in the Two Cases of $\boldsymbol{X}$ Scaled and Non-scaled

Case of both $\boldsymbol{y}$ and $\boldsymbol{X}$ Scaled											
r	m	CR	LS	PC	JS	RPM	RRM	RP	RR	EB	$\overline{y}$
30	6	99.99	3.287	2.604	1.573	1.773	1.776	1.623	1.542	0.263	1.051
20	6	99.99	3.287	2.365	1.555	1.640	1.626	1.551	1.340	0.162	1.051
19	6	99.99	3.287	2.336	1.595	1.579	1.550	1.559	1.339	0.117	1.051
16	5	99.99	3.287	2.033	1.489	1.531	1.541	1.485	1.270	0.256	1.051
10	5	99.99	3.287	1.558	1.390	1.412	1.401	1.368	1.273	0.488	1.051
6	2	99.98	3.287	1.267	1.201	1.267	1.267	1.257	1.190	0.701	1.051
Case of $\boldsymbol{y}$ Scaled but $\boldsymbol{X}$ Non-scaled											
r	m	CR	LS	PC	JS	RPM	RRM	RP	RR	EB	$\overline{y}$
37	6	99.99	4.296	4.206	1.923	2.201	2.109	1.750	1.433	0.329	1.051
30	6	99.99	4.296	3.282	1.966	1.984	1.947	1.809	1.571	0.348	1.051
20	6	99.99	4.296	2.564	1.835	1.805	1.796	1.726	1.412	0.396	1.051
10	5	99.99	4.296	1.826	1.653	1.662	1.613	1.627	1.271	0.531	1.051
6	2	99.98	4.296	1.632	1.502	1.632	1.632	1.527	1.316	0.701	1.051

5 Concluding Remarks

In this paper we consider a linear multiple regression model with fewer observations than the number of predictor variables. We have shown that the least squares estimator (LSE) of the regression parameter $\boldsymbol{\beta}$ is not only biased but unstable in the sense of having a very large prediction error. When the observation vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ on the predictor variables are centered and assumed to have an empirical distribution $\widehat{F}_x$ under which $E[\boldsymbol{x}|\widehat{F}_x] = \boldsymbol{0}$ and $E[\boldsymbol{x}\boldsymbol{x}^t|\widehat{F}_x] = n^{-1}\boldsymbol{X}^t\boldsymbol{X}$, then it is shown that the empirical Bayes estimators $\widehat{\boldsymbol{\beta}}^{JS}$, $\widehat{\boldsymbol{\beta}}^{RPM}$, and $\widehat{\boldsymbol{\beta}}^{RRM}$ have smaller risks than the LSE $\widehat{\boldsymbol{\beta}}$, where the risk is defined by $E[(\hat{y} - y)^2]$, and $\hat{y}$ is the predicted value of y. Although the theoretical performance of the estimator $\widehat{\boldsymbol{\beta}}^{EB}$ given in (2.15) is not known, it has been found to perform better than all the six empirical Bayes estimators and the least squares. However, when both $\boldsymbol{y}$ and $\boldsymbol{X}$ are not scaled, such a superiority does not exist. We therefore recommend that both $\boldsymbol{y}$ and $\boldsymbol{X}$ be centered and scaled before using any of the six empirical Bayes estimators. However, the use of least squares estimator is not recommended.

6 Appendix

We first define the Moore-Penrose inverse $\boldsymbol{A}^+$ of the matrix $\boldsymbol{A}$, which is unique. That is $\boldsymbol{A}^+$ satisfies the following four conditions (i) $\boldsymbol{A}\boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{A}$, (ii) $\boldsymbol{A}^+\boldsymbol{A}\boldsymbol{A}^+ = \boldsymbol{A}^+$, (iii) $\boldsymbol{A}^+\boldsymbol{A} = (\boldsymbol{A}^+\boldsymbol{A})^t$ and (iv) $\boldsymbol{A}\boldsymbol{A}^+ = (\boldsymbol{A}\boldsymbol{A}^+)^t$.

Table 4: Relative Efficiencies $R(\boldsymbol{\beta}, 1, \boldsymbol{\delta})/R(\boldsymbol{\beta}, 1, \widetilde{\boldsymbol{\beta}})$ for $\sigma^2 = 1$ in Estimation Errors by Simulation under (4.2) where $\widetilde{\boldsymbol{\beta}}$ is the LS Estimator (4.1)

$p = 300$, $n = 40$, $r = 30$, $m = 10$							
k	PC	JS	RPM	RRM	RP	RR	EB
1	0.7686	0.3183	0.3289	0.2416	0.3282	0.1297	0.1295
2	0.7826	0.3486	0.3599	0.2811	0.3591	0.1689	0.1279
3	0.8138	0.4190	0.4315	0.3706	0.4307	0.2618	0.1232
4	0.8475	0.4986	0.5123	0.4709	0.5115	0.3683	0.1119
5	0.8756	0.5697	0.5841	0.5565	0.5835	0.4633	0.0957
6	0.8970	0.6278	0.6424	0.6232	0.6418	0.5414	0.0780
7	0.9127	0.6736	0.6878	0.6739	0.6874	0.5995	0.0634
$p = 300$, $n = 40$, $r = 15$, $m = 3$							
k	PC	JS	RPM	RRM	RP	RR	EB
1	0.3865	0.1214	0.17843	0.1710	0.1368	0.0783	0.0808
2	0.4099	0.1534	0.20872	0.2022	0.1685	0.1108	0.0851
3	0.4637	0.2270	0.27880	0.2730	0.2411	0.1864	0.0928
4	0.5221	0.3067	0.35504	0.3501	0.3197	0.2692	0.0927
5	0.5712	0.3735	0.41888	0.4146	0.3857	0.3395	0.0878
6	0.6086	0.4240	0.46715	0.4639	0.4357	0.3932	0.0792
7	0.6361	0.4608	0.50250	0.4999	0.4722	0.4329	0.0701
$p = 300$, $n = 40$, $r = 5$, $m = 2$							
k	PC	JS	RPM	RRM	RP	RR	EB
1	0.1285	0.0943	0.1285	0.1271	0.0948	0.0476	0.0681
2	0.1594	0.1263	0.1594	0.1581	0.1268	0.0799	0.0661
3	0.2294	0.1992	0.2294	0.2282	0.1997	0.1536	0.0606
4	0.3052	0.2780	0.3052	0.3041	0.2785	0.2335	0.0544
5	0.3688	0.3442	0.3688	0.3678	0.3447	0.3011	0.0489
6	0.4171	0.3943	0.4171	0.4162	0.3949	0.3528	0.0425
7	0.4526	0.4311	0.4526	0.4518	0.4317	0.3912	0.0372

Table 5: Relative Efficiencies $R(\boldsymbol{\beta}, 1, \boldsymbol{\delta})/R(\boldsymbol{\beta}, 1, \widetilde{\boldsymbol{\beta}})$ for $\sigma^2 = 1$ in Estimation Errors by Simulation under (4.3) where $\widetilde{\boldsymbol{\beta}}$ is the LS Estimator (4.1)

$p = 300$, $n = 40$, $r = 20$, $m = 8$							
k	PC	JS	RPM	RRM	RP	RR	EB
1	0.5174	0.2526	0.2631	0.2087	0.2565	0.0948	0.1016
2	0.5195	0.2558	0.2662	0.2120	0.2597	0.0989	0.1016
3	0.5260	0.2656	0.2760	0.2223	0.2695	0.1109	0.1001
4	0.5363	0.2811	0.2914	0.2391	0.2849	0.1294	0.0977
5	0.5495	0.3009	0.3109	0.2603	0.3046	0.1527	0.0949
6	0.5647	0.3236	0.3334	0.2849	0.3273	0.1793	0.0928
7	0.5809	0.3477	0.3574	0.3109	0.3513	0.2080	0.0927
$p = 300$, $n = 40$, $r = 10$, $m = 5$							
k	PC	JS	RPM	RRM	RP	RR	EB
1	0.2561	0.1677	0.1850	0.1649	0.1686	0.0623	0.0734
2	0.2677	0.1809	0.1978	0.1780	0.1817	0.0768	0.0742
3	0.2998	0.2171	0.2331	0.2141	0.2178	0.1168	0.0763
4	0.3426	0.2654	0.2803	0.2621	0.2660	0.1697	0.0794
5	0.3869	0.3152	0.3290	0.3116	0.3157	0.2242	0.0844
6	0.4273	0.3604	0.3732	0.3567	0.3609	0.2736	0.0877
7	0.4616	0.3987	0.4107	0.3951	0.3991	0.3150	0.0906
$p = 300$, $n = 40$, $r = 5$, $m = 2$							
k	PC	JS	RPM	RRM	RP	RR	EB
1	0.1285	0.0943	0.1285	0.1271	0.0948	0.0476	0.0681
2	0.1803	0.1488	0.1803	0.1791	0.1494	0.1029	0.0648
3	0.2839	0.2572	0.2839	0.2828	0.2580	0.2114	0.0579
4	0.3728	0.3502	0.3728	0.3718	0.3512	0.3046	0.0515
5	0.4335	0.4136	0.4335	0.4326	0.4148	0.3675	0.0421
6	0.4730	0.4548	0.4730	0.4722	0.4561	0.4082	0.0326
7	0.4990	0.4819	0.4990	0.4982	0.4833	0.4348	0.0242

Next, we provide the proof of Theorem 3.1.

Proof of Theorem 3.1. Letting $R(\boldsymbol{\beta}, \sigma^2, \widehat{\boldsymbol{\beta}}) = E[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t \boldsymbol{X}^t \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})]/\sigma^2$, we show that $\Delta = R(\boldsymbol{\beta}, \sigma^2, \widehat{\boldsymbol{\beta}}) - R(\boldsymbol{\beta}, \sigma^2, \widehat{\boldsymbol{\beta}}_c^{EB}(\widehat{\boldsymbol{\gamma}}_1)) \geq 0$ for any $\boldsymbol{\beta}$ and σ^2. We note that $\widehat{\boldsymbol{\gamma}} \sim \mathcal{N}_r(\boldsymbol{\gamma}, \sigma^2 \boldsymbol{D}^{-2})$ and $\widehat{\boldsymbol{\gamma}}_2 \sim \mathcal{N}_{r-m}(\boldsymbol{\gamma}_2, \sigma^2 \boldsymbol{D}_2^{-2})$. We also note that from the properties of A_1 and A_2 matrices and the fact that $\boldsymbol{\gamma} = \boldsymbol{A}^t \boldsymbol{\beta}$.

$$\begin{aligned}
\boldsymbol{A}^t \widehat{\boldsymbol{\beta}}_c^{EB}(\widehat{\boldsymbol{\gamma}}_1) &= \widehat{\boldsymbol{\gamma}} - \left[\boldsymbol{I}_r - (\boldsymbol{I}_r + \hat{\lambda}^{-1} \boldsymbol{D}^{-c})^{-1}\right] \left(\widehat{\boldsymbol{\gamma}} - \begin{pmatrix} \widehat{\boldsymbol{\gamma}}_1 \\ \boldsymbol{0} \end{pmatrix}\right) \\
&= \begin{pmatrix} \widehat{\boldsymbol{\gamma}}_1 \\ \widehat{\boldsymbol{\gamma}}_2^{EB} \end{pmatrix}
\end{aligned}$$

where

$$\widehat{\boldsymbol{\gamma}}_2^{EB} = \left[\boldsymbol{I}_{r-m} - (\boldsymbol{I}_{r-m} + \hat{\lambda} \boldsymbol{D}_2^c)^{-1}\right] \widehat{\boldsymbol{\gamma}}_2,$$

and $\boldsymbol{D}_2 = \mathrm{diag}\,(d_{m+1}, \ldots, d_r)$. The equation (2.7) expressed in terms of $\boldsymbol{\gamma}$ and $\widehat{\boldsymbol{\gamma}}$ can be written as

$$\widehat{\boldsymbol{\gamma}}_2^t \left[\boldsymbol{D}^{-2} + \lambda_* \boldsymbol{D}^{c-2}\right] \widehat{\boldsymbol{\gamma}}_2 = a\hat{\sigma}^2. \tag{6.1}$$

The risk difference of the two estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_c^{EB}(\widehat{\boldsymbol{\gamma}}_1)$ is written by

$$\Delta = E\left[(\widehat{\boldsymbol{\gamma}}_2 - \boldsymbol{\gamma}_2)^t \boldsymbol{D}^2 (\widehat{\boldsymbol{\gamma}}_2 - \boldsymbol{\gamma}_2)\right] - E\left[(\widehat{\boldsymbol{\gamma}}_2^{EB} - \boldsymbol{\gamma}_2)^t \boldsymbol{D}^2 (\widehat{\boldsymbol{\gamma}}_2^{EB} - \boldsymbol{\gamma}_2)\right],$$

where

$$\widehat{\boldsymbol{\gamma}}_2^{EB} = \left(\boldsymbol{I}_{r-m} - (\boldsymbol{I}_{r-m} + \hat{\lambda} \boldsymbol{D}_2^c)^{-1}\right) \widehat{\boldsymbol{\gamma}}_2,$$

$\boldsymbol{D}_2 = \mathrm{diag}\,(d_{m+1}, \ldots, d_r)$ and the equation (2.7) is expressed by

$$\widehat{\boldsymbol{\gamma}}_2^t \left(\boldsymbol{D}_2^{-2} + \lambda_* \boldsymbol{D}^{c-2}\right)^{-1} \widehat{\boldsymbol{\gamma}}_2 = (r - m - 2)\hat{\sigma}^2. \tag{6.2}$$

Let $\boldsymbol{Z} = (Z_{m+1}, \ldots, Z_r)^t = \boldsymbol{D}_2 \widehat{\boldsymbol{\gamma}}_2/\sigma$, $\boldsymbol{\theta} = (\theta_{m+1}, \ldots, \theta_r)^t = \boldsymbol{D}_2 \boldsymbol{\gamma}_2/\sigma$ and $V = (n - r + 1)\hat{\sigma}^2/\sigma^2$. Then, $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{I}_{r-m})$ and $V \sim \chi^2_{n-r-1}$. The risk difference is further rewritten as

$$\begin{aligned}
\Delta/\sigma^2 =& E\left[\|\boldsymbol{Z} - \boldsymbol{\theta}\|^2\right] - E\left[\|\widehat{\boldsymbol{\theta}}^{EB} - \boldsymbol{\theta}\|^2\right] \\
=& E\left[2 \sum_{i=m+1}^{r} (Z_i - \theta_i) \frac{Z_i}{1 + d_i^c \hat{\lambda}} - \sum_{i=m+1}^{r} \frac{Z_i^2}{(1 + d_i^c \hat{\lambda})^2}\right],
\end{aligned} \tag{6.3}$$

where $\widehat{\boldsymbol{\theta}}^{EB} = \boldsymbol{D}_2 \widehat{\boldsymbol{\gamma}}_2^{EB}/\sigma = (\boldsymbol{I}_{r-m} - (\boldsymbol{I}_{r-m} + \hat{\lambda} \boldsymbol{D}_2^c)^{-1}) \boldsymbol{Z}$ and the equation (6.2) is given by

$$\boldsymbol{Z}^t (\boldsymbol{I}_{r-m} + \lambda_* \boldsymbol{D}_2^c)^{-1} \boldsymbol{Z} = \sum_{i=m+1}^{r} \frac{Z_i^2}{1 + d_i^c \hat{\lambda}} = a^* V$$

for $a^* = a/(n-r+1)$. Applying the implicit function theorem to this equation and noting that $\hat{\lambda} = \max(\lambda_*, 0)$, we observe that

$$\begin{aligned} Z_i \frac{\partial \hat{\lambda}}{\partial Z_i} =& Z_i \frac{\partial \lambda_*}{\partial Z_i} I(\lambda_* > 0) = 2 \frac{Z_i^2/(1+d_i^c \lambda_*)}{\sum_{j=m+1}^r d_j^c Z_j^2/(1+d_j^c \lambda_*)^2} I(\lambda_* > 0) \\ \leq & 2 \frac{Z_i^2/(1+d_i^c \hat{\lambda})}{\sum_{j=m+1}^r d_j^c Z_j^2/(1+d_j^c \hat{\lambda})^2}, \end{aligned} \tag{6.4}$$

and

$$\begin{aligned} \frac{\partial \hat{\lambda}}{\partial V} =& - \frac{a^*}{\sum_{i=m+1}^r d_i^c Z_i^2/(1+d_i^c \lambda_*)^2} I(\lambda_* > 0) \\ \geq & - \frac{a^*}{\sum_{i=m+1}^r d_i^c Z_i^2/(1+d_i^c \hat{\lambda})^2}. \end{aligned} \tag{6.5}$$

By using the Stein identity (Stein (1981)), the cross product term in (6.3) is evaluated as

$$\begin{aligned} E\left[\sum_{i=m+1}^r (Z_i - \theta_i) \frac{Z_i}{1+d_i^c \hat{\lambda}}\right] =& E\left[\sum_{i=m+1}^r \frac{\partial}{\partial Z_i}\left(\frac{Z_i}{1+d_i^c \hat{\lambda}}\right)\right] \\ =& E\left[\sum_{i=m+1}^r \frac{1}{1+d_i^c \hat{\lambda}} - \sum_{i=m+1}^r \frac{d_i^c Z_i}{(1+d_i^c \hat{\lambda})^2} \frac{\partial \hat{\lambda}}{\partial Z_i}\right] \\ \geq & E\left[\sum_{i=m+1}^r \frac{1}{1+d_i^c \hat{\lambda}} - 2 \frac{\sum_{i=m+1}^r d_i^c Z_i^2/(1+d_i^c \hat{\lambda})^3}{\sum_{i=m+1}^r d_i^c Z_i^2/(1+d_i^c \hat{\lambda})^2}\right] \\ \geq & E\left[\sum_{i=m+1}^r \frac{1}{1+d_i^c \hat{\lambda}} - 2 \frac{1}{1+d_r^c \hat{\lambda}}\right], \end{aligned} \tag{6.6}$$

where the first inequality follows from (6.4). Using the chi-square indentity (Efron and Morris (1976)), on the other hand, we see that

$$\begin{aligned} E\left[\sum_{i=m+1}^r \frac{Z_i^2/V}{(1+d_i^c \hat{\lambda})^2} V\right] =& E\left[(n-r-3) \sum_{i=m+1}^r \frac{Z_i^2/V}{(1+d_i^c \hat{\lambda})^2} - 4 \sum_{i=m+1}^r \frac{d_i^c Z_i^2}{(1+d_i^c \hat{\lambda})^3} \frac{\partial \hat{\lambda}}{\partial V}\right] \\ \leq & E\left[(n-r-3) \sum_{i=m+1}^r \frac{Z_i^2/V}{(1+d_i^c \hat{\lambda})^2} + 4a^* \frac{\sum_{i=m+1}^r d_i^c Z_i^2/(1+d_i^c \hat{\lambda})^3}{\sum_{i=m+1}^r d_i^c Z_i^2/(1+d_i^c \hat{\lambda})^2}\right] \\ \leq & E\left[(n-r-3) \sum_{i=m+1}^r \frac{Z_i^2/V}{(1+d_i^c \hat{\lambda})^2} + 4 \frac{a^*}{1+d_r^c \hat{\lambda}}\right] \\ \leq & E\left[\frac{a}{1+d_r^c \hat{\lambda}}\right], \end{aligned} \tag{6.7}$$

where the first inequality follows from (6.5). Combining (6.3), (6.6) and (6.7) yields that

$$\begin{aligned}\Delta/\sigma^2 \geq & E\left[\sum_{i=m+1}^{r}\frac{2}{1+d_i^c\hat{\lambda}}-\frac{a+4}{1+d_r^c\hat{\lambda}}\right]\\ = & E\left[\frac{2}{1+d_r^c\hat{\lambda}}\left\{\sum_{i=m+1}^{r}\frac{1+d_r^c\hat{\lambda}}{1+d_i^c\hat{\lambda}}-\frac{a}{2}-2\right\}\right],\end{aligned}$$

which is non-negative under the condition (3.5) since $d_i^c > d_r^c$ and hence $(1+d_r^c\hat{\lambda})/(1+d_i^c\hat{\lambda}) > d_r^c/d_i^c$. Therefore, the proof of Theorem 3.1 is complete. ■

Acknowledgments. The research of the authors were supported in part by grants from the Ministry of Education, Japan, Nos. 13680371, 15200021, 15200022 and 16500172 and in part by a grant from the 21st Century COE Program at the Faculty of Economics, University of Tokyo.

REFERENCES

Brown, P.J., Fearn, T., and Vannucci, M. (1999). The choice of variables in multivariate regression: A non-conjugate Bayesian decision theory approach. *Biometrika*, **86**, 635-648.

Brown, P.J., Fearn, T., and Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *J. Amer. Statist. Assoc.* **96**, 398-408.

Copas, J.B. (1983). Regression, prediction, shrinkage. *J. Roy. Statist. Soc., B*, **45**, 311-335, (Discussion, 335-354).

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

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.

Kubokawa, T., and Srivastava, M.S. (2004). Improved empirical Bayes ridge regression estimators under multicollinearity. *Communications in Statistics - Theory Methods*, **33**, 1943-1973.

Osborne, B.G., Feran, T., Miller, A.R., and Douglas, S. (1984). Application of near-infrared reflectance spectroscopy to compositional analysis of biscuits and biscuit doughs. *J. Science of Food and Agriculture*, **35**, 99-105.

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.

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

Srivastava, M.S., and Khatri, C.G. (1979). *An Introduction to Multivariate Statistics*. North-Holland, New York.

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

West, M. (2003). Bayesian factor regression models in the 'large p, small n' paradigm. *Bayesian Statistics*, **7**, 723-732.