2015

An iterative approach to minimize the mean squared error in ridge regression

Ka Yiu Wong
Department of Mathematics, Hong Kong Baptist University

Sung Nok Chiu
Department of Mathematics, Hong Kong Baptist University, snchiu@hkbu.edu.hk

An iterative approach to minimize the mean squared error in ridge regression

Ka Yiu Wong · Sung Nok Chiu

Abstract The methods of computing the ridge parameters have been studied for more than four decades. However, there is still no way to compute its optimal value. Nevertheless, many methods have been proposed to yield ridge regression estimators of smaller mean squared errors than the least square estimators empirically. This paper compares the mean squared errors of 26 existing methods for ridge regression in different scenarios. A new approach is also proposed, which minimizes the empirical mean squared errors iteratively. It is found that the existing methods can be divided into two groups: one is those that are better, but only slightly, than the least squares method in many cases, and the other is those that are much better than the least squares method in only some cases but can be (sometimes much) worse than it in many others. The new method, though not uniformly the best, outperforms the least squares method well in many cases and underperforms it only slightly in a few cases.

Keywords Least squares · Multicollinearity · Optimal ridge parameter

1 Introduction

The ordinary least squares (OLS) parameter estimator of a standardized multiple regression model requires the inverse of the correlation matrix of the regressors. Thus, multicollinearity will cause a problem because the determinant of the correlation matrix may be small. The seminal paper by Hoerl and Kennard (1970b) suggests the so-called ridge regression (also known as Tikhonov regularization). The ridge regression estimator is obtained by simply adding an equal amount $k > 0$ to each diagonal element of the correlation matrix in the OLS estimator, and it can be shown that there always exists a ridge parameter k_0 such that weighted sums of coefficient mean square errors of the ridge regression estimator is smaller than those of the OLS estimator (Theobald, 1974; Farebrother, 1976).

K. Y. Wong · S. N. Chiu (✉)
Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong
e-mail: snchiu@hkbu.edu.hk

The ridge regression became extremely popular in the seventies and eighties, see the survey in McDonald (2009), and received increasing attention in applications, especially in biostatistics (Fahrmeir et al., 2013, p. 159). However, there is no explicit formula for the optimal value of this ridge parameter. Many authors proposed different approximations for it. Each new suggestion was compared with and often declared victory over some existing ones, but there did not exist a large scale comparison between all known methods. The conventional wisdom is that no single method would be uniformly better than all the others. As a result, a most widely-adopted approach turned out to be just visual inspection of the ridge trace (Hoerl and Kennard, 1970a), which plots the ridge estimates versus k; the least value of k starting from which the estimates seem stabilized would be chosen.

This paper on one hand gives a survey of existing methods and on the other hand proposes a new approach to approximate the optimal ridge parameter. Simulation will be used to compare their performance in terms of mean squared errors (MSE) and prediction sum of squares (PRESS).

2 Ridge regression model

Consider the standardized multiple linear regression model (Kutner et al., 2005, p. 273) of n observations and p regressors:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{Y}$ and $\boldsymbol{\varepsilon}$, respectively, are $n \times 1$ vectors of observations and errors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters and $\boldsymbol{X}$ is an $n \times p$ matrix of regressors. The distribution assumption of $\boldsymbol{\varepsilon}$ is irrelevant to the computation of the estimates.

By solving the normal equation, the OLS estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}.$$

in which $\boldsymbol{X'X}$ is the correlation matrix of $\boldsymbol{X}$ and $\boldsymbol{X'Y}$ is the vector of correlation coefficients between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

If the determinant of $\boldsymbol{X'X}$ is close to zero, in order to stabilize the parameter estimates, a constant $k > 0$ is added to each diagonal element of $\boldsymbol{X'X}$, leading to the ridge regression estimator of $\boldsymbol{\beta}$ as follows:

$$\tilde{\boldsymbol{\beta}}(k) = (\boldsymbol{X'X} + k\boldsymbol{I})^{-1}\boldsymbol{X'Y}, \tag{2}$$

where $\boldsymbol{I}$ is the $p \times p$ identity matrix, and the OLS estimator is the particular (degenerate) ridge regression estimator corresponding to $k = 0$, i.e.

$$\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(0).$$

The estimator in (2) can also be considered as the result of least squares with penalty $k\boldsymbol{\beta'}\boldsymbol{\beta}$. Replacing this L^2-penalty by the L^1-penalty $k\|\boldsymbol{\beta}\|_1$ leads to the lasso (Tibshirani, 1996). See

Hastie et al. (2009, pp. 61–73) for more details on their relationship. This interpretation suggests that $\tilde{\boldsymbol{\beta}}(k)$ shrinks to zero when $k \to \infty$.

Note that the ridge regression is not invariant under scaling of the variables and some authors did not standardize the variables. See Groß (2003, Section 3.4.4) for a discussion of advantages and disadvantages of standardization. Because Hoerl and Kennard (1970b) established properties of the ridge regression estimator under the standardized case and many software packages, like SAS and MATLAB, compute the ridge estimators using the standardized variables by default, this paper considers the standardized model.

As we can see from Appendix, most of the ridge parameter computations are derived from the canonical form of model (1), which is expressed as follows. Denote by $\boldsymbol{\Lambda}$ the $p \times p$ diagonal matrix with elements $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ which are the eigenvalues of $\boldsymbol{X'X}$ and by $\boldsymbol{Q}$ the matrix containing the corresponding normalized orthogonal eigenvectors, such that $\boldsymbol{X'X} = \boldsymbol{Q\Lambda Q'}$. Let $\boldsymbol{Z} = \boldsymbol{XQ}$ and $\boldsymbol{\alpha} = \boldsymbol{Q'\beta}$. Model (1) can now be expressed in the canonical form

$$\boldsymbol{Y} = \boldsymbol{Z\alpha} + \boldsymbol{\varepsilon}.$$

The OLS estimator and the ridge regression estimator, respectively, of $\boldsymbol{\alpha}$ are

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{\Lambda^{-1}Z'Y} = \boldsymbol{Q'}\hat{\boldsymbol{\beta}},$$
$$\tilde{\boldsymbol{\alpha}}(k) = (\boldsymbol{\Lambda} + k\boldsymbol{I})^{-1}\boldsymbol{Z'Y} = \boldsymbol{Q'}\tilde{\boldsymbol{\beta}}(k).$$

Hoerl and Kennard (1970b) remarked that instead of the same k one may add different values to the diagonal elements, such as adding a large value to a small λ_i and vice versa (see e.g. Groß, 2003, Section 3.6), leading to the so-called general ridge estimator. However, they suggested, based on experience, that using the same k could achieve a better estimate.

The MSE of $\tilde{\boldsymbol{\beta}}(k)$ is given by

$$\mathrm{MSE}_{\tilde{\boldsymbol{\beta}}}(k) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2} \tag{3}$$

where σ^2 is the variance of error term $\boldsymbol{\varepsilon}$ and α_i the ith element in $\boldsymbol{\alpha}$. The minimizer of $\mathrm{MSE}_{\tilde{\boldsymbol{\beta}}}(k)$ will be regarded as the optimal ridge parameter. However, the right-hand side of (3) includes the unknown σ^2 as well as the unknown regression parameter $\boldsymbol{\alpha}$. Thus, the optimal k can never be derived analytically from a given sample.

3 New proposed ridge parameter

We propose an iterative method to approximate the optimal k. The idea is to minimize the empirical values of (3). When the regression coefficients are unknown, a natural way to estimate $\mathrm{MSE}_{\tilde{\boldsymbol{\beta}}}(k)$ is to replace the estimate $\boldsymbol{\alpha}$ and σ by their OLS estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\sigma}$, respectively. However, when the correlation matrix of $\boldsymbol{X}$ is close to singular, not only $\hat{\boldsymbol{\alpha}}$ but also the estimated $\mathrm{MSE}_{\tilde{\boldsymbol{\beta}}}(k)$ are numerically unstable. Here iteration is used to estimate

$\boldsymbol{\alpha}$ and $\mathrm{MSE}_{\tilde{\boldsymbol{\beta}}}(k)$. First, $\mathrm{MSE}_{\tilde{\boldsymbol{\beta}}}(k)$ is still estimated using the OLS $\hat{\boldsymbol{\alpha}}$. Then, the minimizer of the estimated $\mathrm{MSE}_{\tilde{\boldsymbol{\beta}}}(k)$ is computed and denoted by $k_{(1)}$. Considering that the ridge estimator $\tilde{\boldsymbol{\alpha}}(k_{(1)})$ is a more stable estimate than $\hat{\boldsymbol{\alpha}}$, we re-estimate $\mathrm{MSE}_{\tilde{\boldsymbol{\beta}}}(k)$ by plugging in $\tilde{\boldsymbol{\alpha}}(k_{(1)})$ and denote the minimizer of the second estimated $\mathrm{MSE}_{\tilde{\boldsymbol{\beta}}}(k)$ by $k_{(2)}$. The above steps are repeated until the difference between $k_{(j)}$ and $k_{(j-1)}$ is sufficiently small for some $j \geq 1$, with the convention that $k_{(0)} = 0$ (corresponding to OLS). To be more precise, the iterative procedure is summarized as follows.

Algorithm 1 An iterative approach to estimate the optimal ridge parameter

Input: eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ of $\boldsymbol{X'X}$; OLS estimate $\hat{\sigma}$; OLS estimate $\hat{\boldsymbol{\alpha}}$; pre-specified tolerance δ; pre-specified maximum number of iterations J.

Output: k, an approximate solution for optimal ridge parameter.

1: Set $k_{(0)} = 0$.
2: **for** $j = 1, \ldots, J$ **do**
3: Set $k_{(j)} = \arg\min_{x \geqslant 0} \left\{ \hat{\sigma}^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + x)^2} + x^2 \sum_{i=1}^{p} \frac{\hat{\alpha}_i^2 \lambda_i^2}{(\lambda_i + x)^2 (\lambda_i + k_{(j-1)})^2} \right\}$.
4: **if** $\left| k_{(j)} - k_{(j-1)} \right| < \delta k_{(j-1)}$ **then**
5: Set $k = k_{(j)}$ and stop.
6: **end if**
7: **end for**
8: Set $k = k_{(J)}$.

We have to pre-specify a maximum number of iteration because we do not have a proof of the convergence of Algorithm 1.

Even though the above iterative approach is natural and quite straightforward, to the best of our knowledge, it has not been considered in the literature at all. In the following we will compare this approach with 26 other existing methods, see Appendix for their details. We denote our ridge parameter obtained from this iterative approach by k_{27}.

4 Simulation

We use $\delta = 10^{-6}$ and $J = 2000$ (with good luck, the algorithm always converged before j reached 2000 in our simulation for all the cases presented in the next section). The minimizers of operation 3 in Algorithm 1 are searched in the interval $[0, 10]$ by the golden section method with tolerance parameter also equal to 10^{-6}.

Following McDonald and Galarneau (1975), we first generate an $n \times (p+1)$ matrix of i.i.d. standard normal random numbers, denoted as $\boldsymbol{M}$, and then compute $\boldsymbol{X}^*$ by

$$\boldsymbol{X}_i^* = (1 - \gamma^2)^{\frac{1}{2}} \boldsymbol{M}_i + \gamma \boldsymbol{M}_{p+1} \qquad \text{for } i = 1, \ldots, p,$$

where $\boldsymbol{X}_i^*$ and $\boldsymbol{M}_i$ are the ith column of $\boldsymbol{X}^*$ and $\boldsymbol{M}$ respectively and γ^2 is the correlation

between each column of $\boldsymbol{X}^*$. The dependent variable $\boldsymbol{Y}^*$ is obtained by

$$\boldsymbol{Y}^* = \boldsymbol{\beta_0} + \boldsymbol{X}^*\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^* \tag{4}$$

where $\boldsymbol{\varepsilon}^*$ is a vector of i.i.d. zero-mean normal numbers with standard deviation σ^*. Then we standardize (4) to get the standardized model (1). The 27 different ridge parameters and the corresponding estimators are computed from the standardized $\boldsymbol{X}$ and $\boldsymbol{Y}$, and then the estimators are transformed back to the original one to compute the MSE ratios. Note that the proposed iterative approach is to minimize the empirical $\mathrm{MSE}_{\tilde{\boldsymbol{\beta}}}(k)$ of the parameters in the standardized model (1), not the empirical $\mathrm{MSE}_{\tilde{\boldsymbol{\beta}}^*}(k)$ of the parameters in the original model (4), because as mentioned in Section 2, we use the standardized model in the formulation of the ridge estimation. However, when comparing the performance of the different ridge estimators, we believe that it is of practical interest to compare the errors in estimating the parameters $\boldsymbol{\beta}^*$ of the original models. Nevertheless, it is of course also possible to compare the estimation errors of the parameter estimates of the standardized models and/or to minimize the empirical MSE of the parameters in the original model but we do not consider these variations here.

For simplicity, $\boldsymbol{\beta_0}$ is set to be zero for all cases simulated. Three different scenarios for $\boldsymbol{\beta^*} = (\beta_1^*, \ldots, \beta_p^*)'$ will be considered in the simulation. The first one is that for each generated $\boldsymbol{X}^*$, $\boldsymbol{\beta^*}$ is the normalized eigenvector of the largest eigenvalue λ_1 of the correlation matrix $\boldsymbol{X'X}$. Newhouse and Oman (1971) showed that if the MSE of the ridge estimator is regarded as a function of $\boldsymbol{\beta^*}$, while σ^*, k and $\boldsymbol{X}$ are kept fixed, then the MSE attains its minimum when $\boldsymbol{\beta^*}$ is the normalized eigenvector corresponding to λ_1. This is a very typical choice of $\boldsymbol{\beta^*}$ in the literature. The second scenario uses random numbers between 0 and 10 for β_j^* with $\beta_1^* = 1$ and $\beta_p^* = 10$. The last one is similar but instead of random numbers from $[0, 10]$, random numbers from $[0, 100]$ with $\beta_1^* = 1$ and $\beta_p^* = 100$ are used.

For each fixed σ^*, thirty six cases of the model, using $p = 2, 4, 12$, $n = 20, 100$ and $\gamma = 0.9999, 0.999, 0.99, 0.8, 0.4, 0.01$ are considered, and we take $\sigma^* = 0.01, 1, 10$. Each case with the same arbitrary but fixed $\boldsymbol{X^*}$ and $\boldsymbol{\beta^*}$ is repeated 1000 times with independently generated $\boldsymbol{\varepsilon^*}$ to get an average MSE and an average PRESS. Although we consider one γ in each model only, in real situation, both low and high correlations between the regressors may occur in a model at the same time. To see what will happen when the ridge estimation is applied to low correlation cases, $\gamma = 0.01, 0.4$ are also considered. The performance of the ridge parameters of each case will be compared with OLS by using the MSE (PRESS, respectively) ratio, which is the ratio of the average MSE (average PRESS) of $\tilde{\boldsymbol{\beta}}^*(k)$ to the average MSE (average PRESS) of $\tilde{\boldsymbol{\beta}}^*(0)$.

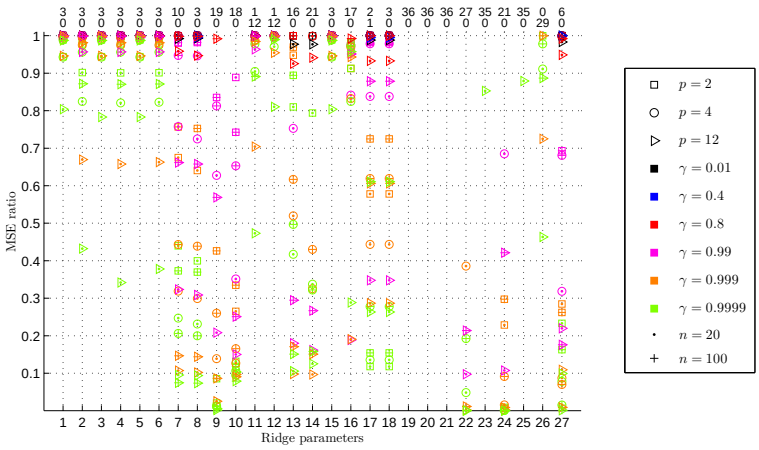


Figure 1: MSE ratios for the 36 cases when $\sigma^* = 0.01$.

5 Empirical Results

5.1 Using normalized eigenvector of λ_1 as $\boldsymbol{\beta}^*$

5.1.1 Model with $\sigma^* = 0.01$

Figure 1 shows the MSE ratios when $\sigma^* = 0.01$. The first row of numbers at the top indicate how many cases that the ridge estimators have MSE ratios greater than 1 by at least 10^{-6} (cases worse than OLS), whilst the second row how many cases that the MSE ratios are between 1 ± 10^{-6} (cases not better than OLS). A case with MSE ratio not larger than $1 - 10^{-6}$ (a case better than OLS) is indicated explicitly by a colored (green, orange, magenta, red, blue or black, corresponding to a different correlation parameter γ) marker ($\triangleright$, $\circ$ or $\square$, corresponding to a different number p of regressors) with either a bullet or a plus sign (corresponding to a different sample size n) inside. On the x-axis are the indices of the ridge parameters (see Appendix; our proposed approach corresponds to the last one, k_{27}).

When $\gamma < 0.99$, most of the ridge estimators are at most as good as OLS estimator. When $p = 2$, only a few ridge estimators have MSE ratios smaller than 1. As p increases, the MSE ratios of most of the ridge estimators decrease. For the model with $\gamma \geqslant 0.99$, we can observe that k_7, k_8, k_9, k_{10}, k_{13}, k_{14}, k_{16}, k_{17}, k_{18}, k_{24} and k_{27} perform well in most cases considered. Among these 11 ridge parameters, k_8, k_{17}, k_{18} and k_{27} are worse than OLS in only 2 to 6 cases in the 36 cases considered. Table 1 shows particularly the numerical values of the MSE ratios for $\gamma = 0.9999$ and $\sigma^* = 0.01$.

Among these 36 cases, although k_1, k_2, k_3, k_4, k_5, k_6 and k_{15} are worse than OLS only 3 times, their MSE ratios, as good as those of k_{11}, k_{12} and k_{26}, are never below 0.9 even when $\gamma = 0.99$ or $p = 2$. Meaningful improvement can be achieved by some of them only in very extreme cases, namely in small-sample cases with $\gamma = 0.9999$ and $p = 12$.

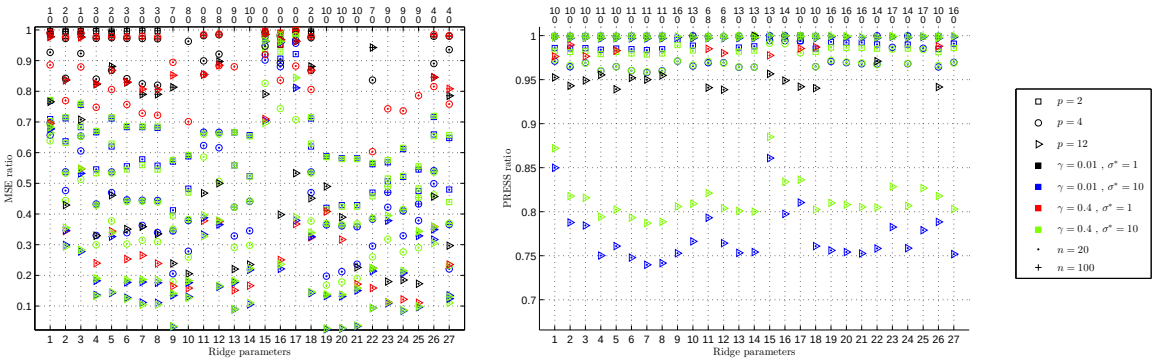
The MSE ratios of k_{19}, k_{20} and k_{21} are larger than that of OLS in all the 36 cases, and k_{23} and k_{25} are better than OLS only when $p = 12$, $n = 20$ and $\gamma = 0.9999$.

The PRESS ratios are not shown for this scenario because even the smallest PRESS ratio among these 36 cases is 0.9998, in an extreme case where $p = 12$, $n = 20$ and $\gamma = 0.99$ (while the largest ratio is 2.0720). That is, the PRESS ratios of all the ridge parameters are at most as good as OLS in this simulation study.

Table 1: MSE ratios of the models with $\gamma = 0.9999$ and $\sigma^* = 0.01$.

	$p = 2$		$p = 4$		$p = 12$	
k	$n = 20$	$n = 100$	$n = 20$	$n = 100$	$n = 20$	$n = 100$
1	0.943984	0.988785	0.942941	0.988443	0.803780	0.987197
2	0.901587	0.978108	0.824541	0.956243	0.432160	0.871633
3	0.943984	0.988785	0.942941	0.988443	0.783089	0.987197
4	0.900799	0.978100	0.820784	0.956189	0.341855	0.870513
5	0.943981	0.988784	0.942936	0.988442	0.783067	0.987195
6	0.901150	0.978059	0.822557	0.956207	0.378063	0.871043
7	0.439584	0.373012	0.247056	0.206299	0.095473	0.074740
8	0.399815	0.369733	0.231497	0.199876	0.093741	0.073639
9	0.085544	0.020689	0.011789	0.012998	0.001915	0.006012
10	0.131714	0.118205	0.106507	0.095486	0.103424	0.079190
11	1.000000	1.000000	0.904680	0.977738	0.473308	0.891055
12	1.000000	1.000000	0.970705	0.994195	0.810343	0.989310
13	0.810157	0.894370	0.417050	0.496920	0.105611	0.151005
14	0.793818	1.342583	0.337479	0.322127	0.125325	0.158294
15	0.943987	0.988786	0.942942	0.988444	0.803781	0.987197
16	0.912786	0.980490	0.824741	0.974141	0.288432	0.957526
17	0.117999	0.154356	0.135577	0.277844	0.263333	0.610535
18	0.117999	0.154356	0.135577	0.277844	0.263278	0.610535
19	> 10	> 10	> 10	> 10	1.192958	> 10
20	> 10	> 10	> 10	> 10	1.191175	> 10
21	> 10	> 10	> 10	> 10	1.188491	> 10
22	1.194441	7.000298	0.048502	0.192574	0.000075	0.000817
23	> 10	> 10	> 10	> 10	0.852779	> 10
24	0.008964	0.006310	0.000669	0.001253	0.000104	0.000198
25	> 10	> 10	> 10	> 10	0.879185	> 10
26	0.999640	1.000000	0.911094	0.978081	0.463502	0.886983
27	0.233215	0.163042	0.078165	0.015392	0.099056	0.001503

5.1.2 Model with $\sigma^* = 1, 10$ and $\gamma = 0.01, 0.4$



(a) (b)

Figure 2: (a) MSE ratios and (b) PRESS ratios for the 24 cases when $\gamma = 0.01$ and 0.4.

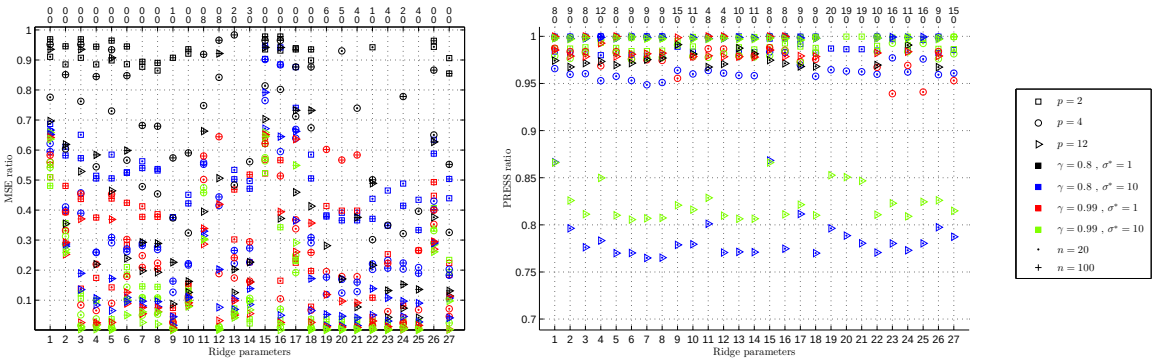
When we increase σ^* substantially from 0.01 to 1 and 10 and compare Figures 1 and 2(a), we can see that the MSE ratios are getting smaller. Figure 2(a) shows that k_{15}, k_{16} and k_{17} outperform OLS in all these cases but are usually far from the best choice in each case. The parameters k_4, k_5, k_6, k_7, k_8, k_{27} are worse than OLS in 2 to 4 cases only and their MSE ratios are not far from the smallest ones in most cases. Although k_9, k_{10}, k_{13}, k_{14}, k_{19}, k_{20}, k_{21}, k_{22}, k_{23}, k_{24} and k_{25} are close to the best choice in the cases when $\sigma^* = 10$, their MSE ratios are greater than that of OLS in most cases when $\sigma^* = 1$.

Figure 2(b) shows for these 24 cases the PRESS ratios, which decrease as σ^* increases, and the performance of most of the ridge parameters are similar. When $\sigma^* = 1$, none of the ridge parameters has PRESS ratio smaller than 0.9. The parameters k_7 and k_8 may be considered slightly better than the others when $\sigma^* = 10$ and $n = 20$, but no one is uniformly better than the others in these cases.

5.1.3 Model with $\sigma^* = 1, 10$ and $\gamma = 0.8, 0.99$

Comparing Figures 2(a) and 3(a), we can see that (which may be true in general) as γ increases, the MSE ratios decrease. Because of the high correlation, as expected, many ridge estimators outperform OLS in the cases considered. The parameters k_9, k_{18}, k_{19}, k_{20}, k_{21}, k_{22}, k_{23}, k_{24} and k_{25} are close to the best choices in most of the cases. However, they still have higher MSE ratios than that of OLS in a few cases. Among those ridge parameters with MSE ratios smaller than OLS in all the 24 cases, k_3, k_4, k_5, k_6, k_7, k_8, k_{10} and k_{27} perform very well and their MSE ratios are not far from the smallest ones.

Different from the MSE ratios, Figures 2(b) and 3(b) suggest that the PRESS ratios do not decrease as γ increases, and we still can only conclude that the ridge parameters perform similarly in terms of the PRESS ratios and no one stands out here.



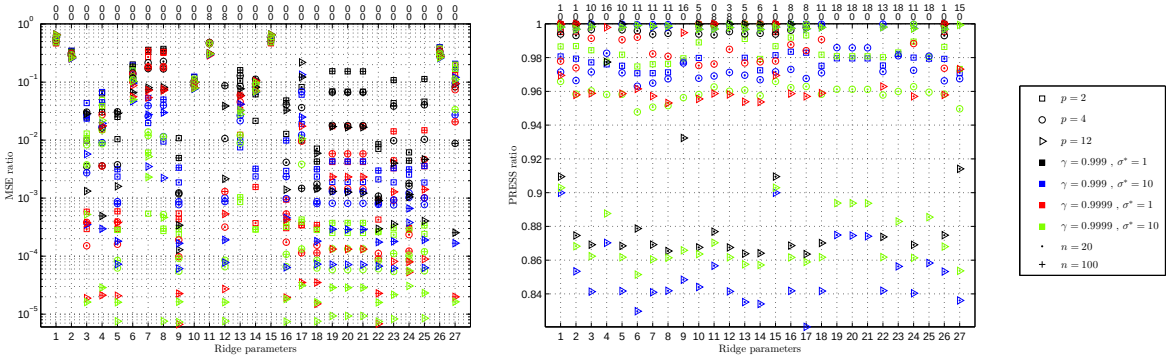
(a) (b)

Figure 3: (a) MSE ratios and (b) PRESS ratios for the 24 cases when $\gamma = 0.8$ and 0.99.

5.1.4 Model with $\sigma^* = 1, 10$ and $\gamma = 0.999, 0.9999$

The logarithmic scale is used on the y-axis for the MSE ratios in Figure 4(a), because they could be very small when the correlation is close to one, as OLS could be really bad. Except k_{11} and k_{12}, which are only equally good as OLS when $p = 2$, all the ridge estimators have less MSE than OLS in these cases. In general, the MSE ratios of k_3, k_5, k_9, k_{18}, k_{22} and k_{24} are lower than those of the others in the cases considered. The parameters k_{16}, k_{19}, k_{20}, k_{21}, k_{23} and k_{25} are slightly worse than the best one in a few cases with $\sigma^* = 1$ and $\gamma = 0.999$, but their performances are generally very good. Although k_4, k_7, k_8, k_{14}, k_{17} and k_{27} are worse than the best one in some cases, their MSE ratios are still very low in all the cases.

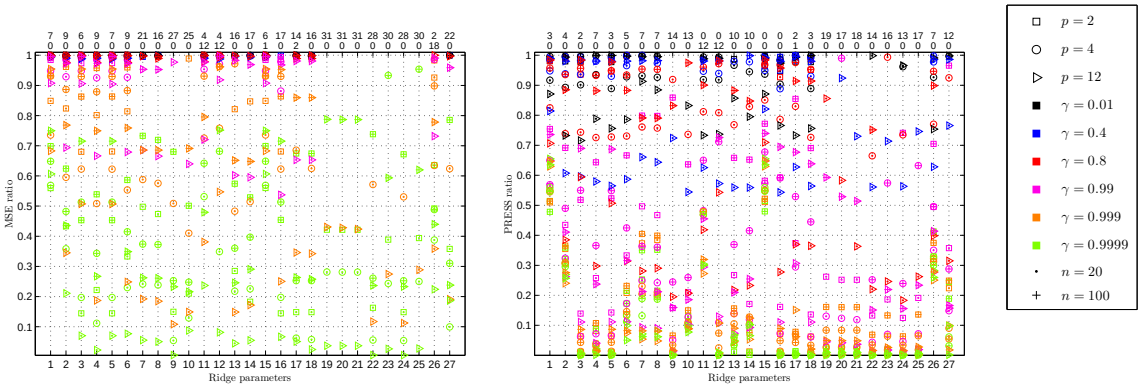
Consider the PRESS. In these 24 cases, none of the ridge parameters gives a ratio smaller than 0.8 and only k_{11} is never worse than OLS. The parameter k_{17} is slightly better than the others in many cases when $n = 20$. However, the differences are not substantial.



(a) (b)

Figure 4: (a) MSE ratios and (b) PRESS ratios for the 24 cases when $\gamma = 0.999$ and 0.9999.

5.2 Using random numbers between 0 and 10 as β_j^*



(a) (b)

Figure 5: MSE ratios for the 36 cases when (a) $\sigma^* = 1$, (b) $\sigma^* = 10$

In this scenario, when $\sigma^* = 0.01$, most the ridge estimators do not outperform OLS. The minimum MSE ratio is 0.997180, which appears in the case $p = 2$, $n = 20$ and $\gamma = 0.9999$.

From Figure 5(a) we can see that for the 24 cases with $\sigma^* = 1$, although k_1, k_2, k_3, k_4, k_5 k_6, k_{11}, k_{12}, k_{15} and k_{26} are worse than OLS only in 2 to 9 cases, their MSE ratios are often larger than 0.9 when $\gamma > 0.999$. The parameters k_9, k_{10}, k_{19}, k_{20}, k_{21}, k_{22}, k_{23}, k_{24} and k_{25} underperform OLS in 25 to 31 out of the 36 cases.

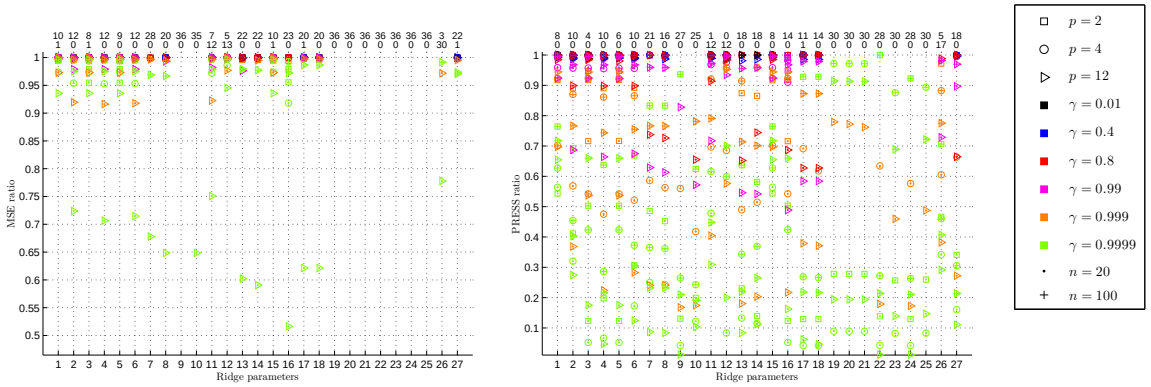
Figure 5(b) shows that when $\sigma^* = 10$, k_{15} and k_{16} are better than OLS in all the cases considered, while k_{19}, k_{20}, k_{21}, k_{22}, k_{23}, and k_{25} are worse than OLS in many cases. When $\gamma > 0.99$, we notice that k_9, k_{19}, k_{20}, k_{21}, k_{22}, k_{23}, k_{24}, k_{25} outperform many other ridge parameters.

The smallest PRESS ratio among all the cases using random number between 0 and 10 as β_j^* is 0.9570, which occurs in the case $p = 4$, $n = 20$, $\gamma = 0.99$ and $\sigma^* = 0.01$. That is, the PRESS ratios of all the ridge parameters are not really smaller than that of OLS in the cases considered here.

5.3 Using random numbers between 0 and 100 as β_j^*

In this scenario, when $\sigma^* = 0.01$, again most the ridge estimators do not outperform OLS. The minimum MSE ratio is 0.999269, which appears in the case $p = 2$, $n = 20$ and $\gamma = 0.9999$. In general, the MSE ratios are getting larger when the possible range of β_j^* is getting wider.

Figure 6(a) shows that when $\sigma^* = 1$, all the ridge parameters are at most as good as OLS in most cases. In spite of the well performance of k_{16} in the case $p = 12$, $n = 20$ and $\gamma = 0.9999$, its MSE ratios in the other models are not much smaller than that of OLS. The parameters k_9, k_{10}, k_{19}, k_{20}, k_{21}, k_{22}, k_{23}, k_{24} and k_{25} underperform OLS in 35 to 36 out of the 36 cases.



(a) (b)

Figure 6: MSE ratios for the 36 cases when (a) $\sigma^* = 1$, (b) $\sigma^* = 10$

Figure 6(b) shows the 36 cases with $\sigma^* = 10$. Even though k_{11} is worse than OLS in 1 case only and k_{12} outperforms OLS in all cases, their MSE ratios are not close to the smallest ones in some cases. When $\gamma \leqslant 0.99$, except the case with $p = 12$, $\gamma = 0.99$ and $n = 20$, the MSE ratios of all the ridge estimators are close to 0.9, i.e. the improvement is mostly minute. The same as in the previous scenario, k_{10}, k_{19}, k_{20}, k_{21}, k_{22}, k_{23}, k_{24} and k_{25} underperform OLS in most of the cases and so are not good choices here.

The smallest PRESS ratio among all the cases using random number between 0 and 100 as β_j^* is 0.9967, which occurs in the case $p = 12$, $n = 20$, $\gamma = 0.01$ and $\sigma^* = 10$. Thus, in terms of the PRESS ratios, again none of the ridge parameters is really better than OLS in these cases.

6 Real data

We consider the Hald (1952, pp. 647) data. There are $n = 13$ observations. The $p = 4$ regressors are percentages of four chemicals in the composition of samples of Portland cement. The dependent variable is the heat evolved in calories per gram of cement. The OLS estimate $\hat{\sigma}^*$ is 2.4460 and the correlations between regressors are given in Table 2. The highest absolute correlation is 0.9730 while the lowest is 0.0295.

Table 3 shows the PRESS ratios of the 27 ridge parameters when the ridge regression was applied to the Hald data. None of these values is greater than 1, and the PRESS ratio of the newly proposed k_{27} is the smallest (0.394653), followed closely by k_9. The parameters k_4, k_6, k_{10}, k_{14}, k_{19}, k_{20}, k_{21}, k_{22}, k_{23} and k_{25} led to PRESS ratios less than 0.45, i.e. they are doing much better than OLS, in this example. The parameters k_7 and k_8, though not worse than, do not really outperform OLS for the Hald data at all.

Because we do not know the 'true' parameter values, we are not able to calculate the MSE ratios. The smallest PRESS ratio given by k_{27} suggests that among all the ridge estimation methods considered, our iterative approach may be the best choice for these data.

Table 2: Correlation matrix of the Hald data

	x_1	x_2	x_3	x_4
x_1	1.0000	0.2286	−0.8241	−0.2454
x_2	0.2286	1.0000	−0.1392	−0.9730
x_3	−0.8241	−0.1392	1.0000	0.0295
x_4	−0.2454	−0.9730	0.0295	1.0000

Table 3: PRESS ratios of the 27 ridge parameters for the Hald data

k_1:	0.727019	k_2:	0.569398	k_3:	0.531504	k_4:	0.449103	k_5:	0.476819
k_6:	0.445442	k_7:	0.999999	k_8:	0.999999	k_9:	0.395702	k_{10}:	0.447298
k_{11}:	0.756867	k_{12}:	0.633927	k_{13}:	0.485180	k_{14}:	0.432585	k_{15}:	0.822830
k_{16}:	0.717988	k_{17}:	0.771695	k_{18}:	0.476819	k_{19}:	0.400976	k_{20}:	0.405132
k_{21}:	0.407829	k_{22}:	0.426245	k_{23}:	0.419211	k_{24}:	0.455442	k_{25}:	0.434654
k_{26}:	0.590518	k_{27}:	0.394653						

7 Conclusion

From the previous sections, we can have some general observations as follows.

1. The performance of the ridge parameters depends on the value of p, n, σ^* and γ. The MSE ratios are usually smaller when p is larger, n is smaller, σ^* is larger, or γ is larger. The PRESS ratios are usually smaller when n is smaller, but no much difference in the PRESS ratios between these 27 parameters can be observed in the simulated cases considered.

2. All these ridge parameters seem not working well when the range of β_j^* is too wide.

3. When β_j^* are just arbitrarily chosen numbers, if the standard deviation σ^* of the error term is

   (a) small, then all ridge parameters are not doing better than OLS;

   (b) immediate, then (i) k_1, k_2, k_3, k_4, k_5 k_6, k_{11}, k_{12}, k_{15} and k_{26} can outperform OLS in many cases but the improvement may not be noteworthy, and (ii) k_9, k_{10}, k_{19}, k_{20}, k_{21}, k_{22}, k_{23}, k_{24} and k_{25} are worse than OLS in many cases;

   (c) large, then (i) k_{15} and k_{16} are doing better than OLS, and (ii) k_{19}, k_{20}, k_{21}, k_{22}, k_{23}, and k_{25} are worse than OLS in many cases.

4. When $\boldsymbol{\beta}^*$ is the normalized eigenvector, if σ^* is

(a) small, then the ridge parameters are often not doing better than OLS unless the correlation parameter γ is very close to 1, under which (i) k_1, k_2, k_3, k_4, k_5, k_6, k_{11}, k_{12}, k_{15} and k_{26} may outperform OLS in many cases but the improvement may not be noteworthy, (ii) k_8, k_{17}, k_{18} and k_{27} are good, and (iii) the rest are not better than OLS in many cases;

(b) immediate, then (i) k_1, k_2, k_3, k_4, k_5, k_6, k_7, k_8, k_{15}, k_{16}, k_{17} and k_{18} outperform OLS in many cases but are usually far from the best choice in each case, (ii) however, the best in one case may be close to the worst in another, (iii) nevertheless, k_{27} is not far from the best in many cases;

(c) large, all the ridge parameters considered are good.

In terms of MSE ratios, among all the ridge estimation methods considered, the newly proposed ridge parameters k_{27} are doing well in many cases in the following sense: When the standard deviation σ^* is very small but the correlation parameter γ is high, k_{27} is a good (and sometimes the best) choice, no matter whether the number of parameters p and the sample size n are large or small; it is not the case for many other ridge parameters. For immediate standard deviation, k_{27} is usually among the best performed group. For large standard deviation, often k_{27} is only slightly worse than the best ridge parameter in each case. Moreover, the largest MSE ratio of k_{27} in all the cases is smaller than 2, while some others can be as high as 100, meaning that even in the worst scenario, k_{27} will not lead to catastrophically wrong estimates but some others may. Finally, k_{27} gives the smallest PRESS ratio when applied to the Hald data.

In conclusion, none of the ridge parameter are uniformly better than the others in all situations. Some can quite consistently make improvement that are unfortunately too little to be practically noteworthy, while some others on one hand can make big improvement in some cases but on the other hand can also make big mistakes in others. The proposed k_{27} succeeds in offering consistently good, though not necessarily the best, improvement in many cases.

Acknowledgements

We thank the two referees for their helpful suggestions. Research supported by a GRF grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. HKBU200710).

Appendix: List of the ridge parameters considered

Denote by $e_i(k)$ the residual of the ith observation in the fitted model with ridge parameter k, $H(k) = [h_{ij}(k)] = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'$, r the rank of $\boldsymbol{X}$, $\lambda_{\max} = \lambda_1$ the largest eigenvalue of $\boldsymbol{X}'\boldsymbol{X}$, and $\hat{\alpha}_{\max}$ the maximum among $\hat{\alpha}_i$.

1. $k = \hat{\sigma}^2/\hat{\alpha}^2_{\max}$ — Hoerl and Kennard (1970b)
2. $k = p\hat{\sigma}^2/(\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\alpha}})$ — Hoerl et al. (1975)
3. $k = \hat{\sigma}^2(\sum \lambda_i^2\hat{\alpha}_i^2)/\sum(\lambda_i\hat{\alpha}_i^2)^2$ — Hocking et al. (1976)
4. $k_{(-1)} = 0$ and for $i \geq 0$, compute iteratively $k_{(i)} = p\hat{\sigma}^2/\{\tilde{\boldsymbol{\alpha}}(k_{(i-1)})'\tilde{\boldsymbol{\alpha}}(k_{(i-1)})\}$ until $(k_{(i)} - k_{(i-1)})/k_{(i-1)} \leqslant \delta$, and finally choose $k = k_{(i)}$, where $\delta = 20 \cdot \mathrm{tr}((\boldsymbol{X}'\boldsymbol{X})^{-1}/p)^{-1.3}$ — Hoerl and Kennard (1976)
5. $k = p\hat{\sigma}^2/(\sum \lambda_i\hat{\alpha}_i^2)$ — Lawless and Wang (1976)
6. k satisfies $\sum \hat{\alpha}_i^2/(\hat{\sigma}^2/k + \hat{\sigma}^2/\lambda_i) = p$ — Dempster et al. (1977)
7. $k = \arg\min_{u\geq 0} \frac{1}{n}\sum e_i(u)^2/\{1 - h_{ii}(u)\}^2$ — Allen (1974)
8. $k = \arg\min_{u\geq 0} n\sum e_i(u)^2/[\sum\{1 - h_{ii}(u)\}]^2$ — Golub et al. (1979)
9. For the jth bootstrap sample of size n, chosen randomly with replacement from the observations, ridge estimates are computed for each member in a pre-selected set Θ of ridge parameter values, $1 \leq j \leq B$. Let $\hat{\boldsymbol{Y}}_j(u)$ be the prediction vector for the unchosen observations $\boldsymbol{Y}_j$ from the ridge estimates with ridge parameter value u. Choose — Delaney and Chatterjee (1986)

$$k = \arg\min_{u\in\Theta} \frac{\sum_{j=1}^{B}(\hat{\boldsymbol{Y}}_j(u) - \boldsymbol{Y}_j)'(\hat{\boldsymbol{Y}}_j(u) - \boldsymbol{Y}_j)}{\sum_{j=1}^{B} \#\ \{\text{elements in } \boldsymbol{Y}_j\}}$$

10. $k = p\hat{\sigma}^2/[\sum\{\hat{\alpha}_i^2/(1 + \sqrt{1 + \lambda_i\hat{\alpha}_i^2/\hat{\sigma}^2})\}]$ — Nomura (1988)
11. $k = (r-2)\hat{\sigma}^2/(\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\alpha}})$ — Brown (1994)
12. $k = (r-2)\hat{\sigma}^2\mathrm{tr}(\boldsymbol{X}'\boldsymbol{X})/(r\hat{y}'\hat{y})$ — Brown (1994)
13. $k = \hat{\sigma}^2/(\prod \hat{\alpha}_i^2)^{\frac{1}{p}}$ — Kibria (2003)
14. $k = \mathrm{median}\,\{\hat{\sigma}^2/\hat{\alpha}_i^2\}$ — Kibria (2003)
15. $k = \lambda_{\max}\hat{\sigma}^2/\{\lambda_{\max}\hat{\alpha}^2_{\max} + (n-p)\hat{\sigma}^2\}$ — Khalaf and Shukur (2005)
16. $k = \max\{\lambda_i\hat{\sigma}^2/[(n-p)\hat{\sigma}^2 + \lambda_i\hat{\alpha}_i^2]\}$ — Alkhamisi et al. (2006)
17. $k = \arg\min_{u\geq 0} \mathrm{ICOMP}(u)$ — Clark and Troskie (2006)

where

$$\text{ICOMP}(u) = -2\log L(\tilde{\boldsymbol{\beta}}(u)) + d\log\left(\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+u)^2}\right) - d\log(d) - \sum_{i=1}^{p}\log\left(\frac{\lambda_i}{(\lambda_i+u)^2}\right)$$

in which $L(\cdot)$ is the likelihood function and

$$d = \text{rank of diag}\left\{\frac{\lambda_1}{(\lambda_1+k)^2},\ldots,\frac{\lambda_p}{(\lambda_p+k)^2}\right\}$$

18. $k = \begin{cases} k_5 & \text{if } k_{17} < k_5 \\ k_{17} & \text{otherwise} \end{cases}$ — Clark and Troskie (2006)
19. $k = \max\{\hat{\sigma}^2/\hat{\alpha}_i^2 + 1/\lambda_i\}$ — Alkhamisi and Shukur (2007)
20. $k = \{\sum(\hat{\sigma}^2/\hat{\alpha}_i^2 + 1/\lambda_i)\}/p$ — Alkhamisi and Shukur (2007)
21. $k = \text{median}\{\hat{\sigma}^2/\hat{\alpha}_i^2 + 1/\lambda_i\}$ — Alkhamisi and Shukur (2007)
22. $k = p\hat{\sigma}^2/(\sum\lambda_i\hat{\alpha}_i^2) + 1/\lambda_{\max}$ — Alkhamisi and Shukur (2007)
23. $k = \left(\prod\sqrt{\hat{\alpha}_i^2/\hat{\sigma}^2}\right)^{\frac{1}{p}}$ — Muniz and Kibria (2009)
24. $k = \left(\prod\sqrt{\hat{\sigma}^2/\hat{\alpha}_i^2}\right)^{\frac{1}{p}}$ — Muniz and Kibria (2009)
25. $k = \text{median}\left\{\sqrt{\hat{\alpha}_i^2/\hat{\sigma}^2}\right\}$ — Muniz and Kibria (2009)
26. $k = \max\{0, p\hat{\sigma}^2/(\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\alpha}}) - 1/(n\text{VIF}_{\max})\}$, where $\text{VIF}_{\max}$ is the maximum among the variance inflation factors of the p regressors — Dorugade and Kashid (2010)

References

Alkhamisi, M., Khalaf, G., and Shukur, G. (2006). Some modifications for choosing ridge parameters. *Communications in Statistics — Theory and Methods* **35,** 2005–2020.

Alkhamisi, M. A. and Shukur, G. (2007). A Monte Carlo study of recent ridge parameters. *Communications in Statistics — Simulation and Computation* **36,** 535–547.

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics* **16,** 125–127.

Brown, P. J. (1994). *Measurement, Regression, and Calibration.* Oxford University Press, New York.

Clark, A. E. and Troskie, C. G. (2006). Ridge regression — a simulation study. *Communications in Statistics — Simulation and Computation* **35,** 605–619.

Delaney, N. J. and Chatterjee, S. (1986). Use of the bootstrap and cross-validation in ridge regression. *Journal of Business & Economic Statistics* **4,** 255–262.

Dempster, A. P., Schatzoff, M., and Wermuth, N. (1977). A simulation study of alternatives to ordinary least squares. *Journal of the American Statistical Association* **72,** 77–91.

Dorugade, A. V. and Kashid, D. N. (2010). Alternative method for choosing ridge parameter for regression. *Applied Mathematical Sciences* **4,** 447–456.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression. Models, Methods and Applications*. Springer-Verlag, Berlin.

Farebrother, R. W. (1976). Further results on the mean square error of ridge regression. *Journal of the Royal Statistical Society Series B* **38,** 248–250.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21,** 215–223.

Groß, J. (2003). *Linear Regression*. Lecture Notes in Statistics **175**. Springer-Verlag, Berlin.

Hald, A. (1952). *Statistical Theory with Engineering Applications*. John Wiley & Sons, New York.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer-Verlag, New York, 2nd edition.

Hocking, R. R., Speed, F. M., and Lynn, M. J. (1976). A class of biased estimators in linear regression. *Technometrics* **18,** 425–437.

Hoerl, A. E., Kannard, R. W., and Baldwin, K. F. (1975). Ridge regression: some simulations. *Communications in Statistics — Theory and Methods* **4,** 105–123.

Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: Applications to nonorthogonal problems. *Technometrics* **12,** 69–82.

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

Hoerl, A. E. and Kennard, R. W. (1976). Ridge regression iterative estimation of the biasing parameter. *Communications in Statistics — Theory and Methods* **5,** 77–88.

Khalaf, G. and Shukur, G. (2005). Choosing ridge parameter for regression problems. *Communications in Statistics — Theory and Methods* **34,** 1177–1182.

Kibria, B. M. G. (2003). Performance of some new ridge regression estimators. *Communications in Statistics — Simulation and Computation* **32,** 419–435.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). *Applied Linear Statistical Models*. McGraw-Hill/Irwin, Boston, 5th edition.

Lawless, J. F. and Wang, P. (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics — Theory and Methods* **5,** 307–323.

McDonald, G. C. (2009). Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics* **1,** 93–100.

McDonald, G. C. and Galarneau, D. I. (1975). A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association* **70,** 407–416.

Muniz, G. and Kibria, B. M. G. (2009). On some ridge regression estimators: An empirical comparisons. *Communications in Statistics — Simulation and Computation* **38,** 621–630.

Newhouse, J. P. and Oman, S. D. (1971). An evaluation of ridge estimators. Technical Report R-716-PR, The RAND Corporation.

Nomura, M. (1988). On the almost unbiased ridge regression estimator. *Communications in Statistics — Simulation and Computation* **17,** 729–743.

Theobald, C. M. (1974). Generalizations of mean square error applied to ridge regression. *Journal of the Royal Statistical Society Series B* **36,** 103–106.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B* **58,** 267–288.