Advances and Applications in Statistical Sciences
Volume 1, Issue 2, 2010, Pages 381-398

TRACKING MSE EFFICIENCIES IN RIDGE REGRESSION

D. R. JENSEN and D. E. RAMIREZ

Department of Mathematics
University of Virginia
Charlottesville, VA 22904-4137, U.S.A.
E-mails: djensen@vt.edu
der@virginia.edu

Abstract

Ridge regression is often favored in the analysis of ill-conditioned systems. A canonical form identifies regions in the parameter space where Ordinary Least Squares (*OLS*) is problematic. The objectives are two-fold: To reexamine the view that ill-conditioning necessarily degrades essentials of *OLS*; and to reassess ranges of the ridge parameter k where ridge is efficient in mean squared error (*MSE*) relative to *OLS*; and conversely. In particular, ridge is intended to ameliorate effects of ill-conditioning over a wide range of k. Contrary to conventional wisdom, ridge often must be abandoned in favor of *OLS* for k sufficiently large.

1. Introduction

In a full-rank model $\{\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\}$, the p equations $\{\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}\}$ yield *OLS* solutions $\hat{\boldsymbol{\beta}}_L$, unbiased with minimal dispersion $V(\hat{\boldsymbol{\beta}}_L) = \sigma^2\boldsymbol{V} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ under conventional errors. Ill-conditioning, as near-dependency among columns of $\boldsymbol{X}$, "causes crucial elements of $\boldsymbol{X}'\boldsymbol{X}$ to be large and unstable", and $\hat{\boldsymbol{\beta}}_L$ to be "very sensitive to small changes in $\boldsymbol{X}$" with "inflated variances" [2, p. 119]. Specifically, the *Variance Inflation Factors* (*VIF*s) of $\hat{\boldsymbol{\beta}}_L = [\hat{\beta}_{L1}, \ldots, \hat{\beta}_{Lp}]'$ are ratios of actual to "ideal" variances had the columns of $\boldsymbol{X}$ been orthogonal, i.e., $\{VIF(\hat{\beta}_{Lj}) = v_{jj} / w_{jj}^{-1}; 1 \le j \le p\}$ with $\boldsymbol{W} = \boldsymbol{X}'\boldsymbol{X}$. Since $\{VIF(\hat{\beta}_{Lj}) \ge 1.0;\ 1 \le j$

2000 Mathematics Subject Classification: Primary 62J07; Secondary 62J20.

Keywords: Ill-conditioned systems, least squares, ridge regression, *MSE* efficiencies of *OLS* and ridge.

Received June 13, 2009

$\leq p\}$, and often much greater, ill-conditioning is widely held to degrade essentials of *OLS*: In fact, $V_1 = \max\{VIF(\hat{\beta}_{Lj});\ 1 \leq i \leq p\}$ is identified in [10] as "the best single measure of the conditioning of the data". Standard remedies include *ridge estimators* as solutions of $\{(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)\,\hat{\boldsymbol{\beta}}_{Rk} = \boldsymbol{X}'\boldsymbol{Y};\ k \geq 0\}$; see [4-6] and numerous subsequent citations. A basic canonical form is $\{\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{D}_\xi\boldsymbol{\theta} + \boldsymbol{\epsilon}\}$, where $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}_p$; $\boldsymbol{Q}$ is orthogonal; $\boldsymbol{D}_\xi = \text{Diag}(\xi_1, \ldots, \xi_p)$, ordered as $\{\xi_1 \geq \xi_2 \geq \ldots \geq \xi_p\}$; and $\boldsymbol{\theta} = \boldsymbol{Q}'\boldsymbol{\beta}$. The *OLS* solutions $\hat{\boldsymbol{\theta}}_L = [\hat{\theta}_{L1}, \ldots, \hat{\theta}_{Lp}]'$, with variances $\{\sigma^2/\xi_1^2 \leq \sigma^2/\xi_2^2 \leq \ldots \leq \sigma^2/\xi_p^2\}$, thus identify "… linear combinations of parameters about which the data at hand are most/least informative" [13]. Moreover, these variances escalate for values of ξ_i^2 near zero, a characteristic of ill-conditioning. To validate ridge as a rational alternative to *OLS* in mean squared error (*MSE*); Hoerl and Kennard [6] gave $k < \sigma^2/\theta_M^2$ as sufficient for ridge to dominate *OLS* with lesser *MSE*, where $\theta_M^2 = \max\{\theta_1^2, \ldots, \theta_p^2\}$.

Our purposes here are two-fold: (i) To reexamine the extent to which *OLS* solutions are degraded through ill-conditioning, and (ii) to reassess ranges of the ridge parameter k where ridge is *MSE*-efficient with respect to *OLS*, and conversely. An outline follows.

Conventions for notation and supporting materials are in Section 2. Section 3 sets forth the principal findings, to include (i) a decomposition of the parameter space according to comparative subspace efficiencies of *OLS*; (ii) a reexamination of conditions sufficient for ridge to dominate *OLS* in *MSE*, and conversely, as k evolves; and (iii) elementary but illuminating examples. Section 4 revisits a highly ill-conditioned case study treated elsewhere and Section 5 reprises essentials of the present study.

2. Preliminaries

2.1. Notation

Designate by $\mathbb{R}^p$ the Euclidean p-space, and by $\mathbb{R}_+^p$ its positive orthant; matrices and vectors are set in bold type; the transpose, inverse and trace of $\boldsymbol{A}$ are $\boldsymbol{A}'$, $\boldsymbol{A}^{-1}$ and $\text{tr}(\boldsymbol{A})$; special arrays are the identity $\boldsymbol{I}_p$, the unit vector

$\mathbf{1}_p = [1, 1, \ldots, 1]' \in \mathbb{R}^p$, and the diagonal matrix $\boldsymbol{D_a} = \boldsymbol{D}(a_i) = \mathrm{Diag}(a_1, \ldots, a_p)$. Let $\boldsymbol{X}$, of order $(n \times p)$, have rank $p < n$; its *singular decomposition* is $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{D}_\xi\boldsymbol{Q}'$, where $\boldsymbol{D}_\xi = \mathrm{Diag}(\xi_1, \ldots, \xi_p)$ comprise its ordered *singular values* $\{\xi_1 \geq \xi_2 \geq \ldots \geq \xi_p > 0\}$; and the columns of $\boldsymbol{P} = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_p]$ and of $\boldsymbol{Q} = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_p]$ are the *left-* and *right-singular vectors* of $\boldsymbol{X}$, such that $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}_p$ and $\boldsymbol{Q}$ is orthogonal. By $Sp(\boldsymbol{Z})$ is meant the linear column span of $\boldsymbol{Z}(p \times k) = [z_1, \ldots, z_k]$ in $\mathbb{R}^p$.

2.2. Stochastic concepts

The expectation, dispersion, and law of distribution of $\boldsymbol{Z} \in \mathbb{R}^p$ are $\mathrm{E}(\boldsymbol{Z})$, $V(\boldsymbol{Z})$ and $\mathcal{L}(\boldsymbol{Z})$. Given $\hat{\boldsymbol{\beta}}$ for estimating $\boldsymbol{\beta} \in \mathbb{R}^p$ having $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}_0$ and second moments, its *MSE* is $M_{SE}(\hat{\boldsymbol{\beta}}) = \mathrm{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = \sum_{i=1}^{p} \mathrm{Var}(\hat{\beta}_i) + (\boldsymbol{\beta}_0 - \boldsymbol{\beta})'(\boldsymbol{\beta}_0 - \boldsymbol{\beta})$, to quantify the trade-off of bias for variance. Since $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \boldsymbol{Q}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ and $\boldsymbol{Q}$ is orthogonal in canonical form, it suffices that $M_{SE}(\hat{\boldsymbol{\beta}}) = M_{SE}(\hat{\boldsymbol{\theta}}) = \mathrm{E}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$. Moreover, if $M_{SE}(\hat{\boldsymbol{\theta}}) < M_{SE}(\tilde{\boldsymbol{\theta}})$, then $\hat{\boldsymbol{\theta}}$ is said to be M_{SE}*-admissible* with respect to $\tilde{\boldsymbol{\theta}}$. Specifically, in regard to estimators $\hat{\alpha}$ and $\tilde{\alpha}$ for $\alpha \in \mathbb{R}^1$ having second moments, the Fisher efficiency of $\hat{\alpha}$ relative to $\tilde{\alpha}$ is defined as $E_F(\hat{\alpha} : \tilde{\alpha}) = \mathrm{Var}(\tilde{\alpha}) / \mathrm{Var}(\hat{\alpha})$.

3. The Principal Findings

3.1. Efficiencies of *OLS*

Despite that $VIF(\hat{\beta}_{Lj}) \geq 1.0$, we reexamine the ostensible degradation of *OLS* in ill-conditioned models having second moments. Take $\boldsymbol{X}'\boldsymbol{X}$ to be centered and scaled as in [6, 9, 10], and elsewhere, such that $\mathrm{tr}(\boldsymbol{X}'\boldsymbol{X}) = p$ but $\boldsymbol{X}'\boldsymbol{X} \neq \boldsymbol{I}_p$. Write $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{D}_\xi^2\boldsymbol{Q}'$; then there are $r \geq 1$ eigenvalues greater than unity; $s \geq 0$ equal to unity; and $t = p - r - s \geq 1$ eigenvalues less than unity. Designate these as $\{\xi_1^2 \geq \ldots \geq \xi_r^2 > 1 = \xi_{r+1}^2 = \ldots = \xi_{r+s}^2 > \xi_{r+s+1}^2 \geq \ldots \geq \xi_p^2 > 0\}$. Conformably partition $\boldsymbol{Q} = [\boldsymbol{Q}_1, \boldsymbol{Q}_2, \boldsymbol{Q}_3]$ and $\boldsymbol{\theta}' = [\boldsymbol{\theta}_1', \boldsymbol{\theta}_2', \boldsymbol{\theta}_3']$,

such that $\boldsymbol{Q}_1 = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_r]$, $\boldsymbol{Q}_2 = [\boldsymbol{q}_{r+1}, \ldots, \boldsymbol{q}_{r+s}]$, and $\boldsymbol{Q}_3 = [\boldsymbol{q}_{r+s+1}, \ldots, \boldsymbol{q}_p]$; $\boldsymbol{\theta}_1' = [\theta_1, \ldots, \theta_r]$, $\boldsymbol{\theta}_2' = [\theta_{r+1}, \ldots, \theta_{r+s}]$, and $\boldsymbol{\theta}_3' = [\theta_{r+s+1}, \ldots, \theta_p]$, and similarly $\hat{\boldsymbol{\theta}}' = [\hat{\boldsymbol{\theta}}_1', \hat{\boldsymbol{\theta}}_2', \hat{\boldsymbol{\theta}}_3']$; and observe from $\boldsymbol{\theta} = \boldsymbol{Q}'\boldsymbol{\beta}$ that $\boldsymbol{\beta} = \boldsymbol{Q}\boldsymbol{\theta} = \boldsymbol{Q}_1\boldsymbol{\theta}_1 + \boldsymbol{Q}_2\boldsymbol{\theta}_2 + \boldsymbol{Q}_3\boldsymbol{\theta}_3$. Identify $L_1 = Sp(\boldsymbol{Q}_1)$, $L_2 = Sp(\boldsymbol{Q}_2)$, and $L_3 = S_p(\boldsymbol{Q}_3)$, as orthogonal complements in $\mathbb{R}^p$ of dimensions $\{r, s, t\}$.

We next identify $\{L_1, L_2, L_3\}$ as subspaces of parameters in $\mathbb{R}^p$ where *OLS* solutions have greater, equal and lesser Fisher efficiencies, in comparison with an orthogonal reference model having $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_p$, $\{\xi_1^2 = \ldots = \xi_p^2 = 1.0\}$, and solution $\tilde{\boldsymbol{\beta}}_0$.

Theorem 1. *Consider OLS estimators* $\{\boldsymbol{c}'\hat{\boldsymbol{\beta}}_L;\ \boldsymbol{c} \in \mathbb{R}^p\}$ *for* $\{\boldsymbol{c}'\boldsymbol{\beta};\ \boldsymbol{c} \in \mathbb{R}^p\}$, *and similarly* $\{\boldsymbol{c}'\tilde{\boldsymbol{\beta}}_0;\ \boldsymbol{c} \in \mathbb{R}^p\}$ *under a fully orthogonal model with* $\{\xi_1^2 = \ldots = \xi_p^2 = 1.0\}$ *and solution* $\tilde{\boldsymbol{\beta}}_0$.

Then

(i) $\dfrac{\sigma^2}{\xi_1^2} \le \dfrac{\mathrm{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}_L)}{\boldsymbol{c}'\boldsymbol{c}} \le \dfrac{\sigma^2}{\xi_r^2} < 1.0\sigma^2$ *and* $E_F(\boldsymbol{c}'\hat{\boldsymbol{\beta}}_L : \boldsymbol{c}'\tilde{\boldsymbol{\beta}}_0) > 1.0$ *for* $\boldsymbol{c} \in L_1 = Sp(\boldsymbol{Q}_1)$;

(ii) $\dfrac{\mathrm{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}_L)}{\boldsymbol{c}'\boldsymbol{c}} = 1.0\sigma^2$ *and* $E_F(\boldsymbol{c}'\hat{\boldsymbol{\beta}}_L : \boldsymbol{c}'\tilde{\boldsymbol{\beta}}_0) = 1.0$ *for* $\boldsymbol{c} \in L_2 = Sp(\boldsymbol{Q}_2)$; *and*

(iii) $1.0\sigma^2 < \dfrac{\sigma^2}{\xi_{r+s+1}^2} \le \dfrac{\mathrm{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}_L)}{\boldsymbol{c}'\boldsymbol{c}} \le \dfrac{\sigma^2}{\xi_p^2}$ *and* $E_F(\boldsymbol{c}'\hat{\boldsymbol{\beta}}_L : \boldsymbol{c}'\tilde{\boldsymbol{\beta}}_0) < 1.0$ *for* $\boldsymbol{c} \in L_3 = Sp(\boldsymbol{Q}_3)$.

Proof. In canonical form write $V(\hat{\boldsymbol{\theta}}) = \sigma^2\mathrm{Diag}\left(\dfrac{1}{\xi_1^2}, \ldots, \dfrac{1}{\xi_p^2}\right) = \sigma^2\Xi$, so that $\mathrm{V}(\hat{\boldsymbol{\theta}}_1', \hat{\boldsymbol{\theta}}_2', \hat{\boldsymbol{\theta}}_3') = \sigma^2\mathrm{Diag}(\Xi_1, \Xi_2, \Xi_3)$, say, and $\mathrm{V}(\hat{\boldsymbol{\theta}}_1) = \sigma^2\Xi_1 = \sigma^2\mathrm{Diag}\left(\dfrac{1}{\xi_1^2}, \ldots, \dfrac{1}{\xi_r^2}\right)$. Variational properties of Rayleigh quotients assure that $\left\{\dfrac{1}{\xi_1^2}\right.$

$\leq \dfrac{\boldsymbol{c}_1'\Xi_1\boldsymbol{c}_1}{\boldsymbol{c}_1'\boldsymbol{c}_1} \leq \dfrac{1}{\xi_r^2};\ \boldsymbol{c}_1 \in \mathbb{R}^r \Big\}$. That these transfer to $\hat{\boldsymbol{\beta}}$ is seen from $\boldsymbol{c}'\hat{\boldsymbol{\beta}} = \boldsymbol{c}'(\boldsymbol{Q}_1\hat{\boldsymbol{\theta}}_1 + \boldsymbol{Q}_2\hat{\boldsymbol{\theta}}_2 + \boldsymbol{Q}_3\hat{\boldsymbol{\theta}}_3) = \boldsymbol{c}'\boldsymbol{Q}_1\hat{\boldsymbol{\theta}}_1 = \boldsymbol{c}_1\hat{\boldsymbol{\theta}}_1$ since $\boldsymbol{c} \in Sp(\boldsymbol{Q}_1)$, so that $\boldsymbol{c}'\boldsymbol{Q}_2 = \boldsymbol{0} = \boldsymbol{c}'\boldsymbol{Q}_3$. Moreover, $E_F(\boldsymbol{c}'\hat{\boldsymbol{\beta}}_L : \boldsymbol{c}'\widetilde{\boldsymbol{\beta}}_0) = \mathrm{Var}(\boldsymbol{c}'\widetilde{\boldsymbol{\beta}}_0)/\mathrm{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}_L) > 1.0$, since $[\mathrm{Var}(\boldsymbol{c}'\widetilde{\boldsymbol{\beta}}_0)/\boldsymbol{c}'\boldsymbol{c}] = 1.0\sigma^2$, to give conclusion (i). Conclusions (ii) and (iii) follow similarly, taking in turn $\Xi_2 = \boldsymbol{I}_s$ and $\Xi_3 = \mathrm{Diag}\left(\dfrac{1}{\xi_{r+s+1}^2}, \ldots, \dfrac{1}{\xi_p^2}\right)$. □

In short, $\boldsymbol{c}'\hat{\boldsymbol{\beta}}_L$ exhibits (i) enhanced efficiency for $\boldsymbol{c} \in Sp(\boldsymbol{Q}_1)$; (ii) equal efficiency for $\boldsymbol{c} \in Sp(\boldsymbol{Q}_2)$; and (iii) degraded efficiency for $\boldsymbol{c} \in Sp(\boldsymbol{Q}_3)$; all relative to an orthogonal model preserving $\mathrm{tr}(\boldsymbol{X}'\boldsymbol{X}) = p$.

Example 1. To illustrate, suppose that $p = 3$; $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3]'$; and $\boldsymbol{X}'\boldsymbol{X}$, its inverse $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, and the matrix $\boldsymbol{Q}$ are given in succession by

$$\begin{bmatrix} 2.97 & 0 & 0 \\ 0 & 0.015 & 0.005 \\ 0 & 0.005 & 0.015 \end{bmatrix}, \begin{bmatrix} \dfrac{1}{2.97} & 0 & 0 \\ 0 & 75.00 & -25.00 \\ 0 & -25.00 & 75.00 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & \dfrac{1}{\sqrt{2}} & \dfrac{1}{\sqrt{2}} \\ 0 & \dfrac{1}{\sqrt{2}} & -\dfrac{1}{\sqrt{2}} \end{bmatrix}. \quad (3.1)$$

Verify directly that $VIF(\hat{\beta}_1) = 1.0$, and $VIF(\hat{\beta}_2) = 1.125 = VIF(\hat{\beta}_3)$. The spectral decomposition gives $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{D}_\xi^2\boldsymbol{Q}'$; $\boldsymbol{D}_\xi^2 = \mathrm{Diag}(2.97, 0.02, 0.01)$; $\boldsymbol{Q}$ is displayed; the canonical parameters $\boldsymbol{\theta} = \boldsymbol{Q}'\boldsymbol{\beta}$ are $\theta_1 = \beta_1$, $\theta_2 = (\beta_2 + \beta_3)/\sqrt{2}$, and $\theta_3 = (\beta_2 - \beta_3)/\sqrt{2}$; and $\mathrm{V}(\hat{\boldsymbol{\theta}}_L) = \sigma^2\mathrm{Diag}(0.3367, 50, 100)$. Clearly $\mathrm{Var}(\hat{\theta}_1) = 0.3367\sigma^2$ is diminished by the factor $1/2.97$ from the orthogonal case, but at the expense of grossly inflated variances for $\hat{\theta}_2$ and $\hat{\theta}_3$. Conventional diagnostics fail to convey the reduction in variance for $\hat{\theta}_1$, nor do the relatively benign *VIF*s of $\{1.0, 1.125, 1.125\}$ anticipate the critically inflated variances $\mathrm{Var}(\hat{\theta}_2) = 50\sigma^2$ and $\mathrm{Var}(\hat{\theta}_3) = 100\sigma^2$. On the other hand, the condition number $c_1(\boldsymbol{X}'\boldsymbol{X}) = \xi_1^2/\xi_3^2 = 297.0$ is the ratio of variances of the least to the most precisely estimated linear functions.

To continue, partition $\boldsymbol{Q}$ as $[\boldsymbol{q}_1, \boldsymbol{Q}_3]$ with $\boldsymbol{Q}_3 = [\boldsymbol{q}_2, \boldsymbol{q}_3]$ from (3.1), where $s = 0$ and $L_2 = Sp(\boldsymbol{Q}_2)$ is empty. Clearly $Sp(\boldsymbol{Q}_3)$ identifies the subspace where *OLS* is problematic. Theorem 1(iii) now gives $\left\{\dfrac{\sigma^2}{0.02} = 50\sigma^2 \le \dfrac{\mathrm{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}})}{\boldsymbol{c}'\boldsymbol{c}} \le \dfrac{\sigma^2}{0.01} = 100\sigma^2;\ \boldsymbol{c} \in Sp(\boldsymbol{Q}_3)\right\}$. In particular, at $\boldsymbol{x}_0 = [0, 1, 2]' \in Sp(\boldsymbol{Q}_3)$, the predicted response has variance $\mathrm{Var}(\hat{\beta}_2 + 2\hat{\beta}_3) = \sigma^2\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0 = 275\sigma^2$, so that $\mathrm{Var}(\boldsymbol{x}_0'\hat{\boldsymbol{\beta}})/\boldsymbol{x}_0'\boldsymbol{x}_0 = 275\sigma^2/5 = 55\sigma^2 \in \sigma^2[50, 100]$, the interval asserted in conclusion (iii) for every $\boldsymbol{x}_0 \in Sp(\boldsymbol{Q}_3)$.

We next revisit the comparative *MSE*-efficiencies of *OLS* and ridge solutions. In order that ridge might better *OLS* as gauged by *MSE*; i.e., that $M_{SE}(\hat{\boldsymbol{\beta}}_R(k)) < M_{SE}(\hat{\boldsymbol{\beta}}_L)$, it is sufficient that $k < \sigma^2/\theta_M^2$, with $\theta_M^2 = \max\{\theta_1^2, \theta_2^2, \ldots, \theta_p^2\}$, as noted earlier from [6]. In what follows we sharpen this condition. Of equally compelling interest, heretofore neglected, is to seek values for k where the reversal $M_{SE}(\hat{\boldsymbol{\beta}}_R(k)) > M_{SE}(\hat{\boldsymbol{\beta}}_L)$ might obtain. The latter goes substantially beyond the work of [6] and numerous subsequent studies, in characterizing larger values for k where ridge necessarily is usurped by *OLS*. In what follows write $\hat{\boldsymbol{\theta}}_{Rk} = [\hat{\boldsymbol{\theta}}_{Rk}^1, \hat{\boldsymbol{\theta}}_{Rk}^2, \ldots, \hat{\boldsymbol{\theta}}_{Rk}^p]'$, with typical element $\hat{\boldsymbol{\theta}}_{Rk}^i$.

3.2. MSE comparisons

3.2.1. Basics

Let $M_{SE}(\hat{\boldsymbol{\beta}}_{Rk}) = \gamma_1(k) + \gamma_2(k)$, where $\gamma_1(k) = \mathrm{tr}[\mathrm{V}(\hat{\boldsymbol{\beta}}_{Rk})]$ and $\gamma_2(k) = (\boldsymbol{\beta}_k^0 - \boldsymbol{\beta})'(\boldsymbol{\beta}_k^0 - \boldsymbol{\beta})$, with $\boldsymbol{\beta}_k^0 = \mathrm{E}(\hat{\boldsymbol{\beta}}_{Rk})$. Then $\gamma_1(k)$ is the total variance and $\gamma_2(k)$ is the sum of squared biases. Since $\boldsymbol{\theta} = \boldsymbol{Q}'\boldsymbol{\beta}$ and $\boldsymbol{Q}$ is orthogonal, it suffices to take

$$\gamma_1(k) = \sum_{i=1}^{p} \mathrm{Var}(\hat{\boldsymbol{\theta}}_{Rk}^i) \text{ and } \gamma_2(k) = \sum_{i=1}^{p} B^2(\hat{\boldsymbol{\theta}}_{Rk}^i) \tag{3.2}$$

with $B^2(\hat{\theta}^i_{Rk}) = [\mathrm{E}(\hat{\theta}^i_{Rk}) - \theta_i]^2$. It is shown in [6] that $\gamma_1(k)$ and $\gamma_2(k)$ are continuous monotone functions, decreasing and increasing, respectively. Their sum, $M_{SE}(\hat{\boldsymbol{\theta}}_{Rk})$, decreases to a minimum and increases thereafter, as shown graphically in Figure 1 of [6, p. 61]. So, too, for the component functions and their partial sum, namely, $\{\mathrm{Var}(\hat{\theta}^i_{Rk}), B^2(\hat{\theta}^i_{Rk}), M_{SE}(\hat{\theta}^i_{Rk}); 1 \le i \le p\}$. Specifically, the condition $k < \sigma^2/\theta^2_M$ of [6] is chosen so that each term in the expansion of $dM_{SE}(\hat{\boldsymbol{\theta}}_{Rk})/dk$, and thus their sum, is negative. Looking ahead, let k^i_c be the crossing value such that $M_{SE}(\hat{\theta}^i_{Rk})$ is equal to $M_{SE}(\hat{\boldsymbol{\theta}}_{Li}) = \sigma^2/\xi^2_i$ not depending on k, where $M_{SE}(\hat{\theta}^i_{Rk})$ clearly crosses from below. Then in order that $M_{SE}(\hat{\theta}^i_{Rk}) < \mathrm{M}_{SE}(\hat{\boldsymbol{\theta}}_{Li})$, it is necessary and sufficient that $k < k^i_c$; whereas $M_{SE}(\hat{\theta}^i_{Rk}) > M_{SE}(\hat{\boldsymbol{\theta}}_{Li})$, holds if and only if $k > k^i_c$. Details follow.

Theorem 2. *Consider OLS solutions* $\hat{\boldsymbol{\theta}}_L = [\hat{\theta}_{L1}, \ldots, \hat{\theta}_{Lp}]'$, *together with canonical ridge estimators* $\hat{\boldsymbol{\theta}}_{Rk} = [\hat{\theta}^1_{Rk}, \hat{\theta}^2_{Rk}, \ldots, \hat{\theta}^p_{Rk}]'$, *solving* $\{(\boldsymbol{D}^2_\xi + k\boldsymbol{I}_p)\hat{\boldsymbol{\theta}}_{Rk} = \boldsymbol{D}_\xi \boldsymbol{P}'\boldsymbol{Y}\}$.

(i) *The typical element* $\hat{\theta}^i_{Rk}$ *has* $M_{SE}(\hat{\theta}^i_{Rk}) = \dfrac{\sigma^2\xi^2_i + k^2\theta^2_i}{(\xi^2_i + k)^2}$.

(ii) $M_{SE}(\hat{\theta}^i_{Rk})$ *achieves its minimal value* $\sigma^2\theta^2_i/(\sigma^2 + \xi^2_i\theta^2_i)$ *at* $k_i = \sigma^2/\theta^2_i$.

(iii) *The cross-over value from below, where equality is attained in* $M_{SE}(\hat{\theta}^i_{Rk}) \le M_{SE}(\hat{\boldsymbol{\theta}}_{Li})$, *occurs at*

$$k^i_c = \frac{2\sigma^2\xi^2_i}{\xi^2_i\theta^2_i - \sigma^2} = \frac{2\sigma^2}{\theta^2_i - \dfrac{\sigma^2}{\xi^2_i}} \tag{3.3}$$

for each $\{i = 1, 2, \ldots, p\}$.

(iv) *If* $k_0 \le \min\{k^1_c, k^2_c, \ldots, k^p_c\}$, *then* $\hat{\boldsymbol{\theta}}_R(k_0)$ *is* M_{SE}*-admissible with respect to* $\hat{\boldsymbol{\theta}}_L$.

(v) *If* $k^{\dagger} \geq \max\{k_c^1, k_c^2, \ldots, k_c^p\}$, *then* $\hat{\boldsymbol{\theta}}_L$ *is* M_{SE}*-admissible with respect to* $\hat{\boldsymbol{\theta}}_R(k^{\dagger})$.

Proof. It follows directly that $\mathrm{E}(\hat{\theta}^i_{Rk}) = [\xi_i^2 / (\xi_i^2 + k)]\theta_i$, with bias $B_i = -k\theta_i/(\xi_i^2 + k)$, and $\mathrm{Var}(\hat{\theta}^i_{Rk}) = \sigma^2\xi_i^2/(\xi_i^2 + k)^2$, giving $M_{SE}(\hat{\theta}^i_{Rk}) = \mathrm{Var}(\hat{\theta}^i_{Rk}) + B^2(\hat{\theta}^i_{Rk})$ from (3.2) as in conclusion (i). Denoting $M_{SE}(\hat{\theta}^i_{Rk})$ as $M_{SEi}(k)$, we have

$$\frac{dM_{SEi}(k)}{dk} = \frac{2k(\xi_i^2 + k)^2\,\theta_i^2 - 2(\sigma^2\xi_i^2 + k^2\theta_i^2)(\xi_i^2 + k)}{(\xi_i^2 + k)^4}. \tag{3.4}$$

Setting the numerator to zero and solving gives $\xi_i^2(k\theta_i^2 - \sigma^2) = 0$, i.e., $k_i = \sigma^2/\theta_i^2$. The second derivative is positive; the solution is minimizing; the minimal value is by substitution; and thus follows conclusion (ii). Equating $M_{SE}(\hat{\theta}^i_{Rk})$ to $M_{SE}(\hat{\theta}_{Li}) = \sigma^2/\xi_i^2$ and solving, gives k_c^i as in conclusion (iii). Since components of $M_{SE}(\hat{\boldsymbol{\theta}}_{Rk}) - M_{SE}(\hat{\boldsymbol{\theta}}_L) = \sum_{i=1}^{p}[M_{SE}(\hat{\theta}^i_{Rk}) - M_{SE}(\hat{\theta}_{Li})]$ are all negative for $k_0 \leq \min\{k_c^1, k_c^2, \ldots, k_c^p\}$, from (iii), conclusion (iv) follows directly. Similarly, that all components are positive for $k^{\dagger} \geq \max\{k_c^1, k_c^2, \ldots, k_c^p\}$, gives conclusion (v), to complete our proof. □

Remark 1. Observe that $k_c^i = 2\sigma^2/(\theta_i^2 - \sigma^2/\xi_i^2)$ at (3.3) is finite and positive, if and only if $\theta_i^2 > \sigma^2/\xi_i^2 = \mathrm{Var}(\hat{\theta}_{Li})$. Otherwise, $\hat{\theta}^i_{Rk}$ is M_{SE}-admissible with respect to $\hat{\theta}_{Li}$ for all values of the ridge parameter k.

Remark 2. Clearly k_c^i is scale-invariant, since $k_c^i = 2\sigma^2/(\theta_i^2 - \sigma^2/\xi_i^2)$ $= 2/(\tau_i^2 - 1/\xi_i^2)$ under rescaling $\{\theta_i = \sigma\tau_i\}$. Accordingly, it often suffices to take $\sigma^2 = 1.0$.

Remark 3. With ξ_i^2 fixed in (3.3), k_c^i decreases as $\tau_i^2 = \theta_i^2/\sigma^2$ increases; for k_c^i fixed at k_c, the parameters θ_i^2 and ξ_i^2 are inversely related.

Remark 4. The condition $k < \sigma^2/\theta_M^2$ of [6] exhibits excessive conservatism. From Theorem 2(ii) this value is to the left of all the minima for the $M_{SE}(\hat{\theta}_{Rk}^i)$ curves, whereas conclusion (iii) adjoins to these the rising branch on the $M_{SE}(\hat{\theta}_{Rk}^i)$ curves up to their crossing with the constant *OLS* line. Further insight regarding conclusion (ii) is seen in the initial rates of change, namely $[dM_{SEi}(k)/dk]|_{k=0} = -\sigma^2/\xi_i^4 < 0$.

Conclusions (iv) and (v) specialize further as follows, where crossings now are taken to coincide at the common value $\{k_c = k_c^1 = k_c^2 = \cdots = k_c^p\}$.

Corollary 1. *If crossings all coincide at* k_c, *then for* $M_{SE}(\hat{\boldsymbol{\theta}}_{Rk}) < M_{SE}(\hat{\boldsymbol{\theta}}_L)$, *it is necessary and sufficient that* $k < k_c$. *Similarly,* $M_{SE}(\hat{\boldsymbol{\theta}}_{Rk}) > M_{SE}(\hat{\boldsymbol{\theta}}_L)$, *if and only if* $k > k_c$.

Example 2. To illustrate, suppose that $p = 3$; $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3]'$; and $\boldsymbol{X'X}$ and its inverse are

$$\boldsymbol{X'X} = \begin{bmatrix} 2.900 & 0 & 0 \\ 0 & 0.050 & 0.005 \\ 0 & 0.005 & 0.050 \end{bmatrix}, (\boldsymbol{X'X})^{-1} = \begin{bmatrix} 0.344828 & 0 & 0 \\ 0 & 20.2020 & -2.0202 \\ 0 & -2.0202 & 20.2020 \end{bmatrix} \quad (3.5)$$

and $\boldsymbol{Q}$ carries over from (3.1). Further computations give $\boldsymbol{D}_\xi^2 = \text{Diag}(2.900, 0.055, 0.045)$, $\text{V}(\hat{\boldsymbol{\theta}}_L) = \sigma^2\text{Diag}(0.3448, 18.1818, 22.2222)$, and the condition number $c_1(\boldsymbol{X'X}) = 2.900/0.045 = 64.4444$. Essential features are demonstrated on taking $\sigma^2 = 1.0$, in which case θ_i and τ_i are interchangeable as in Remark 2.

Table 1. Values $[\theta_1^2, \theta_2^2, \theta_3^2] = [2.8448, 20.6818, 24.7222]$ with common crossover $k_c = 0.8$; $\{[\mathrm{Var}(\hat{\theta}_{Rk}^i), B^2(\hat{\theta}_{Rk}^i)]; 1 \le i \le 3\}$ as variances and squared biases standardized to $\sigma^2 = 1.0$; and $M_{SE}(\hat{\boldsymbol{\theta}}_{Rk})$, for various k.

k	$\mathrm{Var}(\hat{\theta}_{Rk}^1)$	$B^2(\hat{\theta}_{Rk}^1)$	$\mathrm{Var}(\hat{\theta}_{Rk}^2)$	$B^2(\hat{\theta}_{Rk}^2)$	$\mathrm{Var}(\hat{\theta}_{Rk}^3)$	$B^2(\hat{\theta}_{Rk}^3)$	$M_{SE}(\hat{\boldsymbol{\theta}}_{Rk})$
0.0	0.34483	0	18.18182	0	22.22222	0	40.74887
0.2	0.30177	0.01184	0.84583	12.72238	0.74969	16.47462	31.10612
0.4	0.26630	0.04180	0.26567	15.98402	0.22724	19.97503	36.76006
0.6	0.23674	0.08360	0.12820	17.35436	0.10817	21.39295	39.30401
0.8	0.21183	0.13299	0.07524	18.10658	0.06302	22.15920	40.74887
1.0	0.19066	0.18704	0.04942	18.58163	0.04121	22.63888	41.68883

To illustrate its Corollary, Theorem 2(iii) is solved at $k_c = 0.8$ as the common crossover point, giving solutions $[\theta_1^2, \theta_2^2, \theta_3^2] = [2.8448, 20.6818, 24.7222]$ that increase with decreasing ξ_i^2 as in Remark 3. Variances and squared biases for the component curves, namely $\{[\mathrm{Var}(\hat{\theta}_{Rk}^i), B^2(\hat{\theta}_{Rk}^i)]; 1 \le i \le 3\}$, as well as $M_{SE}(\hat{\boldsymbol{\theta}}_{Rk})$, are reported in Table 1 for various k. It follows constructively that $M_{SE}(\hat{\boldsymbol{\theta}}_{Rk}) = M_{SE}(\hat{\boldsymbol{\theta}}_L)$ at $k_c = 0.8$. Moreover, the values $\sum_{i=1}^3 \mathrm{Var}(\hat{\theta}_{Rk}^i) = 0.35009$ and $\sum_{i=1}^3 B^2(\hat{\theta}_{Rk}^i) = 40.39878$, at $k = 0.8$ in the table, sum to $M_{SE}(\hat{\boldsymbol{\theta}}_{Rk}) = 40.74887$. Specifically, in achieving the same *MSE* as the unbiased $\hat{\boldsymbol{\theta}}_L$, the ridge solution $\hat{\boldsymbol{\theta}}_R(0.8)$ encounters excessive biases, with ratio $\sum_{i=1}^3 B^2(\hat{\theta}_{Rk}^i) / \sum_{i=1}^3 \mathrm{Var}(\hat{\theta}_{Rk}^i) = 115.40$. Corollary 1 is demonstrated numerically in the table, as the $M_{SE}(\hat{\boldsymbol{\theta}}_{Rk})$ curve crosses the line $M_{SE}(\hat{\boldsymbol{\theta}}_L)$ from below at $k_c = 0.8$. On the other hand, the extreme conservatism of $k < 1/\theta_M^2 = 1/24.7222 = 0.0404$ from [6] is clear in comparison with our $k_c = 0.8$.

To exemplify Theorem 2(iv, v), we instead fix $[\theta_1^2, \theta_2^2, \theta_3^2]$ at $[3.2, 21.5, 25.0]$, and solve (3.3) for $[k_c^1, k_c^2, k_c^3]$. Entries for $\{[\mathrm{Var}(\hat{\theta}_{Rk}^i), B^2(\hat{\theta}_{Rk}^i)]; 1 \le i \le 3\}$ and $M_{SE}(\hat{\boldsymbol{\theta}}_{Rk})$ for various k, including crossover values $[k_c^1, k_c^2,$

$k_c^3] = [0.7005, 0.6027, 0.7200]$, are reported in Table 2. Clearly $\min\{k_c^1, k_c^2, k_c^3\} = 0.6027$ and $\max\{k_c^1, k_c^2, k_c^3\} = 0.7200$. Theorem 2(iv) assures that $\hat{\boldsymbol{\theta}}_R(k_0)$ is M_{SE}-admissible with respect to $\hat{\boldsymbol{\theta}}_L$ for every $k_0 \le 0.6027$, whereas Theorem 2(v) asserts that $\hat{\boldsymbol{\theta}}_L$ is M_{SE}-admissible with regard to $\hat{\boldsymbol{\theta}}_R(k^\dagger)$ for every $k^\dagger \ge 0.7200$. These assertions are borne out in Table 2.

Table 2. Component values $\{[\mathrm{Var}(\hat{\theta}_{Rk}^i), B^2(\hat{\theta}_{Rk}^i)]; 1 \le i \le 3\}$ and $M_{SE}(\hat{\boldsymbol{\theta}}_{Rk})$ with $\sigma^2 = 1.0$ for various k, including crossover values $[k_c^1, k_c^2, k_c^3] = [0.7005, 0.6027, 0.7200]$ corresponding to $[\theta_1^2, \theta_2^2, \theta_3^2] = [3.2, 21.5, 25.0]$.

k	$\mathrm{Var}(\hat{\theta}_{Rk}^1)$	$B^2(\hat{\theta}_{Rk}^1)$	$\mathrm{Var}(\hat{\theta}_{Rk}^2)$	$B^2(\hat{\theta}_{Rk}^2)$	$\mathrm{Var}(\hat{\theta}_{Rk}^3)$	$B^2(\hat{\theta}_{Rk}^3)$	$M_{SE}(\hat{\boldsymbol{\theta}}_{Rk})$
0.2	0.30177	0.01332	0.84583	13.22568	0.74969	16.65973	31.79601
0.4	0.26630	0.04702	0.26567	16.61635	0.22724	20.19947	37.62205
0.6027	0.23637	0.09475	0.12713	18.05469	0.10725	21.64704	40.26723
0.7005	0.22371	0.12112	0.09636	18.48350	0.08097	22.07292	41.07858
0.7200	0.22130	0.12659	0.09157	18.55667	0.07689	22.14533	41.21835
0.8	0.21183	0.14960	0.07524	18.82289	0.06302	22.40818	41.73076

3.2.2. Connections to other work

An attempt at Theorem 2(iii) is reported in [12], but with the absolute value $|\theta_i^2 - \dfrac{\sigma^2}{\xi_i^2}|$ in the denominator of (3.3) instead. Note that the "minus" sign in equation (10) of those authors should be "plus;" their α^2 should be σ^2; their α_i is our θ_i; and their λ_i is our ξ_i^2. Moreover, with $M_{SE}(k) = M_{SE}(\hat{\boldsymbol{\theta}}_{Rk}) = \sum_{i=1}^{p} (\sigma^2\xi_i^2 + k^2\theta_i^2) / (\xi_i^2 + k)^2$ from Theorem 2(i), those authors allege that the optimal k, minimizing $M_{SE}(\hat{\boldsymbol{\theta}}_{Rk})$, is the root of the fourth degree polynomial $p_4(k) = c_0k^4 + c_1k^3 + c_2k^2 + c_3k + c_4$ from the numerator of $dM_{SE}(k)/dk = 0$, with coefficients

$$c_0 = (p-1)\sum \xi_i^2\theta_i^2$$

$$c_1 = 3\left(\sum \xi_i^2\theta_i^2 \sum \xi_i^2 - \sum \xi_i^4\,\theta_i^2\right) - (p-1)\sigma^2 \sum \xi_i^2$$

$$c_2 = 3\Big[\sum \xi_i^2\theta_i^2 \sum \xi_i^4 - \sum \xi_i^6\,\theta_i^2 - \sigma^2\Big(\sum \theta_i \sum \xi_i^2 - \sum \xi_i^4\Big)\Big]$$

$$c_3 = \sum \xi_i^2\theta_i^2 \sum \xi_i^6 - \sum \xi_i^8\theta_i^2 - 3\sigma^2\Big(\sum \xi_i^2 \sum \xi_i^4 - \sum \xi_i^6\Big)$$

$$c_4 = -\sigma^2\Big(\sum \xi_i^2 \sum \xi_i^6 - \sum \xi_i^8\Big) \tag{3.6}$$

as given following their equation (13). Here the sums all range over $\{i = 1, 2, \ldots, p\}$, and we correct the second parentheses in c_2 to read $\Big(\sum \xi_i^2 \sum \xi_i^2 - \sum \xi_i^4\Big)$. Unfortunately, this solution applies only in the special case $p = 2$. As the rank goes from $p = 2$ to $p = 3$, the polynomial from the numerator of $dM_{SE}(k)/dk$ will increase by 3 in degree to 7; similarly, the case $p = 4$ entails solving a polynomial of degree 10, the degree and intractability growing exponentially with order of the model. The higher polynomials emerge on combining terms to a common denominator.

Example 3. As indicated, the optimal value for k in the case $p = 2$ entails a polynomial of degree 4. Here

$$M_{SE}(k) = \frac{k^2\theta_1^2 + \sigma^2\xi_1^2}{(\xi_1^2 + k)^2} + \frac{k^2\theta_2^2 + \sigma^2\xi_2^2}{(\xi_2^2 + k)^2} \tag{3.7}$$

and

$$\frac{dM_{SE}(k)}{dk} = \sum_{i=1}^{2}\left[\frac{2k\theta_i^2}{(\xi_i^2 + k)^2} - \frac{2(k^2\theta_i^2 + \sigma^2\xi_i^2)}{(\xi_i^2 + k)^3}\right], \tag{3.8}$$

which is zero at the root of the fourth degree polynomial $p_4(k) = c_0k^4 + c_1k^3 + c_2k^2 + c_3k + c_4$ from the numerator of $dM_{SE}(k)/dk$, with coefficients specializing from (3.6) as

$$c_0 = \xi_1^2\theta_1^2 + \xi_2^2\theta_2^2$$

$$c_1 = 3\xi_1^2\xi_2^2(\theta_1^2 + \theta_2^2) - \sigma^2(\xi_1^2 + \xi_2^2)$$

$$c_2 = 3(\xi_1^2\xi_2^4\theta_1^2 + \xi_1^4\xi_2^2\theta_2^2 - 2\sigma^2\xi_1^2\xi_2^2)$$

$$c_3 = \xi_1^2\xi_2^6\theta_1^2 + \xi_1^6\xi_2^2\theta_2^2 - 3\sigma^2(\xi_1^2\xi_2^4 + \xi_1^4\xi_2^2)$$

$$c_4 = -\sigma^2(\xi_1^2\xi_2^6 + \xi_1^6\xi_2^2). \tag{3.9}$$

4. Case Studies

4.1. The setting

As reported in Table 3.8 of Myers [11, pp. 132-133], the Hospital Manpower Data consist of $n = 17$ records at U. S. Naval Hospitals, to include: monthly manhours (Y); Average daily patient load (X_1); monthly X-ray exposures (X_2); monthly occupied bed days (X_3); eligible population in the area $\div 1000$ (X_4); and average length of patients' stay in days (X_5). The working model is

$$\{Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon_i;\ 1 \le i \le n\}. \tag{4.1}$$

Following convention, we center and scale to $\{\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}\}$ with $\boldsymbol{Z}'\boldsymbol{Z}$ in correlation form, our focus being the rates of change $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3, \beta_4, \beta_5]'$. Computations reported here utilize the SAS Programming System. The *OLS* solutions are $\hat{\boldsymbol{\beta}}_L = [-10210.40, 4760.50, 31198.64, -1822.11, -2498.58]'$, with $S = 642.0771$ as the square root of the residual mean square. The data are remarkably ill-conditioned: Singular values of $\boldsymbol{Z}$ are $\boldsymbol{D}_\xi = \text{Diag}(2.048687, 0.816997, 0.307625, 0.201771, 0.007347)$; the condition number is $c_1(\boldsymbol{Z}'\boldsymbol{Z}) = 77{,}754.86$; and the *VIF*s in *OLS* estimation are $\{9595.685, 7.941, 8931.449, 23.289, 4.279\}$.

4.2. Canonical analysis

Defects in *OLS* bear further scrutiny. Returning to Section 3.1, we relate $\boldsymbol{\theta} = \boldsymbol{Q}'\boldsymbol{\beta}$, where $\boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{q}_3, \boldsymbol{q}_4, \boldsymbol{q}_5]$ is displayed as

$$\boldsymbol{Q} = \begin{bmatrix} 0.485286 & -0.002030 & -0.166230 & -0.468190 & -0.719480 \\ 0.453235 & -0.335600 & 0.804239 & 0.187470 & -0.001160 \\ 0.484977 & -0.000850 & -0.153960 & -0.509260 & 0.694081 \\ 0.460969 & -0.310800 & -0.537200 & 0.633861 & 0.023436 \\ 0.333737 & 0.889251 & 0.115239 & 0.290733 & 0.006780 \end{bmatrix}. \tag{4.2}$$

Canonical solutions are $\hat{\boldsymbol{\theta}}_L = \boldsymbol{Q}'\hat{\boldsymbol{\beta}}_L = [10659.49, -3259.16, 1413.40, -12096.70, 28935.42]'$. Clearly $\theta_1 = \boldsymbol{q}_1'\boldsymbol{\beta}$ is close to a scaled average of $\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$; moreover, θ_1 is estimated by *OLS* with variance $\sigma^2/2.048687^2 = 0.238259\sigma^2$,

less than $1.0\sigma^2$ for an orthogonal design. Thus prediction along the equiangular line, with scaled variance $0.398435\sigma^2$, is more precise than $1.0\sigma^2$ in the corresponding orthogonal design. At the other extremity, $\theta_5 = \boldsymbol{q}_5'\boldsymbol{\beta}$ is a near linear contrast between $\{\beta_1, \beta_2\}$ and $\{\beta_3, \beta_4, \beta_5\}$, since the inner product $\mathbf{1}_5'\boldsymbol{q}_5 = 0.003657$ is approximately zero; but this is estimated by *OLS* with variance $\sigma^2/0.007347^2 = 18{,}525.71\sigma^2$. This also resembles the highly variable contrast $(\hat{\beta}_1 - \hat{\beta}_3)$, with $\mathrm{Var}((\hat{\beta}_1 - \hat{\beta}_3)/\sqrt{2}) = 18{,}508.77\sigma^2$. The regressors, X_1: average daily patient load and X_3: monthly occupied bed days, exhibit near collinearity with entry 0.999904 in the $\boldsymbol{Z}'\boldsymbol{Z}$ matrix. A variation on Theorem 1 shows that $\{10.5671\sigma^2 \le \mathrm{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}_L)/\boldsymbol{c}'\boldsymbol{c} \le 24.5631\sigma^2\}$ for $\boldsymbol{c} \in Sp(\boldsymbol{q}_3, \boldsymbol{q}_4)$ under *OLS*, from $\sigma^2/0.307625^2 = 10.5671\sigma^2$ and $\sigma^2/0.201771^2 = 24.5631\sigma^2$.

It follows similarly that $\{0.238259\sigma^2 \le \mathrm{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}_L)/\boldsymbol{c}'\boldsymbol{c} \le 10.5671\sigma^2\}$ for $\boldsymbol{c} \in Sp(\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{q}_3)$. In short, ill-conditioning here essentially devastates *OLS* estimation of the near contrast $\theta_5 = \boldsymbol{q}_5'\boldsymbol{\beta}$, and the actual contrast $(\beta_1 - \beta_3)$, but considerably less for functions $\boldsymbol{c}'\boldsymbol{\beta}$ orthogonal to $\boldsymbol{q}_5'\boldsymbol{\beta}$, i.e., $\{0.238259\sigma^2 \le \mathrm{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}_L)/\boldsymbol{c}'\boldsymbol{c} \le 24.5631\sigma^2\}$ for $\{\boldsymbol{c} \in Sp(\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{q}_3, \boldsymbol{q}_4)\}$. In summary, *OLS* is conspicuously deficient in the one-dimensional subspace spanned by $\boldsymbol{q}_5$. This fact is anticipated neither by $c_1(\boldsymbol{Z}'\boldsymbol{Z}) = 77{,}754.86$, nor by the maximal $V_1 = VIF(\hat{\beta}_1) = 9{,}595.685$, and it becomes apparent only through a canonical analysis as reported.

4.3. Ridge regression

Ill-conditioning renders $\hat{\theta}_{L1}$ as super-efficient, its variance $\mathrm{Var}(\hat{\theta}_{L1}) = \sigma^2/4.197118 = 0.238259\sigma^2$ deflated relative to $1.0\sigma^2$ under "ideal" orthogonality, as in Theorem 1. Since ridge is intended to ameliorate ill effects of ill-conditioning, and $\hat{\theta}_{L1}$ is not among them, it remains to ask whether $\hat{\theta}_{L1}$ should be retained *per se*, with ridge adjustments deferred to

the remaining estimators. Accordingly, we seek effects of ridge adjustments, in the event these were made to $\hat{\theta}^1_{Rk}$ *in lieu of* $\hat{\theta}_{L1}$. Essential features again follow at $\sigma^2 = 1.0$, where θ_i and τ_i become interchangeable from Remark 1. As in Theorem 2(iii), we seek crossing values in pairs (k^1_c, θ^2_1) at which $\hat{\theta}^1_{Rk}$ achieves the same *MSE* as $\hat{\theta}_{L1}$ under *OLS*. Table 3 lists selected values for (k^1_c, θ^2_1), together with the corresponding variance and squared bias for $\hat{\theta}^1_{Rk}$. Entries in the final row are identical, namely $M_{SE}(\hat{\theta}_{L1}) = \mathrm{Var}(\hat{\theta}_{L1}) = \sigma^2/4.197118 = 0.23826$ at $\sigma^2 = 1.0$, since each $M_{SE}(\hat{\theta}^1_{Rk})$ curve is required to cross the *OLS* line. In particular, at $k^1_c = 0.5$ and $\theta^2_1 = 4.23826$, the ridge estimator $\hat{\theta}^1_{Rk}$ achieves the same *MSE* as *OLS*, namely, $M_{SE}(\hat{\theta}^1_{Rk}) = \mathrm{Var}(\hat{\theta}^1_{Rk}) + B^2(\hat{\theta}^1_{Rk}) = (0.19023 + 0.04803) = 0.23826$, but biased with bias ratio $B^2(\hat{\theta}^1_{Rk})/\mathrm{Var}(\hat{\theta}^1_{Rk}) = 0.2525$, as opposed to $\hat{\theta}_{L1}$. Further evaluation from (3.3) shows that $\theta^2_1 > 2.23826$ implies the corresponding $k^1_c < 1.0$. This in turn implies the *MSE*-inadmissibility of $\hat{\boldsymbol{\theta}}_{Rk}$ with respect to *OLS* for some values $k \in (0, 1)$, where *OLS* necessarily is superior to ridge for those values.

Table 3. Values for k^1_c in the Hospital Manpower Data and corresponding values for $\{\mathrm{Var}(\hat{\theta}^1_{Rk}), B^2(\hat{\theta}^1_{Rk}), M_{SE}(\hat{\theta}^1_{Rk})\}$, standardized to $\sigma^2 = 1.0$, such that $M_{SE}(\hat{\theta}^1_{Rk})$ crosses the $M_{SE}(\hat{\theta}_{L1})$ line, where $k^1_c = 0.0$ corresponds to *OLS*.

k^1_c	0.0	0.1	0.3	0.5	0.7	0.9
θ^2_1		20.23826	6.90493	4.23826	3.09540	2.46048
$\mathrm{Var}(\hat{\theta}^1_{Rk})$	0.23826	0.22730	0.20753	0.19023	0.17501	0.16155
$B^2(\hat{\theta}^1_{Rk})$	0.00	0.01096	0.03073	0.04803	0.06325	0.07671
$M_{SE}(\hat{\theta}^1_{Rk})$	0.23826	0.23826	0.23826	0.23826	0.23826	0.23826

Deeply divergent criteria have been advocated for choosing k. Five choices in common usage are identified in Table 4, together with their values for the Hospital Manpower Data. These include $DF_k = \sum_{i=1}^{p} \xi_i^2/(\xi_i^2 + k)$; $PRESS_k$

as the cross-validation rule of [1]; GCV_k as the *Generalized Cross Validation* rule of [3]; C_k as in [8] to achieve a variance-bias trade-off; and HKB_k as in [7] from simulation studies. For further details see [11, pp. 392-411], including numerical values for DF_k, C_k and $PRESS_k$ as reported here. For each criterion, the corresponding θ_1^2 is determined from (3.3), such that $M_{SE}(\hat{\theta}_{Rk}^1)$ crosses the *OLS* line $M_{SE}(\hat{\theta}_{L1})$. Also listed are the variance, the squared bias, and $M_{SE}(\hat{\theta}_{Rk}^1)$ for these (k, θ_1^2) pairs. Note that at $C_k = 0.0050$, the ridge estimator $\hat{\theta}_{Rk}^1$ achieves the same *MSE* as *OLS*, namely, $M_{SE}(\hat{\theta}_{Rk}^1) = \mathrm{Var}(\hat{\theta}_{Rk}^1) + B^2(\hat{\theta}_{Rk}^1) = (0.23769 + 0.00057) = 0.23826$, but with negligible bias ratio $B^2(\hat{\theta}_{Rk}^1)/\mathrm{Var}(\hat{\theta}_{Rk}^1) = 0.002398$. On the other hand, at $HKB_k = 0.61696$, the corresponding crossing values are $M_{SE}(\hat{\theta}_{Rk}^1) = (0.18110 + 0.05716) = 0.23826$ with bias ratio $B^2(\hat{\theta}_{Rk}^1)/\mathrm{Var}(\hat{\theta}_{Rk}^1) = 0.3156$.

Table 4. Choices for k in the Hospital Manpower Data at conventional criteria $\{DF_k, GCV_k, C_k, PRESS_k, HKB_k\}$; and values for $\{\mathrm{Var}(\hat{\theta}_{Rk}^1), B^2(\hat{\theta}_{Rk}^1), M_{SE}(\hat{\theta}_{Rk}^1)\}$, standardized to $\sigma^2 = 1.0$, where $M_{SE}(\hat{\theta}_{Rk}^1)$ crosses the $M_{SE}(\hat{\theta}_{L1})$ line for each k.

Name	DF_k	GCV_k	C_k	$PRESS_k$	HKB_k
Value for k	0.0004	0.004787	0.0050	0.2300	0.61696
θ_1^2	5000.24	418.0365	400.2383	8.93391	3.47994
$\mathrm{Var}(\hat{\theta}_{Rk}^1)$	0.23821	0.23772	0.23769	0.21415	0.18110
$B^2(\hat{\theta}_{Rk}^1)$	0.00005	0.00054	0.00057	0.02411	0.05716
$M_{SE}(\hat{\theta}_{Rk}^1)$	0.23826	0.23826	0.23826	0.23826	0.23826

To continue, suppose that $C_k = 0.0050$ were the actual cross-over point. Then Theorem 2(iv) asserts that $\hat{\theta}_{Rk}^1$ is M_{SE}-admissible with regard to $\hat{\theta}_{L1}$ at $DF_k = 0.0004$ and $GCV_k = 0.004787$. In contrast, Theorem 2(v) shows $\hat{\theta}_{L1}$ to be M_{SE}-admissible with respect to $\hat{\theta}_{Rk}^1$ for $k > 0.0050$. In particular, at $PRESS_k = 0.2300$, further computations show that $M_{SE}(\hat{\theta}_{Rk}^1) = \mathrm{Var}(\hat{\theta}_{Rk}^1) + B^2(\hat{\theta}_{Rk}^1) = (0.21415 + 1.08027) = 1.29442$, grossly exceeding $M_{SE}(\hat{\theta}_{L1}) =$

0.23826 by the factor 5.4328, for which ridge estimation is counter-productive in excess.

Finally compute the standardized estimator $\hat{\theta}_{L1}/S = 10{,}659.49/642.0771 = 16.60157$ from earlier estimates, and treat its square as $\hat{\tau}_1^2$. Substituting this into (3.3) gives $\hat{k}_c^1 = 0.007263$ to estimate the cross-over value where $M_{SE}(\hat{\theta}_{Rk}^1) = M_{SE}(\hat{\theta}_{L1})$. This in turn suggests $\hat{\theta}_{L1}$ to be *MSE*-superior to $\hat{\theta}_{Rk}^1$ for $k \in (0.0073, 1.0000)$ and, in particular, at $PRESS_k = 0.2300$ and $HKB_k = 0.61696$.

5. Conclusions

In summary, further insight is offered regarding a large and continuing literature on the use of ridge regression in ill-conditioned systems. We first reexamine *OLS via* a canonical form identifying parametric functions where *OLS* is problematic, and where it is not. Specifically, subspaces of parameters are characterized where *OLS* has greater, equal, or lesser Fisher efficiency, in comparison with an orthogonal design as "ideal". These results suggest that nonorthogonal regressor designs, properly constituted, could be preferred over orthogonal designs, if the information sought is not of uniform consequence in all directions.

Further studies here address not only the superiority of ridge over *OLS*, but of equal importance, the superiority of *OLS* over ridge, despite the perennial neglect of this topic. These are concerned with circumstances for use of either methodology in lieu of the other. The rule $k < \sigma^2/\theta_M^2$ of [6], given as sufficient for ridge to dominate *OLS* in its smaller *MSE*, is shown to be conservative to the point of excess, and the rule is expanded considerably. Cross-over values k_c are determined where ridge and *OLS* agree in *MSE*, values smaller than k_c favoring ridge over *OLS*, values larger than k_c favoring *OLS* over ridge. Contrary to conventional wisdom, ridge often must be abandoned in favor of *OLS* for k sufficiently large. Extended numerical studies are reported to illustrate the new findings.

References

[1] D. M. Allen, The relationship between variable selection and data augmentation and a method for prediction, Technometrics 16 (1974), 125-127.

[2] D. A. Belsley, Centering, the Constant, First-Differencing and Assessing Conditioning, In Model Reliability, Chapter 5, D. A. Belsley and E. Kuh, Eds., MIT Press, 1986, pp. 117-153.

[3] G. H. Golub, C. G. Heath and G. Wahba, Generalized cross validation as a method for choosing a good ridge parameter, Technometrics 21 (1979), 215-223.

[4] A. E. Hoerl, Application of ridge analysis to regression problems, Chemical Engineering Progress 58 (1962), 54-59.

[5] A. E. Hoerl, Ridge analysis, Chemical Engineering Progress Symposium Series 60 (1964), 67-77.

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

[7] A. E. Hoerl, R. W. Kennard and K. F. Baldwin, Ridge regression: some simulations, Commun. Statist. 4 (1975), 105-123.

[8] C. L. Mallows, Some comments on C_p: Technometrics 15 (1973), 661-675.

[9] D. W. Marquardt, Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation, Technometrics 12 (1970), 591-612.

[10] D. W. Marquardt and R. D. Snee, Ridge regression in practice, The Amer. Statist. 29 (1975), 3-20.

[11] R. H. Myers, Classical and Modern Regression with Applications, Second ed., PWS-KENT Publishing Co., Boston, MA, 1990.

[12] H. Myoken and Y. Uchida, The generalized ridge estimator and improved adjustments for regression parameters, Metrika 24 (1977), 113-124.

[13] R. A. Thisted, Comment, Statistical Science 2 (1987), 91-93.