Journal of Statistical Planning and Inference 140 (2010) 2069–2077

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Surrogate models in ill-conditioned systems[☆]

D.R. Jensen[a], D.E. Ramirez[b,*]

[a] Department of Statistics, Virginia Tech, Blacksburg, VA 24061, USA
[b] Department of Mathematics, University of Virginia, Charlottesville, VA 22904-4137, USA

ARTICLE INFO

Article history:
Received 27 April 2009
Received in revised form
21 October 2009
Accepted 1 February 2010
Available online 6 February 2010

MSC:
62J07
62J20

Keywords:
Ill-conditioned models
Ridge regression: properties
Anomalies
Surrogate models
Marques of orthogonality
Asymptotics

ABSTRACT

Ridge versions of an ill-conditioned system are alleged to "act more like an orthogonal system" than the system itself. Alternatives, called *surrogates* and based on the conditioning of linear systems, are shown to yield smaller expected mean squares than *OLS*, and uniformly smaller residual sums of squares than ridge. Ridge and surrogate solutions are compared on several marques of orthogonality to include conditioning of dispersion arrays, variance inflation factors, isotropy of variances, and sphericity of contours of the estimators. For these, ridge typically exhibits erratic divergence from orthogonality as the ridge scalar evolves, often reverting back to *OLS* in the limit. In contrast, surrogate solutions converge monotonically to those from orthogonal systems. Invariance considerations constrain the computations to models in canonical form. Case studies serve to illustrate the central issues.

1. Introduction

In a full-rank model $\{\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}\}$ with uncorrelated, zero-mean, and homoscedastic errors, the p equations $\{\boldsymbol{X'X\beta}=\boldsymbol{X'Y}\}$ yield the Ordinary Least Squares (*OLS*) solutions $\hat{\boldsymbol{\beta}}_L$ as unbiased, having dispersion matrix $\mathrm{V}(\hat{\boldsymbol{\beta}}_L)=\sigma^2\boldsymbol{V}$ with $\boldsymbol{V}=[v_{ij}]=(\boldsymbol{X'X})^{-1}$. *Ill-conditioning,* as near-dependency among columns of $\boldsymbol{X}$, "causes crucial elements of $\boldsymbol{X'X}$ to be large and unstable," "creating inflated variances," and elements of $\hat{\boldsymbol{\beta}}_L$ that are "very sensitive to small changes in $\boldsymbol{X}$" Belsley (1986, p. 119). Here the *condition number* $c_1(\boldsymbol{X'X})$ is the ratio of largest to smallest eigenvalues; the *Variance Inflation Factors* (*VIF* s) of $\hat{\boldsymbol{\beta}}_L=[\hat{\beta}_L^1,\ldots,\hat{\beta}_L^p]'$ are $\{VIF(\hat{\beta}_L^j)=v_{jj}/w_{jj}^{-1};1\le j\le p\}$ with $\boldsymbol{W}=\boldsymbol{X'X}$, i.e., ratios of actual to "ideal" variances had the columns of $\boldsymbol{X}$ been orthogonal. Scaling columns of $\boldsymbol{X}$ to unit lengths and rearranging $\{VIF(\hat{\beta}_L^j)=v_{jj};1\le j\le p\}$ in order as $\{V_1\ge V_2\ge\cdots\ge V_p\}$, V_1 is identified in Marquardt and Snee (1975) as "the best single measure of the conditioning of the data," whereas $V_1\le c_1(\boldsymbol{X'X})\le p(V_1+\cdots+V_p)$ is shown in Berk (1977);see also Beaton et al. (1976), Davies and Hutton (1975), Marquardt (1970), and Stewart (1987). The *singular decomposition* $\boldsymbol{X}=\boldsymbol{P}_1\boldsymbol{D}_\xi\boldsymbol{Q}'$ yields a *canonical form* $\{\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}\}$ → $\{\boldsymbol{P}_1'\,\boldsymbol{Y}=\boldsymbol{D}_\xi\boldsymbol{\theta}+\boldsymbol{\eta}\}$; $\boldsymbol{\theta}=\boldsymbol{Q}'\boldsymbol{\beta}$ and $\boldsymbol{\eta}=\boldsymbol{P}_1'\,\boldsymbol{\varepsilon}$; such that $\boldsymbol{D}_\xi=\mathrm{Diag}(\xi_1,\ldots,\xi_p)$ is diagonal with $\{\xi_1\ge\cdots\ge\xi_p>0\}$; $\boldsymbol{Q}\in\mathcal{O}(p)$, the

[☆] Research supported in part by the Department of Mathematics, University of Virginia.

[*] Corresponding author.
E-mail addresses: djensen@vt.edu (D.R. Jensen), der@virginia.edu (D.E. Ramirez).

0378-3758/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2010.02.001

orthogonal group; and $\boldsymbol{P}_1' \boldsymbol{P}_1 = \boldsymbol{I}_p$, the identity matrix. This reduction, a staple in the literature, is central to much that follows, supporting by invariance a number of properties otherwise obscured.

Remedies for ill-conditioning include the *ridge system* $\{(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I}_p)\boldsymbol{\beta}=\boldsymbol{X}'\boldsymbol{Y}; k \geq 0\}$ of Hoerl and Kennard (1970), with solutions $\hat{\boldsymbol{\beta}}_{Rk}$. For a class of condition numbers including $c_1(\cdot)$, Riley (1955) has shown that $\{\boldsymbol{X}'\boldsymbol{X}\to(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I}_p)\}$ improves conditioning on the left of the ridge equations, and is cited in Marshall and Olkin (1979) as a rationale for ridge regression over *OLS*. Nonetheless, $\boldsymbol{X}'\boldsymbol{Y}$ remains ill-conditioned on the right. This prompts our concept of *surrogate models* $\{\boldsymbol{Y}=\boldsymbol{X}_k\boldsymbol{\beta}+\boldsymbol{\varepsilon}; k\geq 0\}$, designed to improve conditioning on the left *and* right of the *surrogate system* $\{\boldsymbol{X}_k' \boldsymbol{X}_k\boldsymbol{\beta}=\boldsymbol{X}_k' \boldsymbol{Y}; k\geq 0\}$, with solutions $\hat{\boldsymbol{\beta}}_{Sk}$. Details are given subsequently, together with generalizations. In support of ridge regression, Hoerl and Kennard (1970, p. 65) make the further claim: "At a certain value of k the system will stabilize and have the general characteristics of an orthogonal system," these not having been specified. Our purposes here are (i) to identify features characteristic of orthogonal systems; (ii) to reexamine the foregoing conjecture despite unaddressed ill-conditioning on the right of the ridge equations; and (iii) to investigate the corresponding conjecture for surrogate systems. An outline follows; conventional notation is used throughout.

Section 2 develops surrogate systems alternative to ridge through conditioning. These improve over *OLS* in mean square under conditions of Theorem 1, having by Theorem 2 uniformly smaller residual sums of squares than ridge for each $k>0$. Section 3 identifies useful marques of orthogonality to include key condition numbers and ellipticity indices in Theorem 3, and variance inflation factors in Theorem 4. Of these, ridge is seen to diverge erratically from orthogonality as k evolves, often reverting back to *OLS* in the limit. In contrast, surrogate solutions increasingly resemble those from orthogonal systems, monotonically as k evolves. Case studies in Section 4 illustrate the central issues; Section 5 concludes with a brief summary; and essential proofs are deferred to an Appendix.

2. Ridge and surrogate models

2.1. Generalized and ordinary models

Make the transition $\{\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}=\boldsymbol{X}'\boldsymbol{Y}\}\ \to\ \{(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{C})\boldsymbol{\beta}=\boldsymbol{X}'\boldsymbol{Y}\}$, taking $\boldsymbol{C}$ to be positive definite, commuting with $\boldsymbol{X}'\boldsymbol{X}$, and orthogonally congruent to $\boldsymbol{K}=\text{Diag}(k_1,\ldots,k_p)$. This is *generalized ridge regression*; see Bingham and Larntz (1977), Hoerl and Kennard (1970), Hoerl et al. (1975), and Lowerre (1974), for example. But this amounts to modifying the singular decomposition $\boldsymbol{X}=\boldsymbol{P}_1\boldsymbol{D}_\xi\boldsymbol{Q}'\to\boldsymbol{X}_{\boldsymbol{K}}=\boldsymbol{P}_1\boldsymbol{D}((\xi_i^2+k_i)^{1/2})\boldsymbol{Q}'$ with $\boldsymbol{D}((\xi_i^2+k_i)^{1/2})=\text{Diag}((\xi_1^2+k_1)^{1/2},\ldots,(\xi_p^2+k_p)^{1/2})$, for then $\boldsymbol{X}'\boldsymbol{X}\to\boldsymbol{X}_{\boldsymbol{K}}' \boldsymbol{X}_{\boldsymbol{K}}=\boldsymbol{Q}(\boldsymbol{D}_\xi^2+\boldsymbol{K})\boldsymbol{Q}'=(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{C})$. Ordinary ridge regression specializes at $\boldsymbol{K}=k\boldsymbol{I}_p$, with $\boldsymbol{X}_k=\boldsymbol{P}_1\boldsymbol{D}((\xi_i^2+k)^{1/2})\boldsymbol{Q}'$ and $\boldsymbol{X}_k' \boldsymbol{X}_k=(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I}_p)$. To improve conditioning on both sides of $\{\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}=\boldsymbol{X}'\boldsymbol{Y}\}$, we take the *model* $\{\boldsymbol{Y}=\boldsymbol{X}_{\boldsymbol{K}}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\}$ as a *surrogate* for the ill-conditioned model $\{\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\}$ itself, as in the following.

Definition 1. Given an ill-conditioned model $\{\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\}$, its generalized surrogate is $\{\boldsymbol{Y}=\boldsymbol{X}_{\boldsymbol{K}}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\}$. The generalized surrogate estimator $\hat{\boldsymbol{\beta}}_{SK}$, solving $\{(\boldsymbol{X}_{\boldsymbol{K}}' \boldsymbol{X}_{\boldsymbol{K}})\hat{\boldsymbol{\beta}}_{SK}=\boldsymbol{X}_{\boldsymbol{K}}' \boldsymbol{Y}\}$, is *OLS* in the generalized surrogate model. Specifically, with $\boldsymbol{K}=k\boldsymbol{I}_p$, the *ordinary surrogate* is $\{\boldsymbol{Y}=\boldsymbol{X}_k\boldsymbol{\beta}+\boldsymbol{\varepsilon}\}$, and the solution $\hat{\boldsymbol{\beta}}_{Sk}$ of $\{(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I}_p)\hat{\boldsymbol{\beta}}_{Sk}=\boldsymbol{X}_k' \boldsymbol{Y}\}$ is *OLS* for the ordinary surrogate approximation.

By convention the condition number $c_1(\boldsymbol{X})$ is the ratio of its largest to smallest singular values, namely, $c_1(\boldsymbol{X})=\xi_1/\xi_p$. Specifically, $c_1(\boldsymbol{X})\to c_1(\boldsymbol{X}_k)=(\xi_1+k)/(\xi_p+k)$ reflects improved conditioning on the right of the ordinary surrogate equations, and similarly for their generalized version. In summary, essentials of the generalized ridge and surrogate models are

$$\textit{Ridge}:\quad \boldsymbol{Q}(\boldsymbol{D}_\xi^2+\boldsymbol{K})\boldsymbol{Q}'\hat{\boldsymbol{\beta}}_{RK}=\boldsymbol{Q}\boldsymbol{D}_\xi\boldsymbol{P}_1' \boldsymbol{Y}, \tag{1}$$

$$\textit{Surrogate}:\quad \boldsymbol{Q}(\boldsymbol{D}_\xi^2+\boldsymbol{K}^*)\boldsymbol{Q}'\hat{\boldsymbol{\beta}}_{SK^*}=\boldsymbol{Q}\boldsymbol{D}\Big((\xi_i^2+k_i^*)^{1/2}\Big)\boldsymbol{P}_1' \boldsymbol{Y}. \tag{2}$$

Here $\boldsymbol{K}=\text{Diag}(k_1,\ldots,k_p)$ and $\boldsymbol{K}^*=\text{Diag}(k_1^*,k_2^*,\ldots,k_p^*)$ are distinguished for reasons to follow, showing (i) equivalence, and (ii) that each generalized model specializes to the ordinary version of the other.

Theorem 1. *Given the generalized ridge and surrogate systems, consider the respective solutions $\hat{\boldsymbol{\beta}}_{RK}$ and $\hat{\boldsymbol{\beta}}_{SK^*}$ and, for $\boldsymbol{K}=k\boldsymbol{I}_p$ and $\boldsymbol{K}^*=k^*\boldsymbol{I}_p$, the ordinary solutions $\hat{\boldsymbol{\beta}}_{Rk}$ and $\hat{\boldsymbol{\beta}}_{Sk^*}$.*

(i) *The generalized systems* (1) *and* (2) *are equivalent; elements of $\boldsymbol{K}$ and $\boldsymbol{K}^*$ relate one-to-one through $\{k_i^*=2k_i+k_i^2/\xi_i^2; 1\leq i\leq p\}$.*

(ii) *Ordinary surrogate, with solution $\hat{\boldsymbol{\beta}}_{Sk^*}$, is generalized ridge with parameters $\boldsymbol{K}=\text{Diag}(k_1,\ldots,k_p)$ on solving $\{k^*=2k_i+k_i^2/\xi_i^2; 1\leq i\leq p\}$.*

(iii) *Ordinary ridge, with solution $\hat{\boldsymbol{\beta}}_{Rk}$, is generalized surrogate with $\boldsymbol{K}^*=\text{Diag}(k_1^*,\ldots,k_p^*)$ on solving $\{k_i^*=2k+k^2/\xi_i^2; 1\leq i\leq p\}$.*

Proof. See Appendix A.

2.2. Residual comparisons

That solutions are optimal can be claimed for neither $\{\hat{\boldsymbol{\beta}}_{Rk}; k>0\}$ nor $\{\hat{\boldsymbol{\beta}}_{Sk}; k>0\}$, except that $Q(\boldsymbol{\beta})=(\boldsymbol{Y}-\boldsymbol{X}_k\boldsymbol{\beta})'(\boldsymbol{Y}-\boldsymbol{X}_k\boldsymbol{\beta})$ is minimized at $\hat{\boldsymbol{\beta}}_{Sk}$. Nonetheless, the residual sum of squares (R_{SS}) often is at issue in considering alternatives to *OLS*. The following demonstration is germane.

Theorem 2. *Consider the ordinary ridge* $\{(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I}_p)\hat{\boldsymbol{\beta}}_{Rk}=\boldsymbol{X}'\boldsymbol{Y}; k\geq 0\}$ *and surrogate* $\{(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I}_p)\hat{\boldsymbol{\beta}}_{Sk}=\boldsymbol{X}_k'\,\boldsymbol{Y}; k\geq 0\}$ *systems. Then their residual sums of squares are ordered, for each* $k>0$, *as* $R_{SS}(\hat{\boldsymbol{\beta}}_{Sk})<R_{SS}(\hat{\boldsymbol{\beta}}_{Rk})$.

Proof. See Appendix A.

In short, residual sums of squares are smaller for surrogate than for ridge, point-wise for each $k>0$.

2.3. Admissible values

In estimating $\boldsymbol{\beta}$ using $\tilde{\boldsymbol{\beta}}$ with bias $B(\tilde{\boldsymbol{\beta}})=\mathrm{E}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})=(\boldsymbol{\beta}_0-\boldsymbol{\beta})$, its mean square error is $M_{SE}(\tilde{\boldsymbol{\beta}})=\mathrm{E}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})=\mathrm{tr}\,\mathrm{V}(\tilde{\boldsymbol{\beta}})$ + $(\boldsymbol{\beta}_0-\boldsymbol{\beta})'(\boldsymbol{\beta}_0-\boldsymbol{\beta})$, effecting the variance-bias trade-off under squared error loss. Equivalently, taking $\boldsymbol{\theta}=\boldsymbol{Q}'\boldsymbol{\beta}$ in canonical form with $\boldsymbol{Q}$ orthogonal, thus $\tilde{\boldsymbol{\theta}}=\boldsymbol{Q}'\tilde{\boldsymbol{\beta}}$, it suffices that $M_{SE}(\tilde{\boldsymbol{\beta}})=M_{SE}(\tilde{\boldsymbol{\theta}})=\mathrm{E}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta})'(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta})$. Values k_0 sufficient for $M_{SE}[\hat{\boldsymbol{\beta}}_R(k_0)]<M_{SE}(\hat{\boldsymbol{\beta}}_L)$ are given in Hoerl and Kennard (1970), i.e., ridge solutions M_{SE}-efficient relative to *OLS*, to be designated here as *M_{SE}-admissible* and thus germane to the study. A two-fold void in ridge regression, as widely understood, is that (i) k_0 from Hoerl and Kennard (1970) is unnecessarily conservative and (ii) there is an upper limit for k beyond which *OLS* is M_{SE}-efficient relative to $\hat{\boldsymbol{\beta}}_{Rk}$, the latter to be supplanted by $\hat{\boldsymbol{\beta}}_L$ thereafter. Details follow.

Lemma 1. *Consider the OLS solutions* $\hat{\boldsymbol{\theta}}_L=[\hat{\theta}_L^1,\ldots,\hat{\theta}_L^p]'$ *and ridge estimators* $\hat{\boldsymbol{\theta}}_{Rk}=[\hat{\theta}_{Rk}^1,\hat{\theta}_{Rk}^2,\ldots,\hat{\theta}_{Rk}^p]'$.

(i) *The typical element* $M_{SE}(\hat{\theta}_{Rk}^i)=(\sigma^2\xi_i^2+k^2\theta_i^2)/(\xi_i^2+k)^2$ *achieves its minimal value* $\sigma^2\theta_i^2/(\sigma^2+\xi_i^2\theta_i^2)$ *at* $k_i=\sigma^2/\theta_i^2$.

(ii) *The cross-over value from below, achieving equality in* $M_{SE}(\hat{\theta}_{Rk}^i)\leq M_{SE}(\hat{\theta}_L^i)$, *occurs at* $k_c^i=2\sigma^2\xi_i^2/(\xi_i^2\theta_i^2-\sigma^2)$, *provided that* $\xi_i^2\theta_i^2>\sigma^2$.

(iii) *If* $k_m\leq\min\{k_c^1,k_c^2,\ldots,k_c^p\}$, *then* $\hat{\boldsymbol{\beta}}_R(k_m)$ *is M_{SE}-admissible with respect to* $\hat{\boldsymbol{\beta}}_L$.

(iv) *If* $k_M\geq\max\{k_c^1,k_c^2,\ldots,k_c^p\}$, *then* $\hat{\boldsymbol{\beta}}_L$ *is M_{SE}-admissible with respect to* $\hat{\boldsymbol{\beta}}_R(k_M)$, *and OLS continues thereafter.*

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

Proof. See Jensen and Ramirez (2010). □

Parallel results for $\hat{\boldsymbol{\beta}}_{Sk}$, if available, are equally germane. To these ends write $M_{SEi}(k)$ for either $M_{SE}(\hat{\theta}_{Rk}^i)$ or $M_{SE}(\hat{\theta}_{Sk}^i)$ in context; observe that $M_{SE}(\hat{\theta}_{Rk}^i)$ and, from what follows $M_{SE}(\hat{\theta}_{Sk}^i)$, both originate at $M_{SE}(\hat{\theta}_L^i)=\sigma^2/\xi_i^2$; then decrease to a minimum, increasing thereafter to θ_i^2 which explains the condition of $\xi_i^2\theta_i^2>\sigma^2$, and crossing the constant *OLS* line σ^2/ξ_i^2 from below. The initial rates of change $[dM_{SEi}(k)/dk]|_{k=0}$ are $-2\sigma^2/\xi_i^4$ and $-\sigma^2/\xi_i^4$ for the ridge and surrogate estimators, respectively, but then $M_{SE}(\hat{\theta}_{Rk}^i)$ rises faster than $M_{SE}(\hat{\theta}_{Sk}^i)$, with a smaller cross-over value. Developments corresponding to Lemma 1 apply also in the case of $\hat{\boldsymbol{\beta}}_{Sk}$; essential results follow.

Theorem 3. *Consider the OLS estimators* $\hat{\boldsymbol{\theta}}_L=[\hat{\theta}_L^1,\ldots,\hat{\theta}_L^p]'$ *and the surrogate solutions* $\hat{\boldsymbol{\theta}}_{Sk}=[\hat{\theta}_{Sk}^1,\ldots,\hat{\theta}_{Sk}^p]'$.

(i) *The typical element* $M_{SE}(\hat{\theta}_{Sk}^i)=(\theta_i^2k+2\theta_i^2\xi_i^2-2\theta_i^2\xi_i\sqrt{\xi_i^2+k}+\sigma^2)/(\xi_i^2+k)$ *achieves its minimal value* $\sigma^2\theta_i^2/(\sigma^2+\xi_i^2\theta_i^2)$ *at* $k_i^*=k_i[\sigma^2/\xi_i^2\theta_i^2+2]>2\,k_i$ *with* k_i *as in Lemma* 1.

(ii) *The cross-over value from below, attaining equality in* $M_{SE}(\hat{\theta}_{Sk}^i)\leq M_{SE}(\hat{\theta}_L^i)$, *occurs at* $k_c^{i*}=2k_c^i(\xi_i^2\theta_i^2/(\xi_i^2\theta_i^2-\sigma^2))>2k_c^i$.

(iii) *If* $k_m^*\leq\min\{k_c^{1^*},k_c^{2^*},\ldots,k_c^{p*}\}$, *then* $\hat{\boldsymbol{\beta}}_S(k_m^*)$ *is M_{SE}-admissible with respect to* $\hat{\boldsymbol{\beta}}_L$.

(iv) *If* $k_M^*\geq\max\{k_c^{1^*},k_c^{2^*},\ldots,k_c^{p*}\}$, *then* $\hat{\boldsymbol{\beta}}_L$ *is M_{SE}-admissible with respect to* $\hat{\boldsymbol{\beta}}_S(k_M^*)$ *thereafter.*

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

Proof. See Appendix A.

Hoerl and Kennard (1970) gave $k_0<\min\{k_1,\ldots,k_p\}$ from Lemma 1 as sufficient for $M_{SE}(\hat{\boldsymbol{\beta}}_{Rk})<M_{SE}(\hat{\boldsymbol{\beta}}_L)$. An analog sufficient for $M_{SE}(\hat{\boldsymbol{\beta}}_{Sk})<M_{SE}(\hat{\boldsymbol{\beta}}_L)$ from Theorem 3(i) is $k_0^*<\min\{k_1^*,\ldots,k_p^*\}$. Unfortunately, both choices are conservative in excess, lying to the left of all the minima for the $M_{SE}(\hat{\theta}_{Rk}^i)$ and $M_{SE}(\hat{\theta}_{Sk}^i)$ curves, whereas Lemma 1(iii) and Theorem 3(iii) adjoin the rising branches on these curves up to their crossing with the constant *OLS* line. The user is assured of improvement in *MSE* of ridge estimators for $k \le \min\{k^i_{Rc}; 1 \le i \le p\}$,and of improvement in *MSE* for surrogate estimators for $k^* \le \min\{k^{i*}_c; 1 \le i \le p\}$. For a value $k \le \min\{k^i_c; 1 \le i \le p\}$ from Lemma 1, the value $2k$ will suffice in order that $M_{SE}(\hat{\boldsymbol{\beta}}_{Sk}) < M_{SE}(\hat{\boldsymbol{\beta}}_L)$ from Theorem 3(ii). Accordingly, the surrogate model offers a wider range of admissible values than ridge.

We turn next to concepts allied with the notion of orthogonality of linear systems and its consequences.

3. Orthogonal systems

3.1. Overview

The assertion, "At a certain value of k the system will stabilize and have the general characteristics of an orthogonal system" Hoerl and Kennard (1970, p. 65), remains vague in its failure to stipulate the "orthogonal characteristics" intended. We take this mandate to mean *properties* of solutions; specifically, point estimators, since hypothesis tests and confidence sets are argued in Obenchain (1977) to revert back to *OLS* when k is deterministic.

For reference, the system $\{\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\beta} = \boldsymbol{Z}'\boldsymbol{Y}\}$, with solution $\hat{\boldsymbol{\beta}}$ and dispersion $\mathrm{V}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\Sigma}$, is fully orthogonal if and only if the following properties are met:

P1: The singular values of $\boldsymbol{Z}$ are equal and $\boldsymbol{Z}'\boldsymbol{Z}$ is a scalar matrix.
P2: The condition number $c_1(\boldsymbol{Z}'\boldsymbol{Z}) = c_1(\boldsymbol{\Sigma})$ = 1.0.
P3: The *VIF*s are $\{VIF(\hat{\beta}_i) = 1.0;\ 1 \le i \le p\}$.
P4. Variances are *isotropic*, i.e., $\{\mathrm{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{c}'\boldsymbol{c};\ \boldsymbol{c} \in \mathbb{R}^p\}$; specifically, the fractions $\{\mathrm{Var}(\hat{\beta}_i)/\mathrm{tr}\,\mathrm{V}(\hat{\boldsymbol{\beta}}) = 1/p\,;1 \le i \le p\}$ are uniform.
P5: The probability contours of the distribution $\mathcal{L}(\hat{\boldsymbol{\beta}})$, if Gaussian, are spherical.

Ridge, and now surrogate regression, are intended to ameliorate cited flaws of *OLS* if ill-conditioned. Their merits necessarily rest on progress towards those ends. Since neither system can be construed to be orthogonal under ill-conditioning, users may expect, to some degree, discrepancies from the aforementioned benchmarks of orthogonal systems. These matters are considered in detail throughout the remainder of this study.

3.2. Conditioning and ellipticity

Let $\hat{\gamma} \in \mathbb{R}^p$ be centered at $\gamma_0 \in \mathbb{R}^p$ having $\mathrm{V}(\hat{\gamma}) = \boldsymbol{\Omega}$ with eigenvalues $\{\omega_1^2, \omega_2^2, \ldots, \omega_p^2\}$. Choose c such that

$$R(\gamma) = \{\gamma \in \mathbb{R}^p : (\hat{\gamma}-\gamma)'\boldsymbol{\Omega}^{-1}(\hat{\gamma}-\gamma) \le c^2\} \tag{3}$$

has unit volume. Then $R(\gamma)$ is an *ellipsoid of concentration* of Cramér (1947, p. 300ff.) having uniform measure, a distribution-free concept based on first and second moments, useful for gauging concentration efficiencies of vector estimators. If, in addition, $\mathcal{L}(\hat{\gamma})$ is Gaussian, then $\boldsymbol{\Omega}$ determines its elliptical density contours in $\mathbb{R}^p$. Rotating to standard position gives $\{\omega_1, \omega_2, \ldots, \omega_p\}$ as lengths of the semiprincipal axes. An ellipticity index, $W(\boldsymbol{\Omega}) = [\mathrm{tr}(\boldsymbol{\Omega})]^p/p^p|\boldsymbol{\Omega}|$ as in Mauchly (1940), serves to gauge the nonsphericity of contours of $R(\gamma)$, and of $\mathcal{L}(\hat{\gamma})$ if Gaussian, where $W(\boldsymbol{\Omega}) = 1.0$ at $\boldsymbol{\Omega} = \boldsymbol{I}_p$, and larger values quantify increasing divergence from sphericity. As these quantities figure prominently as marques of orthogonality, they are examined next with regard to surrogate and ridge systems. To these ends designate $c_1(\hat{\boldsymbol{\beta}}) = [c_1(\mathrm{V}(\hat{\boldsymbol{\beta}})]$ and $c_1^*(\hat{\boldsymbol{\beta}}) = [c_1(\hat{\boldsymbol{\beta}})]^{1/2}$;$W(\hat{\boldsymbol{\beta}}) = [W(\mathrm{V}(\hat{\boldsymbol{\beta}}))]$ and $El(\hat{\boldsymbol{\beta}}) = [W(\hat{\boldsymbol{\beta}})]^{1/2}$ as properties of $\hat{\boldsymbol{\beta}}$; and recall that $\{\mathrm{V}(\hat{\boldsymbol{\beta}}_{Sk}) = \boldsymbol{\Sigma}^S_k; k \ge 0\}$.

Theorem 4. *Consider surrogate estimators* $\{\hat{\boldsymbol{\beta}}_{Sk}; k \ge 0\}$, *together with root condition numbers* $c_1^*(\hat{\boldsymbol{\beta}}_{Sk})$ *and ellipticity indices* $El(\hat{\boldsymbol{\beta}}_{Sk})$, *and similarly for ridge estimators* $\{\hat{\boldsymbol{\beta}}_{Rk}; k \ge 0\}$. *Then*

(i) *Condition numbers* $\{c_1^*(\hat{\boldsymbol{\beta}}_{Sk}); k \ge 0\}$ *are monotone decreasing with increasing k and satisfy* $\lim_{k\to\infty} c_1^*(\hat{\boldsymbol{\beta}}_{Sk}) = 1$.
(ii) *Ellipticity indices* $\{El(\hat{\boldsymbol{\beta}}_{Sk}); k \ge 0\}$ *are monotone decreasing with increasing k and satisfy* $\lim_{k\to\infty} El(\hat{\boldsymbol{\beta}}_{Sk}) = 1$.
(iii) *Condition numbers* $\{c_1^*(\hat{\boldsymbol{\beta}}_{Rk}); k \ge 0\}$ *satisfy* $\lim_{k\to\infty} c_1^*(\hat{\boldsymbol{\beta}}_{Rk}) = c_1^*(\hat{\boldsymbol{\beta}}_L)$.
(iv) *Ellipticity indices* $\{El(\hat{\boldsymbol{\beta}}_{Rk}); k \ge 0\}$ *satisfy* $\lim_{k\to\infty} El(\hat{\boldsymbol{\beta}}_{Rk}) = [W(\boldsymbol{V}^{-1})]^{1/2} = [W(\boldsymbol{X}'\boldsymbol{X})]^{1/2}$.

Proof. See Appendix A.

With regard to condition numbers for dispersion, and to concentration and density contours, surrogate estimators increasingly resemble those from an orthogonal system as k evolves. These salutary properties fail for ridge; further evidence accrues through the case studies of Section 4, as listed subsequently in Table 4.

3.3. Variance inflation

Since neither system, ridge nor surrogate, is orthogonal if ill-conditioned, their *VIF*s necessarily exceed unity. Nonetheless, it is instructive to ask whether, as k evolves, *VIF*s proceed towards those from an orthogonal system. An affirmative answer follows in part, where *VIF*s for surrogate, but not ridge, are seen to decrease monotonically towards 1.0 with increasing k. That *VIF*s for $\hat{\boldsymbol{\beta}}_{Rk}$ are not monotone is demonstrated in the section following. For reference, *VIF*s for $\{(\hat{\beta}_S^i, \hat{\beta}_R^i), 1 \le i \le p\}$ are computed as

$$VIF(\hat{\beta}_S^i) = \frac{\sum_{j=1}^p q_{ij}^2/(\xi_j^2+k)}{1/\sum_{j=1}^p(\xi_j^2+k)q_{ij}^2}, \tag{4}$$

and

$$VIF(\hat{\beta}_R^i) = \sum_{j=1}^p \frac{\xi_j^2}{(\xi_j^2+k)^2} q_{ij}^2 \sum_{j=1}^p \frac{(\xi_j^2+k)^2}{\xi_j^2} q_{ij}^2, \tag{5}$$

where $\{\xi_1 \ge \cdots \ge \xi_p > 0\}$ are not all equal, and $\boldsymbol{Q} = [q_{ij}]$ is from the singular decomposition $\boldsymbol{X} = \boldsymbol{P}_1\boldsymbol{D}_\xi\boldsymbol{Q}'$. The principal findings follow.

Theorem 5. *Consider surrogate estimators* $\hat{\boldsymbol{\beta}}_{Sk}$ *with elements* $\{\hat{\beta}_S^i(k); 1 \le i \le p\}$, *together with ridge estimators* $\hat{\boldsymbol{\beta}}_{Rk}$ *with elements* $\{\hat{\beta}_R^i(k); 1 \le i \le p\}$.

(i) *The functions* $\{VIF(\hat{\beta}_S^i(k)); 1 \le i \le p\}$ *decrease monotonically with increasing* k *and satisfy* $\lim_{k\to\infty} VIF(\hat{\beta}_S^i(k)) = 1$.
(ii) *VIFs for* $\hat{\boldsymbol{\beta}}_{Rk}$ *satisfy* $\{\lim_{k\to\infty} VIF(\hat{\beta}_R^i(k)) = VIF(\hat{\beta}_L^i);\ 1 \le i \le p\}$.

Proof. See Appendix A.

4. Case studies

4.1. The setting

Records at n=17 U. S. Naval Hospitals comprise the Hospital Manpower Data as reported in Table 3.8 of Myers (1990, pp. 132–133), to include: monthly man-hours (Y); average daily patient load (X_1); monthly X-ray exposures (X_2); monthly occupied bed days (X_3); eligible population in the area ÷ 1000 (X_4); and average length of patients' stay in days (X_5). The basic model is

$$\{Y_i = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \varepsilon_i; 1 \le i \le n\}. \tag{6}$$

Following Hoerl and Kennard (1970), Marquardt (1970), Marquardt and Snee (1975), Myers (1990), and others, $\{\boldsymbol{Y} = \boldsymbol{Z\beta} + \boldsymbol{\varepsilon}\}$ is centered and scaled, with $\boldsymbol{Z}'\boldsymbol{Z}$ in correlation form and $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3, \beta_4, \beta_5]'$. Computations as reported utilize both PROC IML of SAS and the symbolic program Maple. Ill-conditioning is reflected in: (i) $\boldsymbol{D}_\xi = \mathrm{Diag}(2.048687, 0.816997, 0.307625, 0.201771, 0.007347)$; (ii) $c_1(\boldsymbol{Z}'\boldsymbol{Z}) = 77{,}754.86$; (iii) the maximal *VIF* as $V_1 = VIF(\hat{\beta}_1) = 9{,}595.68$; and (iv) other *VIF*s are listed at k=0 in Table 2. We next explore the ridge and surrogate solutions for these data as outlined previously.

4.2. Residuals

Citing residuals in studies of estimation, Theorem 2 registers for each $k > 0$ the critical inequality $R_{SS}(\hat{\boldsymbol{\beta}}_{Sk}) < R_{SS}(\hat{\boldsymbol{\beta}}_{Rk})$. Numerical values are given in Table 1 as k ranges over $[0, \infty)$, where they differ up to a factor of about 2 for $k \in [0, 1]$. Convergence to the same limit follows since $R_{SS}(\hat{\boldsymbol{\beta}}_k) \to \boldsymbol{U}'\boldsymbol{U} = \boldsymbol{Y}'\boldsymbol{Y}$ for both, as the estimators shrink towards zero.

Table 1
Values of $R_{SS}(\hat{\boldsymbol{\beta}}_{Sk}) \times 10^{-7}$ and $R_{SS}(\hat{\boldsymbol{\beta}}_{Rk}) \times 10^{-7}$ in the hospital manpower data as k evolves.

k	0.00	0.10	0.20	0.30	0.40	0.50	0.75	1.00	∞
$R_{SS}(\hat{\boldsymbol{\beta}}_{Sk})$	0.454	0.600	0.704	0.793	0.878	0.963	1.190	1.440	49.500
$R_{SS}(\hat{\boldsymbol{\beta}}_{Rk})$	0.454	0.802	1.010	1.210	1.420	1.650	2.300	3.040	49.500

4.3. Variance inflation

Section 3.1 identifies $\{VIF(\hat{\beta}_i)=1.0; 1\le i\le 5\}$ as characteristic of orthogonality. At issue is whether the ridge or surrogate system might stabilize towards unity of *VIF*s as k evolves. Table 2 tracks these quantities for $k\in[0,\infty)$. These factors clearly evolve erratically for ridge, but monotonically towards unity for surrogate systems, as certified in Theorem 5. It appears pathological that *VIF*s for ridge should turn back towards those of *OLS* as k becomes large, nor is there evidence otherwise that ridge solutions tend to stabilize in *VIF*s towards orthogonality.

4.4. Uniformity

Let $\{F_i=\mathrm{Var}(\hat{\beta}_i)/\mathrm{tr}\,\mathrm{V}(\hat{\boldsymbol{\beta}}); 1\le i\le 5\}$; Section 3.1 identifies their uniformity as characteristic of orthogonality. Fractions are displayed in Table 3 for the ridge and surrogate solutions, together with the uniformity index $UI(\hat{\boldsymbol{\beta}}_k)=\sum_{i=1}^{5}(F_i-1/5)^2/(\frac{4}{5})$ with range $[0,1]$. The value $\frac{4}{5}$ is the maximum value of the numerator at $\{1, 0, 0, 0, 0\}$. It is seen that $UI(\hat{\boldsymbol{\beta}}_{Rk})$ evolves erratically; its minimum for $k\in[0,1]$ is 0.037090 at k = 0.0019286. On the other hand, a graph in Maple software shows that $UI(\hat{\boldsymbol{\beta}}_{Sk})$ decreases monotonically for $k\in[0,1]$; the curves cross at k = 0.083651 with common value $UI(\hat{\boldsymbol{\beta}}_{Sk})=UI(\hat{\boldsymbol{\beta}}_{Rk})=0.049181$; and $UI(\hat{\boldsymbol{\beta}}_{Sk})<UI(\hat{\boldsymbol{\beta}}_{Rk})$ for $k\in(0.083651,1]$.

4.5. Conditioning and contours

Further characteristics in Section 3.1 are the condition number of $\mathrm{V}(\hat{\boldsymbol{\beta}})$; and spherical concentration ellipsoids of $\mathcal{L}(\hat{\boldsymbol{\beta}})$ under second moments, or spherical probability contours if Gaussian. Designate $c_1^*(\hat{\boldsymbol{\beta}})=[c_1(\mathrm{V}(\hat{\boldsymbol{\beta}}))]^{1/2}$ and $El(\hat{\boldsymbol{\beta}})=[W(\mathrm{V}(\hat{\boldsymbol{\beta}}))]^{1/2}$ as before. These are listed in Table 4 as k ranges over $[0,\infty)$. The index $El(\hat{\boldsymbol{\beta}}_L)=641{,}006$ at k=0

Table 2

Variance inflation factors $\{VIF_{R1},\ldots,VIF_{R5}\}$ for elements of $\hat{\boldsymbol{\beta}}_{Rk}=[\hat{\beta}_{R1},\ldots,\hat{\beta}_{R5}]'$, and $\{VIF_{S1},\ldots,VIF_{S5}\}$ for $\hat{\boldsymbol{\beta}}_{Sk}=[\hat{\beta}_{S1},\ldots,\hat{\beta}_{S5}]'$, as k ranges over $[0,\infty)$.

k	VIF_{R1}	VIF_{R2}	VIF_{R3}	VIF_{R4}	VIF_{R5}	VIF_{S1}	VIF_{S2}	VIF_{S3}	VIF_{S4}	VIF_{S5}
0.00	9596	7.941	8931	23.289	4.279	9596	7.941	8931	23.289	4.279
0.10	55.97	2.383	58.70	2.443	1.402	7.621	4.144	7.518	4.971	1.898
0.20	90.90	1.502	93.17	1.699	1.256	4.375	3.021	4.343	3.373	1.600
0.40	159.22	1.119	157.97	1.628	1.263	2.658	2.153	2.648	2.286	1.377
0.60	242.62	1.086	235.97	1.856	1.346	2.070	1.788	2.065	1.862	1.277
0.80	341.11	1.136	327.76	2.152	1.443	1.774	1.589	1.772	1.637	1.218
1.00	451.58	1.218	430.60	2.474	1.539	1.598	1.463	1.596	1.498	1.178
4.00	2367	2.724	2211	7.199	2.438	1.096	1.079	1.096	1.083	1.036
∞	9596	7.941	8931	23.289	4.279	1.000	1.000	1.000	1.000	1.000

Table 3

Fractions $\{F_i=\mathrm{Var}(\hat{\beta}_i)/\mathrm{trV}(\hat{\boldsymbol{\beta}}); 1\le i\le 5\}$ of total variance and their uniformity index $UI(\hat{\boldsymbol{\beta}})$ for ridge $\hat{\boldsymbol{\beta}}_{Rk}=[\hat{\beta}_{R1},\ldots,\hat{\beta}_{R5}]'$ and surrogate $\hat{\boldsymbol{\beta}}_{Sk}=[\hat{\beta}_{S1},\ldots,\hat{\beta}_{S5}]'$ solutions as k ranges over [0, 1].

k	F_1	F_2	F_3	F_4	F_5	$UI(\hat{\boldsymbol{\beta}}_k)$
	Ridge solutions					
0.00	0.5169	0.0004	0.4812	0.0013	0.0002	0.3734
0.10	0.0973	0.3146	0.1095	0.2880	0.1906	0.0496
0.20	0.0814	0.3017	0.0896	0.2514	0.2758	0.0562
0.40	0.0750	0.2638	0.0799	0.2133	0.3679	0.0781
0.60	0.0786	0.2402	0.0821	0.1963	0.4028	0.0892
0.80	0.0849	0.2262	0.0876	0.1880	0.4133	0.0903
1.00	0.0919	0.2175	0.0941	0.1838	0.4128	0.0859
	Surrogate solutions					
0.00	0.5169	0.0004	0.4812	0.0013	0.0002	0.3734
0.10	0.2914	0.1585	0.2875	0.1902	0.0726	0.0426
0.20	0.2618	0.1808	0.2599	0.2018	0.0957	0.0233
0.40	0.2390	0.1935	0.2381	0.2056	0.1238	0.0111
0.60	0.2284	0.1973	0.2279	0.2055	0.1409	0.0064
0.80	0.2221	0.1989	0.2217	0.2049	0.1524	0.0041
1.00	0.2179	0.1996	0.2176	0.2043	0.1607	0.0027

Table 4
Root condition numbers $c_1^*(\hat{\boldsymbol{\beta}})=[c_1(\mathrm{V}(\hat{\boldsymbol{\beta}}))]^{1/2}$ and elliptical index $El(\hat{\boldsymbol{\beta}})=[W(\mathrm{V}(\hat{\boldsymbol{\beta}}))]^{1/2}$ for ridge and surrogate solutions as k ranges over $[0,\infty)$.

k	$c_1^*(\hat{\boldsymbol{\beta}}_{Rk})$	$c_1^*(\hat{\boldsymbol{\beta}}_{Sk})$	$El(\hat{\boldsymbol{\beta}}_{Rk})$	$El(\hat{\boldsymbol{\beta}}_{Sk})$
0.00	278.846	278.846	641,006	641,006
0.10	21.5242	6.5535	18.0611	4.6869
0.20	28.4299	4.6882	18.1365	3.0124
0.40	41.6741	3.3899	22.4827	2.0820
0.60	52.6449	2.8275	28.7035	1.7406
0.80	60.6254	2.4992	35.7648	1.5598
1.00	66.6915	2.2797	43.4168	1.4469
4.00	136.079	1.4315	193.5504	1.0859
∞	278.8456	1.0000	1301.24	1.0000

reflects that elliptical contours in $\mathbb{R}^5$ for *OLS* are highly elongated and nearly degenerate, its semi-principal axes given by $\boldsymbol{D}_\xi=\mathrm{Diag}(2.048687, 0.816997, 0.307625, 0.201771, 0.007347)$.

It again appears pathological that $c_1^*(\hat{\boldsymbol{\beta}}_{Rk})$ for ridge should turn back towards those of *OLS* as k becomes large, nor is there evidence otherwise that ridge solutions stabilize towards orthogonality. Further computations show that $\min\{c_1^*(\hat{\boldsymbol{\beta}}_{Rk})\}=7.4463$ at k = 0.015. On the other hand, $c_1^*(\hat{\boldsymbol{\beta}}_{Sk})$ decreases monotonically for $k\in[0,\infty)$ as in Theorem 3; the curves cross at k = 0.03045 with common value $c_1^*(\hat{\boldsymbol{\beta}}_{Sk})=c_1^*(\hat{\boldsymbol{\beta}}_{Rk})=11.772$; and $c_1^*(\hat{\boldsymbol{\beta}}_{Sk})<c_1^*(\hat{\boldsymbol{\beta}}_{Rk})$ for $k\in(0.03045,\infty)$. Similarly, $\min\{El(\hat{\boldsymbol{\beta}}_{Rk})\}=7.9791$ occurs at k = 0.00268; $El(\hat{\boldsymbol{\beta}}_{Sk})$ decreases for $k\in[0,\infty)$ by Theorem 3; whereas the curves cross at k = 0.02135 with value $El(\hat{\boldsymbol{\beta}}_{Sk})=El(\hat{\boldsymbol{\beta}}_{Rk})=17.317$, and $El(\hat{\boldsymbol{\beta}}_{Sk})<El(\hat{\boldsymbol{\beta}}_{Rk})$ for $k\in(0.02135,\infty)$. It should be noted that the limit $El(\hat{\boldsymbol{\beta}}_{R\infty})=1301.24=[W((\mathrm{V}(\hat{\boldsymbol{\beta}}_L))^{-1})]^{1/2}$ as in Appendix A.

5. Conclusions

The ridge system $\{(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I}_p)\hat{\boldsymbol{\beta}}_{Rk}=\boldsymbol{X}'\boldsymbol{Y}\}$ is reexamined on the claim: "At a certain value of k the system will stabilize and have the general characteristics of an orthogonal system" Hoerl and Kennard (1970, p. 65), despite vagueness on the intended characteristics. Properties of *solutions* of orthogonal systems are identified here to include (i) unit condition numbers for dispersion matrices; (ii) unit values for Variance Inflation Factors; (iii) isotropic variances; and (iv) spherical contours for ellipsoids of concentration, or their densities if Gaussian. Conditioning arguments, set to account for ill-conditioning of $\boldsymbol{X}$ on both sides of the *OLS* equations, prompt the *generalized surrogate* system $\{(\boldsymbol{X}_K'\boldsymbol{X}_K)\hat{\boldsymbol{\beta}}_{SK}=\boldsymbol{X}_K'\boldsymbol{Y}\}$, and its study at $\boldsymbol{K}=k\boldsymbol{I}_p$ as alternative to ridge. Since neither the ridge nor surrogate system is orthogonal when $\boldsymbol{X}$ is ill-conditioned, this study examines whether their solutions do stabilize and tend towards orthogonality as k evolves.

Duality of generalized ridge and surrogate systems is demonstrated in Theorem 1. Theorem 2 asserts smaller residual sums of squares for $\hat{\boldsymbol{\beta}}_{Sk}$ than $\hat{\boldsymbol{\beta}}_{Rk}$ for each $k>0$. Theorem 3 identifies not only $k>0$ such that $M_{SE}(\hat{\boldsymbol{\beta}}_{Sk})<M_{SE}(\hat{\boldsymbol{\beta}}_L)$, but also values reverting back eventually to *OLS* where $M_{SE}(\hat{\boldsymbol{\beta}}_L)<M_{SE}(\hat{\boldsymbol{\beta}}_{Sk})$ for k sufficiently large. Moreover, if $\hat{\boldsymbol{\beta}}_R(k)$ is M_{SE}-admissible at $k=k_0$, then $\hat{\boldsymbol{\beta}}_S(k_0)$ is M_{SE}-admissible. Critical properties of $\{\hat{\boldsymbol{\beta}}_{Sk};\, k\geq 0\}$ decrease monotonically as k increases, to include conditioning of dispersion and ellipticity indices in Theorem 4, and Variance Inflation Factors in Theorem 5. Of these, none is monotone in the case of $\{\hat{\boldsymbol{\beta}}_{Rk};k\geq 0\}$, tending instead to diverge erratically with k, often reverting back towards values for *OLS*. In short, ridge solutions ultimately appear to exhibit the precise pathologies of *OLS* that they are intended to remedy. Section 4 illustrates these findings for the highly ill-conditioned Hospital Manpower Data.

Ridge regression continues apace. A burgeoning field of application is calibration in chemical engineering and analytical chemistry. Recent review articles include Geladi (2002), Kalivas (2005), and Sundberg (1999), to cite three sources.

Appendix A

A.1. Canonical form

Take $\boldsymbol{P}=[\boldsymbol{P}_1,\boldsymbol{P}_2]\in\mathcal{O}(n)$; partition $\boldsymbol{U}=\boldsymbol{P}'\boldsymbol{Y}$ as $\boldsymbol{U}=[\boldsymbol{U}_1',\boldsymbol{U}_2']'$, with $\boldsymbol{U}_1=\boldsymbol{P}_1'\boldsymbol{Y}\in\mathbb{R}^p$ and $\boldsymbol{U}_2=\boldsymbol{P}_2'\boldsymbol{Y}\in\mathbb{R}^{n-p}$; and let $\boldsymbol{\theta}=\boldsymbol{Q}'\boldsymbol{\beta}$ and $\boldsymbol{P}'\boldsymbol{\varepsilon}=\boldsymbol{\eta}=[\boldsymbol{\eta}_1',\boldsymbol{\eta}_2']'$. Then $\{\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\}$ transfers one-to-one into

$$\begin{bmatrix}\boldsymbol{U}_1\\ \boldsymbol{U}_2\end{bmatrix}=\begin{bmatrix}\boldsymbol{D}_\xi\boldsymbol{\theta}\\ \boldsymbol{0}\end{bmatrix}+\begin{bmatrix}\boldsymbol{\eta}_1\\ \boldsymbol{\eta}_2\end{bmatrix}, \tag{7}$$

and its errors into $\mathrm{E}(\boldsymbol{\eta})=\boldsymbol{0}$ and $\mathrm{V}(\boldsymbol{\eta})=\sigma^2\boldsymbol{I}_n$, preserving essential structure. Moreover, at $\hat{\boldsymbol{\beta}}=\boldsymbol{Q}\hat{\boldsymbol{\theta}}$, the residual sum of squares $R_{SS}(\hat{\boldsymbol{\beta}})$ becomes

$$R_{SS}(\hat{\boldsymbol{\beta}})=(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})=(\boldsymbol{U}_1-\boldsymbol{D}_\xi\hat{\boldsymbol{\theta}})'(\boldsymbol{U}_1-\boldsymbol{D}_\xi\hat{\boldsymbol{\theta}})+\boldsymbol{U}_2'\boldsymbol{U}_2. \tag{8}$$

Here $(\boldsymbol{U}_1 - \boldsymbol{D}_\xi \hat{\boldsymbol{\theta}}_L)$ vanishes identically at $\hat{\boldsymbol{\theta}}_L = \boldsymbol{D}_\xi^{-1}\boldsymbol{U}_1$, so that $R_{SS}(\hat{\boldsymbol{\beta}}_L) = \boldsymbol{U}_2' \boldsymbol{U}_2$, the minimum, and $S^2 = \boldsymbol{U}_2' \boldsymbol{U}_2/(n-p)$ is the *OLS* residual mean square.

A.2. Proofs: the principal results

Proof (*Theorem 1*). To connect generalized ridge and surrogate models, reexpress (1) and (2) in canonical form as $\{(\boldsymbol{D}_\xi + \boldsymbol{D}_\xi^{-1}\boldsymbol{K})\hat{\boldsymbol{\theta}}_{RK} = \boldsymbol{U}_1; (1')\}$ and $\{\boldsymbol{D}((\xi_i^2 + k_i^*)^{1/2})\hat{\boldsymbol{\theta}}_{SK} = \boldsymbol{U}_1; (2')\}$, respectively; equate matrices on the left and solve $\{k_i^* = 2k_i + k_i^2/\xi_i^2; 1 \le i \le p\}$. Clearly (k_i, k_i^*) correspond one-to-one since $\{k_i, k_i^*, \xi_i^2\}$ are positive, to give conclusion (i), so that k_i also may be solved in terms of k_i^*. Setting $\{k_1^* = k_2^* = \cdots = k_p^* = k^*\}$ identifies the ordinary surrogate model specializing generalized ridge, with parameters as in conclusion (ii). Conversely, ordinary ridge follows on setting $\{k_1 = k_2 = \cdots = k_p = k\}$, to give conclusion (iii). □

Proof (*Theorem 2*). From expression (8) it suffices to compare the quadratic forms $Q_1(\hat{\boldsymbol{\theta}}_{Sk}) = (\boldsymbol{U}_1 - \boldsymbol{D}_\xi\hat{\boldsymbol{\theta}}_{Sk})'(\boldsymbol{U}_1 - \boldsymbol{D}_\xi\hat{\boldsymbol{\theta}}_{Sk})$ and $Q_2(\hat{\boldsymbol{\theta}}_{Rk}) = (\boldsymbol{U}_1 - \boldsymbol{D}_\xi\hat{\boldsymbol{\theta}}_{Rk})'(\boldsymbol{U}_1 - \boldsymbol{D}_\xi\hat{\boldsymbol{\theta}}_{Rk})$, since $\boldsymbol{U}_2' \boldsymbol{U}_2$ is fixed. But $(\boldsymbol{U}_1 - \boldsymbol{D}_\xi\hat{\boldsymbol{\theta}}_{Sk}) = (\boldsymbol{U}_1 - \boldsymbol{D}(\xi_i/(\xi_i^2 + k)^{1/2})\boldsymbol{U}_1$, whereas $(\boldsymbol{U}_1 - \boldsymbol{D}_\xi\hat{\boldsymbol{\theta}}_{Rk}) = (\boldsymbol{U}_1 - \boldsymbol{D}(\xi_i^2/(\xi_i^2 + k))\boldsymbol{U}_1$. Components of $Q_1(\hat{\boldsymbol{\theta}}_{Sk})$ and $Q_2(\hat{\boldsymbol{\theta}}_{Rk})$ are $[1 - \xi_i/(\xi_i^2 + k)^{1/2}]U_i^2$ and $[1 - \xi_i^2/(\xi_i^2 + k)]U_i^2$, respectively. But since $0 < \xi_i^2/(\xi_i^2 + k) < \xi_i/(\xi_i^2 + k)^{1/2} < 1$ for $k > 0$, it follows that $Q_1(\hat{\boldsymbol{\theta}}_{Sk}) < Q_2(\hat{\boldsymbol{\theta}}_{Rk})$, to complete our proof. □

Proof (*Theorem 3*). $M_{SE}(\hat{\theta}_{Sk}^i)$ as given follows directly; setting $dM_{SE}(\hat{\theta}_{Sk}^i)/dk$ to zero and solving gives k^*, minimizing since $d^2M_{SE}(\hat{\theta}_{Sk}^i)/dk^2 > 0$; its minimal value is by substitution, to give conclusion (i). Equating $M_{SE}(\hat{\theta}_{Sk}^i)$ to $M_{SE}(\hat{\theta}_L^i) = \sigma^2/\xi_i^2$ and solving, gives k_c^{i*} as in conclusion (ii). Since components of $M_{SE}(\hat{\boldsymbol{\theta}}_{Sk}) - M_{SE}(\hat{\boldsymbol{\theta}}_L) = \sum_{i=1}^p [M_{SE}(\hat{\theta}_{Sk}^i) - M_{SE}(\hat{\theta}_L^i)]$ are all negative for $k_m^* \le \min\{k_c^{1^*}, k_c^{2^*}, \ldots, k_c^{p*}\}$ from (ii), conclusion (iii) follows directly. Similarly, that all components are positive for $k_M^* \ge \max\{k_c^{1^*}, k_c^{2^*}, \ldots, k_c^{p*}\}$, gives conclusion (iv). Conclusion (v) follows directly, to complete our proof. □

Proof (*Theorem 4*). Recall our conventions $c_1(\hat{\boldsymbol{\beta}}) = c_1(\mathrm{V}(\hat{\boldsymbol{\beta}}))$ and $W(\hat{\boldsymbol{\beta}}) = W(\mathrm{V}(\hat{\boldsymbol{\beta}}))$ as properties of $\hat{\boldsymbol{\beta}}$ and, by equivalence, that $c_1(\hat{\boldsymbol{\beta}}) = c_1(\hat{\boldsymbol{\theta}})$ and $W(\hat{\boldsymbol{\beta}}) = W(\hat{\boldsymbol{\theta}})$ in the corresponding canonical forms. Conclusion (i) follows from $c_1(\hat{\boldsymbol{\beta}}_{Sk}) = (\xi_1^2 + k)/(\xi_p^2 + k)$; and (iii) from

$$c_1(\hat{\boldsymbol{\beta}}_{Rk}) = \frac{\max\left\{\dfrac{\xi_i^2}{(\xi_i^2 + k)^2}; 1 \le i \le p\right\}}{\min\left\{\dfrac{\xi_i^2}{(\xi_i^2 + k)^2}; 1 \le i \le p\right\}} = \frac{\max\left\{\dfrac{k^2\xi_i^2}{(\xi_i^2 + k)^2}; 1 \le i \le p\right\}}{\min\left\{\dfrac{k^2\xi_i^2}{(\xi_i^2 + k)^2}; 1 \le i \le p\right\}} \to \frac{\xi_1^2}{\xi_p^2}.$$

The limit in (ii) follows from

$$W(\hat{\boldsymbol{\beta}}_{Sk}) = \frac{\left(\sum_{i=1}^p \dfrac{1}{\xi_i^2 + k}\right)^p}{p^p \prod_{i=1}^p \dfrac{1}{\xi_i^2 + k}} = \frac{\left(\sum_{i=1}^p \dfrac{k}{\xi_i^2 + k}\right)^p}{p^p \prod_{i=1}^p \dfrac{k}{\xi_i^2 + k}} \to 1.0 \quad \text{as } k \to \infty,$$

and the limit in (iv) from

$$W(\hat{\boldsymbol{\beta}}_{Rk}) = \frac{\left(\sum_{i=1}^p \dfrac{\xi_i^2}{(\xi_i^2 + k)^2}\right)^p}{p^p \prod_{i=1}^p \dfrac{\xi_i^2}{(\xi_i^2 + k)^2}} = \frac{\left(\sum_{i=1}^p \dfrac{k^2\xi_i^2}{(\xi_i^2 + k)^2}\right)^p}{p^p \prod_{i=1}^p \dfrac{k^2\xi_i^2}{(\xi_i^2 + k)^2}} \to \frac{(\sum_{i=1}^p \xi_i^2)^p}{p^p \prod_{i=1}^p \xi_i^2} = W(\boldsymbol{V}^{-1}).$$

Monotonocity in (ii) rests on showing that $dW(\hat{\boldsymbol{\beta}}_{Sk})/dk < 0$. Accordingly, let $\{c_i = 1/(\xi_i^2 + k); 1 \le i \le p\}$, and write

$$\frac{d}{dk}\log[W(\hat{\boldsymbol{\beta}}_{Sk})] = \frac{-1}{(c_1 + \cdots + c_p)^p}[p(c_1^2 + \cdots + c_p^2) - (c_1 + \cdots + c_p)^2],$$

which is negative, as the term in the brackets is positive by the Cauchy–Schwarz inequality. □

Proof (*Theorem 5*). $VIF(\hat{\beta}_S^i) = v_{ii}/w_{ii}^{-1}$ with $\boldsymbol{V} = (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I}_p)^{-1}$, and $\boldsymbol{W} = (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I}_p) = \boldsymbol{Q}\boldsymbol{D}((\xi_i^2+k))\boldsymbol{Q}'$, where $\{\xi_1 \geq \cdots \geq \xi_p > 0\}$ are not all equal. Expression (4) follows directly. Rewriting (4) as $VIF(\hat{\beta}_S^i) = \sum_{j,m=1}^{p}(q_{ij}^2/(\xi_j^2+k))((\xi_m^2+k)q_{im}^2)$, we compute

$$\frac{d}{dk}[VIF(\hat{\beta}_S^i)] = \sum_{j<m}^{p} q_{ij}^2 q_{im}^2 \left(\frac{(\xi_j^2-\xi_m^2)}{(\xi_j^2+k)^2} + \frac{(\xi_m^2-\xi_j^2)}{(\xi_m^2+k)^2} \right) = \sum_{j<m}^{p} q_{ij}^2 q_{im}^2 (\xi_j^2-\xi_m^2) \left(\frac{1}{(\xi_j^2+k)^2} - \frac{1}{(\xi_m^2+k)^2} \right) < 0;$$

each $VIF(\hat{\beta}_S^i(k))$ thus is monotone decreasing with increasing k. Limit properties follow on writing (4) as $VIF(\hat{\beta}_S^i) = (\sum_{i=1}^{p} kq_{ij}^2/(\xi_j^2+k))(\sum_{i=1}^{p}(\xi_j^2+k)q_{ij}^2/k) \to 1.0$ as $k \to \infty$ since $\boldsymbol{Q}$ is an orthogonal matrix, to establish conclusion (i). Conclusion (ii) follows on expressing (5) as

$$VIF(\hat{\beta}_R^i) = \sum_{i=1}^{p} \frac{k^2 \xi_j^2 q_{ij}^2}{(\xi_j^2+k)^2} \sum_{j=1}^{p} \frac{(\xi_j^2+k)^2 q_{ij}^2}{k^2 \xi_j^2} \to \sum_{j=1}^{p} \xi_j^2 q_{ij}^2 \sum_{j=1}^{p} \frac{q_{ij}^2}{\xi_j^2} = VIF(\hat{\beta}_L^i)$$

as $k \to \infty$, to complete our proof. □

References

Beaton, A.D., Rubin, D., Barone, J., 1976. The acceptability of regression solutions: another look at computational accuracy. J. Amer. Statist. Assoc. 71, 158–168.

Belsley, D.A., 1986. Centering the constant first-differencing and assessing conditioning. In: Belsley, D.A., Kuh, E. (Eds.), Model Reliability. MIT Press, Cambridge, pp. 117–153 (Chapter 5).

Berk, K.N., 1977. Tolerance and condition in regression computations. J. Amer. Statist. Assoc. 72, 863–866.

Bingham, C., Larntz, K., 1977. Comment. J. Amer. Statist. Assoc. 72, 97–102.

Cramér, H., 1947. Mathematical Methods of Statistics. Princeton University Press, New Jersey.

Davies, R.B., Hutton, B., 1975. The effect of errors in the independent variables in regression. Biometrika 62, 383–392.

Geladi, P., 2002. Some recent trends in the calibration literature. Chemomet. Intell. Lab. 60, 211–224.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12, 55–67.

Hoerl, A.E., Kennard, R.W., Baldwin, K.F., 1975. Ridge regression: some simulations. Commun. Statist. 4, 105–123.

Jensen, D.R., Ramirez, D.E., 2010. Tracking MSE efficiencies in ridge regression. Adv. Appl. Statist. Sci. 1, 381–398.

Kalivas, J.H., 2005. Multivariate calibration: an overview. Anal. Lett. 38, 2259–2279.

Lowerre, J.M., 1974. On the mean square error of parameter estimates for some biased estimators. Technometrics 16, 461–464.

Marquardt, D.W., 1970. Generalized inverses, ridge regression biased linear estimation and nonlinear estimation. Technometrics 12, 591–612.

Marquardt, D.W., Snee, R.D., 1975. Ridge regression in practice. J. Amer. Statist. Assoc. 29, 3–20.

Marshall, A.W., Olkin, I., 1979. Inequalities: Theory of Majorization and its Applications. Academic Press, New York.

Mauchly, J.W., 1940. Significance test for sphericity of a normal *n*-variate distribution. Ann. Math. Statist. 11, 204–209.

Myers, R.H., 1990. Classical and Modern Regression with Applications. PWS-KENT, Boston, MA.

Obenchain, R.L., 1977. Classical *F*-tests and confidence regions for ridge regression. Technometrics 19, 429–439.

Riley, J., 1955. Solving systems of linear equations with a positive definite, symmetric but possibly ill-conditioned matrix. Math. Tables Other Aids Comput. 9, 96–101.

Stewart, G.W., 1987. Collinearity and least squares regression. Statist. Sci. 2, 68–84.

Sundberg, R., 1999. Multivariate calibration—direct and indirect regression methodology. Scand. J. Statist. 26, 161–207.