

Nonorthogonality in Ill-Conditioned Systems[☆]

D. R. Jensen, D. E. Ramirez^{a,b}

^a*Department of Mathematics, Virginia Tech, Blacksburg, VA 24060*

^b*Department of Mathematics, University of Virginia, Charlottesville, VA 22904-4137*

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 motivated by 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 marks of orthogonality, to include conditioning of dispersion parameters, variance inflation factors, isotropy of variances, and sphericity of contours of the estimators. On 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 in the ridge scalar to those from orthogonal systems. Invariance considerations constrain the computations to models in canonical form. Case studies serve to illustrate the central issues.

Key words: Ill-conditioned models, ridge regression: properties, anomalies, surrogate models, hallmarks of orthogonality, asymptotics

2000 MSC: 62J07, 62J20

1. Introduction

Given $\{\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\}$ of full rank with zero-mean, uncorrelated, and homoscedastic errors, the p equations $\{\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}\}$ yield the Ordinary Least Squares (*OLS*) estimators $\widehat{\boldsymbol{\beta}}_L$ as unbiased with dispersion matrix $V(\widehat{\boldsymbol{\beta}}_L) = \sigma^2\mathbf{V}$, where $\mathbf{V} = [v_{ij}] = (\mathbf{X}'\mathbf{X})^{-1}$. Near-dependency among the

[☆]Research supported in part by the Department of Mathematics, University of Virginia.
Email address: djensen@vt.edu, der@virginia.edu (D. R. Jensen, D. E. Ramirez)

columns of \mathbf{X} , as *ill-conditioning*, “causes crucial elements of $\mathbf{X}'\mathbf{X}$ to be large and unstable,” “creating inflated variances,” and elements of $\widehat{\boldsymbol{\beta}}_L$ that are “very sensitive to small changes in \mathbf{X} ,” [2; p.119]. Specifically, the *condition number* $c_1(\mathbf{X}'\mathbf{X})$ is the ratio of largest to smallest eigenvalues, and the *Variance Inflation Factors* (VIFs) of $\widehat{\boldsymbol{\beta}}_L = [\widehat{\beta}_{L1}, \dots, \widehat{\beta}_{Lp}]'$ are $\{VIF(\widehat{\beta}_{Lj}) = v_{jj}/w_{jj}^{-1}; 1 \leq j \leq p\}$ with $\mathbf{W} = \mathbf{X}'\mathbf{X}$, *i.e.*, ratios of actual to “ideal” variances had the columns of \mathbf{X} been orthogonal. Scaling columns of \mathbf{X} to unit lengths and ordering $\{VIF(\widehat{\beta}_{Lj}) = v_{jj}; 1 \leq j \leq p\}$ as $\{V_1 \geq V_2 \geq \dots \geq V_p\}$, V_1 is identified in [17] as “the best single measure of the conditioning of the data,” whereas the connection $V_1 \leq c_1(\mathbf{X}'\mathbf{X}) \leq p(V_1 + \dots + V_p)$ is drawn in [3]; see also [1], [6], [16], and [23]. The *singular decomposition* $\mathbf{X} = \mathbf{P}_1 \mathbf{D}_\xi \mathbf{Q}'$ yields a *canonical* form taking $\{\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\} \rightarrow \{\mathbf{P}'_1 \mathbf{Y} = \mathbf{D}_\xi \boldsymbol{\theta} + \boldsymbol{\eta}\}$ where $\boldsymbol{\theta} = \mathbf{Q}'\boldsymbol{\beta}$, and $\boldsymbol{\eta} = \mathbf{P}'_1 \boldsymbol{\epsilon}$, such that \mathbf{D}_ξ is diagonal, \mathbf{Q} is orthogonal, and $\mathbf{P}'_1 \mathbf{P}_1 = \mathbf{I}_p$. This reduction is central to much that follows, supporting by invariance a number of properties that otherwise might be obscured.

Among palliatives for ill-conditioning, the *ridge system* $\{(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}; k \geq 0\}$ traces to Hoerl and Kennard [9] and their two-fold assertions: (i) *In lieu* of ill-conditioned *OLS* estimators having excessive lengths, the ridge solutions $\{\widehat{\boldsymbol{\beta}}_{Rk}; k > 0\}$ are LaGrange minimizers subject to length constraints. (ii) “At a certain value of k the system will stabilize and have the general characteristics of an orthogonal system” [9; p.65]. Reprinted as [11] in the “Special 40th Anniversary Issue” of *Technometrics*, this work is considered one of several “monumental contributions of the regression classics,” having “dramatically changed the practice of regression analysis as it is advocated today.” Nonetheless, the companion articles [9], [10] are deemed to be “among the most controversial written on the practice of regression analysis,” and despite wide usage, remain in 2000 “as controversial as when it was first introduced.” Quotations are from Gunst [8; p.62].

In partial explanation, it is shown in [13] that $\{\widehat{\boldsymbol{\beta}}_{Rk}; k > 0\}$ are *not* constrained LaGrange solutions, thereby undermining tenets of a substantial literature. The present article offers further insight: At issue are stability and the semblance to orthogonality of the ridge system itself, and whether, in contrast to *OLS*, such semblance might confer desirable features onto ridge solutions.

In perspective, the reach of ridge regression is seen in a September, 2008, Keyword/Title search of the *Current Index to Statistics (CIS)*, giving 487 “Hits.” The six five-year average numbers of citations for 1976–2005 showed

remarkable stability, with a sample mean of 14.97 and standard deviation of 1.54. In contrast, 24 citations are given for 1970–1975, and 13 for 2006–September 2008. A burgeoning field of application, largely missed by *CIS*, is calibration in chemical engineering and analytical chemistry. Here ridge regression often is advocated as the method of choice among biased estimators if ill-conditioned. Recent review articles include [7], [14], and [24], for example. An outline follows.

Supporting developments make up Section 2. Section 3 cites the conditioning of linear systems to motivate *surrogate* estimators $\{\widehat{\beta}_{\text{sk}}; k \geq 0\}$ alternative to ridge, improving *OLS* in mean square under conditions of Theorem 1; and for each $k > 0$, having uniformly smaller residual sums of squares than ridge by Theorem 2. Section 4 identifies marks of orthogonality, to include key condition numbers and ellipticity indices in Theorem 3, and variance inflation factors in Theorem 4. Of these, ridge exhibits erratic behavior, typically diverging 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 5 illustrate the central issues; Section 6 concludes with a brief summary; and essential asymptotics are deferred to an Appendix.

2. Preliminaries

2.1. Notation

Spaces of note include \mathbb{R}^n , \mathbb{R}_+^n , \mathbb{F}_{np} , and \mathbb{S}_p as Euclidean n -space, its positive orthant, the real $(n \times p)$ matrices of rank $p < n$, and the real symmetric $(p \times p)$ matrices; and $\mathcal{O}(n)$ is the real orthogonal group acting on \mathbb{R}^n . The term *unitary invariance* refers to functions of $\mathbf{X} \in \mathbb{F}_{np}$ invariant under left- and right-unitary operators. The transpose, inverse, trace, and determinant of $\mathbf{A} \in \mathbb{F}_{pp}$ are \mathbf{A}' , \mathbf{A}^{-1} , $\text{tr}(\mathbf{A})$, and $|\mathbf{A}|$. Special arrays are the null matrix $\mathbf{0}$; the identity \mathbf{I}_p ; and the diagonal matrix $\mathbf{D}_a = \mathbf{D}(a_i) = \text{Diag}(a_1, \dots, a_p)$. The *singular decomposition* of $\mathbf{X} \in \mathbb{F}_{np}$ is $\mathbf{X} = \mathbf{P}\mathbf{D}\mathbf{Q}'$, where $\mathbf{P} \in \mathcal{O}(n)$, $\mathbf{D}' = [\mathbf{D}_\xi, \mathbf{0}]$ with $\mathbf{D}_\xi = \text{Diag}(\xi_1, \dots, \xi_p)$, and $\mathbf{Q} \in \mathcal{O}(p)$. Equivalently, partition $\mathbf{P} = [\mathbf{P}_1, \mathbf{P}_2]$ with $\mathbf{P}_1 \in \mathbb{F}_{np}$ such that $\mathbf{P}_1' \mathbf{P}_1 = \mathbf{I}_p$, and write $\mathbf{X} = \mathbf{P}_1 \mathbf{D}_\xi \mathbf{Q}'$. Here elements of \mathbf{D}_ξ are the ordered *singular values* $\{\xi_1 \geq \xi_2 \geq \dots \geq \xi_p > 0\}$ of \mathbf{X} ; and the columns of $\mathbf{P}_1 = [\mathbf{p}_1, \dots, \mathbf{p}_p]$ and of $\mathbf{Q} = [\mathbf{q}_1, \dots, \mathbf{q}_p]$ comprise the *left-* and *right-singular vectors* of \mathbf{X} . By $\mathcal{S}p(\mathbf{Z}) \subset \mathbb{R}^n$ is meant the linear column span of $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_k]$.

The mean, dispersion matrix, and distribution of $\mathbf{Y} \in \mathbb{R}^n$ are designated as $E(\mathbf{Y}) = \boldsymbol{\mu}$, $V(\mathbf{Y}) = \boldsymbol{\Sigma}$, and $\mathcal{L}(\mathbf{Y})$. Of note on \mathbb{R}_+^1 are chi-squared distributions $\chi^2(\nu, \lambda)$, having (ν, λ) as degrees of freedom and noncentrality parameter; and on \mathbb{R}^n , the Gaussian law $\mathcal{L}(\mathbf{Y}) = N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Here *iid* refers to independent and identically distributed random elements. For models having second moments, identify $\{V(\widehat{\boldsymbol{\beta}}_{Rk}) = \boldsymbol{\Sigma}_k^R; k \geq 0\}$ and $\{V(\widehat{\boldsymbol{\beta}}_{Sk}) = \boldsymbol{\Sigma}_k^S; k \geq 0\}$. On occasion $\widehat{\boldsymbol{\beta}}_{Rk}$ and $\widehat{\boldsymbol{\beta}}_{Sk}$ are designated also as $\widehat{\boldsymbol{\beta}}_R(k)$ and $\widehat{\boldsymbol{\beta}}_S(k)$. Taking $\widetilde{\boldsymbol{\beta}}$ to estimate $\boldsymbol{\beta}$ with bias $B(\widetilde{\boldsymbol{\beta}}) = E(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (\boldsymbol{\beta}_0 - \boldsymbol{\beta})$, its Mean Square Error (M_{SE}) is $M_{SE}(\widetilde{\boldsymbol{\beta}}) = \text{tr } V(\widetilde{\boldsymbol{\beta}}) + (\boldsymbol{\beta}_0 - \boldsymbol{\beta})'(\boldsymbol{\beta}_0 - \boldsymbol{\beta})$, to quantify the trade-off of variance for bias under squared error loss.

2.2. Canonical Form

Specify $\{\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; E(\boldsymbol{\epsilon}) = \mathbf{0}, V(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}_n\}$, where the span $\mathcal{S}p(\mathbf{X})$, and $\mathcal{S}p^\perp(\mathbf{X})$, are known as *Regressor* and *Error* spaces in \mathbb{R}^n . A canonical version is expeditious. The singular decomposition $\mathbf{X} = \mathbf{P}_1\mathbf{D}_\xi\mathbf{Q}'$, with $\mathbf{P} = [\mathbf{P}_1, \mathbf{P}_2] \in \mathcal{O}(n)$ and $\mathbf{Q} \in \mathcal{O}(p)$, determines that $\mathbf{X} \in \mathcal{S}p(\mathbf{P}_1)$ and $\mathcal{S}p^\perp(\mathbf{P}_1) = \mathcal{S}p(\mathbf{P}_2)$ in \mathbb{R}^n . Since the Residual Sum of Squares (R_{SS}) for $\widehat{\boldsymbol{\beta}}$ is $R_{SS}(\widehat{\boldsymbol{\beta}}) = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'\mathbf{P}\mathbf{P}'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$, it suffices to consider the canonical residuals $(\mathbf{P}'\mathbf{Y} - \mathbf{D}_\xi\mathbf{Q}'\widehat{\boldsymbol{\beta}})$. Accordingly, partition $\mathbf{U} = \mathbf{P}'\mathbf{Y}$ as $\mathbf{U} = [\mathbf{U}_1', \mathbf{U}_2']'$, with $\mathbf{U}_1 = \mathbf{P}_1'\mathbf{Y} \in \mathbb{R}^p$ and $\mathbf{U}_2 = \mathbf{P}_2'\mathbf{Y} \in \mathbb{R}^{n-p}$; and let $\boldsymbol{\theta} = \mathbf{Q}'\boldsymbol{\beta}$ and $\mathbf{P}'\boldsymbol{\epsilon} = \boldsymbol{\eta} = [\boldsymbol{\eta}_1', \boldsymbol{\eta}_2']'$. Then $\{\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\}$ transfers one-to-one into

$$\begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{D}_\xi\boldsymbol{\theta} \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{bmatrix}, \quad (1)$$

and its errors into $E(\boldsymbol{\eta}) = \mathbf{0}$ and $V(\boldsymbol{\eta}) = \sigma^2\mathbf{I}_n$, preserving essential structure. Moreover, at $\widehat{\boldsymbol{\beta}} = \mathbf{Q}\widehat{\boldsymbol{\theta}}$, $R_{SS}(\widehat{\boldsymbol{\beta}})$ becomes

$$R_{SS}(\widehat{\boldsymbol{\beta}}) = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = (\mathbf{U}_1 - \mathbf{D}_\xi\widehat{\boldsymbol{\theta}})'(\mathbf{U}_1 - \mathbf{D}_\xi\widehat{\boldsymbol{\theta}}) + \mathbf{U}_2'\mathbf{U}_2, \quad (2)$$

where $(\mathbf{U}_1 - \mathbf{D}_\xi\widehat{\boldsymbol{\theta}})$ vanishes identically at $\widehat{\boldsymbol{\theta}}_L = \mathbf{D}_\xi^{-1}\mathbf{U}_1$, so that $R_{SS}(\widehat{\boldsymbol{\beta}}_L) = \mathbf{U}_2'\mathbf{U}_2$, the minimum. Moreover, $S^2 = \mathbf{U}_2'\mathbf{U}_2/(n-p)$ is the *OLS* residual mean square; $E(S^2) = \sigma^2$; and $\mathcal{L}(\mathbf{U}_2'\mathbf{U}_2/\sigma^2)$ is central $\chi^2(n-p, 0)$ under Gaussian errors, since $\{U_{p+1}, \dots, U_n\}$ are *iid* $N_1(0, \sigma^2)$; whereas $\{\widehat{\boldsymbol{\beta}}_L, \widehat{\boldsymbol{\beta}}_{Rk}, \widehat{\boldsymbol{\beta}}_{Sk}\}$, as functions of \mathbf{U}_1 , are independent of S^2 .

3. Surrogate Models

3.1. Basics

Condition numbers here conform to the *unitarily invariant* matrix norms of von Neumann, and thus in particular, $c_1(\mathbf{X}'\mathbf{X})$; see [18] and related references. That these are diminished for $k > 0$ through $\mathbf{X}'\mathbf{X} \rightarrow (\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)$, as shown in [22], is cited in [18; p.273] to justify ridge regression, *i.e.*, taking $\{\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}\} \rightarrow \{(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}\}$. Moreover, the enhanced conditioning of $c_1(\cdot)$ follows directly on comparing $c_1(\mathbf{X}'\mathbf{X}) = \xi_1^2/\xi_p^2$ with $c_1(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p) = (\xi_1^2 + k)/(\xi_p^2 + k)$, decreasing with increasing k . Unfortunately, such arguments are incomplete in linear inference. For while the mapping $\{\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}\} \rightarrow \{(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}\}$ does indeed serve to ameliorate ill-conditioning of $\mathbf{X}'\mathbf{X}$ on the left, ill-conditioning intrinsic to \mathbf{X} itself persists on the right. We seek to address this oversight as follows.

Take $\{\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}\} \rightarrow \{(\mathbf{X}'\mathbf{X} + \mathbf{C})\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}\}$, such that \mathbf{C} is positive definite, commuting with $\mathbf{X}'\mathbf{X}$, and orthogonally congruent to $\mathbf{K} = \text{Diag}(k_1, \dots, k_p)$. This is *generalized* ridge regression, taking the canonical OLS equations $\{\mathbf{D}_\xi^2\boldsymbol{\theta} = \mathbf{D}_\xi\mathbf{P}'_1\mathbf{Y}\}$ into $\{(\mathbf{D}_\xi^2 + \mathbf{K})\boldsymbol{\theta} = \mathbf{D}_\xi\mathbf{P}'_1\mathbf{Y}\}$. See [4], [9], [12], and [15], for example. But this amounts to modifying the singular decomposition $\mathbf{X} = \mathbf{P}_1\mathbf{D}_\xi\mathbf{Q}' \rightarrow \mathbf{X}_K = \mathbf{P}_1\mathbf{D}((\xi_i^2 + k_i)^{\frac{1}{2}})\mathbf{Q}'$ with $\mathbf{D}((\xi_i^2 + k_i)^{\frac{1}{2}}) = \text{Diag}((\xi_1^2 + k_1)^{\frac{1}{2}}, \dots, (\xi_p^2 + k_p)^{\frac{1}{2}})$, for then $\mathbf{X}'_K\mathbf{X}_K = \mathbf{Q}(\mathbf{D}_\xi^2 + \mathbf{K})\mathbf{Q}' = (\mathbf{X}'\mathbf{X} + \mathbf{C})$. To further our goal to improve conditioning on both sides of $\{\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}\}$, we take the *model* $\{\mathbf{Y} = \mathbf{X}_K\boldsymbol{\beta} + \boldsymbol{\epsilon}\}$ as a *surrogate* for the ill-conditioned model $\{\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\}$ itself, as in the following.

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

In canonical form, the generalized ridge and surrogate models are

$$\{(\mathbf{D}_\xi^2 + \mathbf{K})\widehat{\boldsymbol{\theta}}_{RK} = \mathbf{D}_\xi\mathbf{U}_1\} \text{ and} \quad (3)$$

$$\{(\mathbf{D}_\xi^2 + \mathbf{K}^*)\widehat{\boldsymbol{\theta}}_{SK^*} = \mathbf{D}((\xi_i^2 + k_i^*)^{\frac{1}{2}})\mathbf{U}_1\}, \quad (4)$$

respectively, where $\mathbf{K} = \text{Diag}(k_1, \dots, k_p)$ and $\mathbf{K}^* = \text{Diag}(k_1^*, k_2^*, \dots, k_p^*)$ are distinguished for reasons to follow. Properties of $\widehat{\boldsymbol{\beta}}_{RK}$ and $\widehat{\boldsymbol{\beta}}_{SK}$ follow directly

through $\widehat{\boldsymbol{\beta}}_{RK} = \mathbf{Q}\widehat{\boldsymbol{\theta}}_{RK}$ and $\widehat{\boldsymbol{\beta}}_{SK^*} = \mathbf{Q}\widehat{\boldsymbol{\theta}}_{SK^*}$ from Section 2.2, specializing at $\mathbf{K} = k\mathbf{I}_p$ and $\mathbf{K}^* = k^*\mathbf{I}_p$ for the ordinary ridge and surrogate models.

3.2. Validation

Values k_0 such that $M_{SE}[\widehat{\boldsymbol{\beta}}_R(k_0)] < M_{SE}(\widehat{\boldsymbol{\beta}}_L)$ are given in [9], *i.e.*, ridge solutions M_{SE} -efficient relative to *OLS*, and thus worthy of further pursuit. Designate these as M_{SE} -admissible. Parallel results for $\widehat{\boldsymbol{\beta}}_{S_k}$, if available, are equally germane. We next demonstrate equivalence of the generalized ridge and surrogate models; each ordinary model specializes from the generalized version of the other; and conditions for $\widehat{\boldsymbol{\beta}}_{S_k}$ to be M_{SE} -admissible are given. Since $(\widehat{\boldsymbol{\beta}}_{SK} - \boldsymbol{\beta}) = \mathbf{Q}(\widehat{\boldsymbol{\theta}}_{SK} - \boldsymbol{\theta})$ and \mathbf{Q} is orthogonal, it suffices that $M_{SE}(\widehat{\boldsymbol{\beta}}_{SK}) = M_{SE}(\widehat{\boldsymbol{\theta}}_{SK}) = \text{E}(\widehat{\boldsymbol{\theta}}_{SK} - \boldsymbol{\theta})'(\widehat{\boldsymbol{\theta}}_{SK} - \boldsymbol{\theta})$, and similarly for $M_{SE}(\widehat{\boldsymbol{\beta}}_{RK})$. Essential results are summarized as follows.

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

(i) *The generalized systems (3) and (4) are equivalent; elements of \mathbf{K} and \mathbf{K}^* relate one-to-one through $\{k_i^* = 2k_i + \frac{k_i^2}{\xi_i^2}; 1 \leq i \leq p\}$.*

(ii) (a) *Ordinary surrogate, with solution $\widehat{\boldsymbol{\theta}}_{SK^*}$, is generalized ridge with parameters $\mathbf{K} = \text{Diag}(k_1, \dots, k_p)$ from $\{k^* = 2k_i + \frac{k_i^2}{\xi_i^2}; 1 \leq i \leq p\}$; whereas (b) ordinary ridge, with solution $\widehat{\boldsymbol{\theta}}_{Rk}$, is generalized surrogate with $\mathbf{K}^* = \text{Diag}(k_1^*, \dots, k_p^*)$ from $\{k_i^* = 2k + \frac{k^2}{\xi_i^2}; 1 \leq i \leq p\}$.*

(iii) *Solutions $\mathbf{K}^\dagger = \text{Diag}(k_1^\dagger, \dots, k_p^\dagger)$ minimizing $M_{SE}(\widehat{\boldsymbol{\beta}}_{SK})$ are given by*

$$\left\{ k_i^\dagger = \frac{\sigma^2}{\theta_i^2} \left[\frac{\sigma^2}{\xi_i^2 \theta_i^2} + 2 \right]; 1 \leq i \leq p \right\}. \quad (5)$$

(iv) *There are values $k^\dagger > 0$ such that $M_{SE}(\widehat{\boldsymbol{\beta}}_{SK^\dagger})$ is less than for *OLS*, namely, $k^\dagger \leq \min\{k_1^\dagger, \dots, k_p^\dagger\}$ as given in (5), so that $\widehat{\boldsymbol{\beta}}_{SK^\dagger}$ is M_{SE} -admissible.*

(v) *$\widehat{\boldsymbol{\beta}}_R(k)$ is M_{SE} -admissible at $k_1 \leq k_0 = \min\{\frac{\sigma^2}{\theta_i^2}; 1 \leq i \leq p\}$; then $\widehat{\boldsymbol{\beta}}_S(k_1)$ is M_{SE} -admissible.*

Proof. To connect generalized ridge and surrogate models, rewrite (3) and (4) as

$$\{(\mathbf{D}_\xi + \mathbf{D}_\xi^{-1}\mathbf{K})\widehat{\boldsymbol{\theta}}_{RK} = \mathbf{U}_1\} \text{ and} \quad (6)$$

$$\{(\mathbf{D}((\xi_i^2 + k_i^*)^{1/2}))\widehat{\boldsymbol{\theta}}_{SK} = \mathbf{U}_1\}; \quad (7)$$

equate matrices on the left of (6) and (7) to give

$$\{k_i^* = 2k_i + \frac{k_i^2}{\xi_i^2}; 1 \leq i \leq p\}; \quad (8)$$

and verify that (k_i, k_i^*) correspond one-to-one since $\{k_i, k_i^*, \xi_i^2\}$ are positive, as in conclusion (i). Thus k_i likewise may be solved in terms of k_i^* . Setting $\{k_1^* = k_2^* = \dots = k_p^* = k^*\}$ in (8) identifies the ordinary surrogate model on specializing generalized ridge, with parameters as in (ii)(a). Conversely, ordinary ridge follows on setting $\{k_1 = k_2 = \dots = k_p = k\}$ in (8), as in conclusion (ii)(b). To continue, M_{SE} –“optimal” choices for the ridge scalars in (3) are known to be $\{k_i = \sigma^2/\theta_i^2; 1 \leq i \leq p\}$, and $\widehat{\beta}_R(k_0)$ to be M_{SE} –admissible for $k_0 \leq \min\{k_1, \dots, k_p\}$; see [9] for details. Since the models correspond one-to-one, surrogate values are found on substituting σ^2/θ_i^2 for k_i into (8), to give conclusion (iii). Conclusion (iv) follows directly as in [9]. Finally, since $k_0 \leq \min\{k_1, \dots, k_p\}$ implies $k_0 \leq \min\{k_1^\dagger, \dots, k_p^\dagger\}$ from conclusion (iv), the M_{SE} –admissibility of $\widehat{\beta}_R(k_1)$, as shown in [9], implies that of $\widehat{\beta}_S(k_1)$ as in conclusion (v), to complete our proof. \square

Direct computations show that $M_{SEi}(\widehat{\theta}_{RK})$ and $M_{SEi}(\widehat{\theta}_{SK})$ achieve identical minima of $\sigma^2\theta_i^2/(\sigma^2 + \xi_i^2\theta_i^2)$ at the respective values $k_i = \sigma^2/\theta_i^2$ and $k_i^* = k[(\sigma^2/\xi_i^2\theta_i^2) + 2]$. Initial rates of change $[dM_{SEi}(k)/dk]|_{k=0} = -\sigma^2/\xi_i^4 < 0$ offer further insight. We subsequently consider ordinary surrogate models as alternatives to ridge where, on occasion, we may compare $\widehat{\beta}_R(k)$ and $\widehat{\beta}_S(k)$ point-wise at $k = k_1$. Conclusion (v) then assures that if $\widehat{\beta}_R(k_1)$ is M_{SE} –admissible from the considerable literature on ridge regression, then $\widehat{\beta}_S(k_1)$ cannot be supplanted by a superior *OLS* estimator.

3.3. Residual Comparisons

Claims that solutions are optimal can be registered neither for $\{\widehat{\beta}_{Rk}; k > 0\}$ nor for $\{\widehat{\beta}_{Sk}; k > 0\}$, except that $Q(\beta) = (\mathbf{Y} - \mathbf{X}_k\beta)'(\mathbf{Y} - \mathbf{X}_k\beta)$ is minimized at $\widehat{\beta}_{Sk}$. Nonetheless, R_{SS} often is at issue in considering alternatives to *OLS*. The following demonstration is germane.

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

Proof. It suffices to compare the quadratic form $Q_1(\widehat{\theta}_{Sk}) = (\mathbf{U}_1 - \mathbf{D}_\xi\widehat{\theta}_{Sk})'(\mathbf{U}_1 - \mathbf{D}_\xi\widehat{\theta}_{Sk})$ with $Q_2(\widehat{\theta}_{Rk}) = (\mathbf{U}_1 - \mathbf{D}_\xi\widehat{\theta}_{Rk})'(\mathbf{U}_1 - \mathbf{D}_\xi\widehat{\theta}_{Rk})$ at (2), since $\mathbf{U}'_2\mathbf{U}_2$

is fixed. But $(\mathbf{U}_1 - \mathbf{D}_\xi \widehat{\boldsymbol{\theta}}_{S_k}) = (\mathbf{U}_1 - \mathbf{D}(\xi_i/(\xi_i^2 + k)^{\frac{1}{2}}))\mathbf{U}_1$, whereas $(\mathbf{U}_1 - \mathbf{D}_\xi \widehat{\boldsymbol{\theta}}_{Rk}) = (\mathbf{U}_1 - \mathbf{D}(\xi_i^2/(\xi_i^2 + k)))\mathbf{U}_1$. Components of $Q_1(\widehat{\boldsymbol{\theta}}_{S_k})$ and $Q_2(\widehat{\boldsymbol{\theta}}_{Rk})$ are $[1 - \xi_i/(\xi_i^2 + k)^{\frac{1}{2}}]U_i^2$ and $[1 - \xi_i^2/(\xi_i^2 + k)]U_i^2$, respectively. But since $0 < \xi_i^2/(\xi_i^2 + k) < \xi_i/(\xi_i^2 + k)^{\frac{1}{2}} < 1$ for $k > 0$, it follows that $Q_1(\widehat{\boldsymbol{\theta}}_{S_k}) < Q_2(\widehat{\boldsymbol{\theta}}_{Rk})$, to complete our proof. \square

In short, residual sums of squares are smaller for surrogate than for ridge, point-wise for each $k > 0$. We turn next to concepts allied with the notion of orthogonality of linear systems and its consequences.

4. Orthogonal Systems

4.1. Overview

The assertion, “At a certain value of k the system will stabilize and have the general characteristics of an orthogonal system” [9; 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 [21] to revert back to *OLS* when k is deterministic.

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

P1: The singular values of \mathbf{Z} are equal and $\mathbf{Z}'\mathbf{Z}$ is a scalar matrix;

P2: The condition number $c_1(\mathbf{Z}'\mathbf{Z}) = c_1(\boldsymbol{\Sigma}) = 1.0$;

P3: The *VIFs* are $\{VIF(\widehat{\beta}_i) = 1.0; 1 \leq i \leq p\}$;

P4. Variances are *isotropic*, i.e., $\{\text{Var}(\mathbf{c}'\widehat{\boldsymbol{\beta}}) = \sigma^2 \mathbf{c}'\mathbf{c}; \mathbf{c} \in \mathbb{R}^p\}$; specifically, the fractions $\{\text{Var}(\widehat{\beta}_i)/\text{tr}V(\widehat{\boldsymbol{\beta}}) = 1/p; 1 \leq i \leq p\}$ are uniform; and

P5: The probability contours of the distribution $\mathcal{L}(\widehat{\boldsymbol{\beta}})$, 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.

4.2. Conditioning and Ellipticity

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

$$R(\gamma) = \{\gamma \in \mathbb{R}^p : (\hat{\gamma} - \gamma)' \mathbf{\Omega}^{-1} (\hat{\gamma} - \gamma) \leq c^2\} \quad (9)$$

has unit volume. Then $R(\gamma)$ is an *ellipsoid of concentration* of Cramér [5, 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 $\mathbf{\Omega}$ determines its elliptical density contours in \mathbb{R}^p . Rotating to standard position gives $\{\omega_1, \omega_2, \dots, \omega_p\}$ as lengths of the semiprincipal axes. An ellipticity index, $W(\mathbf{\Omega}) = [\text{tr}(\mathbf{\Omega})]^p / p^p |\mathbf{\Omega}|$ as in [19], serves to gauge the nonsphericity of contours of $R(\gamma)$, and of $\mathcal{L}(\hat{\gamma})$ if Gaussian, where $W(\mathbf{\Omega}) = 1.0$ at $\mathbf{\Omega} = \mathbf{I}_p$, and larger values quantify increasing divergence from sphericity. As these quantities figure prominently as marks of orthogonality, they are examined next with regard to surrogate systems. To these ends designate $c_1^*(\hat{\beta}) = c_1^{\frac{1}{2}}[V(\hat{\beta})]$ and $El(\hat{\beta}) = W^{\frac{1}{2}}[V(\hat{\beta})]$ as properties of $\mathcal{L}(\hat{\beta})$, and recall that $\{V(\hat{\beta}_{S_k}) = \mathbf{\Sigma}_k^S; k \geq 0\}$.

Theorem 3. *Consider surrogate estimators $\{\hat{\beta}_{S_k}; k \geq 0\}$, together with root condition numbers $c_1^*(\hat{\beta}_{S_k}) = c_1^{\frac{1}{2}}(\mathbf{\Sigma}_k^S)$ and ellipticity indices $El(\hat{\beta}_{S_k}) = W^{\frac{1}{2}}(\mathbf{\Sigma}_k^S)$. Then*

- (i) *Condition numbers $\{c_1^*(\hat{\beta}_{S_k}); k \geq 0\}$ are monotone decreasing with increasing k ; and*
- (ii) *Ellipticity indices $\{El(\hat{\beta}_{S_k}); k \geq 0\}$ decrease monotonically with increasing k .*

Proof: Conclusion (i) follows directly since the eigenvalues of $V(\hat{\beta}_{S_k}) = \sigma^2(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1}$ are proportional to $\{1/(\xi_1^2 + k), \dots, 1/(\xi_p^2 + k)\}$, so that $c_1[V(\hat{\beta}_{S_k})] = (\xi_1^2 + k)/(\xi_p^2 + k)$ decreases with increasing k . To continue, it suffices to consider ellipticity indices for $\{\hat{\theta}_{S_k}; k \geq 0\}$. Accordingly, take $\{\alpha_i = \xi_i^2; 1 \leq i \leq p\}$ and let

$$W(k) = \frac{1}{p^p} \left(\frac{1}{\alpha_1 + k} + \dots + \frac{1}{\alpha_p + k} \right)^p [(\alpha_1 + k) \cdots (\alpha_p + k)]$$

with $\{\alpha_1 \geq \dots \geq \alpha_p > 0\}$ and $k \geq 0$. The goal is to show that $dW(k)/dk < 0$.

It is more expedient to work with $\log[W(k)]$, for which

$$\begin{aligned} \frac{d}{dk} \log[W(k)] &= \frac{p}{\frac{1}{\alpha_1+k} + \dots + \frac{1}{\alpha_p+k}} \left[\frac{-1}{(\alpha_1+k)^2} + \dots + \frac{-1}{(\alpha_p+k)^2} \right] \\ &+ \left(\frac{1}{\alpha_1+k} + \dots + \frac{1}{\alpha_p+k} \right). \end{aligned}$$

Next multiply both terms on the right by $\left(\frac{1}{\alpha_1+k} + \dots + \frac{1}{\alpha_p+k}\right) > 0$ to obtain

$$\left(\frac{1}{\alpha_1+k} + \dots + \frac{1}{\alpha_p+k} \right)^2 - p \left[\frac{1}{(\alpha_1+k)^2} + \dots + \frac{1}{(\alpha_p+k)^2} \right].$$

Letting $\{a_i = 1/(\alpha_i+k); 1 \leq i \leq p\}$ and $\{b_i \equiv 1\}$, we see by the Cauchy–Schwarz inequality $(a_1b_1 + \dots + a_pb_p)^2 \leq (a_1^2 + \dots + a_p^2)(b_1^2 + \dots + b_p^2)$ that the derivative is negative and the function decreasing, to establish conclusion (ii) and complete our proof. \square

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 5, as listed subsequently in Table 6.

4.3. Variance Inflation

Since neither ridge nor surrogate system is orthogonal under ill-conditioning, their *VIF*s necessarily exceed unity. Nonetheless, it is instructive to ask, as k evolves, whether *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 .

Theorem 4. *For the surrogate estimators $\widehat{\beta}_{S_k}$ with elements $\{\widehat{\beta}_{S_i}(k); 1 \leq i \leq p\}$, the functions $\{VIF(\widehat{\beta}_{S_i}); 1 \leq i \leq p\}$ decrease monotonically with increasing k , and are computed by*

$$VIF(\widehat{\beta}_{S_i}) = \frac{\sum_{j=1}^p q_{ij}^2 / (\xi_j^2 + k)}{1 / \sum_{j=1}^p (\xi_j^2 + k) q_{ij}^2} \quad (10)$$

where $\{\xi_1 \geq \dots \geq \xi_p > 0\}$ are not all equal, and $\mathbf{Q} = [q_{ij}]$ comes from the singular decomposition $\mathbf{X} = \mathbf{P}_1 \mathbf{D}_\xi \mathbf{Q}'$.

Proof. $VIF(\widehat{\beta}_{Si}) = v_{ii}/w_{ii}^{-1}$ with $\mathbf{V} = (1/\sigma^2)\mathbf{V}(\widehat{\beta}_{Sk}) = (\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1}$, and $\mathbf{W} = (\mathbf{X}'\mathbf{X} + k\mathbf{I}_p) = \mathbf{Q}\mathbf{D}((\xi_i^2 + k))\mathbf{Q}'$. Expression (10) follows directly. We compute

$$\begin{aligned} & \frac{d}{dk} \left(\sum_{j=1}^p \frac{q_{ij}^2}{\xi_j^2 + k} \sum_{j=1}^p (\xi_j^2 + k) q_{ij}^2 \right) = \frac{d}{dk} \left(\sum_{j,m=1}^p \frac{q_{ij}^2}{\xi_j^2 + k} (\xi_m^2 + k) q_{im}^2 \right) \\ &= \frac{d}{dk} \left(\sum_{j=1}^p q_{ij}^4 + \sum_{j \neq m}^p q_{ij}^2 q_{im}^2 \frac{(\xi_m^2 + k)}{\xi_j^2 + k} \right) = \sum_{j < m}^p q_{ij}^2 q_{im}^2 \left(\frac{(\xi_j^2 - \xi_m^2)}{(\xi_j^2 + k)^2} + \frac{(\xi_m^2 - \xi_j^2)}{(\xi_m^2 + k)^2} \right) \\ &= \sum_{j < m}^p q_{ij}^2 q_{im}^2 (\xi_j^2 - \xi_m^2) \left(\frac{1}{(\xi_j^2 + k)^2} - \frac{1}{(\xi_m^2 + k)^2} \right) < 0; \end{aligned}$$

demonstrating that each $\{VIF(\widehat{\beta}_{Si}(k)); 1 \leq i \leq p\}$ does decrease monotonically with increasing k . \square

In comparison, for ridge estimates we have

$$\{VIF(\widehat{\beta}_{Ri}) = \sum_{j=1}^p \frac{\xi_j^2}{(\xi_j^2 + k)^2} q_{ij}^2 \sum_{j=1}^p \frac{(\xi_j^2 + k)^2}{\xi_j^2} q_{ij}^2; 1 \leq i \leq p\}. \quad (11)$$

That these functions need not be monotone is demonstrated subsequently in Table 4 for the case studies of Section 5.

4.4. Summary Properties

Both $\{\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\}$, and $\{\mathbf{U} = \mathbf{D}\boldsymbol{\theta} + \boldsymbol{\eta}\}$ at (1) as its canonical equivalent, support *OLS*, ridge, and surrogate versions, related through $[\widehat{\boldsymbol{\theta}}_L, \widehat{\boldsymbol{\theta}}_{Rk}, \widehat{\boldsymbol{\theta}}_{Sk}] = \mathbf{Q}'[\widehat{\boldsymbol{\beta}}_L, \widehat{\boldsymbol{\beta}}_{Rk}, \widehat{\boldsymbol{\beta}}_{Sk}]$, and their moments through $\mathbf{E}(\widehat{\boldsymbol{\theta}}) = \mathbf{Q}'\mathbf{E}(\widehat{\boldsymbol{\beta}})$ and $\mathbf{V}(\widehat{\boldsymbol{\theta}}) = \mathbf{Q}'\mathbf{V}(\widehat{\boldsymbol{\beta}})\mathbf{Q}$ for each pair. For easy access, essentials of $\{\widehat{\boldsymbol{\beta}}_L, \widehat{\boldsymbol{\beta}}_{Rk}, \widehat{\boldsymbol{\beta}}_{Sk}\}$ are summarized in Table 1, where expressions for *VIF*'s support properties asymptotic in k as in Appendix A.

Several concepts, to include condition numbers and the index of ellipticity, are summarized in Table 2 for $\{\widehat{\boldsymbol{\theta}}_L, \widehat{\boldsymbol{\theta}}_{Rk}, \widehat{\boldsymbol{\theta}}_{Sk}\}$. In explanation, it is seen from $\widehat{\boldsymbol{\beta}}_L = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ that sensitivity of $\widehat{\boldsymbol{\beta}}_L$ to small changes in \mathbf{X} is quantified through conditioning of the composition $[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']$. But since $\widehat{\boldsymbol{\beta}}_L = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{Q}\mathbf{D}_\xi^{-2}\mathbf{Q}'\mathbf{Q}\mathbf{D}_\xi\mathbf{P}'_1\mathbf{Y} = \mathbf{Q}\widehat{\boldsymbol{\theta}}_L$ and \mathbf{Q} is orthogonal, it suffices to consider conditioning of $\widehat{\boldsymbol{\theta}}_L = \mathbf{D}_\xi\mathbf{U}_1$ as a data transformation.

Table 1: Properties of $\{\widehat{\beta}_L, \widehat{\beta}_{Rk}, \widehat{\beta}_{Sk}\}$ under Gauss–Markov assumptions, where $\mathbf{X}_k = \mathbf{P}_1 \text{Diag}((\xi_1^2 + k)^{\frac{1}{2}}, \dots, (\xi_p^2 + k)^{\frac{1}{2}}) \mathbf{Q}'$ and $\mathbf{A}_k = (\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)$.

Estimator	$E(\widehat{\beta})$	$V(\widehat{\beta})$	$VIF(\widehat{\beta}_i)$
$\widehat{\beta}_L$	β	$\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$	$\sum_{j=1}^p \frac{1}{\xi_j^2} q_{ij}^2 \sum_{j=1}^p \xi_j^2 q_{ij}^2$
$\widehat{\beta}_{Rk}$	$\mathbf{A}_k^{-1} \mathbf{X}'\mathbf{X}\beta$	$\sigma^2 \mathbf{A}_k^{-1} \mathbf{X}'\mathbf{X} \mathbf{A}_k^{-1}$	$\sum_{j=1}^p \frac{\xi_j^2}{(\xi_j^2 + k)^2} q_{ij}^2 \sum_{j=1}^p \frac{(\xi_j^2 + k)^2}{\xi_j^2} q_{ij}^2$
$\widehat{\beta}_{Sk}$	$\mathbf{A}_k^{-1} \mathbf{X}'\mathbf{X}\beta$	$\sigma^2 \mathbf{A}_k^{-1}$	$\sum_{j=1}^p \frac{1}{\xi_j^2 + k} q_{ij}^2 \sum_{j=1}^p (\xi_j^2 + k) q_{ij}^2$

This is listed in the second column of Table 2, where $c_1[\widehat{\beta}_L(\mathbf{Y})] = c_1[\widehat{\theta}_L(\mathbf{U}_1)] = c_1(\mathbf{D}_\xi) = \xi_1/\xi_p$. Similarly, $E(\widehat{\beta}_{Rk}) = \mathbf{A}_k^{-1} \mathbf{X}'\mathbf{X}\beta = T(\beta)$, as a parameter transformation, is subject to the conditioning of $(\mathbf{A}_k^{-1} \mathbf{X}'\mathbf{X})$, as listed for its canonical equivalent $\widehat{\theta}_{Rk}$ in the third column of Table 2. Owing to invariance under orthogonal congruence, $c_1[V(\widehat{\beta})] = c_1[V(\widehat{\theta})]$ and $W[V(\widehat{\beta})] = W[V(\widehat{\theta})]$ apply in turn for *OLS*, ridge, and surrogate solutions. In short, quantities listed in Table 2 apply verbatim for $\{\widehat{\beta}_L, \widehat{\beta}_{Rk}, \widehat{\beta}_{Sk}\}$.

Regarding the ridge $\{\widehat{\beta}_{Rk}; k \geq 0\}$ and surrogate $\{\widehat{\beta}_{Sk}; k \geq 0\}$ estimators, both shrink stochastically towards the origin with increasing k , as do their means and variances from Table 1. Similar comments apply for $\{\widehat{\theta}_{Rk}; k \geq 0\}$ and $\{\widehat{\theta}_{Sk}; k \geq 0\}$. Moreover, it is seen for each $k > 0$ that $\widehat{\beta}_{Sk}$ achieves lesser shrinkage, both in expectation and variance, than $\widehat{\beta}_{Rk}$. On occasion, taking limits as $k \rightarrow \infty$ garners evidence regarding solutions for moderate to large, but finite k . Although $V(\widehat{\beta}_{Rk}) = \sigma^2 \mathbf{A}_k^{-1} \mathbf{X}'\mathbf{X} \mathbf{A}_k^{-1}$ and $V(\widehat{\beta}_{Sk}) = \sigma^2 \mathbf{A}_k^{-1}$ both diminish towards the zero matrix, selected quantities of interest here are scale-invariant. Accordingly, quantities such as $c_1^*(\cdot)$, $El(\cdot)$, and *VIFs*, as scale-invariants of those objects, are examined legitimately as $k \rightarrow \infty$. Details appear in Appendix A.

Table 2: Condition numbers for $\widehat{\boldsymbol{\theta}}(\mathbf{U}_1)$ as a data transformation, for $E(\widehat{\boldsymbol{\theta}}) = T(\boldsymbol{\theta})$ as a parameter transformation, and for $V(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\Sigma}$; and an ellipticity index $W(\boldsymbol{\Sigma})$ for contours of $\widehat{\boldsymbol{\theta}}$; for each of $\{\widehat{\boldsymbol{\theta}}_L, \widehat{\boldsymbol{\theta}}_{Fk}, \widehat{\boldsymbol{\theta}}_{Sk}\}$ and, by equivalence, for $\{\widehat{\boldsymbol{\beta}}_L, \widehat{\boldsymbol{\beta}}_{Fk}, \widehat{\boldsymbol{\beta}}_{Sk}\}$, respectively.

$\widehat{\boldsymbol{\theta}}$	$c_1[\widehat{\boldsymbol{\theta}}(\mathbf{U}_1)]$	$c_1[T(\boldsymbol{\theta})]$	$c_1(\boldsymbol{\Sigma})$	$W(\boldsymbol{\Sigma})$
$\widehat{\boldsymbol{\theta}}_L$	$\frac{\xi_1}{\xi_p}$	1.00	$\frac{\xi_1^2}{\xi_p^2}$	$\frac{\left(\sum_{i=1}^p \xi_i^{-2}\right)^p}{p^p \prod_{i=1}^p \xi_i^{-2}}$
$\widehat{\boldsymbol{\theta}}_{Fk}$	$\frac{\max\left\{\frac{\xi_i}{(\xi_i^2+k)}\right\}}{\min\left\{\frac{\xi_i}{(\xi_i^2+k)}\right\}}$	$\frac{\xi_1^2(\xi_p^2+k)}{\xi_p^2(\xi_1^2+k)}$	$\frac{\max\left\{\frac{\xi_i^2}{(\xi_i^2+k)^2}\right\}}{\min\left\{\frac{\xi_i^2}{(\xi_i^2+k)^2}\right\}}$	$\frac{\left(\sum_{i=1}^p \frac{\xi_i^2}{(\xi_i^2+k)^2}\right)^p}{p^p \prod_{i=1}^p \frac{\xi_i^2}{(\xi_i^2+k)^2}}$
$\widehat{\boldsymbol{\theta}}_{Sk}$	$\frac{\sqrt{\xi_1^2+k}}{\sqrt{\xi_p^2+k}}$	$\frac{\xi_1 \sqrt{\xi_p^2+k}}{\xi_p \sqrt{\xi_1^2+k}}$	$\frac{\xi_1^2+k}{\xi_p^2+k}$	$\frac{\left(\sum_{i=1}^p \frac{1}{\xi_i^2+k}\right)^p}{p^p \prod_{i=1}^p \frac{1}{\xi_i^2+k}}$

5. Case Studies

5.1. The Setting

Consider the Hospital Manpower Data reported in [20], where records at $n = 17$ U. S. Naval Hospitals 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 $\div 1000$ (X_4); and Average length of patients' stay in days (X_5). The basic model is

$$\{Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon_i; 1 \leq i \leq n\}. \quad (12)$$

Following [9], [16], [17], [20], and others, we instead take $\{\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}\}$ to be centered and scaled, with $\mathbf{Z}'\mathbf{Z}$ in correlation form, our focus being the rates of change $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3, \beta_4, \beta_5]'$. The data appear as Table 3.8 of [20; pp.132–133], and our computations utilize both PROC IML of SAS and the symbolic program Maple. The data are remarkably ill-conditioned, with $\mathbf{D}_\xi = \text{Diag}(2.048687, 0.816997, 0.307625, 0.201771, 0.007347)$ and $c_1(\mathbf{Z}'\mathbf{Z}) = 77,754.86$. The maximal VIF for OLS is $V_1 = VIF(\widehat{\beta}_1) = 9,595.68$; other VIF s are listed at $k = 0$ in Table 4. In this section we undertake a comparative

study of the ridge and surrogate solutions for these data as in preceding sections.

5.2. Residuals

Citing residuals in studies of estimation, Theorem 1 registers for each $k > 0$ the critical inequality $R_{SS}(\hat{\beta}_{S_k}) < R_{SS}(\hat{\beta}_{R_k})$. Numerical values are given

Table 3: Values of $R_{SS}(\hat{\beta}_{S_k}) \times 10^{-7}$ and $R_{SS}(\hat{\beta}_{R_k}) \times 10^{-7}$ 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	∞
$R_{SS}(\hat{\beta}_{S_k})$	0.454	0.600	0.704	0.793	0.878	0.963	1.190	1.440	49.500
$R_{SS}(\hat{\beta}_{R_k})$	0.454	0.802	1.010	1.210	1.420	1.650	2.300	3.040	49.500

in Table 3 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 $R_{SS}(\hat{\beta}_k) \rightarrow \mathbf{U}'\mathbf{U} = \mathbf{Y}'\mathbf{Y}$ for both, as the estimators shrink towards zero.

5.3. Variance Inflation

Property **P3**, Section 4.1, identifies $\{VIF(\hat{\beta}_i) = 1.0; 1 \leq i \leq 5\}$ as marks of an orthogonal system. At issue here is whether ridge or surrogate systems might stabilize in VIF 's towards unity as k evolves. Table 4 tracks these quantities for $k \in [0, \infty)$. These factors clearly evolve erratically for ridge, but monotonically towards unity for surrogate solutions, as certified in Theorem 4. It is a pathology 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 those from an orthogonal system.

5.4. Uniformity

Let $\{F_i = \text{Var}(\hat{\beta}_i)/\text{trV}(\hat{\beta}); 1 \leq i \leq 5\}$; **P4** of Section 4.1 identifies their uniformity as characteristic of orthogonality. Fractions are displayed in Table 5 for the ridge and surrogate solutions, together with $UI(\hat{\beta}_k) = \sum_{i=1}^5 (F_i - 1/5)^2 / (4/5)$ as a uniformity index which has range $[0, 1]$. The value $4/5$ is the maximum value of the numerator at $\{1, 0, 0, 0, 0\}$. It is seen that $UI(\hat{\beta}_{R_k})$ 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 $UI(\hat{\beta}_{S_k})$

Table 4: Variance inflation factors $\{VIF_{R1}, \dots, VIF_{R5}\}$ for elements of $\widehat{\beta}_{Rk} = [\widehat{\beta}_{R1}, \dots, \widehat{\beta}_{R5}]'$, and $\{VIF_{S1}, \dots, VIF_{S5}\}$ for $\widehat{\beta}_{Sk} = [\widehat{\beta}_{S1}, \dots, \widehat{\beta}_{S5}]'$, 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

decreases monotonically for $k \in [0, 1]$; the curves cross at $k = 0.083651$ with common value $UI(\widehat{\beta}_{Sk}) = UI(\widehat{\beta}_{Rk}) = 0.049181$; and $UI(\widehat{\beta}_{Sk}) < UI(\widehat{\beta}_{Rk})$ for $k \in (0.083651, 1]$.

5.5. Conditioning and Contours

Further characteristics in Section 4.1 are **P2**: The condition number of $V(\widehat{\beta})$; and **P5**: Spherical concentration ellipsoids of $\mathcal{L}(\widehat{\beta})$ under second moments, or spherical probability contours if Gaussian. Designate $c_1^*(\widehat{\beta}) = c_1^{\frac{1}{2}}[V(\widehat{\beta})]$ and $El(\widehat{\beta}) = W^{\frac{1}{2}}[V(\widehat{\beta})]$ as before. These are listed in Table 6 as k ranges over $[0, \infty)$. The index $El(\widehat{\beta}_L) = 641006$ at $k = 0$ reflects that elliptical contours in \mathbb{R}^5 for *OLS* are highly elongated and nearly degenerate, its semi-principal axes given by $D_\xi = \text{Diag}(2.048687, 0.816997, 0.307625, 0.201771, 0.007347)$.

It again appears pathological that $c_1^*(\widehat{\beta}_{Rk})$ and $El(\widehat{\beta}_{Rk})$ 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 $\min\{c_1^*(\widehat{\beta}_{Rk})\} = 7.4463$ at $k = 0.015$. On the other hand, $c_1^*(\widehat{\beta}_{Sk})$ decreases monotonically for $k \in [0, \infty)$ as in Theorem 3; the curves cross at $k = 0.03045$ with common value $c_1^*(\widehat{\beta}_{Sk}) = c_1^*(\widehat{\beta}_{Rk}) = 11.772$; and $c_1^*(\widehat{\beta}_{Sk}) < c_1^*(\widehat{\beta}_{Rk})$ for $k \in (0.03045, \infty)$. Similarly, $\min\{El(\widehat{\beta}_{Rk})\} = 7.9791$ occurs at $k = 0.00268$; $El(\widehat{\beta}_{Sk})$ decreases for $k \in [0, \infty)$ by Theorem 3; whereas the curves cross at $k = 0.02135$ with value $El(\widehat{\beta}_{Sk}) = El(\widehat{\beta}_{Rk}) = 17.317$, and

Table 5: Fractions $\{F_i = \text{Var}(\hat{\beta}_i)/\text{trV}(\hat{\beta}); 1 \leq i \leq 5\}$ of total variance and their Uniformity Index $UI(\hat{\beta})$ for ridge $\hat{\beta}_{Rk} = [\hat{\beta}_{R1}, \dots, \hat{\beta}_{R5}]'$ and surrogate $\hat{\beta}_{Sk} = [\hat{\beta}_{S1}, \dots, \hat{\beta}_{S5}]'$ solutions as k ranges over $[0, 1]$.

k	F_1	F_2	F_3	F_4	F_5	$UI(\hat{\beta}_k)$
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

$El(\hat{\beta}_{Sk}) < El(\hat{\beta}_{Rk})$ for $k \in (0.02135, \infty)$. It should be noted that the limit $El(\hat{\beta}_{R\infty}) = 1301.24 = W^{\frac{1}{2}}[(V(\hat{\beta}_L)^{-1})]$ as in Appendix A.

6. Conclusions

The ridge system $\{(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)\hat{\beta}_{Rk} = \mathbf{X}'\mathbf{Y}\}$, often but incorrectly venerated as yielding constrained LaGrange minimizers, is reexamined on the claim: “At a certain value of k the system will stabilize and have the general characteristics of an orthogonal system” [9; p.65]. This remains vague in failing to articulate 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 under second moments, or for their densities if Gaussian. Conditioning arguments, set to account for ill-conditioning of \mathbf{X} on both sides of the *OLS*

Table 6: Root condition numbers $c_1^*(\hat{\beta}) = c_1^{\frac{1}{2}}[V(\hat{\beta})]$ and elliptical index $El(\hat{\beta}) = W^{\frac{1}{2}}[V(\hat{\beta})]$ for ridge and surrogate solutions as k ranges over $[0, \infty)$.

k	$c_1^*(\hat{\beta}_{Rk})$	$c_1^*(\hat{\beta}_{Sk})$	$El(\hat{\beta}_{Rk})$	$El(\hat{\beta}_{Sk})$
0.00	278.846	278.846	641006	641006
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

equations, prompt the *generalized surrogate* system $\{(\mathbf{X}'_K \mathbf{X}_K) \hat{\beta}_{SK} = \mathbf{X}'_K \mathbf{Y}\}$, and its study at $\mathbf{K} = k\mathbf{I}_p$ as alternative to ridge. Since neither the ridge nor surrogate system is orthogonal when \mathbf{X} is ill-conditioned, this study examines whether their solutions stabilize and tend towards orthogonality as k evolves.

Theorem 1 identifies $k > 0$ such that $\hat{\beta}_{Sk}$ is M_{SE} -admissible under squared error loss, and thus a credible alternative to *OLS*. Moreover, if $\hat{\beta}_R(k)$ is M_{SE} -admissible at $k = k_0$, then $\hat{\beta}_S(k_0)$ is M_{SE} -admissible. Theorem 2 asserts for each $k > 0$ that $\hat{\beta}_{Sk}$ has smaller residual sum of squares than $\hat{\beta}_{Rk}$. Critical properties of $\{\hat{\beta}_{Sk}; k \geq 0\}$ are seen to decrease monotonically as k increases, to include conditioning of dispersion and ellipticity indices in Theorem 3, and Variance Inflation Factors in Theorem 4. On the other hand, none of these characteristics is monotone for $\{\hat{\beta}_{Rk}; k \geq 0\}$. Instead, they tend to diverge erratically with k , often reverting back towards values for *OLS* for moderate to large k . In short, ridge solutions ultimately appear to exhibit precisely the same pathologies of *OLS* that they are intended to rectify. Case studies of Section 5 illustrate these findings for the highly ill-conditioned Hospital Manpower Data.

Ordinary models, having a single scalar k , specialize from generalized ridge and surrogate models. Theorem 1 shows that (i) the generalized models correspond one-to-one; (ii) each ordinary model specializes from the generalized version of the other; (iii) despite the remarkable transition of ordi-

nary surrogate systems to orthogonality, generalized surrogate systems need not exhibit such properties, as demonstrated for the special case of ordinary ridge; nonetheless, (iv) generalized ridge systems can exhibit such properties, as seen in the special case of ordinary surrogate systems.

A. Appendix

Here we establish limit results as stated in the text. Tables 1 and 2 show the required expressions.

Theorem 5. *Given surrogate estimators $\{\widehat{\beta}_{S_k}; k \geq 0\}$ with elements $\{\widehat{\beta}_{S_i}(k); 1 \leq i \leq p\}$ and $V(\widehat{\beta}_{S_k}) = \Sigma_k^S$. Then*

- (i) $\lim_{k \rightarrow \infty} c_1(\Sigma_k^S) = 1$,
- (ii) $\lim_{k \rightarrow \infty} W(\Sigma_k^S) = 1$, and
- (iii) $\lim_{k \rightarrow \infty} VIF(\widehat{\beta}_{S_i}(k)) = 1$ for each $\{i = 1, \dots, p\}$.

Proof. The condition number $c_1(\Sigma_k^S) = \frac{\xi_1^2 + k}{\xi_p^2 + k} \rightarrow 1$ as $k \rightarrow \infty$, as asserted in conclusion (i). The ellipticity index

$$W(\Sigma_k^S) = \frac{\left(\sum_{i=1}^p \frac{1}{\xi_i^2 + k}\right)^p}{p^p \prod_{i=1}^p \frac{1}{\xi_i^2 + k}} = \frac{\left(\sum_{i=1}^p \frac{k}{\xi_i^2 + k}\right)^p}{p^p \prod_{i=1}^p \frac{k}{\xi_i^2 + k}} \rightarrow 1 \text{ as } k \rightarrow \infty$$

in support of (ii). The *VIF*s for each $\{i = 1, \dots, p\}$ satisfy

$$VIF(\widehat{\beta}_{S_i}) = \sum_{j=1}^p \frac{q_{ij}^2}{\xi_j^2 + k} \sum_{j=1}^p (\xi_j^2 + k) q_{ij}^2 = \left(\sum_{i=1}^p \frac{k q_{ij}^2}{\xi_j^2 + k}\right) \left(\sum_{j=1}^p \frac{(\xi_j^2 + k) q_{ij}^2}{k}\right)$$

which tends to 1 as $k \rightarrow \infty$ since \mathbf{Q} is an orthogonal matrix, as in (iii), to complete our proof. \square

Theorem 6. *Given ridge estimators $\{\widehat{\beta}_{R_k}; k \geq 0\}$ with elements $\{\widehat{\beta}_{R_i}(k); 1 \leq i \leq p\}$ and $V(\widehat{\beta}_{R_k}) = \Sigma_k^R$, and the OLS estimators $\widehat{\beta}_L = [\widehat{\beta}_{L1}, \dots, \widehat{\beta}_{Lp}]'$ with $V(\widehat{\beta}_L) = \sigma^2 \mathbf{V} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$. Then*

- (i) $\lim_{k \rightarrow \infty} c_1(\Sigma_k^R) = \frac{\xi_1^2}{\xi_p^2} = c_1(\mathbf{V})$ for $\widehat{\beta}_L$ under OLS,
- (ii) $\lim_{k \rightarrow \infty} W(\Sigma_k^R) = W(\mathbf{V}^{-1}) = W(\mathbf{X}'\mathbf{X})$, and
- (iii) $\lim_{k \rightarrow \infty} VIF(\widehat{\beta}_{R_i}(k)) = VIF(\widehat{\beta}_{L_i})$ under OLS, for each $\{i = 1, \dots, p\}$.

Proof. The condition number

$$c_1(\Sigma_k^R) = \frac{\max\left\{\frac{\xi_i^2}{(\xi_i^2+k)^2}; 1 \leq i \leq p\right\}}{\min\left\{\frac{\xi_i^2}{(\xi_i^2+k)^2}; 1 \leq i \leq p\right\}} = \frac{\max\left\{\frac{k^2 \xi_i^2}{(\xi_i^2+k)^2}; 1 \leq i \leq p\right\}}{\min\left\{\frac{k^2 \xi_i^2}{(\xi_i^2+k)^2}; 1 \leq i \leq p\right\}} \rightarrow \frac{\xi_1^2}{\xi_p^2}$$

as $k \rightarrow \infty$, as in conclusion (i). The ellipticity index

$$W(\Sigma_k^R) = \frac{\left(\sum_{i=1}^p \frac{\xi_i^2}{(\xi_i^2+k)^2}\right)^p}{p^p \prod_{i=1}^p \frac{\xi_i^2}{(\xi_i^2+k)^2}} = \frac{\left(\sum_{i=1}^p \frac{k^2 \xi_i^2}{(\xi_i^2+k)^2}\right)^p}{p^p \prod_{i=1}^p \frac{k^2 \xi_i^2}{(\xi_i^2+k)^2}} \rightarrow \frac{\left(\sum_{i=1}^p \xi_i^2\right)^p}{p^p \prod_{i=1}^p \xi_i^2} = W(\mathbf{V}^{-1})$$

as $k \rightarrow \infty$, as asserted in (ii). The VIFs for each $\{i = 1, \dots, p\}$ satisfy

$$\begin{aligned} VIF(\widehat{\beta}_{Ri}) &= \sum_{j=1}^p \frac{\xi_j^2 q_{ij}^2}{(\xi_j^2 + k)^2} \sum_{j=1}^p \frac{(\xi_j^2 + k)^2 q_{ij}^2}{\xi_j^2} \\ &= \left(\sum_{i=1}^p \frac{k^2 \xi_j^2 q_{ij}^2}{(\xi_j^2 + k)^2} \right) \left(\sum_{j=1}^p \frac{(\xi_j^2 + k)^2 q_{ij}^2}{k^2 \xi_j^2} \right) \\ &\rightarrow \left(\sum_{j=1}^p \xi_j^2 q_{ij}^2 \right) \left(\sum_{j=1}^p \frac{q_{ij}^2}{\xi_j^2} \right) = VIF(\widehat{\beta}_{Li}) \end{aligned}$$

as $k \rightarrow \infty$, to establish (iii) and complete our proof. \square

References

- [1] Beaton, A. D., Rubin, D., and Barone, J. (1976). The acceptability of regression solutions: Another look at computational accuracy. *J. Amer. Statist. Assoc.* **71**: 158–168.
- [2] Belsley, D.A. (1986). Centering, the constant, first-differencing, and assessing conditioning. In: D.A. Belsley and E. Kuh (Eds.), *Model Reliability*, MIT Press, Boston, MA 117–153.
- [3] Berk, K.N. (1977). Tolerance and condition in regression computations. *J. Amer. Statist. Assoc.* **72**: 863–866.

- [4] Bingham, C. and Larntz, K. (1977). Comment. *J. Amer. Statist. Assoc.* **72**: 97–102.
- [5] Cramér, H. (1947). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ.
- [6] Davies, R.B and Hutton, B. (1975). The effect of errors in the independent variables in regression. *Biometrika* **62**: 383–392.
- [7] Geladi, P. (2002). Some recent trends in the calibration literature. *Chemometrics and Intelligent Laboratory Systems* **60**: 211–224.
- [8] Gunst, R.F. (2000). Classical studies that revolutionized the practice of regression analysis. *Technometrics* **42**: 62–64.
- [9] Hoerl, A.E. and Kennard, R.W. (1970a). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**: 55–67.
- [10] Hoerl, A.E. and Kennard, R.W. (1970b). Ridge regression: Applications to nonorthogonal problems. *Technometrics* **12**: 69–82.
- [11] Hoerl, A.E. and Kennard, R.W. (2000). Ridge regression: Applications to nonorthogonal problems. *Technometrics* **42**: 80–86.
- [12] Hoerl, A.E., Kennard, R.W. and Baldwin, K.F. (1975). Ridge regression: Some simulations. *Communications in Statistics* **4**: 105–123.
- [13] Jensen, D.R. and Ramirez, D.E. (2008). Anomalies in the foundations of ridge regression. *International Statistical Review* **76**: 89–105.
- [14] Kalivas, J.H. (2005). Multivariate calibration: An overview. *Analytical Letters* **38**: 2259–2279.
- [15] Lowerre, J.M. (1974). On the mean square error of parameter estimates for some biased estimators. *Technometrics* **16**: 461–464.
- [16] Marquardt, D.W. (1970). Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics* **12**: 591–612.
- [17] Marquardt, D.W. and Snee, R.D. (1975). Ridge regression in practice. *The Amer. Statist.* **29**: 3–20.

- [18] Marshall, A.W. and Olkin, I. (1979). *Inequalities: Theory of Majorization and Its Applications*. Academic Press, New York.
- [19] Mauchly, J.W. (1940). Significance test for sphericity of a normal n -variate distribution. *Ann. Math. Statist.* **11**: 204–209.
- [20] Myers, R.H. (1990). *Classical and Modern Regression with Applications*, Second ed. PWS-KENT, Boston, MA,
- [21] Obenchain, R.L. (1977). Classical F-tests and confidence regions for ridge regression. *Technometrics* **19**: 429–439.
- [22] Riley, J. (1955). Solving systems of linear equations with a positive definite, symmetric but possibly ill-conditioned matrix. *Mathematical Tables and Other Aids to Computation* **9**: 96–101.
- [23] Stewart, G.W. (1987). Collinearity and least squares regression. *Statist. Sci.* **2**: 68–84.
- [24] Sundberg, R. (1999). Multivariate calibration – Direct and indirect regression methodology. *Scandinavian J. Statist.* **26**: 161–207.