ANOMALIES IN THE FOUNDATIONS OF RIDGE REGRESSION

D. R. JENSEN AND D. E. RAMIREZ

Abstract. Anomalies persist in the foundations of ridge regression as set forth in Hoerl and Kennard (1970) and subsequently. Conventional ridge estimators and their properties do not follow on constraining lengths of solution vectors using LaGrange's method, as claimed. Estimators so constrained have singular distributions; the proposed solutions are not necessarily minimizing; and heretofore undiscovered bounds are exhibited for the ridge parameter. None of the considerable literature on estimation, prediction, cross–validation, choice of ridge parameter, and related issues, collectively known as *ridge regression,* is consistent with constrained optimization, nor with corresponding inequality constraints. The problem is traced to a misapplication of LaGrange's principle, failure to recognize the singularity of distributions, and misplaced links between constraints and the ridge parameter. Other principles, based on condition numbers, are seen to validate both conventional ridge and *surrogate ridge* regression to be defined. Numerical studies illustrate that ridge analysis often exhibits some of the same pathologies it is intended to redress.

1. Introduction

Given the full–rank model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with zero–mean, homoscedastic, and uncorrelated errors, the ordinary least squares *(OLS)* estimators $\widehat{\boldsymbol{\beta}}_L$ solve the k equations $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$ on minimizing $Q(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$. Ill–conditioned models long have posed special challenges, in that $\widehat{\boldsymbol{\beta}}_L$ often exhibits excessive length, inflated variances, instability, and other intrinsic difficulties. Noting these, Hoerl (1962, 1964) considered *ad hoc* solutions $\widehat{\boldsymbol{\beta}}_R = \{\widehat{\boldsymbol{\beta}}_{R_\lambda} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_k)^{-1}\boldsymbol{X}'\boldsymbol{Y}; \lambda \geq 0\}$ and noted their successful applications in chemical engineering. Analyses built around these have been labeled *ridge regression* in statistics, although Levenberg (1944) and Riley (1955) earlier posed such solutions in numerical analysis. Noting that *OLS* "does not have built into it a method for portraying sensitivity of the solutions to the estimation criterion," Hoerl and Kennard (1970) sought mathematical foundations beyond Gauss's principle with its inherent limitations. Specifically, they asserted that $\widehat{\boldsymbol{\beta}}_R$ are solutions minimizing $Q(\boldsymbol{\beta})$ subject to the constraint $\{\boldsymbol{\beta}'\boldsymbol{\beta} = c^2\}$. Others identify ridge regression instead with the constraints $\{\boldsymbol{\beta}'\boldsymbol{\beta} \leq c^2\}$ of Balakrishnan (1963); however, Hoerl and Kennard (1970), p. 64, specifically relegate this to approaches other than ridge regression.

1991 *Mathematics Subject Classification.* Primary: 62J07; Secondary: 62J20.

Key words and phrases. Constrained optimization, incomplete use of LaGrange's method, nonsingular distributions, alternative foundations.

Ridge estimators abound, based on estimative, predictive, cross–validative, and numerous other criteria, typically giving disparate choices for λ. Even the early simulations of Dempster, Schatzoff, and Wermuth (1977) identified 57 ridge and related shrinkage estimators. An expository survey and numerical examples are provided in Myers (1990). In short, a considerable literature, spanning the past thirty–six years, rests on the foundations of Hoerl and Kennard (1970), ostensibly the mathematics of constrained optimization, to remedy defects of *OLS* in ill–conditioned systems.

In fact, little of the collective literature known as ridge regression is consistent with the constrained optimization of Hoerl and Kennard (1970), nor with corresponding inequality constraints. Here the problem is traced to (i) a misapplication of LaGrange's principle, (ii) failure to identify singular distributions, and (iii) invalid links between the constraints and the ridge parameters. These errors are evident also in Marquardt (1970), Marquardt and Snee (1975), Golub, Heath and Wahba (1979), van Nostrand (1980), and elsewhere throughout the literature. In consequence, much that is known about ridge regression rests on a false premise. By analogy, Hoerl and Kennard (1970) considered generalized ridge regression as solving the modified equations $(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Lambda})\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$, with nonnegative ridge parameters $\boldsymbol{\Lambda} = \mathrm{Diag}(\lambda_1, \ldots, \lambda_k)$. As noted later, these solutions again are inconsistent with LaGrange minimization. In summary, not to denigrate its usefulness in practice, the collective body of ridge regression rests on little more than heuristics. To the contrary, aspects of ridge regression have proven useful enough, often enough, to deserve sound rationale for their implementation. In this spirit we seek to supplant the missing foundations with alternatives based on conditioning of the linear system $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$. An outline follows.

Supporting developments comprise Section 2, to include notation and the basics of invariance and condition numbers. Section 3 reexamines LaGrange optimization in linear inference. Section 4 develops supporting rationale for ridge regression as currently practiced, and an alternative approach using *surrogate ridge* models. A case study in Section 5 revisits an ill–conditioned data set considered elsewhere. Section 6 concludes with a brief summary.

2. Preliminaries

2.1. **Notation.** The symbols $\mathbb{R}^k$ and $\mathbb{R}^k_+$ designate Euclidean k–space and its positive orthant; $\mathbb{F}_{nk}$ and $\mathbb{F}^{\mathbb{C}}_{nk}$ comprise the real and complex $(n \times k)$ matrices of rank $k \le n$; and $\mathbb{S}_k$ and $\mathbb{S}^+_k$ designate the real symmetric $(k \times k)$ matrices and their positive definite varieties. The transpose, inverse, trace, and determinant of $\boldsymbol{A} \in \mathbb{F}_{kk}$ are $\boldsymbol{A}'$, $\boldsymbol{A}^{-1}$, $\mathrm{tr}(\boldsymbol{A})$, and $|\,\boldsymbol{A}\,|$, and $\boldsymbol{V}^*$ is the conjugate transpose of $\boldsymbol{V} \in \mathbb{F}^{\mathbb{C}}_{nk}$. Groups of note include $\mathcal{U}(k)$ as the unitary $(k \times k)$ matrices, and $\mathcal{O}(k)$ as the real orthogonal group. Special arrays are the $(k \times k)$ identity $\boldsymbol{I}_k$, the unit vector $\boldsymbol{1}_k = [1, 1, \ldots, 1]' \in \mathbb{R}^k$, and the diagonal matrix $\boldsymbol{D}_a = \mathrm{Diag}(a_1, \ldots, a_k)$. The mapping $\sigma(\boldsymbol{X}) = [\xi_1, \ldots, \xi_k]'$ takes $\boldsymbol{X} \in \mathbb{F}^{\mathbb{C}}_{nk}$ into its ordered singular values $\{\xi_1 \ge \ldots \ge \xi_k > 0\}$. The *singular decomposition* is $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^*$, such that $\boldsymbol{D} = \mathrm{Diag}(\boldsymbol{D}_\xi, \boldsymbol{0})$ of order $(n \times k)$, $\boldsymbol{D}_\xi = \mathrm{Diag}(\xi_1, \ldots, \xi_k)$, $\boldsymbol{U} \in \mathcal{U}(n)$, and $\boldsymbol{V} \in \mathcal{U}(k)$, where the columns of $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n]$ and $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k]$ comprise the *left–* and *right–singular vectors* of $\boldsymbol{X}$. Equivalently write $\boldsymbol{X} = \boldsymbol{U}_1\boldsymbol{D}_\xi\boldsymbol{V}^* = \sum_{i=1}^k \xi_i \boldsymbol{u}_i\boldsymbol{v}_i^*$ with $\boldsymbol{U}_1 = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k]$, and its Moore–Penrose inverse as $\boldsymbol{X}^\dagger = \boldsymbol{V}\boldsymbol{D}^\dagger\boldsymbol{U}^*$, with $\boldsymbol{D}^\dagger = \mathrm{Diag}(\boldsymbol{D}_\xi^{-1}, \boldsymbol{0})$ of order $(k \times n)$. Specifically, the real model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ in canonical form becomes $\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{D}_\xi\boldsymbol{\theta} + \boldsymbol{\epsilon}$, where $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{D}_\xi\boldsymbol{Q}'$ and $\boldsymbol{\theta} = \boldsymbol{Q}'\boldsymbol{\beta}$ is an orthogonal reparametrization. For $\boldsymbol{Y} \in \mathbb{R}^n$ random, designate its mean vector, its dispersion and correlation matrices as $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{\mu}$, $\mathrm{V}(\boldsymbol{Y}) = \boldsymbol{\Sigma}$, and $\boldsymbol{C}(\boldsymbol{Y}) = \boldsymbol{R}$, and its law of distribution as $\mathcal{L}(\boldsymbol{Y})$.

2.2. **Invariance and Conditioning.** A function $\psi(\cdot)$ on $\mathbb{F}_{nk}^{\mathbb{C}}$ is called *unitarily invariant* if, for each $\boldsymbol{G} \in \mathbb{F}_{nk}^{\mathbb{C}}$ and any unitary matrices $\boldsymbol{U} \in \mathcal{U}(n)$ and $\boldsymbol{V} \in \mathcal{U}(k)$, it follows that $\psi(\boldsymbol{G}) = \psi(\boldsymbol{U}\boldsymbol{G}\boldsymbol{V}^*)$. Then $\psi(\boldsymbol{G})$ depends on $\boldsymbol{G}$ only through its ordered singular values $\sigma(\boldsymbol{G}) = [\gamma_1, \ldots, \gamma_k]'$. Let Φ comprise the *symmetric gauge functions* on $\mathbb{R}^k$ such that for each $\phi(\cdot) \in \Phi$, (i) $\phi(u_1, \ldots, u_k)$ is symmetric under the $2^k k!$ permutations and reflections about the origin; (ii) $\phi(\boldsymbol{u}) > 0$ when $\boldsymbol{u} \neq \boldsymbol{0}$; (iii) $\phi(\cdot)$ is homogeneous, *i.e.,* $\phi(c\boldsymbol{u}) = |c|\ \phi(\boldsymbol{u})$ for $c \neq 0$; and (iv) $\phi(\boldsymbol{u} + \boldsymbol{v}) \leq \phi(\boldsymbol{u}) + \phi(\boldsymbol{v})$. Let Ψ comprise the unitarily invariant matrix norms on $\mathbb{F}_{nk}^{\mathbb{C}}$; von Neumann (1937) demonstrated that these are generated as $\{\| \cdot \|_\phi;\ \phi \in \Phi\}$ with $\| \boldsymbol{G} \|_\phi = \phi(\gamma_1, \ldots, \gamma_k)$. Corresponding norms on $\mathbb{F}_{nk}$ are invariant under $\boldsymbol{X} \to \boldsymbol{P}\boldsymbol{X}\boldsymbol{Q}'$, with $(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{O}(n) \times \mathcal{O}(k)$; see also Schatten (1970) and Marshall and Olkin (1979). In particular, the Frobenius norm on $\mathbb{F}_{nk}$ is $\| \boldsymbol{X} \|_F = [\mathrm{tr}(\boldsymbol{X}'\boldsymbol{X})]^{1/2} = (\sum_{i=1}^k \xi_i^2)^{1/2}$ in terms of the singular decomposition $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{D}_\xi\boldsymbol{Q}'$, with $\| \cdot \|$ as the Euclidean norm on $\mathbb{R}^k$.

Two types of conditioning are germane to the present study:

Type A Conditioning: Stability of the solution $\boldsymbol{z}$ of the linear system $\boldsymbol{A}\boldsymbol{z} = \boldsymbol{b}$, when the coefficients $\boldsymbol{A} \in \mathbb{F}_{kk}$ are subjected to small perturbations, is gauged by the *condition number* $c_g(\boldsymbol{A}) = g(\boldsymbol{A})g(\boldsymbol{A}^{-1})$, where $g(\cdot)$ ordinarily is a norm. The system is well conditioned at $\boldsymbol{A} = \boldsymbol{I}_k$ with $c_g(\boldsymbol{I}_k) = 1.0$, larger values reflecting greater ill-conditioning. Specifically, with $g(\boldsymbol{A}) = \| \boldsymbol{A} \|_\phi$, then $\{c_\phi(\cdot);\ \phi \in \Phi\}$ comprise the *unitarily invariant Type A condition numbers,* so that $\{c_\phi(\boldsymbol{A}) = \|\boldsymbol{A}\|_\phi \|\boldsymbol{A}^{-1}\|_\phi;\ \phi \in \Phi\}$, as treated in Marshall and Olkin (1979), Horn and Johnson (1985), and elsewhere. In particular, take $c_1(\boldsymbol{A}) = \alpha_1/\alpha_k$, where $\{\alpha_1 \geq \ldots \geq \alpha_k\}$ are the ordered eigenvalues of $\boldsymbol{A}$.

Type B Conditioning: The concept of elasticities is invoked in Belsley, Kuh and Welsch (1980) to link sensitivities of solutions, and of variances of $\widehat{\boldsymbol{\beta}}_L = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$, with disturbances in the data matrix $\boldsymbol{Z} \in \mathbb{F}_{nk}$, as gauged by its condition number $c_1(\boldsymbol{Z}) = \xi_1/\xi_k$, with $\sigma(\boldsymbol{Z}) = [\xi_1, \ldots, \xi_k]'$. Here $\boldsymbol{Z}$ is the result of scaling the columns of $\boldsymbol{X}$ to have (approximately) equal lengths. More generally, the unitarily invariant condition numbers on $\mathbb{F}_{nk}$ are $c_\phi(\boldsymbol{X}) = \phi(\boldsymbol{X})\phi(\boldsymbol{X}^\dagger)$, with $\boldsymbol{X}^\dagger$ as the Moore–Penrose inverse. The system is well conditioned at $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{I}_k\boldsymbol{Q}'$, where $c_\phi(\boldsymbol{P}\boldsymbol{I}_k\boldsymbol{Q}') = c_\phi(\boldsymbol{I}_k) = 1.0$, larger values reflecting greater ill-conditioning. In summary, Belsley *et al.* (1980) proceed to scale the columns of $\boldsymbol{X} \to \boldsymbol{Z}$ to have approximately equal lengths, and to focus on $\|\boldsymbol{Z}\|_{\phi_1} = \xi_1$, so that $c_1(\boldsymbol{Z}) = \xi_1/\xi_k$.

3. The Principal Issues

3.1. **LaGrange's Method.** Given differentiable functions $f(u_1, \ldots, u_k)$ and $g(u_1, \ldots, u_k)$ such that the gradient $\nabla g(u_1, \ldots, u_k) \neq \mathbf{0}$ on $G_0 = \{\boldsymbol{u} \in \mathbb{R}^k : g(\boldsymbol{u}) = 0\}$, the problem is to minimize $f(u_1, \ldots, u_k)$ subject to the constraint $g(u_1, \ldots, u_k) = 0$. Write $L(u_1, \ldots, u_k, \lambda) = f(u_1, \ldots, u_k) + \lambda[g(u_1, \ldots, u_k) - 0]$. It is necessary that gradient vectors in $\mathbb{R}^k$ be parallel, *i.e,*

$$\nabla f(u_1, \ldots, u_k) = \lambda \nabla g(u_1, \ldots, u_k), \tag{3.1}$$

whereas

$$\partial L(u_1, \ldots, u_k, \lambda)/\partial \lambda = [g(u_1, \ldots, u_k) - 0] \tag{3.2}$$

recovers the constraint. LaGrange's principle requires solving the $k+1$ equations, (3.1) and (3.2), in the $k+1$ unknowns $\{u_1, \ldots, u_k, \lambda\}$. To minimize $f(u_1, \ldots, u_k)$ subject to $g(u_1, \ldots, u_k) \geq 0$, define the Lagrangian $L(\boldsymbol{u}, \lambda) = f(\boldsymbol{u}) - \lambda g(\boldsymbol{u})$. Stuezle[1] has given conditions for $\boldsymbol{u}^*$ to be a solution, namely, (i) $g(\boldsymbol{u}^*) \geq 0$; (ii) $\nabla_{\boldsymbol{u}} L(\boldsymbol{u}^*, \lambda^*) = \mathbf{0}$; (iii) $\lambda^* g(\boldsymbol{u}^*) = \mathbf{0}$; and (iv) $\lambda^* \geq 0$.

For constrained least squares the objective function is now

$$L(\beta_1, \ldots, \beta_k, \lambda) = Q(\boldsymbol{\beta}) + \lambda(\boldsymbol{\beta}'\boldsymbol{\beta} - c^2)$$

with $Q(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta})$ as before. Corresponding to (3.1) and (3.2) are

$$(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_k)\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y} \tag{3.3}$$

$$\boldsymbol{\beta}'\boldsymbol{\beta} = c^2 \tag{3.4}$$

to be solved for the $k+1$ unknowns $(\beta_1, \ldots, \beta_k, \lambda)$. Designate these as $\{\widehat{\boldsymbol{\beta}}_c, \widehat{\lambda}\}$ such that $\widehat{\boldsymbol{\beta}}_c'\widehat{\boldsymbol{\beta}}_c = c^2$, as apparently intended by Hoerl and Kennard (1970). If instead $Q(\boldsymbol{\beta})$ is to be minimized subject to $\{\boldsymbol{\beta}'\boldsymbol{\beta} \leq c^2\}$, then the constrained solution $\widehat{\boldsymbol{\beta}}_0$ satisfies $\widehat{\boldsymbol{\beta}}_0 = \widehat{\boldsymbol{\beta}}_L$ whenever $\widehat{\boldsymbol{\beta}}_L'\widehat{\boldsymbol{\beta}}_L < c^2$, and otherwise $(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_k)\boldsymbol{\beta}_0 = \boldsymbol{X}'\boldsymbol{Y}$ for some $\lambda > 0$ such that $\widehat{\boldsymbol{\beta}}_0'\widehat{\boldsymbol{\beta}}_0 = c^2$, as shown in Balakrishnan (1963). See also the conditions (i)–(iv) of Stuezle (2005) as cited.

3.2. **Ridge Regression: A Survey.** We recall essentials of conventional ridge regression as set forth principally in Hoerl and Kennard (1970), Marquardt (1970), and Marquardt and Snee (1975). For continuity we retain their notation, with their $\{k, p, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}^*\}$ corresponding to our $\{\lambda, k, \widehat{\boldsymbol{\beta}}_L, \widehat{\boldsymbol{\beta}}_{R_\lambda}\}$ and, on occasion, $\widehat{\boldsymbol{\beta}}_{R_\lambda} = \widehat{\boldsymbol{\beta}}_\lambda^* = \widehat{\boldsymbol{\beta}}^*(\lambda)$. Accordingly, write the residual sum of squares as $\phi = (\boldsymbol{Y} - \boldsymbol{Xb})'(\boldsymbol{Y} - \boldsymbol{Xb}) = \phi_{min} + \phi(\boldsymbol{b})$, where $\phi_{min} = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$ and $\phi(\boldsymbol{b}) = (\boldsymbol{b} - \widehat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{b} - \widehat{\boldsymbol{\beta}})$. Various assertions have been set forth, as enumerated here for later reference.

A1. Hoerl and Kennard (1970), p. 57: "$\widehat{\boldsymbol{\beta}}^* = [\boldsymbol{I}_p + k(\boldsymbol{X}'\boldsymbol{X})^{-1}]^{-1}\widehat{\boldsymbol{\beta}}$ (2.3)*."

A2. Hoerl and Kennard (1970), pp. 58–59: "The ridge trace can be shown to be following a path through the sums of squares surface so that for a fixed ϕ a single value for

[1] Stuezle, W., "Chapter 5. Notes on Ridge Regression," online notes for BioStat 538, Winter 2005, University of Washington, at www.stat.washington.edu/wxs/Stat538–w05

$\boldsymbol{b}$ is chosen and that is the one with minimal length." Precisely: "Minimize $\boldsymbol{b}'\boldsymbol{b}$ subject to $(\boldsymbol{b}-\widehat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{b}-\widehat{\boldsymbol{\beta}}) = \phi_0$ (3.2)*." "This reduces to $\boldsymbol{b} = \widehat{\boldsymbol{\beta}}^* = (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ where k is chosen to satisfy the restraint (3.2)*."

A3. Hoerl and Kennard (1970), p. 59: "Of course, in practice it is easier to choose a $k \geq 0$ and then to compute ϕ_0. In terms of $\widehat{\boldsymbol{\beta}}^*$ the residual sum of squares becomes $\phi^*(k) = (\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}^*)'(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}^*) = \phi_{min} + k^2\widehat{\boldsymbol{\beta}}^{*\prime}(\boldsymbol{X}'\boldsymbol{X})^{-1}\widehat{\boldsymbol{\beta}}^*$ (3.6)*."

A4. Hoerl and Kennard (1970), p. 59: "A completely equivalent statement of the problem is this: If the squared length of the regression vector $\boldsymbol{b}$ is fixed at R^2, then $\widehat{\boldsymbol{\beta}}^*$ is the value of $\boldsymbol{b}$ that gives a minimum sum of squares. That is, $\widehat{\boldsymbol{\beta}}^*$ is the value of $\boldsymbol{b}$ that minimizes the function $F_1 = (\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{b})'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{b}) + (1/k)(\boldsymbol{b}'\boldsymbol{b}-R^2)$ (3.7)*."

A5. Marquardt and Snee (1975), p. 5: "If $\widehat{\boldsymbol{\beta}}^*$ is the solution of $(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})\widehat{\boldsymbol{\beta}}^* = \boldsymbol{g}$, then $\widehat{\boldsymbol{\beta}}^*$ minimizes the sum of squares of residuals on the sphere centered at the origin whose radius is the length of $\widehat{\boldsymbol{\beta}}^*$." Here $\boldsymbol{g} = \boldsymbol{X}'\boldsymbol{Y}$.

3.3. **Properties of Solutions.** We next examine distributions of the constrained solutions $\widehat{\boldsymbol{\beta}}_c$, subject to $\widehat{\boldsymbol{\beta}}_c'\widehat{\boldsymbol{\beta}}_c = c^2$, to continue the unfinished work of Hoerl and Kennard (1970), and of $\widehat{\boldsymbol{\beta}}_0$ under inequality constraints. To these ends identify the sphere $S_c = \{\boldsymbol{u} \in \mathbb{R}^k : \boldsymbol{u}'\boldsymbol{u} = c^2\}$ and the open ball $B_c = \{\boldsymbol{u} \in \mathbb{R}^k : \boldsymbol{u}'\boldsymbol{u} < c^2\}$, both of radius c, and the complement $B_c^c = \{\boldsymbol{u} \in \mathbb{R}^k : \boldsymbol{u}'\boldsymbol{u} \geq c^2\}$. Accordingly, let $\mu(\cdot)$ be the probability measure on $\mathbb{R}^k$ induced through $\boldsymbol{Y} \to \widehat{\boldsymbol{\beta}}_L$; let $\mu_{S_c}(\cdot)$ be the measure on $S_c \subset \mathbb{R}^k$ induced through solutions $\widehat{\boldsymbol{\beta}}_c$ of (3.3) and (3.4); and let $\mu_{B_c}(\cdot)$ be the nonsingular measure on $B_c \subset \mathbb{R}^k$ induced through $\mathcal{L}(\widehat{\boldsymbol{\beta}}_L \,|\, \widehat{\boldsymbol{\beta}}_L'\widehat{\boldsymbol{\beta}}_L < c^2)$. Stochastic properties of $\widehat{\boldsymbol{\beta}}_c$ and $\widehat{\boldsymbol{\beta}}_0$ are given next.

Theorem 1. *Let $\widehat{\boldsymbol{\beta}}_c \in \mathbb{R}^k$ be the constrained solution satisfying (3.3) and (3.4), and let $\widehat{\boldsymbol{\beta}}_0 \in \mathbb{R}^k$ minimize $Q(\beta_1, \ldots, \beta_k)$ subject to $\{\boldsymbol{\beta}'\boldsymbol{\beta} \leq c^2\}$, with $\mu_0(\cdot)$ as its probability measure on $\mathbb{R}^k$.*

(i) The joint distribution $\mathcal{L}(\widehat{\boldsymbol{\beta}}_c) = F_c(\boldsymbol{b})$ corresponding to $\mu_{S_c}(\cdot)$ is singular on $\mathbb{R}^k$ of rank $k-1$.

(ii) The measure $\mu_0(\cdot)$ for $\widehat{\boldsymbol{\beta}}_0$ admits the mixture representation

$$\mu_0(A) = \alpha \cdot \mu_{B_c}(A) + \bar{\alpha} \cdot \mu_{S_c}(A) \tag{3.5}$$

with mixing probabilities $\alpha = 1 - \bar{\alpha} \in (0,1)$, such that

(iii) $\mu_{B_c}(A) = [\mu(B_c)]^{-1}\int_A I_{B_c}(\boldsymbol{t})d\mu(\boldsymbol{t})$, where $I_{B_c}(\boldsymbol{t})$ is the indicator function; and

(iv) $\alpha = \mu(B_c)$.

Proof: Conclusion (i) is immediate, since $\widehat{\boldsymbol{\beta}}_c \in \mathbb{R}^k$ constructively lies on the sphere $\widehat{\boldsymbol{\beta}}_c'\widehat{\boldsymbol{\beta}}_c = c^2$. We proceed by conditioning on the exclusive outcomes $\widehat{\boldsymbol{\beta}}_L \in B_c$ and $\widehat{\boldsymbol{\beta}}_L \in B_c^c$. Clearly $\widehat{\boldsymbol{\beta}}_0$ takes the value $\widehat{\boldsymbol{\beta}}_L$ with probability $\alpha = P(\widehat{\boldsymbol{\beta}}_L'\widehat{\boldsymbol{\beta}}_L < c^2) = \mu(B_c)$, where the conditional measure corresponding to $\mathcal{L}(\widehat{\boldsymbol{\beta}}_L \,|\, \widehat{\boldsymbol{\beta}}_L'\widehat{\boldsymbol{\beta}}_L < c^2)$ is $\mu_{B_c}(A) = [\mu(B_c)]^{-1}\int_A I_{B_c}(\boldsymbol{t})d\mu(\boldsymbol{t})$, as asserted, to give conclusion (iii). Similarly, $\widehat{\boldsymbol{\beta}}_0$ takes the value $\widehat{\boldsymbol{\beta}}_c$ with probability $\bar{\alpha} = 1 - \alpha$ as in (iv), its conditional measure as in (i), to complete our proof. □

Observe that the singular distribution $\mathcal{L}(\widehat{\boldsymbol{\beta}}_c)$ of conclusion (i) may be added to the list of distributions arising in the analysis of directional data, to include the von Mises–Fisher distributions, for example. For further reference see Batschelet (1981), Fisher (1993), Fisher, Lewis and Embleton (1993), Evans, Hastings and Peacock (2000), and Mardia and Jupp (2000). Conclusion (ii) for $\widehat{\boldsymbol{\beta}}_0$ complements the work of Balakrishnan (1963) in the context of linear estimation. Moreover, under Gaussian errors, $\alpha = \mu(B_c)$ derives from a weighted sum of k independent noncentral chi–squared random variables, each having a single degree of freedom; see Kotz, Johnson and Boyd (1967).

3.4. **A Critique.** We next reexamine the critical assertions of Section 3.2.

Assertion **A1:** False. As noted, the solution $\widehat{\boldsymbol{\beta}}_c$ necessarily lies on the sphere $\widehat{\boldsymbol{\beta}}_c'\widehat{\boldsymbol{\beta}}_c = c^2$ and thus has a joint singular distribution in $\mathbb{R}^k$ of rank $k-1$. To the contrary, Assertion **A1** implies that $\widehat{\boldsymbol{\beta}}^*(\lambda)$ has a nonsingular distribution for each $\lambda \geq 0$, yet $\widehat{\boldsymbol{\beta}}^*$ clearly refers to the constrained solution throughout Section 3 of Hoerl and Kennard (1970). The assertion is false, applying to solutions of (3.3) only, as there is no one–to–one linear transformation taking $\widehat{\boldsymbol{\beta}}_L$ onto the sphere $\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\beta}} = c^2$. In consequence, expression (3.6)* of Hoerl and Kennard (1970) is in error, as are its implications, since the term $k^2\widehat{\boldsymbol{\beta}}^{*\prime}(\boldsymbol{X}'\boldsymbol{X})^{-1}\widehat{\boldsymbol{\beta}}^*$ derives from the inapplicable Assertion **A1.**

Assertion **A2,** and its dual **A4,** appear to be essentially intact. The exception is that "$1/k$" in expression (3.7)* of Hoerl and Kennard (1970) instead should be "k."

Assertion **A5:** False. This assertion arises as the dual to **A3,** excluding (3.6)* of Hoerl and Kennard (1970). The basic idea is to solve (3.3) as $\widehat{\boldsymbol{\beta}}^*(\lambda)$ for fixed $\lambda > 0$, and then to discover the implied constraint $\{\boldsymbol{\beta}'\boldsymbol{\beta} = c^{*2}\}$ at (3.4) on evaluating $\widehat{\boldsymbol{\beta}}^{*\prime}\widehat{\boldsymbol{\beta}}^* = c^{*2}$. However, the solution $\widehat{\boldsymbol{\beta}}^*(\lambda)$ need not minimize the residual sum of squares $SS(\lambda) = [\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^*(\lambda)]'[\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^*(\lambda)]$, as claimed. This fallacy stems from the tacit but unfounded assumption that λ and c^2 correspond one–to–one. To the contrary, it is demonstrated in Section 5 that multiple solutions may have the same length but different λs, for example, $\|\widehat{\boldsymbol{\beta}}^*(\lambda_1)\| = \|\widehat{\boldsymbol{\beta}}^*(\lambda_2)\|$ with $\lambda_1 < \lambda_2$. But then the solution $\widehat{\boldsymbol{\beta}}^*(\lambda_2)$ cannot be minimizing, as $SS(\lambda_2) > SS(\lambda_1)$ from the monotonicity of $SS(\lambda)$. In this regard Figure 3 of Marquardt and Snee (1975) is particularly misleading. Assertions **A2,** "for a fixed ϕ a single value for $\boldsymbol{b}$ is chosen and that is the one with minimal length," and **A5,** that "$\widehat{\boldsymbol{\beta}}^*$ minimizes the sum of squares of residuals on the sphere centered at the origin whose radius is the length of $\widehat{\boldsymbol{\beta}}^*$," often are misrepresented as equivalent assertions regarding solutions $\widehat{\boldsymbol{\beta}}_{R_\lambda}$ of (3.3) alone. See van Nostrand (1980), for example.

To continue, for fixed c define the equivalence class

$$\Lambda(c) = \{\lambda : \|\widehat{\boldsymbol{\beta}}^*(\lambda)\| = c\}, \tag{3.6}$$

and let $\lambda_c = \min\{\Lambda(c)\}$. Then Assertion **A5** may be corrected as follows.

Assertion **A5*.** If $\widehat{\boldsymbol{\beta}}^*(\lambda)$ is a solution of $(\boldsymbol{X}'\boldsymbol{X}+\lambda\boldsymbol{I})\widehat{\boldsymbol{\beta}}^* = \boldsymbol{X}'\boldsymbol{Y}$ having length $\|\widehat{\boldsymbol{\beta}}^*(\lambda)\| = c^*$, then $\widehat{\boldsymbol{\beta}}^*(\lambda_c^*)$ minimizes the sum of squares of residuals on the sphere centered at the origin whose radius is the length c^* of $\widehat{\boldsymbol{\beta}}^*$, where $\lambda_c^* = \min\{\Lambda(c^*)\}$.

Assertion **A5*** has profound consequences in practice. Of the many schemes devised for choosing the ridge parameter λ, the user then must examine the corresponding equivalence class for each such λ. If it is a singleton set, then the solution thus attained is minimizing. Otherwise the algorithm **A5*** must be implemented to attain the minimizing solution. Further details are provided in Section 5.3.

It is clear that $\widehat{\boldsymbol{\beta}}_c$ is the LaGrange solution minimizing $Q(\boldsymbol{\beta})$ subject to $\{\boldsymbol{\beta}'\boldsymbol{\beta} = c^2\}$. To the contrary, Hoerl and Kennard (1970), Marquardt (1970), Marquardt and Snee (1975), Golub *et al.* (1979), and others concerned with constrained optimization, instead take $\widehat{\boldsymbol{\beta}}_{R_\lambda}$ as the ridge estimator, solving (3.3) alone for some $\lambda > 0$. Together with Assertion **A5,** this is tantamount to asserting that the k linear equations (3.3) somehow embody the constraint (3.4) as well, which they clearly cannot. Yet $\widehat{\boldsymbol{\beta}}_R$, not $\widehat{\boldsymbol{\beta}}_c$, comprise *the* ridge estimators on which essentially all of ridge regression now rests. Assertion **A1** clearly holds for solutions $\widehat{\boldsymbol{\beta}}_{R_\lambda}$ satisfying (3.3) only.

Confusion persists in the meaning of ridge regression. Bunke (1975), Hocking (1976), and Tibshirani (1996), for example, assert that ridge regression embodies the inequality constraint $\{\boldsymbol{\beta}'\boldsymbol{\beta} \leq c^2\}$, despite the disclaimer of Hoerl and Kennard (1970). Yet nowhere do these authors acknowledge the constrained solution $\widehat{\boldsymbol{\beta}}_0$ of Balakrishnan (1963), nor its properties as in Theorem 1, opting instead for the ridge solutions $\{\widehat{\boldsymbol{\beta}}_{R_\lambda}; \lambda \geq 0\}$ of Hoerl (1962, 1964). On the other hand, the inequality–constrained solution $\widehat{\boldsymbol{\beta}}_0$ does have the nonsingular mixture distribution of Theorem 1. However, we are aware of no work in ridge regression that explicitly accounts for the structure of either $\widehat{\boldsymbol{\beta}}_c$ or of $\widehat{\boldsymbol{\beta}}_0$ as in Theorem 1.

In short, ridge regression in its present form rests essentially on $\widehat{\boldsymbol{\beta}}_R$ through an accident of history. Indeed, expressions for variances and biases; solutions for λ purporting to minimize expected mean squares; prediction, validation, and cross–validation; and other aspects of ridge regression; all are predicated on Assertion **A1.** If instead either $\widehat{\boldsymbol{\beta}}_c$ or $\widehat{\boldsymbol{\beta}}_0$ were taken as starting points, as required under the aegis of constrained optimization, then the ensuing "ridge regressions" would differ dramatically from the conventional one based on $\{\widehat{\boldsymbol{\beta}}_{R_\lambda}; \lambda \geq 0\}$, together with the critical but false Assertion **A1,** with c^2 now corresponding to λ. These differences necessarily would include issues such as (i) the stability of the solutions $\widehat{\boldsymbol{\beta}}_c$ or $\widehat{\boldsymbol{\beta}}_0$ instead of $\widehat{\boldsymbol{\beta}}_{R_\lambda}$, in comparison with $\widehat{\boldsymbol{\beta}}_L$; (ii) the inflation of variances, taking into account actual variances to be derived from Theorem 1 as reference; (iii) prediction using $\widehat{\boldsymbol{Y}}_c = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_c$ or $\widehat{\boldsymbol{Y}}_0 = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_0$, instead of $\widehat{\boldsymbol{Y}}_{R_\lambda} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{R_\lambda}$; (iv) the use, meaning, and properties of cross–validative and predictive criteria based on $\widehat{\boldsymbol{Y}}_c$ or $\widehat{\boldsymbol{Y}}_0$, instead of $\widehat{\boldsymbol{Y}}_{R_\lambda}$; (v) ridge traces as modified to take into account $\widehat{\boldsymbol{\beta}}_0$ and singularity of the joint distribution of $\widehat{\boldsymbol{\beta}}_c$; and (vi) the trade–off between bias and variance of the constrained estimators $\widehat{\boldsymbol{\beta}}_c$ and $\widehat{\boldsymbol{\beta}}_0$, as determined using actual moments to be derived from Theorem 1. Other differences may be noted. All such properties would have to be established anew, complicated considerably by the nonstandard distributions encountered in Theorem 1.

By analogy, Hoerl and Kennard (1970) further considered generalized ridge regression invoking the k equations $(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Lambda})\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$, with $\boldsymbol{\Lambda} = \mathrm{Diag}(\lambda_1, \ldots, \lambda_k)$ as nonnegative ridge parameters. Note that this, too, cannot have resulted from LaGrange minimization: Given that $\{\beta_1^2 = c_1^2, \ldots, \beta_k^2 = c_k^2\}$, the only function of the data now would be to determine signs of the roots $\{\widehat{\beta}_1 = \pm c_1, \ldots, \widehat{\beta}_k = \pm c_k\}$. On the other hand, if inequality constraints $\{\beta_1^2 \le c_1^2, \ldots, \beta_k^2 \le c_k^2\}$ are invoked instead, then correct solutions are provided by Myoken and Uchida (1977) akin to those of Balakrishnan (1963) where $\{\lambda_1 = \cdots = \lambda_k = \lambda\}$.

4. Foundations Via Conditioning

We seek substitutes for the failed principle of constrained optimization as a basis for conventional ridge regression. In what follows we consider $\{\widehat{\boldsymbol{\beta}}_{R_\lambda}; \lambda \ge 0\}$ as solutions to (3.3) alone as in Hoerl (1962, 1964), without reference to constrained optimization and discredited assertions thereto as noted. Type A conditioning of the linear system $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$ prompts the modification $(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I}_k)\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$, from the perspective of both numerical analysis (Levenberg (1944) and Riley (1955)) and of statistics (Hoerl (1962, 1964)). A survey is provided subsequently. Moreover, the Type B conditioning of $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is also germane, since the conditioning of $\boldsymbol{X}'\boldsymbol{X}$ depends on that of $\boldsymbol{X}$, and for further reasons to be cited. A new approach to ill conditioned systems, using *surrogate ridge models,* rests essentially on Type B conditioning. Details follow.

4.1. **Background.** Ill–conditioned models typically arise from nonorthogonality of columns of $\boldsymbol{X}$. Let $\boldsymbol{W} = \boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{V} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$. Since $\mathrm{V}(\widehat{\boldsymbol{\beta}}_L) = \sigma^2\boldsymbol{V}$, the *variance inflation factors* (*VIF*s) of $\widehat{\boldsymbol{\beta}}_L = [\widehat{\beta}_{L_1}, \ldots, \widehat{\beta}_{L_k}]'$ are defined as $\{\mathit{VIF}(\widehat{\beta}_{L_j}) = v_{jj}/w_{jj}^{-1}; 1 \le j \le k\}$, *i.e.,* the ratio of the actual variance to the "ideal" variance attained when columns of $\boldsymbol{X}$ are orthogonal, so that $\boldsymbol{W} = \mathrm{Diag}(w_{11}, \ldots, w_{kk})$. Often $\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is taken with $\boldsymbol{Z}'\boldsymbol{Z}$ in "correlation form" having unit diagonal elements; then $\{\mathit{VIF}(\widehat{\beta}_{L_j}) = v_{jj}; 1 \le j \le k\}$ are diagonal elements of $\boldsymbol{V} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ from the scale–invariance of *VIF*s. With $\{V_1 \ge V_2 \ge \ldots \ge V_k\}$ as the *ordered* diagonal elements of $\boldsymbol{V}$, Marquardt and Snee (1975) identify V_1 to be "the best single measure of the conditioning of the data," thus a critical diagnostic tool. See also Marquardt (1970), Beaton, Rubin and Barone (1976), and Davies and Hutton (1975). A basic connection between *VIF*s and condition numbers is due to Berk (1977):

Lemma 1. *Given $\boldsymbol{Z}'\boldsymbol{Z}$ in correlation form, with $\{V_1 \ge V_2 \ge \ldots \ge V_k\}$ as the ordered diagonal elements of $\boldsymbol{V} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$. Then the condition number $c_1(\boldsymbol{Z}'\boldsymbol{Z})$ satisfies*

$$V_1 \le c_1(\boldsymbol{Z}'\boldsymbol{Z}) \le k(V_1 + \cdots + V_k). \tag{4.1}$$

Since $\{c_\phi(\boldsymbol{A}) = c_\phi(\boldsymbol{A}^{-1}); \phi \in \Phi\}$ from Section 2.2, the Type A condition number for $\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\beta} = \boldsymbol{Z}'\boldsymbol{Y}$ is identical to $c_\phi[\mathrm{V}(\widehat{\boldsymbol{\beta}}_L)]$, so that Lemma 1 is really about dispersion parameters in the equivalent form

$$V_1 \le c_1[\mathrm{V}(\widehat{\boldsymbol{\beta}}_L)] \le k(V_1 + \cdots + V_k). \tag{4.2}$$

4.2. **Ridge Regression.** That $\boldsymbol{X}'\boldsymbol{X} \to (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I}_k)$ improves conditioning has been cited by Marshall and Olkin (1979) as a justification for ridge regression. In brief, their Theorem C.3, p. 273, asserts that for any $(\boldsymbol{A}, \boldsymbol{B}) \in \mathbb{S}_k^+$ such that $c_\phi(\boldsymbol{B}) \le c_\phi(\boldsymbol{A})$, with $\{c_\phi(\cdot); \phi \in \Phi\}$ as in Section 2.2, then $c_\phi(\boldsymbol{A} + \boldsymbol{B}) \le c_\phi(\boldsymbol{A})$. Riley (1955) showed that $\boldsymbol{B} = \lambda\boldsymbol{I}_k$ satisfies the hypothesis of the theorem for any $\boldsymbol{A} \in \mathbb{S}_k^+$, where λ depends on numerical considerations.

This holds for any Type A conditioning of $\boldsymbol{Az} = \boldsymbol{b} \to (\boldsymbol{A} + \lambda\boldsymbol{I}_k)\boldsymbol{z} = \boldsymbol{b}$ as in Section 2.2, and thus in particular for $\boldsymbol{X'X\beta} = \boldsymbol{X'Y} \to (\boldsymbol{X'X} + \lambda\boldsymbol{I}_k)\boldsymbol{\beta} = \boldsymbol{X'Y}$, as noted by Marshall and Olkin (1979), p. 273, to give Type A conditioning as a basis for ridge regression. Moreover, using condition numbers $c_1(\cdot)$, the improvement is seen directly on comparing $c_1(\boldsymbol{X'X})$ $= \xi_1^2/\xi_k^2$ with $c_1(\boldsymbol{X'X} + \lambda\boldsymbol{I}_k) = (\xi_1^2 + \lambda)/(\xi_k^2 + \lambda)$, where $\sigma(\boldsymbol{X}) = [\xi_1, \ldots, \xi_k]'$. Essential properties of $\widehat{\boldsymbol{\beta}}_L$ and $\widehat{\boldsymbol{\beta}}_{R_\lambda}$ are summarized in Table 1, along with the *surrogate estimator,* $\widehat{\boldsymbol{\beta}}_{S_\lambda}$, to be defined subsequently.

TABLE 1. Properties of $\{\widehat{\boldsymbol{\beta}}_L, \widehat{\boldsymbol{\beta}}_{R_\lambda}, \widehat{\boldsymbol{\beta}}_{S_\lambda}\}$ under Gauss–Markov assumptions, where $\boldsymbol{X}_\lambda = \boldsymbol{P}\mathrm{Diag}(\sqrt{\xi_1^2 + \lambda}, \ldots, \sqrt{\xi_k^2 + \lambda})\boldsymbol{Q}'$ and $\boldsymbol{A}_\lambda = (\boldsymbol{X'X} + \lambda\boldsymbol{I}_k)$.

Estimator	Definition	$\mathrm{E}(\widehat{\boldsymbol{\beta}})$	$\mathrm{V}(\widehat{\boldsymbol{\beta}})$
$\widehat{\boldsymbol{\beta}}_L$	$(\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$	$\boldsymbol{\beta}$	$\sigma^2(\boldsymbol{X'X})^{-1}$
$\widehat{\boldsymbol{\beta}}_{R_\lambda}$	$\boldsymbol{A}_\lambda^{-1}\boldsymbol{X'Y}$	$\boldsymbol{A}_\lambda^{-1}\boldsymbol{X'X\beta}$	$\sigma^2\boldsymbol{A}_\lambda^{-1}\boldsymbol{X'XA}_\lambda^{-1}$
$\widehat{\boldsymbol{\beta}}_{S_\lambda}$	$\boldsymbol{A}_\lambda^{-1}\boldsymbol{X}_\lambda'\boldsymbol{Y}$	$\boldsymbol{A}_\lambda^{-1}\boldsymbol{X}_\lambda'\boldsymbol{X\beta}$	$\sigma^2\boldsymbol{A}_\lambda^{-1}$

4.3. **Surrogate Models.** Nonetheless, the correspondence $\boldsymbol{Az} = \boldsymbol{b} \longleftrightarrow \boldsymbol{X'X\beta} = \boldsymbol{X'Y}$ is incomplete in the context of linear inference, since both $\boldsymbol{A} = \boldsymbol{X'X}$ and $\boldsymbol{b} = \boldsymbol{X'Y}$ are subject to disturbances in $\boldsymbol{X}$. This has not been taken into account. In particular, ridge solutions satisfying $(\boldsymbol{X'X} + \lambda\boldsymbol{I}_k)\boldsymbol{\beta} = \boldsymbol{X'Y}$, despite improved conditioning on the left, still are subject to the ill–conditioning of $\boldsymbol{X}$ on the right. To correct this oversight, we invoke Type B conditioning from Section 2.2 on observing that $\boldsymbol{X'X} \to (\boldsymbol{X'X} + \lambda\boldsymbol{I}_k)$ is tantamount to modifying $\boldsymbol{X}$ itself as a means to enhanced conditioning. In particular, begin with the singular decomposition $\boldsymbol{X} = \boldsymbol{PD}_\xi\boldsymbol{Q}'$; let $\boldsymbol{X}_\lambda = \boldsymbol{P}\mathrm{Diag}(\sqrt{\xi_1^2 + \lambda}, \ldots, \sqrt{\xi_k^2 + \lambda})\boldsymbol{Q}'$; observe that $(\boldsymbol{X'X} + \lambda\boldsymbol{I}_k) = \boldsymbol{X}_\lambda'\boldsymbol{X}_\lambda$; and note that ridge regression entails $\boldsymbol{X}_\lambda'\boldsymbol{X}_\lambda\boldsymbol{\beta} = \boldsymbol{X'Y}$. Instead, we take $\boldsymbol{Y} = \boldsymbol{X}_\lambda\boldsymbol{\beta} + \boldsymbol{\epsilon}$ as an approximation, or *surrogate,* for the ill–conditioned model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ itself, as in the following.

Definition 1. Given an ill–conditioned model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, its *ridge surrogate* is a modified model $\boldsymbol{Y} = \boldsymbol{X}_\lambda\boldsymbol{\beta} + \boldsymbol{\epsilon}$. The *surrogate estimator* $\widehat{\boldsymbol{\beta}}_{S_\lambda}$, solving $\boldsymbol{X}_\lambda'\boldsymbol{X}_\lambda\boldsymbol{\beta} = \boldsymbol{X}_\lambda'\boldsymbol{Y}$, is *OLS* for the surrogate model.

To continue, the order of approximation of $\boldsymbol{X}_\lambda$ for $\boldsymbol{X}$ may be gauged by the Frobenius distance

$$\| \boldsymbol{X} - \boldsymbol{X}_\lambda \|_F = \left[\sum_{i=1}^{k} \left(\xi_i - \sqrt{\xi_i^2 + \lambda} \right)^2 \right]^{1/2}, \tag{4.3}$$

from the unitary invariance of $\| \cdot \|_F$. Moreover, the conditioning of $\boldsymbol{X}_\lambda'\boldsymbol{X}_\lambda\boldsymbol{\beta} = \boldsymbol{X}_\lambda'\boldsymbol{Y}$ now may be gauged through Type B conditioning as in Section 2.2. For later reference, basic properties of $\{\widehat{\boldsymbol{\beta}}_{S_\lambda}; \lambda \geq 0\}$ are summarized in Table 1. It remains to compare properties of $\{\widehat{\boldsymbol{\beta}}_L, \widehat{\boldsymbol{\beta}}_{R_\lambda}, \widehat{\boldsymbol{\beta}}_{S_\lambda}\}$. Direct comparisons are somewhat obscure; however, these become more transparent on invoking canonical forms to be considered next.

4.4. **Canonical Forms.** The singular decomposition $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{D}_\xi\boldsymbol{Q}'$, with $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}_k$, together with the orthogonal reparametrization $\boldsymbol{\theta} = \boldsymbol{Q}'\boldsymbol{\beta}$, gives $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \rightarrow \boldsymbol{Y} = \boldsymbol{P}\boldsymbol{D}_\xi\boldsymbol{Q}'\boldsymbol{\beta} + \boldsymbol{\epsilon} \rightarrow \boldsymbol{U} = \boldsymbol{P}'\boldsymbol{Y} = \boldsymbol{D}_\xi\boldsymbol{\theta} + \boldsymbol{P}'\boldsymbol{\epsilon}$, such that $\mathrm{E}(\boldsymbol{P}'\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\mathrm{V}(\boldsymbol{P}'\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{P}'\boldsymbol{I}_n\boldsymbol{P} = \sigma^2\boldsymbol{I}_k$ under Gauss–Markov assumptions regarding the errors of $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Accordingly, $\mathrm{E}(\boldsymbol{U}) = \boldsymbol{D}_\xi\boldsymbol{\theta}$ and $\mathrm{V}(\boldsymbol{U}) = \sigma^2\boldsymbol{I}_k$. In canonical form it follows that $\widehat{\boldsymbol{\theta}}_L = (\boldsymbol{D}_\xi^2)^{-1}\boldsymbol{D}_\xi\boldsymbol{U} = \boldsymbol{D}_\xi^{-1}\boldsymbol{U}$, $\mathrm{E}(\widehat{\boldsymbol{\theta}}_L) = \boldsymbol{\theta}$, and $\mathrm{V}(\widehat{\boldsymbol{\theta}}_L) = \sigma^2\boldsymbol{D}_\xi^{-2}$ under *OLS,* as given in Table 2. Similar expressions for the canonical ridge estimators $\{\widehat{\boldsymbol{\theta}}_{R_\lambda}; \lambda \geq 0\}$, and the canonical surrogate ridge estimators $\{\widehat{\boldsymbol{\theta}}_{S_\lambda}; \lambda \geq 0\}$, are reported in Table 2. Since $\widehat{\boldsymbol{\beta}} = \boldsymbol{Q}\widehat{\boldsymbol{\theta}}$, $\mathrm{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{Q}\mathrm{E}(\widehat{\boldsymbol{\theta}})$, and $\mathrm{V}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{Q}\mathrm{V}(\widehat{\boldsymbol{\theta}})\boldsymbol{Q}'$ for all three estimators, Table 1 follows directly from Table 2, and conversely. Moreover, issues regarding the conditioning of $\{\widehat{\boldsymbol{\beta}}_L, \widehat{\boldsymbol{\beta}}_{R_\lambda}, \widehat{\boldsymbol{\beta}}_{S_\lambda}\}$, as linear data transformations, and conditioning of the corresponding dispersion matrices $\{\mathrm{V}(\widehat{\boldsymbol{\beta}}_L), \mathrm{V}(\widehat{\boldsymbol{\beta}}_{R_\lambda}), \mathrm{V}(\widehat{\boldsymbol{\beta}}_{S_\lambda})\}$, are considered subsequently. These can be established directly in terms of those of $\{\widehat{\boldsymbol{\theta}}_L, \widehat{\boldsymbol{\theta}}_{R_\lambda}, \widehat{\boldsymbol{\theta}}_{S_\lambda}\}$, since $\boldsymbol{Q}$ is orthogonal and condition numbers here are unitarily invariant.

Table 2. Properties of $\{\widehat{\boldsymbol{\theta}}_L, \widehat{\boldsymbol{\theta}}_{R_\lambda}, \widehat{\boldsymbol{\theta}}_{S_\lambda}\}$ under standard Gauss–Markov assumptions, where $\boldsymbol{U} = \boldsymbol{P}'\boldsymbol{Y}$ and $\boldsymbol{D}(\omega_i) = \mathrm{Diag}(\omega_1, \ldots, \omega_k)$.

Estimator	Definition	$\mathrm{E}(\widehat{\boldsymbol{\theta}})$	$\mathrm{V}(\widehat{\boldsymbol{\theta}})$
$\widehat{\boldsymbol{\theta}}_L$	$\boldsymbol{D}_\xi^{-1}\boldsymbol{U}$	$\boldsymbol{\theta}$	$\sigma^2\boldsymbol{D}_\xi^{-2}$
$\widehat{\boldsymbol{\theta}}_{R_\lambda}$	$\boldsymbol{D}(\xi_i/(\xi_i^2+\lambda))\boldsymbol{U}$	$\boldsymbol{D}(\xi_i^2/(\xi_i^2+\lambda))\boldsymbol{\theta}$	$\sigma^2\boldsymbol{D}(\xi_i^2/(\xi_i^2+\lambda)^2)$
$\widehat{\boldsymbol{\theta}}_{S_\lambda}$	$\boldsymbol{D}(1/\sqrt{\xi_i^2+\lambda})\boldsymbol{U}$	$\boldsymbol{D}(\xi_i/(\sqrt{\xi_i^2+\lambda}))\boldsymbol{\theta}$	$\sigma^2\boldsymbol{D}(1/(\xi_i^2+\lambda))$

Specifically, in canonical form we have $\boldsymbol{D}_\xi\widehat{\boldsymbol{\theta}}_L = \boldsymbol{U}$, so that the Type B condition number $c_1(\boldsymbol{X}) = c_1(\boldsymbol{D}_\xi) = \xi_1/\xi_k$ properly gauges the sensitivity of the solution $\widehat{\boldsymbol{\beta}}_L$ to disturbances in $\boldsymbol{X}$. Similarly, with $\boldsymbol{D}_\xi^\lambda = \mathrm{Diag}((\xi_1^2+\lambda)/\xi_1^2, \ldots, (\xi_k^2+\lambda)/\xi_k^2)$, observe from $\boldsymbol{D}_\xi^\lambda\widehat{\boldsymbol{\theta}}_{R_\lambda} = \boldsymbol{U}$ that its condition number gauges sensitivity of the solution $\widehat{\boldsymbol{\theta}}_{R_\lambda}$, and thus of $\widehat{\boldsymbol{\beta}}_{R_\lambda} = \boldsymbol{Q}\widehat{\boldsymbol{\theta}}_{R_\lambda}$ to perturbations in $\boldsymbol{X}$, from the orthogonality of $\boldsymbol{Q}$. This underscores the central role of Type B conditioning from Section 2.2, as set forth in Belsley *et al.* (1970).

4.5. **Central Issues.** Several issues, to be examined empirically in Section 5, appear to be open questions not addressed in the voluminous literature on ridge regression. Intrinsic difficulties with *OLS* include (i) nonorthogonality of the columns of $\boldsymbol{X}$, as reflected in $c_\phi(\boldsymbol{X})$ and $c_\phi(\boldsymbol{X}'\boldsymbol{X})$; (ii) instability of solutions linked to the conditioning of the data transformation $\widehat{\boldsymbol{\beta}}_L(\boldsymbol{Y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$, considered as a function of $\boldsymbol{Y}$; and (iii) pathologies in dispersion parameters as reflected in *VIF*s and the ill–conditioning of $\mathrm{V}(\widehat{\boldsymbol{\beta}}_L)$. Moreover, at some level the conditioning of $\mathrm{E}(\widehat{\boldsymbol{\beta}}_{R_\lambda}) = T(\boldsymbol{\beta})$ becomes an issue in transforming the parameter space, as in assessing the trade-off between variance and bias. As ridge regression seeks remedies, it is pertinent to ask how well the ridge solutions progress towards those ends. Regarding item (i), the apparent "correlations" in $\boldsymbol{W} = \boldsymbol{X}'\boldsymbol{X}$, namely $\{w_{ij}/\sqrt{w_{ii}w_{jj}}\}$, are taken into $\{w_{ij}/\sqrt{(w_{ii}+\lambda)(w_{jj}+\lambda)}\}$ as elements of $(\boldsymbol{X}'\boldsymbol{X}+\lambda\boldsymbol{I}_k)$. These in turn decrease in magnitude with increasing λ. Nonetheless, ridge solutions themselves are subject to nonorthogonality, together with attendant difficulties regarding stability, *VIF*s, and conditioning of their dispersion matrices. Improving stability of the solutions thus hinges on the conditioning of $\widehat{\boldsymbol{\beta}}_{R_\lambda}(\boldsymbol{Y})$ when considered as a data transformation. Moreover, the capacity to ameliorate dispersion problems of *OLS* hinges on improving *VIF*s and condition numbers for $\mathrm{V}(\widehat{\boldsymbol{\beta}}_{R_\lambda})$. On the other hand, it is widely known that $\widehat{\boldsymbol{\beta}}_{R_\lambda}$ shrinks stochastically towards the origin, as do its mean and dispersion matrix, with increasing λ. These issues in turn prompt several questions to be considered subsequently.

Q1: Does it follow that stability of $\widehat{\boldsymbol{\beta}}_{R_\lambda}(\boldsymbol{Y})$ necessarily improves with increasing λ?

Q2: Given that $\mathrm{V}(\widehat{\boldsymbol{\beta}}_{R_\lambda}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I}_k)^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I}_k)^{-1}$, does it follow that condition numbers $c_1[\mathrm{V}(\widehat{\boldsymbol{\beta}}_{R_\lambda})]$ decrease with increasing λ?

Q3: With regard to variance inflation, does it follow that *VIF*s for elements of $\widehat{\boldsymbol{\beta}}_{R_\lambda}$ decrease with increasing λ?

Q4: Viewing $\mathrm{E}(\widehat{\boldsymbol{\beta}}_{R_\lambda}) = T(\boldsymbol{\beta})$ as a transformation on the space of parameters, does it follow that its conditioning improves with increasing λ?

For completeness, observe that the foregoing issues pertain not only to the ridge estimators $\{\widehat{\boldsymbol{\beta}}_{R_\lambda}; \lambda \geq 0\}$ themselves, but also to other biased solutions to include $\{\widehat{\boldsymbol{\beta}}_{S_\lambda}; \lambda \geq 0\}$.

We next undertake a comparative study of properties of ridge and surrogate ridge solutions, to be continued in the case studies of Section 5.

4.6. **Some Comparisons.** Regarding the conventional $\{\widehat{\boldsymbol{\beta}}_{R_\lambda}; \lambda \geq 0\}$ and surrogate ridge $\{\widehat{\boldsymbol{\beta}}_{S_\lambda}; \lambda \geq 0\}$ estimators, both shrink stochastically towards the origin with increasing λ, as do their means and variances, and similarly for $\{\widehat{\boldsymbol{\theta}}_{R_\lambda}; \lambda \geq 0\}$ and $\{\widehat{\boldsymbol{\theta}}_{S_\lambda}; \lambda \geq 0\}$. Specifically, for a given λ, it is seen from Table 2 that $\widehat{\boldsymbol{\theta}}_{S_\lambda}$ achieves lesser shrinkage, both in expectation and variance, than $\widehat{\boldsymbol{\theta}}_{R_\lambda}$.

Condition numbers for various arrays are given in Table 3 for the canonical estimators $\{\widehat{\boldsymbol{\theta}}_L, \widehat{\boldsymbol{\theta}}_{R_\lambda}, \widehat{\boldsymbol{\theta}}_{S_\lambda}\}$. These arrays include (i) coefficients defining $\widehat{\boldsymbol{\theta}}(\boldsymbol{U})$ with reference to stability of the solutions; (ii) coefficients defining the parameter transformations $\mathrm{E}(\widehat{\boldsymbol{\theta}}) = T(\boldsymbol{\theta})$; and (iii) the dispersion matrix $\mathrm{V}(\widehat{\boldsymbol{\theta}})$. Entries in Table 3 follow directly from Table 2 and the definition of $c_1(\cdot)$, on recalling that elements of $\boldsymbol{D}_\xi = \mathrm{Diag}(\xi_1, \ldots, \xi_k)$ are ordered as $\{\xi_1 \geq \ldots \geq \xi_k > 0\}$. Observe, moreover, that the rows of Table 3 may be identified equivalently as $\{\widehat{\boldsymbol{\beta}}_L, \widehat{\boldsymbol{\beta}}_{R_\lambda}, \widehat{\boldsymbol{\beta}}_{S_\lambda}\}$, and the columns as $\{c_1[\widehat{\boldsymbol{\beta}}(\boldsymbol{Y})], c_1[T(\boldsymbol{\beta})], c_1[\mathrm{V}(\widehat{\boldsymbol{\beta}})]\}$, respectively. This follows since $\boldsymbol{\beta} = \boldsymbol{Q}\boldsymbol{\theta}$, $\widehat{\boldsymbol{\beta}} = \boldsymbol{Q}\widehat{\boldsymbol{\theta}}$, and $\mathrm{V}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{Q}\mathrm{V}(\widehat{\boldsymbol{\theta}})\boldsymbol{Q}'$, $\boldsymbol{Q}$ is orthogonal, and the condition numbers are unitarily invariant.

Note further that $c_1[\widehat{\boldsymbol{\beta}}_L(\boldsymbol{Y})] = c_1(\boldsymbol{X})$ and $c_1[\mathrm{V}(\widehat{\boldsymbol{\beta}}_L)] = c_1(\boldsymbol{X}'\boldsymbol{X})$, whereas $c_1[\widehat{\boldsymbol{\beta}}_{S_\lambda}(\boldsymbol{Y})] = c_1(\boldsymbol{X}_\lambda)$ and $c_1[\mathrm{V}(\widehat{\boldsymbol{\beta}}_{S_\lambda})] = c_1(\boldsymbol{X}_\lambda'\boldsymbol{X}_\lambda)$, as both are *OLS* in their respective models. Moreover, both condition numbers, $c_1[\widehat{\boldsymbol{\beta}}_{S_\lambda}(\boldsymbol{Y})] = (\sqrt{\xi_1^2 + \lambda}/\sqrt{\xi_k^2 + \lambda})$, and its square $c_1[\mathrm{V}(\widehat{\boldsymbol{\beta}}_{S_\lambda})]$, decrease monotonically with increasing λ, thus assuring improved conditioning for the surrogate estimators. Condition numbers associated with $\widehat{\boldsymbol{\theta}}_{R_\lambda}$, and thus with $\widehat{\boldsymbol{\beta}}_{R_\lambda}$, are more convoluted and will be examined further in Section 5.

TABLE 3. Condition numbers for data transformations $\widehat{\boldsymbol{\theta}}(\boldsymbol{U})$, for parameter transformations $\mathrm{E}(\widehat{\boldsymbol{\theta}}) = T(\boldsymbol{\theta})$, and for $\mathrm{V}(\widehat{\boldsymbol{\theta}})$, for each of $\{\widehat{\boldsymbol{\theta}}_L, \widehat{\boldsymbol{\theta}}_{R_\lambda}, \widehat{\boldsymbol{\theta}}_{S_\lambda}\}$.

Estimator	$c_1[\widehat{\boldsymbol{\theta}}(\boldsymbol{U})]$	$c_1[T(\boldsymbol{\theta})]$	$c_1[\mathrm{V}(\widehat{\boldsymbol{\theta}})]$
$\widehat{\boldsymbol{\theta}}_L$	$\frac{\xi_1}{\xi_k}$	1.00	$\frac{\xi_1^2}{\xi_k^2}$
$\widehat{\boldsymbol{\theta}}_{R_\lambda}$	$\frac{\max\{\xi_i/(\xi_i^2+\lambda)\}}{\min\{\xi_i/(\xi_i^2+\lambda)\}}$	$\frac{\xi_1^2(\xi_k^2+\lambda)}{\xi_k^2(\xi_1^2+\lambda)}$	$\frac{\max\{\xi_i^2/(\xi_i^2+\lambda)^2\}}{\min\{\xi_i^2/(\xi_i^2+\lambda)^2\}}$
$\widehat{\boldsymbol{\theta}}_{S_\lambda}$	$\frac{\sqrt{\xi_1^2+\lambda}}{\sqrt{\xi_k^2+\lambda}}$	$\frac{\xi_1\sqrt{\xi_k^2+\lambda}}{\xi_k\sqrt{\xi_1^2+\lambda}}$	$\frac{\xi_1^2+\lambda}{\xi_k^2+\lambda}$

5. Case Studies

5.1. **The Data.** We reexamine the Hospital Manpower Data as reported in Myers (1990). Records at $n = 17$ U. S. Naval Hospitals include: Y : Monthly man–hours; X_1 : Average daily patient load; X_2 : Monthly X–ray exposures; X_3 : Monthly occupied bed days; X_4 : Eligible population in the area ÷ 1000; and X_5 : Average length of patients' stay in days. The basic model is

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon_i; 1 \le i \le n. \tag{5.1}$$

Following Hoerl and Kennard (1970), Marquardt (1970), Marquardt and Snee (1975), Myers (1990), and others, we center and scale the model, so that $\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{Z}'\boldsymbol{Z}$ in correlation form, the central focus being the rates of change $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3, \beta_4, \beta_5]'$. The data are given in Table 3.8, pp. 132–133, of Myers (1990), and computations were done mostly using PROC IML of the SAS Programming System. The data are exceedingly ill–conditioned: Elements of $\boldsymbol{D}_\xi$ are $\boldsymbol{D}_\xi = \mathrm{Diag}(2.048687, 0.816997, 0.307625, 0.201771, 0.007347)$; $c_1(\boldsymbol{Z}'\boldsymbol{Z})$ = 77,754.86; the maximal *VIF* in *OLS* estimation is $V_1 = \mathit{VIF}(\widehat{\beta}_1) = 9{,}595.685$; and other *VIF*s appear subsequently in Table 8 at $\lambda = 0$.

5.2. **Choices for λ.** Widely diverse criteria have evolved in the choice for λ, with profound consequences regarding ridge estimators, ridge predictors, and their properties. Five criteria in common usage are reported in Table 4, together with definitions and their values as determined for the Hospital Manpower Data. These include $\mathit{DF}_\lambda = \mathrm{tr}(\boldsymbol{H}_\lambda)$ with $\boldsymbol{H}_\lambda = [\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{I}_k)^{-1}\boldsymbol{Z}']$; the cross–validation PRESS_λ statistic of Allen (1974); a rotation–invariant version called Generalized Cross Validation (GCV_λ) by Golub *et al.* (1979); C_λ as a device for variance–bias trade–off as in Mallows (1973); and HKB_λ as recommended by Hoerl, Kennard and Baldwin (1975) based on simulation studies. As listed in Table 4, $SS_{Res,\lambda}$ is the residual sum of squares using ridge regression; $\widehat{\sigma}^2$ is the *OLS* residual mean square; and $\{e^2_{(i,\lambda)}\}$ are the *PRESS* residuals for ridge regression. Further details are given in Myers (1990), pp. 392–411, including numerical values for DF_λ, C_λ, and PRESS_λ as reported in Table 4. Further choices include $\lambda \in \{0.01, 0.03, 0.05, 0.07, 0.09\}$ and others to be noted subsequently.

Table 4. Choices for λ in the Hospital Manpower Data corresponding to conventional criteria DF_λ, GCV_λ, C_λ, $PRESS_\lambda$, and HKB_λ.

Name	Definition	Value for λ
DF_λ	$\mathrm{tr}(\boldsymbol{H}_\lambda) = \sum_{i=1}^{k} \frac{\xi_i^2}{(\xi_i^2+\lambda)}$	0.0004
GCV_λ	$\frac{SS_{Res,\lambda}}{[n-(1+\mathrm{tr}(\boldsymbol{H}_\lambda))]^2}$	0.004787
C_λ	$[\frac{SS_{Res,\lambda}}{\widehat{\sigma}^2} - n + 2 + 2\mathrm{tr}(\boldsymbol{H}_\lambda)]$	0.0050
$PRESS_\lambda$	$\sum_{i=1}^{n} e_{(i,\lambda)}^2$	0.2300
HKB_λ	$\frac{k\widehat{\sigma}^2}{\widehat{\boldsymbol{\beta}}_L'\widehat{\boldsymbol{\beta}}_L}$	0.616964

5.3. **Minimizing Solutions.** Often a definitive value for the constraint $\{\boldsymbol{\beta}'\boldsymbol{\beta} = c^2\}$ is not apparent in a particular study. This motivates the dual Assertions **A3** and **A5** of Section 3.2: (i) choose λ; (ii) solve (3.3) for $\widehat{\boldsymbol{\beta}}_{R_\lambda}$; (iii) evaluate the implied constraint at (3.4) as $\widehat{\boldsymbol{\beta}}_{R_\lambda}'\widehat{\boldsymbol{\beta}}_{R_\lambda} = c^{*2}$; and (iv) assert as in **A5** that the solution so attained "minimizes the sum of squares of residuals on the sphere centered at the origin whose radius is the length" of $\widehat{\boldsymbol{\beta}}_{R_\lambda}$. We have claimed that Assertion **A5** is false. Evidence is provided in Table 5, where lengths

Table 5. Lengths of $\widehat{\boldsymbol{\beta}}_{R_\lambda}$, and square roots of residual sums of squares $R(\lambda) = [(\boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{R_\lambda})'(\boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{R_\lambda})]^{\frac{1}{2}}$, for designated values of λ.

λ	0.00	0.04	0.08	0.12	0.16	0.20	0.24	0.28
$\Vert\widehat{\boldsymbol{\beta}}_{R_\lambda}\Vert$	394.67	137.82	33.14	31.50	70.02	99.19	122.10	140.73
$R(\lambda)$	2129.53	2474.87	2735.75	2914.38	3057.54	3184.84	3305.00	3422.13
λ	0.32	0.36	0.40	0.48	0.56	0.60	0.64	0.68
$\Vert\widehat{\boldsymbol{\beta}}_{R_\lambda}\Vert$	156.25	169.40	180.69	199.00	213.09	218.93	224.11	228.70
$R(\lambda)$	3538.22	3654.22	3770.58	4004.70	4240.27	4358.28	4476.26	4594.06
λ	0.72	0.76	0.80	0.84	0.88	0.92	0.96	1.00
$\Vert\widehat{\boldsymbol{\beta}}_{R_\lambda}\Vert$	232.79	236.42	239.65	242.53	245.08	247.34	249.33	251.09
$R(\lambda)$	4711.56	4828.65	4945.23	5061.20	5176.49	5291.03	5404.77	5517.64

$\Vert\widehat{\boldsymbol{\beta}}_{R_\lambda}\Vert$, and square roots $R(\lambda) = [(\boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{R_\lambda})'(\boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{R_\lambda})]^{\frac{1}{2}}$, are reported as λ ranges systematically over $[0, 1]$. Recall that this range is stipulated by Hoerl and Kennard (1970) and others when $\boldsymbol{Z}'\boldsymbol{Z}$ is in "correlation form." Here $\widehat{\boldsymbol{\beta}}_{R_\lambda} = [\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4, \widehat{\beta}_5]'$ consists of rates of change; similar trends are exhibited when $\boldsymbol{\beta}$ is expanded to include the intercept. It is seen that $\Vert\widehat{\boldsymbol{\beta}}_{R_\lambda}\Vert$ initially decreases to a minimum, then increases beyond $\lambda = 1.0$, but eventually decreases to zero since $\widehat{\boldsymbol{\beta}}_{R_\lambda}$ is a shrinkage estimator.

Greater detail is seen on recalling from Section 4.4 that $\widehat{\boldsymbol{\beta}}_{R_\lambda} = \boldsymbol{Q}\widehat{\boldsymbol{\theta}}_{R_\lambda}$; that $\boldsymbol{Q}$ is orthogonal; and thus, letting $g_{\widehat{\boldsymbol{\beta}}_R}(\lambda) = \Vert\widehat{\boldsymbol{\beta}}_{R_\lambda}\Vert^2$, that $g_{\widehat{\boldsymbol{\beta}}_R}(\lambda) = g_{\widehat{\boldsymbol{\theta}}_R}(\lambda)$. The canonical form of Section

4.4 assures that $g_{\widehat{\boldsymbol{\theta}}_R}(\lambda) = \sum_{i=1}^{k} U_i^2 \xi_i^2/(\xi_i^2 + \lambda)^2$. This is differentiable; its derivative is

$$\partial g_{\widehat{\boldsymbol{\theta}}_R}(\lambda)/\partial\lambda = -2\sum_{i=1}^{k} U_i^2 \xi_i^2 (\xi_i^2 + \lambda)^{-3}; \tag{5.2}$$

and its path traces evolution of the derivative as λ varies. In particular, at $\lambda = 0$ we have $[\partial g_{\widehat{\boldsymbol{\theta}}_R}(\lambda)/\partial\lambda]_{\lambda=0} = -2\sum_{i=1}^{k} U_i^2/\xi_i^4$. This is precipitous for the Hospital Manpower Data in view of the fact that $\xi_k = \xi_5 = 0.007347$.

A detailed local view is provided in Table 6, to include not only $\|\widehat{\boldsymbol{\beta}}_{R_\lambda}\|$ and $R(\lambda)$, but also the ridge estimates $\widehat{\boldsymbol{\beta}}_{R_\lambda} = [\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4, \widehat{\beta}_5]'$ in rows corresponding to various choices for λ. Values of $\widehat{\boldsymbol{\beta}}_{R_\lambda}$ for $\lambda \in \{0.08, 0.11, 0.12\}$ are as in Table 8.9 of Myers (1990), who

TABLE 6. Ridge estimators $\widehat{\boldsymbol{\beta}}_{R_\lambda}$, lengths of $\widehat{\boldsymbol{\beta}}_{R_\lambda}$, and square roots $R(\lambda) = [(\boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{R_\lambda})'(\boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{R_\lambda})]^{\frac{1}{2}}$ of residual sums of squares, for designated values of λ.

λ	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$	$\|\widehat{\boldsymbol{\beta}}_{R_\lambda}\|$	$R(\lambda)$
0.08	10.6354	0.065428	0.359139	6.3206	-30.7471	33.1448	2735.75
0.08095	10.6118	0.065432	0.358279	6.3674	-28.9649	31.5000	2740.68
0.08797	10.4475	0.065444	0.352298	6.6903	-16.4728	20.6250	2775.83
0.0981	10.2378	0.065414	0.344681	7.0942	-0.3156	12.4645	2823.03
0.09829	10.2342	0.065413	0.344548	7.1012	-0.0308	12.4615	2823.89
0.0983	10.2340	0.065413	0.344541	7.1015	-0.0159	12.4615	2823.93
0.11	10.0248	0.065325	0.336955	7.4935	16.3900	20.6251	2874.22
0.12	9.8679	0.065217	0.331280	7.7785	28.8834	31.5000	2914.38

reports ridge estimates for $\lambda \in [0, 0.24]$ by increments of 0.01. It is seen that $\|\widehat{\boldsymbol{\beta}}_{R_\lambda}\|$ takes its minimum value, 12.46150, at $\lambda_{\min} = 0.09829$. To continue, designate $\widehat{\boldsymbol{\beta}}_{R_\lambda}$ as $\widehat{\boldsymbol{\beta}}_R(\lambda)$. It is seen that $\widehat{\boldsymbol{\beta}}_R(0.12)$ and $\widehat{\boldsymbol{\beta}}_R(0.08095)$ have the same length, namely, $\|\widehat{\boldsymbol{\beta}}_R(0.12)\| = 31.500 = \|\widehat{\boldsymbol{\beta}}_R(0.08095)\|$, so that $\Lambda(31.500) = \{0.08095, 0.12\}$ in the notation of (3.6). Suppose a user chooses $\widehat{\boldsymbol{\beta}}_R(0.12)$ as the ridge estimate for the Hospital Manpower Data. Then $\widehat{\boldsymbol{\beta}}_R(0.12)$ is not the minimizing solution of length 31.500; this is seen from $R(0.12) = 2914.38 > 2740.68 = R(0.08095)$. Similarly, it is clear that $\Lambda(20.625) = \{0.08797, 0.11\}$ as in (3.6), and that $\widehat{\boldsymbol{\beta}}_R(0.11)$ of Table 6 is not minimizing, to be supplanted instead by $\widehat{\boldsymbol{\beta}}_R(0.08797)$ from Table 6. A continuum of further examples can be constructed by reflecting λ asymmetrically about $\lambda_{\min} = 0.09829$, the smaller λ of each pair corresponding to the minimizing solution. These clearly constitute counterexamples to Assertion **A5.**

Not only are definitive values for the constraint $\{\boldsymbol{\beta}'\boldsymbol{\beta} = c^2\}$ not evident beforehand, but profound and heretofore undiscovered limits pertain to admissible values for λ in order that solutions of given length c^* be minimizing. To fix ideas, suppose in equation (3.4) that $\{g_{\widehat{\boldsymbol{\beta}}_R}(0.00) > c^{*2} \geq c^2 \geq g_{\widehat{\boldsymbol{\beta}}_R}(0.09829) = 12.46150^2 = 155.2877\}$. Then the only feasible values for λ are those in the interval $[\min g_{\widehat{\boldsymbol{\beta}}_R}^{-1}(c^2), 0.09829]$. For example, if $33.14481^2 =$

$1098.5784 \geq c^2 \geq 155.2877$, then from Table 6 the feasible values are $\lambda \in [0.08,\, 0.09829]$. For $\{g_{\widehat{\boldsymbol{\beta}}_R}(0.00) \geq c^2 \geq g_{\widehat{\boldsymbol{\beta}}_R}(0.09829)\}$, the feasible values are $\lambda \in [0.00,\, 0.09829]$. These are the only feasible values for $\lambda \in [0,\, 1]$. On the other hand, choosing $\{0 < c^2 < g_{\widehat{\boldsymbol{\beta}}_R}(0.09829) = 155.2877\}$ requires λ in the interval $(\min g^{-1}_{\widehat{\boldsymbol{\beta}}_R}(c^2),\, \infty)$, where $\min g^{-1}_{\widehat{\boldsymbol{\beta}}_R}(155.2877) > 158$. For example, if $c^2 < 100$, then the feasible values are $\lambda \in (198,\, \infty)$. As these are far outside the recommended interval $[0,\, 1]$, constraints $c^{*2} \in (0,\, 155.2877)$ must be declared to be inadmissible. Values reported for $\{0 < c^2 < g_{\widehat{\boldsymbol{\beta}}_R}(0.09829) = 155.2877\}$ are supported by the Maple software package. Values reported for $PRESS_\lambda$ and HKB_λ in Table 4 are thus inadmissible in view of Assertion **A5*.**

In short, imbedded in the Hospital Manpower Data are the hidden feasible constraints $\{\boldsymbol{\beta}'\boldsymbol{\beta} = c^2\}$ with $c^2 \geq 155.2877$. These could not have been discerned beforehand short of the foregoing detailed analyses.

To summarize, origins of the anomaly exhibited here may be traced as follows: (i) The ridge trace of $\widehat{\beta}_5(\lambda)$ exhibits a down-up–down character, beginning with $\widehat{\beta}_5(0.00) =$ -394.3280, decreasing to zero between $\lambda = 0.09$ and $\lambda = 0.10$, and increasing thereafter to $\widehat{\beta}_5(1.00) = 250.8307$ and beyond, and eventually decreasing to zero through shrinkage. (ii) $|\widehat{\beta}_5(\lambda)|$ dominates other estimates by orders of magnitude ranging from one to four except near its minimum. (iii) Other estimates exhibit relatively narrow ranges in comparison with $\widehat{\beta}_5(\lambda)$ as λ varies over $[0,\, 1]$. (iv) In consequence, $\|\, \widehat{\boldsymbol{\beta}}_{R_\lambda} \|$ is largely determined by $[\widehat{\beta}_5(\lambda)]^2$ as λ varies. Finally note that $\Lambda(c^*)$ from (3.6) takes on two values in the cases examined, from the down–up–down character of $\|\widehat{\boldsymbol{\beta}}_{R_\lambda}\|$ as λ evolves. It is clear in other circumstances that $\Lambda(c^*)$ may consist of three or more elements. For example, a single dominant estimate may exhibit multiple sign changes, whereas estimates for other coefficients may have one or more sign changes as well. These and related matters are studied in Zhang and McDonald (2005), and references cited therein, under special structure of $\boldsymbol{Z}'\boldsymbol{Z}$ in correlation form. Properties, to include sign changes, crossings, and rates–of–change of individual ridge estimates, as well as bounds on the number of sign changes, are determined by those authors on identifying zeros and derivatives of polynomials in λ of degree $k-1$, under special structure as cited.

These facts alone challenge the meaning of numerous simulation studies purporting to compare alternative criteria for choosing λ, when all such choices have ignored the minimizing constraints on λ. Thus aggregates of minimizing/non–minimizing values are compared with other such aggregates, to the effect of total obfuscation.

We turn next to properties of ridge and surrogate ridge solutions, to include condition numbers and other diagnostics. Computations for the condition numbers proceed as in Table 3, based on equivalence between conditioning for $\widehat{\boldsymbol{\beta}}$ and the canonical estimators $\widehat{\boldsymbol{\theta}}$, as noted in Section 4.6.

5.4. **Properties of $\widehat{\boldsymbol{\beta}}_{R_\lambda}$ and $\widehat{\boldsymbol{\beta}}_{S_\lambda}$.** In summary, the ridge solutions to $(\boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{I}_k)\boldsymbol{\beta} = \boldsymbol{Z}'\boldsymbol{Y}$ account for ill–conditioning of $\boldsymbol{Z}'\boldsymbol{Z}$ on the left of $\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\beta} = \boldsymbol{Z}'\boldsymbol{Y}$, whereas the surrogate solutions to $\boldsymbol{Z}_\lambda'\boldsymbol{Z}_\lambda\boldsymbol{\beta} = \boldsymbol{Z}_\lambda'\boldsymbol{Y}$ account for ill–conditioning on the right as well. It thus is germane to compare $\{\widehat{\boldsymbol{\beta}}_{S_\lambda};\, \lambda \geq 0\}$ with $\{\widehat{\boldsymbol{\beta}}_{R_\lambda};\, \lambda \geq 0\}$ using the data at hand. We next examine critical issues from Section 4.5, applicable both to ridge and to surrogate ridge solutions. Table 7 lists condition numbers and other quantities affiliated with $\{\widehat{\boldsymbol{\beta}}_{R_\lambda}; \lambda \geq 0\}$ and $\{\widehat{\boldsymbol{\beta}}_{S_\lambda}; \lambda \geq 0\}$, under values for λ as listed. Question 1 of Section 4.5 is negated for $\widehat{\boldsymbol{\beta}}_{R_\lambda}$: Stability of

TABLE 7. Condition numbers for $\widehat{\boldsymbol{\beta}}_{R_\lambda}(\boldsymbol{Y})$, $\widehat{\boldsymbol{\beta}}_{S_\lambda}(\boldsymbol{Y})$, $\mathrm{V}(\widehat{\boldsymbol{\beta}}_{R_\lambda})$, and $\mathrm{V}(\widehat{\boldsymbol{\beta}}_{S_\lambda})$; the maximal *VIF*s $V_M(\widehat{\boldsymbol{\beta}}_{R_\lambda})$ and $V_M(\widehat{\boldsymbol{\beta}}_{S_\lambda})$; and the Frobenius distance $D_Z(\boldsymbol{Z}_\lambda) = \|\,\boldsymbol{Z} - \boldsymbol{Z}_\lambda\|_F$, under various choices for λ.

λ	$c_1(\widehat{\boldsymbol{\beta}}_{R_\lambda})$	$c_1(\widehat{\boldsymbol{\beta}}_{S_\lambda})$	$V_M(\widehat{\boldsymbol{\beta}}_{R_\lambda})$	$c_1[\mathrm{V}(\widehat{\boldsymbol{\beta}}_{R_\lambda})]$	$V_M(\widehat{\boldsymbol{\beta}}_{S_\lambda})$	$c_1[\mathrm{V}(\widehat{\boldsymbol{\beta}}_{S_\lambda})]$	$D_Z(\boldsymbol{Z}_\lambda)$
0.0004	33.1584	96.1565	141.5345	1099.4770	1146.399	9246.064	0.0140
0.004787	9.0957	29.4630	10.9688	82.7319	112.6300	868.0653	0.0638
0.005	9.0537	28.8348	10.8874	81.9695	108.0918	831.4473	0.0654
0.010	8.1707	20.4561	9.2481	66.7610	56.6915	418.4530	0.0974
0.030	11.6724	11.8596	21.2905	136.2440	21.2197	140.6508	0.1847
0.050	15.1539	9.2114	34.0995	229.6392	13.6552	84.8507	0.2511
0.070	17.8166	7.8046	42.5990	317.4320	10.2639	60.9119	0.3083
0.090	20.4222	6.8997	51.7827	417.0673	8.3166	47.6061	0.3598
0.230	29.6720	4.3868	100.5675	880.4276	3.9338	19.2438	0.6429
0.616964	53.4183	2.7932	250.4309	2853.5130	2.0374	7.8022	1.1769
1.000	66.6915	2.2797	451.5788	4447.7550	1.5976	5.1968	1.5745

the solutions $\widehat{\boldsymbol{\beta}}_{R_\lambda}$, as gauged by $c_1[\widehat{\boldsymbol{\beta}}_{R_\lambda}(\boldsymbol{Y})]$, initially improves but then erodes. Further computations show that $c_1[\widehat{\boldsymbol{\beta}}_{R_\lambda}(\boldsymbol{Y})]$ takes its minimal value, 7.4463, at $\lambda = 0.015$, and increases thereafter. In contrast, despite higher beginning values than $c_1[\widehat{\boldsymbol{\beta}}_{R_\lambda}(\boldsymbol{Y})]$, the condition numbers $c_1[\widehat{\boldsymbol{\beta}}_{S_\lambda}(\boldsymbol{Y})]$ for surrogate estimators decrease monotonically with increasing λ, the trends $c_1[\widehat{\boldsymbol{\beta}}_{R_\lambda}(\boldsymbol{Y})] = 11.7723 = c_1[\widehat{\boldsymbol{\beta}}_{S_\lambda}(\boldsymbol{Y})]$ crossing at $\lambda = 0.03045$.

Questions 2 and 3 of Section 4.5 are refuted for $\widehat{\boldsymbol{\beta}}_{R_\lambda}$: Computations interpolating those of Table 7 show that $c_1[\mathrm{V}(\widehat{\boldsymbol{\beta}}_{R_\lambda})]$ temporarily decreases over $\lambda \in [0,\, 0.015]$, where its minimum is 55.4470, but it increases thereafter. Similarly, the maximal *VIF*s for $\widehat{\boldsymbol{\beta}}_{R_\lambda}$ initially decrease and then increase. By comparison, both the condition numbers $c_1[\mathrm{V}(\widehat{\boldsymbol{\beta}}_{S_\lambda})]$, and the maximal *VIF*s for $\widehat{\boldsymbol{\beta}}_{S_\lambda}$, decrease with increasing λ. Although initially larger, $V_M(\widehat{\boldsymbol{\beta}}_{S_\lambda})$ approximates $V_M(\widehat{\boldsymbol{\beta}}_{R_\lambda})$ at $\lambda = 0.030$, and the ratio $V_M(\widehat{\boldsymbol{\beta}}_{R_\lambda})/V_M(\widehat{\boldsymbol{\beta}}_{S_\lambda})$ increases markedly thereafter.

Recall that the surrogate $\boldsymbol{Y} = \boldsymbol{Z}_\lambda\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is intended as an approximation to $\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. The order of approximation, as gauged by the Frobenius distance (4.3), is tabulated as the final column of Table 7. Relative changes, given by $\|\,\boldsymbol{Z} - \boldsymbol{Z}_\lambda\,\|_F \,/\, \|\,\boldsymbol{Z}\,\|_F$, are 0.1123 at $\lambda = 0.05$, ranging up to 0.5263 at $\lambda = 0.616964$, where the denominator is $\|\,\boldsymbol{Z}\,\|_F = 2.236068$.

Further details are given in Tables 8 and 9, from which several entries of Table 7 are drawn. Table 8 examines the evolution of *VIF*s, and conditioning of the correlation matrices, for $\widehat{\boldsymbol{\beta}}_{R_\lambda}$ as λ varies. Values for $c_1[\boldsymbol{C}(\widehat{\boldsymbol{\beta}}_{R_\lambda})]$ are included, as Lemma 1 applies in each case. It is found that $c_1[\boldsymbol{C}(\widehat{\boldsymbol{\beta}}_{R_\lambda})]$ achieves its minimum, 61.4449, at $\lambda = 0.0173$. In all

TABLE 8. Variance inflation factors for $\widehat{\boldsymbol{\beta}}_{R_\lambda}$, and condition numbers for $\boldsymbol{C}(\widehat{\boldsymbol{\beta}}_{R_\lambda})$ and $T(\boldsymbol{\beta}) = \mathrm{E}(\widehat{\boldsymbol{\beta}}_{R_\lambda})$, for designated values of λ.

λ	*VIF1*	*VIF2*	*VIF3*	*VIF4*	*VIF5*	$c_1[\boldsymbol{C}(\widehat{\boldsymbol{\beta}}_{R_\lambda})]$	$c_1[T(\boldsymbol{\beta})]$
0.000	9595.68	7.9406	8931.449	23.2887	4.2794	54756.83	1.0000
0.0004	141.5345	7.8481	133.0221	13.0512	3.3997	576.8409	8.4095
0.004787	7.1604	7.1682	7.8840	10.9688	3.0128	90.13222	89.5726
0.005	7.1047	7.1379	7.8349	10.8874	2.9972	89.50392	93.5175
0.010	8.0001	6.4919	8.8456	9.2481	2.6830	75.66936	185.8150
0.030	19.7743	4.7268	21.2905	5.6339	2.0003	109.4703	552.8219
0.050	32.0013	3.6885	34.0995	4.0168	1.6988	177.2545	916.3722
0.070	42.5990	3.0187	45.0695	3.1473	1.5363	227.9178	1276.515
0.090	51.7827	2.5598	54.4589	2.6269	1.4377	267.3171	1633.297
0.230	100.5675	1.3868	102.4723	1.6364	1.2446	441.9639	4040.511
0.616964	250.4309	1.0879	243.2535	1.8791	1.3541	1047.931	9965.795
1.000	451.5788	1.2184	430.5957	2.4738	1.5386	2174.418	14961.96

TABLE 9. Variance inflation factors for $\widehat{\boldsymbol{\beta}}_{S_\lambda}$, and condition numbers for $\boldsymbol{C}(\widehat{\boldsymbol{\beta}}_{S_\lambda})$ and $\mathrm{V}(\widehat{\boldsymbol{\beta}}_{S_\lambda})$, for designated values of λ.

λ	*VIF1*	*VIF2*	*VIF3*	*VIF4*	*VIF5*	$c_1[\boldsymbol{C}(\widehat{\boldsymbol{\beta}}_{S_\lambda})]$
0.000	9595.68	7.9406	8931.449	23.2887	4.2794	54756.83
0.0004	1146.399	7.8846	1068.211	14.2203	3.5089	5091.248
0.004787	112.6300	7.5308	106.0234	12.0987	3.2190	458.9380
0.005	108.0918	7.5147	101.7946	12.0488	3.2099	440.5738
0.010	56.6915	7.1607	53.8459	11.0379	3.0219	233.7461
0.030	21.2197	6.0737	20.5412	8.4506	2.5374	93.1862
0.050	13.6552	5.3181	13.3380	6.9511	2.2606	63.4167
0.070	10.2639	4.7584	10.0777	5.9605	2.0792	48.8365
0.090	8.3166	4.3258	8.1934	5.2538	1.9500	39.9124
0.230	3.9338	2.8218	3.9092	3.1133	1.5493	17.9102
0.616964	2.0374	1.7669	2.0330	1.8380	1.2710	7.5371
1.000	1.5976	1.4635	1.5957	1.4981	1.1781	5.0614

instances each *VIF* initially decreases, then increases, but values of λ at which the changes occur differ across the five estimators. If we view $\mathrm{E}(\widehat{\boldsymbol{\beta}}_{R_\lambda}) = T(\boldsymbol{\beta})$ as a transformation on the parameter space, Question 4 of Section 4.5 asks whether its conditioning improves with increasing λ. To the contrary, the last column of Table 8 shows that condition numbers increase explosively with increasing λ. From Table 3 it is clear that corresponding condition numbers for $\mathrm{E}(\widehat{\boldsymbol{\beta}}_{S_\lambda}) = T(\boldsymbol{\beta})$ are square roots of those listed in Table 8 for $\widehat{\boldsymbol{\beta}}_{R_\lambda}$.

Similar entries in Table 9 give the evolution of *VIF*s and $c_1[\boldsymbol{C}(\widehat{\boldsymbol{\beta}}_{S_\lambda})]$ for $\widehat{\boldsymbol{\beta}}_{S_\lambda}$.

A noted departure from Table 8 is that the maximal *VIF* is $V_M(\widehat{\boldsymbol{\beta}}_{S_\lambda}) = VIF(\widehat{\beta}_1)$ for all cases, independently of λ. Further computations show that the crossing $c_1[\boldsymbol{C}(\widehat{\boldsymbol{\beta}}_{R_\lambda})] = 99.56217 = c_1[\boldsymbol{C}(\widehat{\boldsymbol{\beta}}_{S_\lambda})]$ occurs at $\lambda = 0.02750$.

6. Conclusions

Little of the considerable literature on ridge regression is found to be consistent with the optimization of Hoerl and Kennard (1970) under equality constraints $\{\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\beta}} = c^2\}$, and under the inequality constraints $\{\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\beta}} \le c^2\}$ of Balakrishnan (1963), despite pervasive claims to the contrary.

The problem is traced to (i) a misapplication of LaGrange's principle; (ii) the false claim that the constrained solutions have nonsingular distributions, corresponding one–to–one with $\widehat{\boldsymbol{\beta}}_L$; and (iii) the implied but incorrect assertion that the ridge parameter λ corresponds one–to–one with c^2, and thus the false claim that the solution $\widehat{\boldsymbol{\beta}}_{R_\lambda}$ of $(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_k)\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$ minimizes the residual sum of squares among estimators of length $\widehat{\boldsymbol{\beta}}'_{R_\lambda}\widehat{\boldsymbol{\beta}}_{R_\lambda} = c^{*2}$. Our Theorem 1 supplies the missing distributions appropriate to constrained minimization. Generalized ridge regression, seen as solving the equations $(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Lambda})\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$ with nonnegative ridge parameters $\boldsymbol{\Lambda} = \mathrm{Diag}(\lambda_1, \dots, \lambda_k)$, is also shown to be inconsistent with LaGrange minimization.

LaGrange optimization having failed as a rational foundation for conventional ridge regression, alternatives based on conditioning are developed in Section 4. Limitations in Type A conditioning, on which a justification for $\widehat{\boldsymbol{\beta}}_{R_\lambda}$ rests, prompt the introduction of *surrogate ridge* solutions, $\widehat{\boldsymbol{\beta}}_{S_\lambda}$, to account for ill–conditioning of $\boldsymbol{X}$ on both sides of the *OLS* equations, $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$. Extensive numerical studies, as reported in Section 5, reexamine the Hospital Manpower Data in a manner complementary to the conventional analyses undertaken in Myers (1990). It is demonstrated that none of the conditionings of $\widehat{\boldsymbol{\beta}}_{R_\lambda}(\boldsymbol{Y})$, $\mathrm{E}(\widehat{\boldsymbol{\beta}}_{R_\lambda}) = T(\boldsymbol{\beta})$, and $\mathrm{V}(\widehat{\boldsymbol{\beta}}_{R_\lambda})$, nor the variance inflation factors, as critical properties of the ridge estimators $\{\widehat{\boldsymbol{\beta}}_{R_\lambda}; \lambda \ge 0\}$, is enhanced monotonically on increasing λ. In contrast, for the surrogate solutions $\widehat{\boldsymbol{\beta}}_{S_\lambda}$, all (except $T(\boldsymbol{\beta})$) of these are uniformly enhanced as λ evolves. It is seen that $\widehat{\boldsymbol{\beta}}_{R_\lambda}$ is better within a narrow range for small λ, but its *VIF*s and condition numbers often become excessive within the range of λ often recommended in practice. In short, ridge regression often exhibits some of the very pathologies it is intended to redress.

In summary, there is a vast and expanding compendium on the so–called theory, methodology, and simulation studies surrounding ridge regression. If indeed constrained optimization is to be pivotal, then the bulk of these studies will have to be reworked to take into account the nonstandard distributions of Section 3.3, as well as constraints for the ridge parameter to be minimizing, as documented in Sections 3.4 and 5.3. It is remarkable that this field of applied engineering has thrived for so long, despite critical false assertions and a dearth of sustaining foundation principles.

References

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

[2] Balakrishnan, A.V. (1963). "An operator theoretic formulation of a class of control problems and a steepest descent method of solution." *Journal on Control.* **1:** 109–127.

[3] Batschelet, E. (1981). *Circular Statistics in Biology.* Academic Press, New York.

[4] 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.

[5] Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics.* Wiley, New York.

[6] Bunke, O. (1975). "Minimax linear, ridge and shrunken estimators for linear parameters." *Math. Operationsforsch. u. Statist.* **6:** 697–701.

[7] Davies, R.B. and Hutton, B. (1975). "The effect of errors in the independent variables in regression." *Biometrika* **62:** 383–392.

[8] Dempster, A.P., Schatzoff, M. and Wermuth, N. (1977). "A simulation study of alternatives to ordinary least squares." *J. Amer. Statist. Assoc.* **72:** 77–106.

[9] Evans, M., Hastings, N. and Peacock, B. (2000). "von Mises Distribution." Chapter 41 in *Statistical Distributions, 3rd ed.* John Wiley, New York.

[10] Fisher, N.I. (1993). *Statistical Analysis of Circular Data.* Cambridge University Press, UK.

[11] Fisher, N.I., Lewis, T. and Embleton, B.J.J. (1993). *Statistical Analysis of Spherical Data.* Cambridge University Press, UK.

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

[13] Hoerl, A.E. (1962). "Application of ridge analysis to regression problems." *Chemical Engineering Progress* **58:** 54–59.

[14] Hoerl, A.E. (1964). "Ridge analysis." *Chemical Engineering Progress Symposium Series 60,* 67–77.

[15] Hoerl, A.E. and Kennard, R.W. (1970). "Ridge regression: Biased estimation for Anonorthogonal problems." *Technometrics* **12:** 55–67.

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

[17] Hoerl, A.E., Kennard, R.W. and Baldwin, K.F. (1975). "Ridge regression: Some simulations." *Communications in Statistics* **4:** 105–123.

[18] Horn, R.A. and Johnson, C.R. (1985). *Matrix Analysis.* Cambridge University Press, UK.

[19] Levenberg, K. (1944). "A method for the solution of certain non–linear problems in least squares." *Quarterly of Applied Mathematics* **2:** 164–168.

[20] Mallows, C.L. (1973). "Some comments on C_p." *Technometrics* **15:** 661–675.

[21] Mardia, K.V. and Jupp, P. (2000). *Directional Statistics, 2nd ed.* Wiley, New York.

[22] Marquardt, D.W. (1970). "Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation." *Technometrics* **12:** 591–612.

[23] Marquardt, D.W. and Snee, R.D. (1975). "Ridge regression in practice." *The American Statistician* **29:** 3–20.

[24] Marshall, A.W. and Olkin, I. (1965). "Norms and inequalities for condition numbers." *Pacific J. Math.* **15:** 241-247.

[25] Marshall, A.W. and Olkin, I. (1979). *Inequalities: Theory of Majorization and Its Applications.* Academic Press, New York.

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

[27] Myoken, H. and Uchida, Y. (1977). "The generalized ridge estimator and improved adjustments for regression parameters." *Metrika* **24:** 113–124.

[28] Neumann, J. von (1937) "Some matrix inequalities and metrization of matrix space." *Tomsk University Rev.* **1:** 286-300.

[29] Nostrand, C. van (1980). "Comment." *J. Amer. Statist. Assoc.* **75:** 92–94.

[30] 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.

[31] Schatten, R. (1970). *Norm Ideals of Completely Continuous Operators.* New York: Springer-Verlag.

[32] Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *J. Royal Statist. Soc. Ser. B* **58:** 267–288.

[33] Zhang, R. and McDonald, G.C. (2005). "Characterization of ridge trace behavior." *Commun. Statist.– Theory and Methods* **34:** 1487–1501.

Department of Mathematics, University of Virginia, Charlottesville, VA 22904-4137
E-mail address: `der@virginia.edu`