Statistics and Probability Letters 79 (2009) 2237–2241

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Concentration reversals in ridge regression

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

[a] Department of Statistics, Virginia Tech, Blacksburg, VA 24061, United States
[b] Department of Mathematics, University of Virginia, Charlottesville, VA 22904–4137, United States

ARTICLE INFO

Article history:
Received 20 May 2009
Received in revised form 23 July 2009
Accepted 24 July 2009
Available online 30 July 2009

MSC:
62J07
62J20

ABSTRACT

Ridge regression is often the method of choice for approaching ill-conditioned systems. A canonical form identifies regions in the parameter space where Ordinary Least Squares (*OLS*) is problematic. A curious but unrecognized property of ridge solutions emerges: Under spherical errors with or without moments, the relative concentrations of the canonical estimators reverse as the ridge scalar evolves, the estimators least concentrated under *OLS* being most concentrated under ridge regression, and conversely.

1. Introduction

A full-rank model $\{\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}\}$ yields *OLS* solutions $\widehat{\boldsymbol{\beta}}_L$, unbiased with minimal dispersion $V(\widehat{\boldsymbol{\beta}}_L) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ under conventional errors. Near dependency among columns of $\boldsymbol{X}$, known as ill-conditioning, "causes crucial elements of $\boldsymbol{X}'\boldsymbol{X}$ to be large and unstable", $\widehat{\boldsymbol{\beta}}_L$ to be "very sensitive to small changes in $\boldsymbol{X}$" with "inflated variances" (Belsley, 1986, p. 119). Remedies in wide usage include using *ridge estimators* as solutions of $\{(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)\widehat{\boldsymbol{\beta}}_{Rk} = \boldsymbol{X}'\boldsymbol{Y};\, k \geq 0\}$; see Hoerl (1962, 1964) and Hoerl and Kennard (1970) and a considerable further literature. A canonical form is $\{\boldsymbol{Y} = \boldsymbol{PD}_\xi\boldsymbol{\theta} + \boldsymbol{\epsilon}\}$, where $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}_p$; $\boldsymbol{D}_\xi = \mathrm{Diag}(\xi_1, \ldots, \xi_p)$, ordered as $\{\xi_1 \geq \xi_2 \geq \cdots \geq \xi_p\}$; and $\boldsymbol{\theta} = \boldsymbol{Q}'\boldsymbol{\beta}$, with $\boldsymbol{Q}$ orthogonal. The *OLS* solutions $\widehat{\boldsymbol{\theta}}_L = [\hat{\theta}_{L1}, \ldots, \hat{\theta}_{Lp}]'$ identify "linear combinations of parameters about which the data at hand are most/least informative" (Thisted, 1987), since their variances $\{\sigma^2/\xi_1^2 \leq \sigma^2/\xi_2^2 \leq \cdots \leq \sigma^2/\xi_p^2\}$ escalate for values of ξ_i^2 near zero, a mark of ill-conditioning. In consequence, as gauged by variances through Chebyshev's inequality, elements of $[\hat{\theta}_{L1}, \ldots, \hat{\theta}_{Lp}]$ are successively less concentrated in probability about their means. It is germane in practice to ask whether this ordering is preserved as k evolves in ridge regression. To the contrary, we demonstrate under spherical errors that the canonical estimators reverse in peakedness as in Birnbaum (1948), with or without moments. An outline follows.

Notation and supporting materials are collected in Section 2. Section 3 sets forth the principal findings, to include conditions for reversal in concentration of ridge solutions as k evolves, and elementary but illuminating examples. Section 4 revisits a case study given elsewhere, and Section 5 offers a brief summary.

2. Preliminaries

2.1. Notation

Designate Euclidean p-space as $\mathbb{R}^p$, and its positive orthant as $\mathbb{R}^p_+$; matrices and vectors are set in bold type; the transpose, inverse, and trace of $\boldsymbol{A}$ are $\boldsymbol{A}'$, $\boldsymbol{A}^{-1}$, and $\mathrm{tr}(\boldsymbol{A})$; special arrays are the identity $\boldsymbol{I}_p$ and the diagonal matrix $\boldsymbol{D}_{\boldsymbol{a}} = \boldsymbol{D}(a_i) = \mathrm{Diag}(a_1, \ldots, a_p)$. Let $\boldsymbol{X}$, of order $(n \times p)$, have rank $p < n$; its *singular decomposition* is $\boldsymbol{X} = \boldsymbol{PD}_\xi\boldsymbol{Q}'$, where

[*] Corresponding author. Tel.: +1 434 924 4934; fax: +1 434 982 3084.
E-mail addresses: djensen@vt.edu (D.R. Jensen), der@virginia.edu (D.E. Ramirez).

0167-7152/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2009.07.022

$\boldsymbol{D}_\xi = \mathrm{Diag}(\xi_1, \ldots, \xi_p)$ comprise its ordered *singular values* $\{\xi_1 \geq \xi_2 \geq \cdots \geq \xi_p > 0\}$; and the columns of $\boldsymbol{P} = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_p]$ and of $\boldsymbol{Q} = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_p]$ are the *left-* and *right-singular vectors* of $\boldsymbol{X}$, such that $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}_p$ and $\boldsymbol{Q}$ is orthogonal. This in turn supports the canonical form $\{\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{D}_\xi\boldsymbol{\theta} + \boldsymbol{\epsilon}\}$ for the model $\{\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\}$ of Section 1, with $(\boldsymbol{Y}, \boldsymbol{\epsilon}) \in \mathbb{R}^n$, $\boldsymbol{X}$ of order $(n \times p)$, and $(\boldsymbol{\beta}, \boldsymbol{\theta}) \in \mathbb{R}^p$. In particular, the canonical ridge equations are $\{(\boldsymbol{D}_\xi^2 + k\boldsymbol{I}_p)\widehat{\boldsymbol{\theta}}_{Rk} = \boldsymbol{D}_\xi\boldsymbol{P}'\boldsymbol{Y};\ k \geq 0\}$, with $\widehat{\boldsymbol{\theta}}_{Rk} = [\hat{\theta}_{Rk}^1, \ldots, \hat{\theta}_{Rk}^p]'$.

2.2. Stochastic concepts

The expectation, dispersion, and law of distribution of $\boldsymbol{Z} \in \mathbb{R}^p$ are $\mathrm{E}(\boldsymbol{Z})$, $V(\boldsymbol{Z})$, and $\mathcal{L}(\boldsymbol{Z})$; moreover, $F_Z(\boldsymbol{z})$ and $g_Z(\boldsymbol{t})$ designate its cumulative distribution *(cdf)* and characteristic *(chf)* functions. Following Birnbaum (1948), a probability measure μ on $\mathbb{R}^1$ is said to be *more peaked* about a than is ν about b on $\mathbb{R}^1$ if their translates are ordered as $\mu_a([-c, c]) \geq \nu_b([-c, c])$ for every $c \geq 0$, where $\mu_a([u, v]) = \mu([u - a, v - a])$. Designate this ordering as $\mu \succeq_B \nu$, choosing a and b in context. Further denote by Γ the *order-determining class* on $\mathbb{R}^1$, *i.e.,* functions that are even, continuous, and increasing on $[0, \infty)$. The following essential result is from Jensen and Foutz (1989).

Lemma 1. *Let $\{F((x - \theta)/\sigma);\ (\theta, \sigma) \in \mathbb{R}^1 \times \mathbb{R}_+^1\}$ be a location–scale family of cdfs on $\mathbb{R}^1$, each symmetric about θ, and let (μ, ν) be probability measures on $\mathbb{R}^1$.*

(i) *For fixed θ, the scale family $\{F((x - \theta)/\sigma);\ \sigma \in \mathbb{R}_+^1\}$ decreases monotonically in peakedness about θ as σ increases.*

(ii) *The ordering $\mu \succeq_B \nu$ holds if and only if*

$$\int_{-\infty}^{\infty} \gamma(x)\mathrm{d}\mu_a(x) \leq \int_{-\infty}^{\infty} \gamma(y)\mathrm{d}\nu_b(y) \tag{2.1}$$

for every function $\gamma \in \Gamma$ for which both expectations are defined.

Symmetric errors for $\{\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\}$, not necessarily having moments, arise in practical circumstances to be noted. Specifically, suppose the error *chf* is $g_{\boldsymbol{\epsilon}}(\boldsymbol{t}) = \phi_{\boldsymbol{\epsilon}}(\sigma^2\boldsymbol{t}'\boldsymbol{t})$ defined on $[0, \infty)$. Then the error distribution $\mathcal{L}(\boldsymbol{\epsilon})$ is called *spherical* on $\mathbb{R}^n$, to be designated as $\mathcal{L}(\boldsymbol{\epsilon}) = \mathcal{S}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$, centered at $\boldsymbol{0} \in \mathbb{R}^n$ with $\sigma^2\boldsymbol{I}_n$ as scale parameters. Examples of note include spherical Cauchy and Gaussian errors in $\mathbb{R}^n$. Elementary properties of *chf* s imply that $g_Y(\boldsymbol{t}) = \mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{X}\boldsymbol{\beta}}\phi_{\boldsymbol{\epsilon}}(\boldsymbol{t}'\boldsymbol{t})$ at $\sigma^2 = 1.0$ and, under the affine transformation $\{\boldsymbol{Y} \rightarrow \boldsymbol{Z} = \boldsymbol{A}'\boldsymbol{Y} + \boldsymbol{b}\}$ with $\boldsymbol{A}(n \times k)$ and $\boldsymbol{b} \in \mathbb{R}^k$, that $g_Z(\boldsymbol{s}) = \mathrm{e}^{\mathrm{i}\boldsymbol{s}'(\boldsymbol{A}'\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{b})}\phi_{\boldsymbol{\epsilon}}(\boldsymbol{s}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{s})$. The resulting distribution, having elliptical contours and centered at $\boldsymbol{A}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{b}$, is designated as $\mathcal{L}(\boldsymbol{Z}) = \mathcal{S}_k(\boldsymbol{A}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{b}, \boldsymbol{A}'\boldsymbol{A})$. We return to these error structures subsequently.

3. The principal findings

3.1. Ridge reversals

At issue are concentrations in probability of elements of $\widehat{\boldsymbol{\theta}}$ for estimating $\boldsymbol{\theta} \in \mathbb{R}^p$. Under second moments the ordered variances

$$\{\mathrm{Var}(\hat{\theta}_{L1}) \leq \mathrm{Var}(\hat{\theta}_{L2}) \leq \cdots \leq \mathrm{Var}(\hat{\theta}_{Lp})\}, \tag{3.1}$$

together with Chebyshev inequalities, suffice to order $\{\hat{\theta}_{L1}, \hat{\theta}_{L2}, \ldots, \hat{\theta}_{Lp}\}$ from most to least concentrated about their means, as noted earlier. Since the ridge solutions $\widehat{\boldsymbol{\theta}}_{Rk}$ are *OLS* at $k = 0$, we seek to determine whether this ordering persists among $\{\hat{\theta}_{Rk}^1, \hat{\theta}_{Rk}^2, \ldots, \hat{\theta}_{Rk}^p\}$ as k evolves. Although second moments conform to conventional Gauss–Markov theory, instead we take the error distribution to be *spherical,* with or without moments, *i.e.,* $\mathcal{L}(\boldsymbol{\epsilon}) = \mathcal{S}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$, $g_{\boldsymbol{\epsilon}}(\boldsymbol{t}) = \phi_{\boldsymbol{\epsilon}}(\sigma^2\boldsymbol{t}'\boldsymbol{t})$, and $\mathcal{L}(\boldsymbol{Y}) = \mathcal{S}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$, with $(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$ as location and scale parameters. On occasion it suffices to take $\sigma^2 = 1.0$. It is shown in Jensen (1997) that elements of $\widehat{\boldsymbol{\beta}}_L$ (i) are *median-unbiased* through symmetry, (ii) are *unbiased in mean* under first moments, and (iii) are *modal-unbiased* under unimodal errors. Noting that $\{\hat{\theta}_{L1}, \hat{\theta}_{L2}, \ldots, \hat{\theta}_{Lp}\}$ typically differ in location, as do $\{\hat{\theta}_{Rk}^1, \hat{\theta}_{Rk}^2, \ldots, \hat{\theta}_{Rk}^p\}$, we proceed to compare the Birnbaum (1948) peakedness about their respective centering parameters. Details follow, where $\{\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{D}_\xi\boldsymbol{\theta} + \boldsymbol{\epsilon}\}$ and $\{\boldsymbol{U} = \boldsymbol{P}'\boldsymbol{Y}\}$ give the reduced form $\{\boldsymbol{U} = \boldsymbol{D}_\xi\boldsymbol{\theta} + \boldsymbol{P}'\boldsymbol{\epsilon}\}$, together with $g_Y(t) = \mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{P}\boldsymbol{D}_\xi\boldsymbol{\theta}}\phi_{\boldsymbol{\epsilon}}(\boldsymbol{t}'\boldsymbol{t})$ and $g_U(\boldsymbol{r}) = \mathrm{e}^{\mathrm{i}\boldsymbol{r}'\boldsymbol{D}_\xi\boldsymbol{\theta}}\phi_{\boldsymbol{\epsilon}}(\boldsymbol{r}'\boldsymbol{r})$, with $\boldsymbol{r} \in \mathbb{R}^p$. The pivotal *chf* s for $\widehat{\boldsymbol{\theta}}_{Rk}$ are

$$\{g_{\widehat{\boldsymbol{\theta}}_{Rk}}(\boldsymbol{s}) = \mathrm{e}^{\mathrm{i}\boldsymbol{s}'\boldsymbol{D}(\xi_i^2/(\xi_i^2+k))\boldsymbol{\theta}}\phi_{\boldsymbol{\epsilon}}(\boldsymbol{s}'\boldsymbol{D}(\xi_i^2/(\xi_i^2 + k)^2)\boldsymbol{s});\ k \geq 0\}, \tag{3.2}$$

where $\boldsymbol{D}(\omega_i) = \mathrm{Diag}(\omega_1, \ldots, \omega_p)$; $g_{\widehat{\boldsymbol{\theta}}_L}(\boldsymbol{s})$ follows at $k = 0$; and under second moments, $\{\hat{\theta}_{Rk}^1, \hat{\theta}_{Rk}^2, \ldots, \hat{\theta}_{Rk}^p\}$ are mutually uncorrelated in view of their diagonal second-moment matrix. Moreover, on setting in succession all elements of $\boldsymbol{s}$ but one to zero, we deduce that the marginal *chf* s are

$$\{g_{\hat{\theta}_{Rk}^j}(s_j) = \mathrm{e}^{\mathrm{i}s_j\xi_j^2\theta_j/(\xi_j^2+k)}\phi_{\boldsymbol{\epsilon}}(s_j^2\xi_j^2/(\xi_j^2 + k)^2);\ 1 \leq j \leq p\}. \tag{3.3}$$

Subsequently designate $C_e(\hat{\theta}_{Rk}^j) = \xi_j^2\theta_j/(\xi_j^2 + k)$ and $S_c(\hat{\theta}_{Rk}^j) = \xi_j^2/(\xi_j^2 + k)^2$ as their respective location and scale parameters where, owing to symmetry, $C_e(\hat{\theta}_{Rk}^j)$ serves as the *center of symmetry* or the *centering parameter* of the corresponding marginal distribution.

We next examine the pairwise ordering $\mu \succeq_B \nu$ among $\{\hat{\theta}_{L1}, \hat{\theta}_{L2}, \ldots, \hat{\theta}_{Lp}\}$, and its evolution with k among $\{\hat{\theta}^1_{Rk}, \hat{\theta}^2_{Rk}, \ldots, \hat{\theta}^p_{Rk}\}$, all such comparisons to be taken as peakedness about the respective centers of symmetry.

Theorem 1. *Consider ridge solutions in* $\{(\boldsymbol{D}^2_\xi + k\boldsymbol{I}_p)\widehat{\boldsymbol{\theta}}_{Rk} = \boldsymbol{D}_\xi \boldsymbol{U}; k \geq 0\}$ *under spherical errors having scale parameters* $\{S_c(\hat{\theta}^j_{Rk}) = \xi_j^2/(\xi_j^2 + k)^2; 1 \leq j \leq p\}$ *such that* $\{\xi_1^2 \geq \xi_2^2 \geq \cdots \geq \xi_p^2 > 0\}$.

(i) *OLS solutions are ordered in peakedness as* $\{\mathcal{L}(\hat{\theta}_{L1}) \succeq_B \mathcal{L}(\hat{\theta}_{L2}) \succeq_B \cdots \succeq_B \mathcal{L}(\hat{\theta}_{Lp})\}$.

(ii) *Ridge solutions reverse in peakedness from OLS, i.e.,* $\mathcal{L}(\hat{\theta}^j_R(k)) \succeq_B \mathcal{L}(\hat{\theta}^i_R(k))$ *for* $j > i$, *if and only if* $k > \xi_i \xi_j$.

(iii) *Ridge solutions preserve the order of OLS for* $k \leq \min\{\xi_i \xi_j\}$.

(iv) *Complete reversal* $\{\mathcal{L}(\hat{\theta}^p_{Rk}) \succeq_B \cdots \succeq_B \mathcal{L}(\hat{\theta}^1_{Rk})\}$ *occurs at* $k > \max\{\xi_i \xi_j\}$.

(v) *Let* $(\hat{\theta}_1, \hat{\theta}_2)$ *be any such scalar estimators centered at* (θ_1, θ_2), *under spherical errors having finite moments* $E(\epsilon_i^{2r})$, *such that* $\mathcal{L}(\hat{\theta}_1) \succeq_B \mathcal{L}(\hat{\theta}_2)$. *Then even moments are ordered as*

$$\{E(\hat{\theta}_1 - \theta_1)^{2j} \leq E(\hat{\theta}_2 - \theta_2)^{2j}; 1 \leq j \leq r\};$$

odd moments $\{E(\hat{\theta} - \theta)^{2r-1}; 1 \leq j \leq r\}$ *vanish for both; and absolute moments satisfy*

$$\{E|\hat{\theta}_1 - \theta_1|^{2r-1} \leq E|\hat{\theta}_2 - \theta_2|^{2r-1}; 1 \leq j \leq r\}.$$

Proof. For a typical $\hat{\tau}_j = [\hat{\theta}_j - C_e(\hat{\theta}_j)]$ at issue here, all marginal distributions have *chf* s of the type $g_{\hat{\tau}_j}(s_j) = \phi_\epsilon(\omega_j s_j^2)$. These all belong to the same scale family, determined by $\phi_\epsilon(\cdot)$, symmetric about $0 \in \mathbb{R}^1$ since $\phi_\epsilon(s^2)$ is an even function. Conclusion (i) follows on successive applications of Lemma 1(i), since $\{\omega_1 \leq \omega_2 \leq \cdots \leq \omega_p\}$ with $\{\omega_j = 1/\xi_j^2; 1 \leq j \leq p\}$. To continue, for $i < j$ and $\xi_i^2 > \xi_j^2$, it suffices to compute the difference $S_c(\hat{\theta}^i_R(k)) - S_c(\hat{\theta}^j_R(k))$ in scale parameters as

$$\frac{\xi_i^2}{(\xi_i^2 + k)^2} - \frac{\xi_j^2}{(\xi_j^2 + k)^2} = \frac{\xi_i^2(\xi_j^2 + k)^2 - \xi_j^2(\xi_i^2 + k)^2}{(\xi_i^2 + k)^2(\xi_j^2 + k)^2}. \tag{3.4}$$

The numerator is $\xi_i^2\xi_j^4 - \xi_i^4\xi_j^2 + k^2(\xi_i^2 - \xi_j^2)$ which, on letting $\xi_i^2 = \xi_j^2 + \delta$ with $\delta > 0$, becomes

$$(\xi_j^2 + \delta)\xi_j^4 - (\xi_j^2 + \delta)^2\xi_j^2 + k^2(\xi_j^2 + \delta - \xi_j^2) = \delta[k^2 - \xi_j^2(\xi_j^2 + \delta)]. \tag{3.5}$$

The latter is positive, and so $S_c(\hat{\theta}^i_R(k)) > S_c(\hat{\theta}^j_R(k))$ and $\mathcal{L}(\hat{\theta}^j_R(k)) \succeq_B \mathcal{L}(\hat{\theta}^i_R(k))$, if and only if $k^2 > \xi_i^2\xi_j^2$, as asserted in conclusion (ii). Conclusions (iii) and (iv) follow on repeated applications of (ii). Conclusion (v) follows from Lemma 1(ii) on noting that all central and absolute central moments, as listed, belong to the order-determining class Γ of Lemma 1. All odd central moments vanish by symmetry. □

Remark 1. The ordering $\succeq_B$ substantially transcends the variance ordering of *OLS* as in (3.1). Specifically, under first moments Theorem 1(v) implies the L^1 ordering $E(|\hat{\theta}_1 - C_e(\hat{\theta}_1)|) < E(|\hat{\theta}_2 - C_e(\hat{\theta}_2)|)$ when $\omega_1 < \omega_2$, as well as the L^2 ordering $E[\hat{\theta}_1 - C_e(\hat{\theta}_1)]^2 < E[\hat{\theta}_2 - C_e(\hat{\theta}_2)]^2$ under second moments.

Remark 2. Under third moments the quantity $E(|\hat{\theta}_1 - C_e(\hat{\theta}_1)|^3)$ figures prominently in central limit theory, in constructing Berry–Esseén bounds on rates of convergence. Smaller values ensure tighter bounds.

Remark 3. Spherical Student t error distributions on ν degrees of freedom trace to Zellner (1976), but with no apparent contact with ridge regression. Nonetheless, it is remarkable that the canonical estimators exhibit the peakedness orderings of Theorem 1(i)–(iv), even for spherical Cauchy errors at $\nu = 1$ having heavy tails devoid of first moments.

3.2. Example

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

$$\begin{bmatrix} 2.97 & 0 & 0 \\ 0 & 0.015 & 0.005 \\ 0 & 0.005 & 0.015 \end{bmatrix}, \quad \begin{bmatrix} \frac{1}{2.97} & 0 & 0 \\ 0 & 75.00 & -25.00 \\ 0 & -25.00 & 75.00 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}. \tag{3.6}$$

The spectral decomposition $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{D}^2_\xi\boldsymbol{Q}'$ gives $\boldsymbol{D}^2_\xi = \mathrm{Diag}(2.97, 0.02, 0.01)$ and $\boldsymbol{Q}$ as displayed. The canonical form $\boldsymbol{\theta} = \boldsymbol{Q}'\boldsymbol{\beta}$ specifies $\theta_1 = \beta_1$, $\theta_2 = (\beta_2 + \beta_3)/\sqrt{2}$, and $\theta_3 = (\beta_2 - \beta_3)/\sqrt{2}$. Moreover, $V(\widehat{\boldsymbol{\theta}}_L) = \mathrm{Diag}(0.336700, 50, 100)$

Table 1
Choices for k in the Hospital Manpower Data corresponding to conventional criteria DF_k, GCV_k, C_k, $PRESS_k$, and HKB_k.

Name	DF_k	GCV_k	C_k	$PRESS_k$	HKB_k
Value for k	0.0004	0.004787	0.0050	0.2300	0.616964

Table 2
Products $\{\xi_i \xi_j;\ i < j\}$ for elements of $\boldsymbol{D}_\xi = \mathrm{Diag}(\xi_1, \ldots, \xi_5)$ as singular values for $\boldsymbol{Z}$ in the Hospital Manpower Data.

ξ_i	$\xi_2 = 0.816997$	$\xi_3 = 0.307625$	$\xi_4 = 0.201771$	$\xi_5 = 0.007347$
$\xi_1 = 2.048687$	1.673771	0.630228	0.413365	0.015052
$\xi_2 = 0.816997$		0.251329	0.164846	0.006003
$\xi_3 = 0.307625$			0.062070	0.002260
$\xi_4 = 0.201771$				0.001482

under second moments, and the condition number $c_1(\boldsymbol{X}'\boldsymbol{X}) = \xi_1^2/\xi_3^2 = 297.0$ is the ratio, $\mathrm{Var}(\hat{\theta}_3)/\mathrm{Var}(\hat{\theta}_1)$, of variances of the least to the most precisely estimated linear functions.

If instead we take $\mathcal{L}(\boldsymbol{\epsilon}) = \mathcal{S}_n(\boldsymbol{0}, \boldsymbol{I}_n)$, we may track the comparative peakedness of the estimators *via* Theorem 1. Observe that $k = 0.25 > \xi_1\xi_2 = 0.243721 > \xi_1\xi_3 = 0.172337 > \xi_2\xi_3 = 0.014142$. The scale matrix for ridge estimators at $k = 0.25$ is now $S_c(\widehat{\boldsymbol{\theta}_{Rk}}) = \mathrm{Diag}(0.286447, 0.274348, 0.147929)$. Progressing from $k = 0$ for *OLS*, to $k = 0.25$ for ridge, has seen a complete reversal in peakedness for estimating $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3]'$. In particular, $\hat{\theta}_3$, least precise under *OLS* for the contrast $\theta_3 = (\beta_2 - \beta_3)/\sqrt{2}$, now has greatest peakedness at approximately half the scale parameters of $\hat{\theta}_1$ and $\hat{\theta}_2$, all as ridge estimators at $k = 0.25$. On the other hand, Theorem 1(iii) assures that no reversals in peakedness might occur among ridge estimates for $k \le 0.014142$.

4. Case studies

4.1. The setting

As reported in Myers (1990), the Hospital Manpower Data consist of records at $n = 17$ US Naval Hospitals, to include: Monthly man-hours (Y); Average daily patient load (X_1); Monthly X-ray exposures (X_2); Monthly occupied bed days (X_3); Eligible population in the area ÷ 1000 (X_4); and Average length of patients' stay in days (X_5). The working model is

$$\{Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon_i;\ 1 \le i \le 17\}. \tag{4.1}$$

Following conventional usage, we center and scale to $\{\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}\}$ with $\boldsymbol{Z}'\boldsymbol{Z}$ in correlation form, our focus being the rates of change $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3, \beta_4, \beta_5]'$. The data are given in Table 3.8 of Myers (1990, pp. 132–133), and computations reported here utilize PROC IML of the SAS Programming System. The data are remarkably ill-conditioned: Singular values of $\boldsymbol{Z}$ are $\boldsymbol{D}_\xi = \mathrm{Diag}(2.048687, 0.816997, 0.307625, 0.201771, 0.007347)$, and the condition number is $c_1(\boldsymbol{Z}'\boldsymbol{Z}) = 77{,}754.86$.

4.2. Ridge regression

A striking diversity in choices for k has evolved, reflecting the problem as less than well-posed, often with profound differences among the solutions. Selected choices in wide usage are identified in Table 1, together with their values for the Hospital Manpower Data. These encompass $DF_k = \mathrm{tr}(\boldsymbol{H}_k) = \sum_{i=1}^{p} \xi_i^2/(\xi_i^2 + k)$ with $\boldsymbol{H}_k = [\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z} + k\boldsymbol{I}_p)^{-1}\boldsymbol{Z}']$; $PRESS_k = \sum_{i=1}^{n} e_{(i,\lambda)}^2$ as the cross-validation statistic of Allen (1974); $GCV_k = SS_{Res,k}/[n - (1 + \mathrm{tr}(\boldsymbol{H}_k))]^2$, a rotation-invariant analog called *Generalized Cross Validation* in Golub et al. (1979); $C_k = [(SS_{Res,k}/\widehat{\sigma}^2) - n + 2 + 2\mathrm{tr}(\boldsymbol{H}_k)]$ to achieve a variance–bias trade-off (Mallows, 1973); and $HKB_k = \widehat{\sigma}^2/\widehat{\boldsymbol{\beta}}_L'\widehat{\boldsymbol{\beta}}_L$ as in Hoerl et al. (1975) from simulation studies. Here $SS_{Res,k}$ is the residual sum of squares using ridge regression; $\widehat{\sigma}^2$ is the *OLS* residual mean square; and $\{e_{(i,\lambda)}^2\}$ are the *PRESS* residuals from ridge regression. For further details see Myers (1990, pp. 392–411), including numerical values for DF_k, C_k, and $PRESS_k$ as reported here. The ridge parameter k can be estimated by different methods. For more details on the estimation of k, we refer the reader to Kibria (2003) and Muniz and Kibria (2009), among others.

These choices carry hidden consequences regarding the concentration reversals of Theorem 1. Table 2 displays threshold values as products $\{\xi_i\xi_j;\ i < j\}$. From the matrix $\boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{q}_3, \boldsymbol{q}_4, \boldsymbol{q}_5]$, it is clear that $\theta_1 = \boldsymbol{q}_1'\boldsymbol{\beta}$ resembles a scaled average of $\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$, since $\boldsymbol{q}_1' = [0.485286, 0.453235, 0.484977, 0.460969, 0.333737]$. At the other extremity, $\theta_5 = \boldsymbol{q}_5'\boldsymbol{\beta}$, with $\boldsymbol{q}_5' = [-0.719480, -0.001160, 0.694081, 0.023436, 0.006780]$, is a near linear contrast between $\{\beta_1, \beta_2\}$ and $\{\beta_3, \beta_4, \beta_5\}$, since the inner product $\boldsymbol{1}_5'\boldsymbol{q}_5 = 0.003657$ is approximately zero.

As noted, the *OLS* estimates $\{\hat{\theta}_{L1}, \hat{\theta}_{L2}, \hat{\theta}_{L3}, \hat{\theta}_{L4}, \hat{\theta}_{L5}\}$ exhibit successively diminished precision, as do concentrations of ridge estimates for $k \le 0.001482$ from Theorem 1(iii). Since $GCV_k = 0.004787$ and $C_k = 0.0050$ exceed both thresholds $\xi_3\xi_5 = 0.002260$ and $\xi_4\xi_5 = 0.001482$, we find that ridge solutions for those choices of k now estimate the near contrast $\theta_5 = \boldsymbol{q}_5'\boldsymbol{\beta}$, with concentration greater than of $\{\hat{\theta}_{Rk}^3, \hat{\theta}_{Rk}^4\}$. Similarly, since $PRESS_k = 0.2300$ exceeds all of $\{\xi_1\xi_5 = 0.015052, \xi_2\xi_5 = 0.006003, \xi_2\xi_4 = 0.164846, \xi_3\xi_4 = 0.062070, \xi_3\xi_5 = 0.002260, \xi_4\xi_5 = 0.001482\}$, we may assert that the ridge estimate $\hat{\theta}_{Rk}^{5}$ at $k = 0.2300$ now is most concentrated, with scale parameter smaller than for $\{\hat{\theta}_{Rk}^{1}, \hat{\theta}_{Rk}^{2}, \hat{\theta}_{Rk}^{3}, \hat{\theta}_{Rk}^{4}\}$, and that $\hat{\theta}_{Rk}^{4}$ is more concentrated than $\{\hat{\theta}_{Rk}^{2}, \hat{\theta}_{Rk}^{3}\}$, no further reversals being apparent. If instead the choice were $HKB_k = 0.616964$, then the only non-reversals in peakedness, in comparison with both *OLS* and with ridge estimators at $k \leq 0.001482$, are between $(\hat{\theta}_{Rk}^{1}, \hat{\theta}_{Rk}^{2})$ and $(\hat{\theta}_{Rk}^{1}, \hat{\theta}_{Rk}^{3})$. All such reversals emerge as unintended consequences of the widely divergent values for k as derived from various *ad hoc* desiderata.

Observe that the foregoing phenomena could have been discovered numerically, apart from σ^2, on evaluating $V(\widehat{\boldsymbol{\theta}}_{Rk})$ under second moments, or equivalently $S_c(\widehat{\boldsymbol{\theta}}_{Rk})$ under spherical errors, as cases of note. The advantage of Theorem 1 is that it supports inequality assertions for $\{k > \xi_i\xi_j\}$ generally, and for $\{k \leq \min\{\xi_i\xi_j\}\}$ and $\{k > \max\{\xi_i\xi_j\}\}$ in particular, as in the preceding examples. We see such reversals as unanticipated and heretofore undiscovered consequences of ridge regression as initially formulated.

5. Conclusions

In summary, further insight is offered regarding the use of ridge regression in ill-conditioned models. This study begins with successive concentration orderings of *OLS* estimators for the canonical parameters. The ordering continues to hold for ridge solutions as k evolves, up to a threshold value given in Theorem 1(iii). Exceeding this, individual ridge estimators encounter reversals in their comparative concentrations as k evolves, up to another threshold value in Theorem 1(iv), beyond which complete reversal obtains. Specifically, for k sufficiently large, the estimator least concentrated at the outset achieves greatest concentration in comparison with other estimators at that k. Knowledge of these facts could be useful in choosing k so as to enhance the comparative performance for selected parameters of singular interest to an investigation.

Ridge regression continues in wide usage, prominently now in calibration studies in the chemical engineering and related literature. See Frank and Friedman (1993), Geladi (2002), Kalivas (2005) and Sundberg (1999) for recent surveys. Here we have emphasized spherical error models not necessarily having moments, together with the peakedness ordering of Birnbaum (1948). These encompass spherical Student t distributions on ν degrees of freedom as in Zellner (1976), to include heavy-tailed Cauchy errors at $\nu = 1$ devoid of moments. Connections to calibration problems are found in Jensen and Ramirez (2009), where it is seen that linearly calibrated data, subject to errors of calibration, may arise as ratios of Gaussian variables, thus having Cauchy distributions.

Acknowledgement

The authors thank the reviewer for a quick response and a contribution with additional references.

References

Allen, D.M., 1974. The relationship between variable selection and data augmentation and a method for prediction. Technometrics 16, 125–127.
Belsley, D.A., 1986. Centering, the constant, first-differencing, and assessing conditioning. In: Belsey, D.A., Kuh, E. (Eds.), Model Reliability. MIT Press, Cambridge, pp. 117–153 (Chapter 5).
Birnbaum, Z.W., 1948. On random variables with comparable peakedness. Ann. Math. Statist. 19, 76–81.
Frank, I.E., Friedman, J.H., 1993. A statistical view of some chemometrics regression tools. Technometrics 35, 109–135.
Geladi, P., 2002. Some recent trends in the calibration literature. Chemometr. Intell. Lab. Syst. 60, 211–224.
Golub, G.H., Heath, C.G., Wahba, G., 1979. Generalized cross validation as a method for choosing a good ridge parameter. Technometrics 21, 215–223.
Hoerl, A.E., 1962. Application of ridge analysis to regression problems. Chem. Eng. Prog. 58, 54–59.
Hoerl, A.E., 1964. Ridge analysis. Chem. Eng. Prog. Symp. Ser. 60, 67–77.
Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67.
Hoerl, A.E., Kennard, R.W., Baldwin, K.F., 1975. Ridge regression: Some simulations. Commun. Statist. 4, 105–123.
Jensen, D.R., 1997. Symmetry and unimodality in linear inference. J. Multivariate Anal. 60, 188–202.
Jensen, D.R., Foutz, R.V., 1989. The structure and analysis of spherical time-dependent processes. SIAM J. Appl. Math. 49, 1834–1844.
Jensen, D.R., Ramirez, D.E., 2009. Anomalies in the analysis of calibrated data. J. Statist. Comput. Simul. 79, 299–314.
Kalivas, J.H., 2005. Multivariate calibration: An overview. Anal. Lett. 38, 2259–2279.
Kibria, B.M.G., 2003. Performance of some new ridge regression estimators. Commun. Statist. Simul. Comput. 32, 419–435.
Mallows, C.L., 1973. Some comments on C_p. Technometrics 15, 661–675.
Muniz, G., Kibria, B.M.G., 2009. On some ridge regression estimators: An empirical comparisons. Commun. Statist. Simul. Comput. 38, 621–630.
Myers, R.H., 1990. Classical and Modern Regression with Applications, second ed. PWS-KENT Publishing Co., Boston, MA.
Sundberg, R., 1999. Multivariate calibration — Direct and indirect regression methodology. Scandinavian J. Statist. 26, 161–207.
Thisted, R.A., 1987. Comment. Statist. Sci. 2, 91–93.
Zellner, A., 1976. Bayesian and non-Bayesian analysis of the regression model with multivariate Student t error terms. J. Amer. Statist. Assoc. 71, 400–405.