Ridge Regression: A Historical Context

Roger W. Hoerl
Union College

Abstract
Two classical papers on Ridge Regression by Arthur Hoerl and Robert Kennard were published in Technometrics in 1970, making 2020 their 50th anniversary. The theory and practice of Ridge Regression, and of related biased shrinkage estimators, have been extensively developed over the years. Further, newer shrinkage estimators, such as the Lasso and the Elastic Net, have become popular more recently. These newer developments have led to renewed interest in the original 1970 papers. What has perhaps been lost since 1970 is the context of these classic papers. That is, who were Art Hoerl and Bob Kennard, and what led two statisticians working in the private sector to develop Ridge Regression in the first place? What are the origins of Ridge Regression? Where did the name come from? The purpose of this paper is to provide this historical context by discussing the men involved, their work at DuPont, and their approach to methodological development. As Art Hoerl was my father, this is admittedly a personal viewpoint.

1. INTRODUCTION

In 1970, Technometrics published two papers by Art Hoerl and Bob Kennard (Hoerl and Kennard 1970a,b) on the topic of Ridge Regression, essentially introducing this methodology to the statistics community. It is fair to say that no one, including the authors, suspected how impactful these articles would ultimately turn out to be. While their proposed estimation approach for collinear data met its share of criticism and resistance (e.g., Draper and Smith 1981), the method not only became common in practice, but also led to further developments in shrinkage estimation, such as Lasso (Tibshirani 1996) and Elastic Net (Zou and Hastie 2005). Research and utilization of these more modern but related methods has renewed interest in the classic Hoerl-Kennard papers. Art Hoerl (hereafter referred to as AH) was my father; I studied under him in graduate school, and subsequently published two articles on Ridge Regression with him (Hoerl et al. 1985, Hoerl et al. 1986). As might be imagined, we discussed the origins of Ridge Regression quite a bit. I also spent two summers as an intern in the Applied Statistics Group at DuPont while in graduate school. By that time, Robert (Bob) Kennard (hereafter referred to as RK) had been promoted outside of the ASG, but was still overseeing it. I met with him several times while working there, in addition to meeting him socially growing up. Ironically, I went to high school with his son Eric. So, I would like to think that I have a unique view into these men's journey to the development of Ridge Regression, which I share below. First, however, let me share some information on these men as individuals.

2. ABOUT THE AUTHORS

2.1 Who Was Art Hoerl (AH)?

AH was technically Arthur Edwin Hoerl, Jr., as his father was Arthur Hoerl. Arthur Hoerl, my grandfather, was the child of German immigrants, and lived much of his life in New York City. He eventually became a writer, and moved his family, including my father, to Los Angeles.

Although my grandfather was a prolific writer in several venues, he is best known professionally for writing the screenplay to the 1936 cult classic *Reefer Madness*. Their house was actually in Beverly Hills, and AH graduated from Beverly Hills High School with, among others, future producer Blake Edwards and actress Rhonda Fleming. I say this somewhat tongue-in-cheek, because their house, which I visited, was quite modest, and AH never fit in with the Hollywood or Beverly Hills crowds.

He received his B.S. in Mechanical Engineering from the University of Southern California (USC) in 1944. Immediately upon graduation he was drafted in the army, and had orders to report to Belgium for what we now know as the "Battle of the Bulge". Because of his engineering background and high scores on the Army math aptitude test, he was reassigned at the last minute to the Manhattan Project at Los Alamos, New Mexico, where he worked on bombing tables. At Los Alamos, he met Enrico Fermi, Robert Oppenheimer, and Klaus Fuchs, among other scientists, and also Marguerite Field, my mother.

Soon after the war, he began working as a mechanical engineer, and was exposed to problems related to data analysis, which intrigued him. He reentered USC, receiving an M.S. in mathematics in 1950. His real interest was in statistics, but at that time USC didn't offer a statistics degree. Upon graduation, he became the first statistician hired by the DuPont Company. In 1967 he left DuPont to join the University of Delaware faculty, in order to spend less time traveling, and more at home with his family, and also to focus more on research. He retired in 1986, and passed away in 1994.

It is my belief that AH's background in engineering significantly impacted the way he approached problems, in particular, the multicollinearity issue in regression. He was fundamentally a creative problem solver who knew statistics, not a statistician per se, at least not in the classical sense. Statistical methods were always a means to an end to him, not the end in themselves. My own undergraduate degree was in mathematics, and early in my career I tended to view problems from a mathematical versus problem-solving lens, so AH and I typically had very different takes on issues and how to approach them.

I do find it reassuring that there has recently been a broader realization of the need for statisticians to take more of an engineering (problem-solving) viewpoint, versus a statistical or mathematical viewpoint, when addressing real problems. I point to the formation of the International Statistical Engineering Association (ISEA - https://isea-change.org/), and also Michael Jordan's recent presentation at the University of Michigan's Symposium on Statistics in the Data Science Era (https://media.rackham.umich.edu/rossmedia/Play/1f811e3d1ad94e4d9d0f1b430cba8a341d). During this talk, Jordan noted the need for developing a problem-solving culture within the statistics discipline. Further, he suggested that we "embrace being engineers," and consider "...what statistical engineering could look like, as a counterpart to statistical science."

2.2 Who Was Bob Kennard (RK)?

The following borrows significantly from Bob's obituary, which can be found in its entirety at https://obittree.com/obituary/us/florida/indian-harbour-beach/beach-funeral-homes---east/robert-kennard/987155/. Robert Wakely Kennard was born January 27, 1923 in Newark, Delaware. RK

graduated from Newark High School in 1940. He was a "local boy who made good", becoming the class President and Valedictorian at Newark High. RK eloped with Helen Elizabeth Staats (Betty) in nearby Elkton, Maryland, which had a reputation for liberal marriage laws. Bob also served in the military (army) in World War II. He sustained a back injury in training and had the first successful disc surgery that was performed at the Walter Reed Army Hospital. Similar to AH's experience, the army noticed RK's mathematical skills, and promoted him to Technical Sergeant in the 2nd Signal Service Battalion of the Signal Corps. He was assigned to the Vint Hill intercept station in Warrenton, Virginia. There he identified and intercepted Japanese military and diplomatic radio signals transmitted at high speed Morse Code. His unit broke the Japanese "purple code" and intercepted the Imperial Command's message to their staff that they were going to surrender to the United States. Clearly, RK's early career involved solving real problems utilizing an engineering viewpoint.

After being discharged from the Army in 1946, he resumed his studies at the University of Delaware. He graduated in 1949 with a B.S. in physics, and M.S. in statistics in 1952. He received his Ph.D. in mathematical statistics at Carnegie Technological University (now Carnegie-Mellon). It is noteworthy both that RK had an undergraduate degree in a physical science – physics, and also that his Ph.D. was in mathematical statistics. Motivated by his background in physics, he retained an interest in physics, astronomy, and medicine throughout his life, often seeking opportunities for lifelong learning in these areas. The Hoerl-Kennard team was, therefore, grounded in engineering problem solving, natural science, and also mathematical statistics. I believe all three viewpoints were required in the development of Ridge Regression.

RK began his career with DuPont in 1955, five years after AH began working there. He moved through various supervisory positions, eventually becoming the manager for the Systems Engineering Division, within which the Applied Statistics Group resided. He retired in 1982, moving to Groveland, Florida. He taught math and statistics at Lake Sumter Community College for ten years, and passed away in 2011.

3. THE DEVELOPMENT AND EXTENSIONS OF RIDGE REGRESSION

3.1 Ridge Analysis: The Origins of Ridge Regression

While employed in the statistical group at DuPont, AH was often asked to optimize industrial processes involving more than the two or three independent variables traditionally seen in response surface literature. Although the method of canonical analysis had been developed by that time (Davies 1956), this was generally inadequate for multidimensional surfaces, for reasons to be discussed shortly. Possessing an engineering background, he felt the need for more than a numerical optimization of the estimated model. That is, he desired engineering insight as to what was going on in the process. Ridge Analysis was the approach he developed for this problem. Since the publication of Hoerl and Kennard (1970a,b), there has been significant confusion between Ridge Analysis and Ridge Regression, but Ridge Analysis was clearly the initial step, and it was the application of Ridge Analysis to the regression sum of squares that later led to the development of Ridge Regression.

The classic paper by Box and Wilson (1951) popularized the use of response surface methodology (RSM) to optimize industrial processes, particularly in the chemical industry. RSM

involves sequential designed experiments, often fitting the subsequent data with second order polynomials, to account for curvature and interaction. Quadratic response surface models of the following form were commonly applied:

$$y = \mathrm{b}_0 + \sum_{i=1}^{p} \mathrm{b}_i x_i + \sum\sum_{1 \le i \le j}^{p} b_{ij} x_i x_j + \mathrm{e}. \qquad (1)$$

Note that this model includes linear, quadratic, and two-factor interaction terms. In matrix notation, this can be written:

$$y = \mathrm{b}_0 + \boldsymbol{b}'\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}'\mathbf{B}\boldsymbol{x} + \mathrm{e}, \qquad (2)$$

where $\boldsymbol{b}$ is the $p \times 1$ vector of linear coefficients from equation (1), $\boldsymbol{x}$ is the $p \times 1$ vector of independent variables, e is the random error, and **B** is the $p \times p$ symmetric matrix whose diagonal elements are twice the quadratic terms, and whose off-diagonal elements are the interaction terms.

If the independent variables are standardized to have zero mean and equal variances, the experimental region can be easily interpreted as a geometric figure with the center point as the origin. For central composite designs (Box et al. 2005), this is roughly the hypersphere defined by $\boldsymbol{x}'\boldsymbol{x} < C^2$, for some distance ($C$) from the origin, depending on the placement of the axial points. By taking partial derivatives with respect to the independent variables and setting them equal to zero, we can find the stationary point (max, min, or "saddle point") at $\boldsymbol{x} = -\mathbf{B}^{-1}\boldsymbol{b}$. This point may or may not be inside the design space. In the case of two or three independent variables, contour plots of the response can be used to reveal promising areas of the design space (maximum or minimum), in addition to the numerical stationary point given above. These plots also reveal when the analyst is extrapolating outside the design space. In higher dimensions, however, contour plots require fixing $p - 2$of the independent variables, which makes interpretation much more difficult and confusing, especially in the case of interaction.

A canonical analysis, noted above, can be performed for any p, but lacks the simplicity of contour plots. It shifts the reference point away from the origin (center point) to the stationary point, which may be well outside the design space. Interpreting a response surface based on extrapolations is obviously of limited value. These were the issues that led AH to develop an alternative approach, which he named Ridge Analysis (Hoerl 1959, 1964). It is noteworthy, given AH's engineering background, that these papers were both published in chemical engineering journals, as was his first mention of Ridge Regression in the literature (Hoerl 1962).

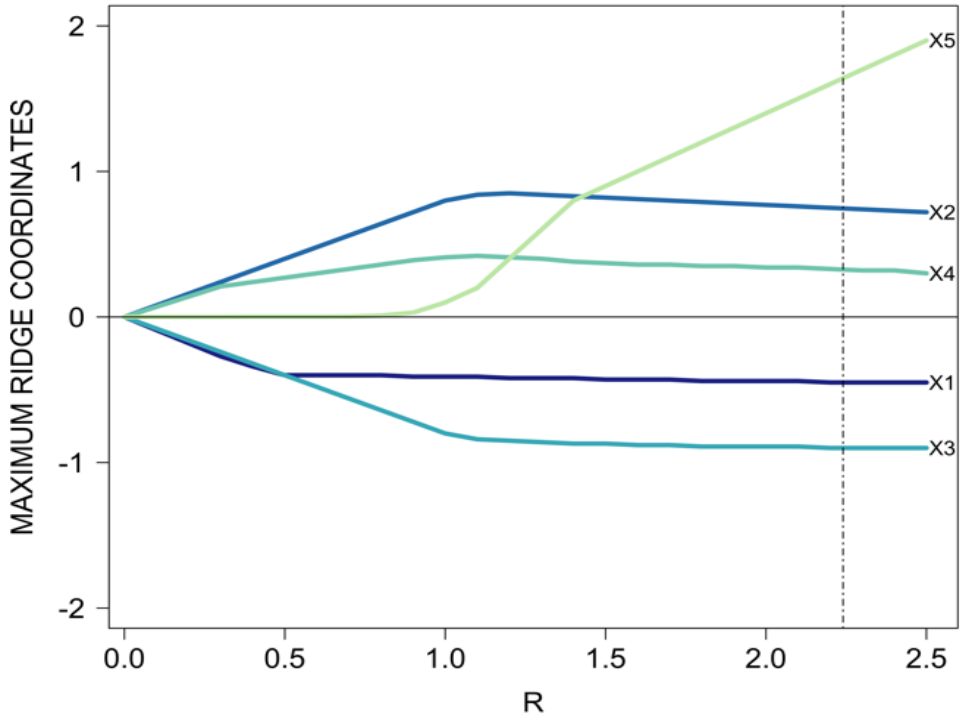
Canonical analysis uses the eigenvalues of **B** to trace response ridges relative to the stationary point. In contrast, Ridge Analysis determines the maximum (or minimum) predicted value of the response on concentric hyperspheres about the origin (center point), defined by $\boldsymbol{x}'\boldsymbol{x}$ = R2, where R is the distance from the origin. This is accomplished through repeatedly solving the following equation for a range of values of λ, based on the eigenvalues of **B**:

$$\mathbf{x} = -(\mathbf{B} - \lambda\mathbf{I})^{-1}\mathbf{b}, \qquad (3)$$

where **I** is the identity matrix, and λ is a Lagrangian multiplier (Hoerl 1959). The maximum "ridge" is defined by using values of $\lambda > \lambda_1$, i.e., where λ_1 is the largest eigenvalue of **B**. The minimum ridge is defined by using values of $\lambda < \lambda_p$, where λ_p is the smallest eigenvalue of **B**. Note that **B** is not typically positive semi-definite, so some eigenvalues may be negative. Secondary ridges, corresponding to secondary optima, are defined for λ values between the max and min eigenvalues. Of course, for designs that roughly form hyperspheres, the factorial and axial points of the design will be roughly the same distance from the center point. Letting this value be r, values of R less than r would obviously be within the design space, and values greater than r would constitute extrapolation.

The coordinates of this constrained maximum track the "ridge" of the maximum response from the center point to the boundary of the design space, hence the term *Ridge Analysis*. The graph of these coordinates versus R is called the "ridge trace", and shows the specific path of the maximum ridge. These terms would be subsequently applied to regression analysis involving collinear variables in HK-70 (Hoerl and Kennard 1970a), as discussed below. Figure 1 shows a sample plot of the ridge trace for a response surface discussed in my own paper, Hoerl (1985), which was of course based on AH's paper (Hoerl 1964). Figure 1 shows the coordinates in x space of the maximum ridge, between the origin (R = 0) to the perimeter of the design space (R = 2.24). My paper was intended to reintroduce Ridge Analysis to the statistics profession, hence the title "Ridge Analysis 25 Years Later".

Figure 1 Maximum Ridge Coordinates



3.2 The Application to Regression: Ridge Regression

A fundamental problem in collinear regression problems using least squares estimates, of course, is that the variances of the regression coefficients become large. Most of the methods that have been proposed over the years to address this problem are shrinkage estimates, which attempt to shrink the coefficients in order to reduce these variances, while adding some bias. The rationale for shrinkage is that the expected value of $\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\beta}}$, the squared magnitude of the estimated coefficient vector (in standardized units), is larger than $\boldsymbol{\beta}'\boldsymbol{\beta}$, the actual squared coefficient vector. That is,

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}'\boldsymbol{\beta} + \sigma^2 \mathrm{Tr}(\mathbf{X}'\mathbf{X})^{-1} = \boldsymbol{\beta}'\boldsymbol{\beta} + \sigma^2 \sum_{i=1}^{q} \lambda_i^{-1},$$

where λ_i is the i-th eigenvalue of $\mathbf{X}'\mathbf{X}$, and q is the number of terms in the regression model. We assume here that $\mathbf{X}$ is in correlation units, so $\mathbf{X}'\mathbf{X}$ is q×q, i.e., there is no β_0 term. Note that the squared magnitude of the coefficient vector is always biased high, but with nearly singular $\mathbf{X}'\mathbf{X}$ matrices, producing small eigenvalues approaching zero, it will be extremely biased on the high side. From a practical point of view, this often results in sign reversals of the coefficients. That is, coefficients that are known to be positive based on subject matter theory may have negative estimates, and vice versa.

Based on the need to shrink (or zero) the coefficient vector from the least squares' solution towards the origin, Hoerl and Kennard's approach (RK was now working alongside AH) was to apply Ridge Analysis to the residual sum of squares. The least squares solution obviously provides the overall minimum residual sum of squares, by definition. However, we can apply Ridge Analysis to the residual sum of squares, as it is a quadratic function of the coefficient vector. That is, we can write the least squares residual sum of squares as:

$$(\boldsymbol{y} - \widehat{\boldsymbol{y}})'(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \boldsymbol{y}'\boldsymbol{y} - 2\,\widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \widehat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X})\widehat{\boldsymbol{\beta}}, \qquad (4)$$

using the fact that $\widehat{\boldsymbol{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$. Equation 4 reveals that the residual sum of squares is a quadratic function of the parameter vector $\widehat{\boldsymbol{\beta}}$, just as in Equation 2 the response is a quadratic function of the independent variables. Ridge Analysis can therefore be applied to trace the coefficient coordinates of the minimum residual sum of squares (minimum ridge) from the origin ($\widehat{\boldsymbol{\beta}}=\mathbf{0}$) to the least squares' solution. Of course, we could calculate the minimum ridge beyond this point, but since the least squares' coefficient vector tends to be inflated in magnitude, especially for collinear data, this would not be of practical value. Equation 3 from Ridge Analysis becomes the familiar Ridge Regression equation:

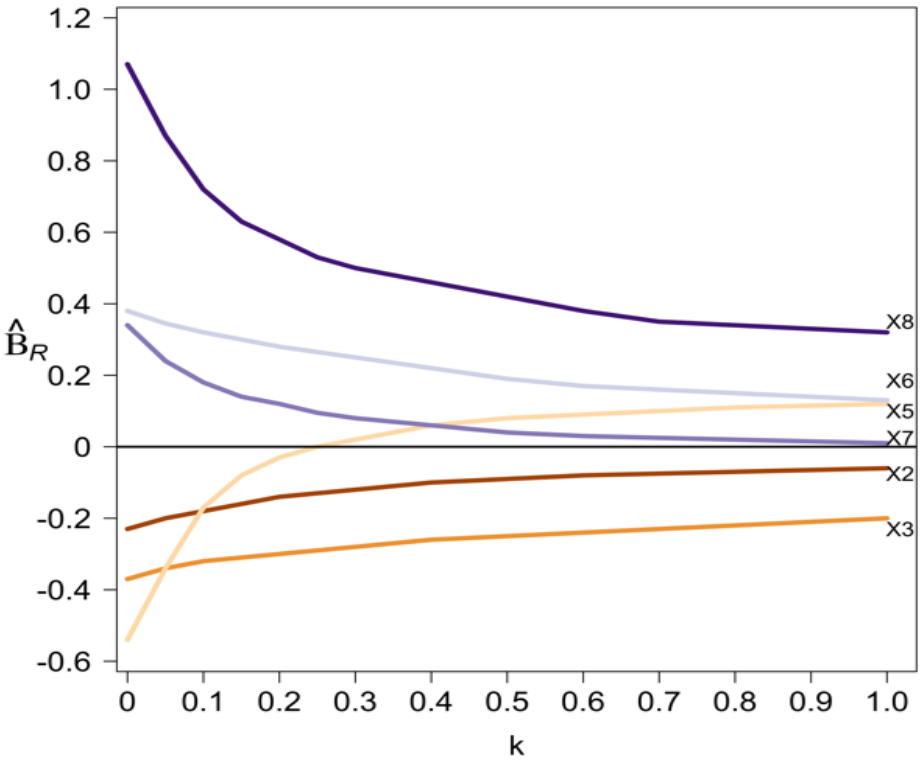
$$\widehat{\boldsymbol{\beta}_R} = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\boldsymbol{y}, \qquad (5)$$

where -k simply replaces λ from Equation 3. In Ridge Regression, positive values of k are used (negative λ). This is because we are interested in the minimum ridge, which tracks the minimum residual sum of squares for any distance from the origin ($\widehat{\boldsymbol{\beta}}=\mathbf{0}$). In a regression context, $\mathbf{X}'\mathbf{X}$ is positive semi-definite, so all eigenvalues are non-negative. Normally in Ridge Analysis we want to use λ values less than the smallest eigenvalue for the minimum ridge, or between 0 and λ_q. A

value of 0 for λ is of course the least squares solution ("stationary point"), and small positive values result in solutions farther away from the origin (larger coefficient vectors), hence negative λ, or positive *k*, are typically applied in Ridge Regression.

In short, Ridge Regression shrinks the estimated coefficient vector towards the origin along a path, a "ridge trace", that finds the coefficient estimates minimizing the residual sum of squares subject to the constraint $\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\beta}} = c^2$, where c is a constant varying from 0 (origin) to the original least squares coefficient vector magnitude. The user then selects the appropriate value of *k*, which corresponds to c, to determine the final estimates, either by algorithm or by looking at the ridge trace to see at what point the coefficients stabilize. The least squares solution is defined for $k = 0$, which typically appears on the left of the graph, and then the coefficients shrink as a function of *k*, moving left to right on the horizontal axis. See Figure 2, which is taken from Hoerl and Kennard (1970b). Note that this shows the ridge trace after removing two variables (x_1 and x_4) from the model.

Figure 2 Ridge Regression Trace



Hoerl and Kennard (1970a) provided proof of an existence theorem, stating in short that there always exists a value of $k > 0$ such that the expected mean square error of the coefficients is lower for Ridge Regression than least squares. That is,

$$E[\widehat{\boldsymbol{\beta}_R} - \boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}_R} - \boldsymbol{\beta})] < \mathrm{E}[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \qquad (6)$$

Of course, finding such a k in practice is the challenge. There has been extensive research over the years investigating algorithms to select *k*, including early work, such as Hoerl et al. (1975), and Lawless and Wang (1976). However, the practical advantage of looking at the ridge trace to gain engineering insight into the problem, which was a key aspect of Hoerl and Kennard (1970b), seems to have been all too often neglected. For example, Hoerl and Kennard (1970b) and Hoerl et al. (1985) illustrated how the ridge trace can be used to determine which variables have "staying power", and warrant being retained in a reduced model. Again, the lack of emphasis on the advantage of the ridge trace in practice may relate to a mathematical or algorithmic perspective, as opposed to the original authors' scientific-engineering perspective.

3.3 Relationship to Other Work

There has been, of course, a variety of other related work in the general area of estimate stabilization and shrinkage estimation, some of which preceded HK-70, some of which was concurrent, and some of which has occurred since. I comment here on only four specific developments, Tikhonov regularization, Stein Shrinkage, the Nonnegative Garrote, and the Lasso/Elastic Net.

Andrey Nikolayevich Tikhonov was a Soviet mathematician (both pure and applied) and geophysicist, who worked on, among other things, the "inverse problem" in geophysics (Tikhonov and Arsenin 1977, Tikhonov et al. 1998). Note that the dates of these publications represent when they were published in English. It is called an inverse problem because it starts with the effects and then calculates the causes, such as calculating the density of the earth from measurements of its gravity. In reality, of course, the density determines the gravity. In several of these problems, and also in solving matrix differential equations, Tikhonov needed to invert a matrix, and found that the matrix was nearly singular. Younger statisticians may not realize how poor the software for matrix inversion was prior to the last few decades, but it was quite unreliable. Tikhonov discovered that by adding a small positive constant to the diagonals of the matrix, numerical challenges such as inverting it, calculation of eigenvalues and vectors, the determinant, and so on, were much easier.

For example, suppose we have the following set of linear equations: $\mathbf{A}\boldsymbol{x} = \boldsymbol{b}$. The standard solution for $\boldsymbol{x}$ would be $\mathbf{A}^{-1}\boldsymbol{b}$. Suppose **A** is nearly singular, and its inverse cannot be accurately calculated, using numerical methods from the 20th century? Tikhonov's solution was to add a "Tikhonov matrix" **Γ** to **A** before inverting. That is, his solution was:

$$\mathbf{x} = (\mathbf{A} + \mathbf{\Gamma})^{-1}\mathbf{b} \qquad (7)$$

If Γ is chosen to be a multiple of the identity matrix, then the inversion of (**A** + **Γ**) is much more stable numerically than inversion of **A**, and Equation (7) is mathematically analogous to Equation (3) in ridge analysis. The use of this small diagonal matrix **Γ** has been referred to in the literature as "Tikhonov regularization", and was a breakthrough in solving numerous applied math and scientific problems of the day. In ridge analysis, of course, the **B** matrix isn't typically near-singular, because it is made up of regression coefficients resulting from designed experiments. So, the objective is quite different; tracing maximum coordinates, rather than addressing numerical instability.

Ridge regression, on the other hand, is typically applied to the near-singularity problem which motivated Tikhonov. The emphasis, however, is on tracing the coordinates of the minimum residual sum of squares ridge through the parameter space, towards the origin. That is, the emphasis is on shrinking the least squares estimates towards the origin, along a specific path - the minimum ridge of the residual sum of squares.

In a series of papers in the early 1960's, Stein (1960, 1962) and James and Stein (1961) proposed a type of shrinkage estimation that is often referred to in the literature as "Stein Shrinkage". Their fundamental approach was quite different from Ridge Regression, and was presented in a more general manner, that is, it was not limited to regression problems. James and Stein (1961) also provided an existence theorem, proving that under the assumption that a vector of random variables is multivariate normal, with mean vector $\boldsymbol{\theta}$ and variance/covariance matrix $\sigma^2\mathbf{I}$, one can reduce the expected mean square error of estimating $\boldsymbol{\theta}$ by shrinking the estimates by a factor of C. That is, there always exists a C such that the mean square error for Stein Shrinkage is lower than for the maximum likelihood estimate. Applied to regression, this implied that there is always a constant C such that the expected mean square error of estimating the regression coefficients (see Equation (6)) of Stein Shrinkage is lower than for least squares. The Stein Shrinkage estimator in a regression context would be of the form $\widehat{\boldsymbol{\beta}}_{ss} = C\widehat{\boldsymbol{\beta}}$, for some C, $0 < C < 1$, where $\widehat{\boldsymbol{\beta}}$ is the least squares estimate, and $\widehat{\boldsymbol{\beta}}_{ss}$ is the Stein Shrinkage estimate. Note that this approach comprises linear shrinkage, in that each coefficient is shrunk by an equal proportion.

As with selection of *k* in Ridge Regression, significant research has been conducted to identify appropriate values of C in practice (Stein 1962). From a practical point of view, a limitation of Stein Shrinkage is that, as a linear shrinkage approach, it is unable to reverse signs of coefficients. One of the original motivations for Ridge Regression is the common practical problem in which coefficients have the wrong sign, based on subject-matter knowledge. Hoerl and Kennard (1970b) provided two case studies where this was occurred, but with appropriate selection of *k*, the signs conformed to the expected sign.

The growth in application of Ridge Regression, as well as the expansion of research conducted on it, has led to more modern approaches to regression. One such method is the Nonnegative (nn) Garrote (Breiman 1995). The motivation for Breiman's work in this area was the desire to combine subset selection – to simplify models, with shrinkage estimation – to further reduce variance in estimation. The nn-Garrote accomplishes both, at least to a degree. Like Stein Shrinkage, it shrinks each parameter linearly, but in the case of the nn-Garrote, there are different shrinkage values for each parameter, so we have a set of $c_{i,}\ i = 1, \ldots, n$. Alternatively, we can define a diagonal matrix $\mathbf{C}$, which has the individual $c_{i,}$ on the diagonal, and zeros on the off diagonals. To obtain the nn-Garrote estimates, we minimize the residual sum of squares, based on the least squares estimates ($\widehat{\boldsymbol{\beta}}$), as a function of $\mathbf{C}$:

$$(\boldsymbol{y} - \mathbf{X}\boldsymbol{C}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \mathbf{X}\boldsymbol{C}\widehat{\boldsymbol{\beta}}), \qquad (8)$$

subject to: $c_i \geq 0$, and $\sum c_i \leq s$, for some constant *s*. Obviously, if $\mathbf{C}$ were the identity matrix (all $c_i = 1$, then we simply have the least squares estimates. However, "as the "garrote is drawn

tighter by decreasing s" (Breiman 1995 p. 374), some c_i are driven to 0, while others are shrunk but remain positive.

While the nn-Garrote achieves the joint objectives of subset selection and shrinkage, like Stein Shrinkage it is unable to reverse the signs of coefficients, as it is a "nonnegative" Garrote, i.e., the c_i cannot be negative. Building on Breiman's work, Tibshirani (1996) proposed a method for direct shrinkage and selection of the parameters, rather than identification of individual shrinkage values. The Least Absolute Shrinkage Selection Operator, or Lasso, uses a similar approach to Ridge Regression, but modifies the least squares estimates through the constraint $\Sigma|\beta_i| \leq c$, that is, by fixing the maximum sum of the absolute values of the estimated coefficients, not their squares. So, the Lasso fixes the sum of the parameters, in absolute value, rather than the sum of the shrinkage values. It obviously shares similarities with both Ridge Regression and the nn-Garotte. Some coefficients are typically forced to zero in the Lasso as well, hence it also provides subset selection as well as shrinkage. As noted previously, the ridge trace provides graphical guidance as to which variables should be retained in the model, but none of the coefficients are forced to zero mathematically. An advantage of the Lasso is that like Ridge Regression, it can reverse the signs of coefficients that appear incorrect relative to subject matter knowledge.

Zhou and Hastie (2005) introduced a related technique, the Elastic Net, to overcome some limitations of the Lasso, particularly when the number of independent variables exceeds the sample size ($q > n$), and with very high degrees of collinearity. The Elastic Net incorporates both Ridge Regression (sum of squared coefficients) and Lasso (sum of absolute value of coefficients) penalties in the estimation process. That is, it minimizes an objective function that is the residual sum of squares plus Ridge and Lasso penalties. The result can therefore be considered an intermediate solution between Ridge Regression and the Lasso. See James et al. (2013) for further details on these modern regression methodologies.

4. Summary

Arthur Hoerl and Robert Kennard published their classic papers on Ridge Regression in 1970 (Hoerl and Kennard 1970a,b), making 2020 the 50th anniversary of their publication. While the underlying theory and practical application of this method have been well-researched since 1970, the story of the motivation behind these papers has, perhaps, been overlooked or forgotten. This is unfortunate, as I feel strongly that it is important for the profession to understand that novel developments such as Ridge Regression do not occur in a vacuum. Rather, there is almost always an important context that provides clues as to why researchers accomplished such breakthroughs. In the case of Hoerl and Kennard, I believe that a critical aspect of this context was the application of scientific and engineering perspectives to the problem of collinearity, to augment a statistical/mathematical perspective. Secondly, the integration of multiple methods, in this case combining the response surface technique Ridge Analysis with least squares regression modeling, is almost always required to address complex problems.

It is an honor to be able to celebrate "HK-50" with the readers of Technometrics, the journal which published these original papers.

References:

Box, G.E.P., Hunter, J.S., and Hunter, W.G. (2005), *Statistics for Experimenters*, 2nd ed., John Wiley and Sons, Hoboken, NJ.

Box, G.E.P., and Wilson, K.B. (1951), "On the Experimental Attainment of Optimum Conditions," *Journal of the Royal Statistical Society*, *Series B*, 13, 1-38.

Brieman, L. (1995), "Better Subset Regression using the Nonnegative Garrote," *Technometrics*, 37, 4, 373-384.

Davies, O.L. (1956), *Design and Analysis of Industrial Experiments*, Hafner, New York.

Hoerl, A.E. (1959), "Optimum Solution of Many Variables Equations," *Chemical Engineering Progress*, 55, 69-78.

Hoerl, A.E. (1962), "Application of Ridge Analysis to Regression Problems," *Chemical Engineering Progress*, 58, 54-59.

Hoerl, A.E. (1964), "Ridge Analysis," *Chemical Engineering Progress Symposium Series*, 60, 67-77.

Hoerl, A.E., and Kennard, R.W. (1970a) "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 1, 55-67.

Hoerl, A.E., and Kennard, R.W. (1970b), "Ridge Regression: Applications to Nonorthogonal Problems," *Technometrics*, 12, 1, 69-82.

Hoerl, A.E., Kennard, R.W., and Hoerl, R.W (1985), "Practical Use of Ridge Regression: A Challenge Met," *Applied Statistics*, 34, 2, 114-120.

Hoerl, A.E., and Kennard, R.W., and Baldwin, R.F. (1975), "Ridge Regression: Some Simulations," *Communications in Statistics, Part A – Theory and Methods*, 4, 105-123.

Hoerl, R.W. (1985), "Ridge Analysis Twenty-Five Years Later," *The American Statistician*, 39, 3, 186-192.

Hoerl, R.W., Schuenemeyer, J.H., and Hoerl, A.E. (1986), "A Simulation of Biased Estimation and Subset Selection Regression Techniques," *Technometrics*, 28, 4, 369-390.

James, W. and Stein, C.M. (1961), "Estimation with Quadratic Loss," *Proceedings of the 4th Berkeley Symposium*, 1, 361-379.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning*, Springer, NY.

Lawless, J.K., and Wang, P. (1967), "A Simulation Study of Ridge and Other Regression Estimators," *Communications in Statistics, Part A – Theory and Methods*, 5, 307-323.

Stein, C.M, (1960), "Multiple Regression," in *Essays in Honor of Harold Hotelling*, Stanford University Press, Palo Alto, CA.

Stein, C.M, (1962), "Confidence Sets of the Mean of a Multivariate Normal Distribution," *Journal of the Royal Statistical Society*, Ser. B, 24, 265-296.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Ser. B, 58, 267–288.

Tikhonov, A.N., and Arsenin, V.Y. (1977). *Solution of Ill-posed Problems*, Winston & Sons, Washington.

Tikhonov, A.N., Leonov, A.S., and Yagola, A.G. (1998). *Non-linear Ill-posed Problems*, Chapman & Hall, London.

Zhou, H, and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society*, Ser. B, 67, 301-320.