mathematics

MDPI

Article

Ridge Regression and the Elastic Net: How Do They Do as Finders of True Regressors and Their Coefficients?

Rajaram Gana

Department of Biochemistry and Molecular & Cellular Biology, School of Medicine, Georgetown University, Washington, DC 20057, USA; rg476@georgetown.edu

Abstract: For the linear model $Y = Xb + error$, where the number of regressors (p) exceeds the number of observations (n), the Elastic Net (EN) was proposed, in 2005, to estimate b. The EN uses *both* the Lasso, proposed in 1996, and ordinary Ridge Regression (RR), proposed in 1970, to estimate b. However, when $p > n$, using *only* RR to estimate b has not been considered in the literature thus far. Because RR is based on the least-squares framework, only using RR to estimate b is computationally much simpler than using the EN. We propose a generalized ridge regression (GRR) algorithm, a superior alternative to the EN, for estimating b as follows: partition X from left to right so that every partition, but the last one, has 3 observations per regressor; for each partition, we estimate Y with the regressors in that partition using ordinary RR; retain the regressors with statistically significant t-ratios and the corresponding RR tuning parameter k, by partition; use the retained regressors and k values to re-estimate Y by GRR across all partitions, which yields b. Algorithmic efficacy is compared using 4 metrics by simulation, because the algorithm is mathematically intractable. Three metrics, with their probabilities of RR's superiority over EN in parentheses, are: the proportion of true regressors discovered (99%); the squared distance, from the true coefficients, of the significant coefficients (86%); and the squared distance, from the true coefficients, of estimated coefficients that are both significant and true (74%). The fourth metric is the probability that none of the regressors discovered are true, which for RR and EN is 4% and 25%, respectively. This indicates the additional advantage RR has over the EN in terms of discovering causal regressors.

Keywords: elastic net; generalized ridge regression; ordinary ridge regression; statistical significance

MSC: 62J05; 62J07

Citation: Gana, R. Ridge Regression and the Elastic Net: How Do They Do as Finders of True Regressors and Their Coefficients? *Mathematics* **2022**, *10*, 3057. https://doi.org/10.3390/math10173057

Academic Editors: Codruta Mare and Ioana Florina Coita

Received: 23 July 2022
Accepted: 19 August 2022
Published: 24 August 2022

1. Introduction

Although half a century has passed since Hoerl and Kennard's [1] discovery of ridge regression (RR), the perception that it cannot set the coefficients of "insignificant" regressors to zero lives on. Because this is a mathematically intractable perception, it has likely limited, more than have other things, the standalone use of RR for regressor subset selection when the number of regressors (p) exceeds the number of observations (n). We address this perception by proposing a simple regressor selection method based on RR in conjunction with the classical concept of statistical significance testing with t-ratios, when $p > n$. We examine our proposal via a simulation study because of its mathematically intractable nature. The use of computer methods for mathematical discovery is fairly well known to mathematicians. For others, the excellent work of Petkovsek, Wilf and Zeilberger [2] may be of interest in this regard.

As background, we recapitulate some key RR concepts first, which are relevant to this paper. These concepts are mathematized using matrix analysis (see Schott [3] and Seber [4]), which is a convenient methodology to work with them.

For the well-known linear model $Y = Xb + \epsilon$, where Y is centered, X is in correlation form (Vinod [5] discuses why this is important in RR), ϵ is a vector of unobservable errors (a.k.a. residuals) and $n > p$, the "ordinary" RR estimator of b, b_{RR}, is the following one:

$$b_{RR} = \left(X^T X + kI\right)^{-1} X^T Y \tag{1}$$

where $k > 0$, I is a $p \times p$ identity matrix and T is the matrix transpose operator. Among several linear estimators, (1) is known to optimally shrink b (see Frank and Friedman [6]).

When $k = 0$, b_{RR} defaults to the well-known least squares (LS) estimator of b, b_{LS}. For non-stochastic k, the variance of b_{RR} is given by:

$$Var(b_{RR}) \equiv s_k^2 = s^2 \left(X^T X + kI\right)^{-1} X^T X \left(X^T X + kI\right)^{-1} \tag{2}$$

where s^2 is the estimated residual variance when b is estimated by LS. The ith element of b_{RR} divided by the square root of the ith diagonal element of s_k^2 is denoted by $t_{RR}^{(i)}$; and it measures the signal-to-noise ratio (SNR) in the ith element of b_{RR}, while making it an approximate t-ratio (see Halawa and El-Bassiouni [7]). Testing the null hypothesis that the ith element of b, b_i, is zero for several values of k, is discussed by Gokpinar and Ebegil [8].

Using classical calculus, Hoerl and Kennard [1] mathematically proved that there always exists a value of $k > 0$ such that the average squared distance of b_{RR} from the true coefficient vector b is strictly less than the average squared distance of b_{LS} from b. However, this RR existence theorem does not lead to determining what an "optimal" value of k should be, given X. Different methods for selecting k are discussed by Muniz et al. [9]. This limitation regarding k, together with the complex distributional properties associated with studying b_{RR}, make several aspects of RR mathematically intractable. Thus, computer methods, and simulation in particular, become invaluable methodologies for discovering RR properties.

What originally motivated RR? Long before the work of Hoerl and Kennard [1], mathematicians had recognized that inverting an ill-conditioned matrix, A, after diagonal incrementation would closely approximate its inverse (see Piegorsch and Casella [10]). That is, it was fairly well-known to mathematicians that, for small k, $(A + kI)^{-1} \cong A^{-1}$. RR, whose origins can be traced back to 1959 (see Hoerl [11]), was originally proposed for accurately estimating b when $X^T X$ is ill-conditioned (see Hoerl [12]). Later, and somewhat surprisingly, Brook and Moore [13] found that even absent multicollinearity in X, b_{LS} tends to be much too long on average, providing additional motivation for the use of RR despite early criticism of it (e.g., by Smith and Campbell [14]).

An alternative mathematical representation of RR is the following optimization problem:

$$\left.\begin{array}{c} Minimize\ (Y - Xb)^T (Y - Xb) \\ subject\ to \\ \sum b_i^2 \leq k \geq 0 \end{array}\right\} \tag{3}$$

Given k, b_{RR} would emerge as the solution to this nonlinear programming (NLP) problem with a quadratic objective function and a single quadratic constraint. Simple matrix analysis will get us from (3) to (1), eliminating the need to enter into the more complex world of trying to solve NLP problems.

In 1996, the Lasso (see Tibshirani [15]) was proposed for estimating b; it is the solution to the following optimization problem:

$$\left.\begin{array}{c} Minimize\ (Y - Xb)^T (Y - Xb) \\ subject\ to \\ \sum |b_i| \leq k \geq 0 \end{array}\right\} \tag{4}$$

The optimization in (4) is done using the procedures of Osborne, Presnell and Turlach [16], and Efron, et al. [17]. As is easy to see, (4) is more complex to solve than is (3), because in (4) we leave the LS framework behind and, thus, require mathematical operations beyond matrix inversion to solve it. Formulation (4) can be seen as a quadratic programming problem with linear constraints. Other linear constraints (on b) can be introduced into (4), but solving it will become more complex (e.g., see Delbos and Gilbert [18]). That is, in the "ordinary" Lasso, the sum of squared errors is minimized, subject to a constraint on the sum of the absolute values of the coefficients. In contrast, RR minimizes the same objective function with a constraint on the sum of the squares of the coefficients. Mathematicians were aware of the idea of the Lasso before the work of Tibshirani [15], as attested to by the works of Taylor, Banks and McCoy [19] and Santosa and Symes [20]. Tibshirani's [15] work popularized the Lasso for subset regressor selection.

To solve for b when $p > n$, the Elastic Net (EN) was proposed in 2005 by Zou and Hastie [21]. The EN was hypothesized to be "*like a stretchable fishing net that retains 'all the big fish'*" and uses *both* the Lasso and RR to find b. The EN is represented as the solution of b to the following optimization problem that mixes the Lasso and RR concepts:

$$\left.\begin{array}{c} Minimize\ (Y - Xb)^T(Y - Xb) \\ subject\ to \\ \sum|b_i| \leq t_1 \geq 0 \\ \sum b_i^2 \leq t_2 \geq 0 \end{array}\right\} \quad (5)$$

Using Lagrange [22] multipliers [23], (5) can be recast as the solution of b to the following optimization problem:

$$Minimize: \ (Y - Xb)^T(Y - Xb) + \delta_1 \sum|b_i| + \delta_2 \sum b_i^2 \quad (6)$$

where $0 < \delta_0 \leq \delta_2 \leq 1$ and δ_0 is a "small" predetermined constant for us. b is estimated after accounting for the double shrinkage inherent in the EN formulation (by rescaling coefficients by $1 + \delta_2$). The optimization is executed per Efron et al. [17]. As is obvious, the EN formulation is also complex.

Thus far in the literature, the question of whether b can be estimated *only* using RR (when $p > n$) has not been raised. The rest of this paper is dedicated to answering this question. We compare RR and the EN, under the assumption of a linear model, by computing four metrics: the proportion of true regressors discovered; the squared distance, from the true coefficients, of the estimated coefficients that are significant (but not necessarily true); the squared distance, from the true coefficients, of estimated coefficients that are both significant and true; and the chance that none of the significant regressors found are true. Results indicate that RR surpasses the EN with regard to all of these metrics. This means there may be simpler ways to combine the Lasso and RR for model discovery, in lieu of the more complicated way of doing so using the EN. RR is much "simpler" than the EN because RR is a variant of the simple, and profound, least squares method independently discovered by Gauss and Legendre in the early 18th century (see the works of Plackett [24] and Stigler [25]). Specifically, Stigler [25] makes the following penetrating observation: "The method of least squares is the automobile of modern statistical analysis: despite its limitations, occasional accidents, and incidental pollution, it and its numerous variations, extensions, and related conveyances carry the bulk of statistical analyses, and are known and valued by nearly all." Thus, RR can provide important feedback on the outputs of other fashionable competitors, such as machine learning, with their pervasive "black-box" focus on prediction, rather than on the process (i.e., causality) generating the data, as an end in itself.

The next section provides an overview of the methods used. Other sections provide granular details on the simulation methodology, its outputs, and the results found. Because scientific replicability is germane to any such study, several tables of results derived and other relevant materials, such as the code used for simulation, are included as supplemen-

tary materials (SM) to this paper. The tables may provide insight into other results that we may have missed observing.

2. Materials and Methods

It is well known that, if $p > n$, then b cannot be estimated by LS. However, if $p > n$, then b can be estimated by RR, which is a "simple" and elegant modification of the LS method. But RR is not used, on a standalone basis, to estimate b because of the perception that RR will estimate *all* of the coefficients of X and thus produce a "meaningless" solution if only a subset of the p regressors, $p^{true} < p$, say, actually generate Y in nature. However, entertaining such a perception is mistaken because the RR-estimated t-ratio for the ith regressor, $t_{RR}^{(i)}$, can be used to eliminate insignificant regressors from the specification of Y. This paper addresses this mathematically intractable perception by showing, via a simulation study, that RR on a standalone basis can yield important insights regarding p^{true}.

2.1. Algorithmic Description of the RR Alternative to the EN

To simplify the concept of the EN, and avoid complex optimization methods, we propose a simple and novel alternative to it, which can be expressed at high-level as follows: Given X with $p > n$, we partition X from left to right such that every partition, but the last one, has between 3 and 4 observations per regressor (OPR). For each partition, we estimate Y with the regressors in that partition using two predefined values of the RR tuning parameter, k: one that of Hoerl, Kennard and Baldwin [26] and the other that of Lawless and Wang [27]. We retain the statistically significant regressors found via both values of k (using $t_{RR}^{(i)}$), at an α level of 15%, in each partition. Then we use these selected values of k and the statistically significant regressors retained to re-estimate Y using generalized RR (GRR), with each set of the predefined k values selected—i.e., Y is re-estimated twice. When "low" levels of collinearity are present in X, we use the set of k values of Hoerl, Kennard and Baldwin (HKB) to estimate b; otherwise, we use the set of k values of Lawless and Wang (LW) to estimate b. With high probability, this simple algorithm will find more of the true regressors than would the EN, and, with high probability, the resultant estimate of b will be more precise than the corresponding estimate yielded by the EN.

The alternative to the EN outlined in the predecessor paragraph is made precise by putting it in algorithmic form next. In particular, it is recast precisely as the following algorithm:

(a) Starting from left to right, we partition X into m sub-matrices such that $3 \leq \lfloor \frac{n}{m} \rfloor < 4$, where $\lfloor \cdot \rfloor$ denotes the greatest integer function. If $\frac{n}{m} = 3$, then there will be m partitions with m regressors in each partition; otherwise, the mth partition will have $p - \lfloor \frac{n}{m} \rfloor \times m$ regressors and all predecessor partitions will have m regressors each. This partitioning is done so that all but the last partition has 3 OPR. We let the partitions be named, from left to right, $X_1, X_2, \ldots, X_m$, respectively. The number 3 OPR is chosen as the lower bound of $\lfloor \frac{n}{m} \rfloor$ because Austin and Steyerberg [28] have shown that an OPR of 2 is enough to detect statistical significance; and we judgmentally increased that by one. This is set as a hard constraint.

(b) We define a "concatenation" operator $\cup$ that will be applied to matrices having the same number of rows but not necessarily the same number of columns. The operator $\cup$ will concatenate two matrices by creating a new matrix containing all of the columns of the two matrices while retaining duplicate columns only once. Concatenation is done from left to right; when duplicate columns exist, the leftmost one will be the only one (among the duplicates) retained. When such an operator operates over multiple matrices, we will index them, following familiar practice with other well-known operators such as the summation operator (e.g., $\sum_{i=1}^{i=5} w_i$). For example, under this convention, $X = \cup_{j=1}^{j=m} X_j \equiv \cup_1^m X_j$.

(c) We define another concatenation operator V that will concatenate matrices by retaining all columns, including duplicate columns, as is. The following matrices (where the lower-case letters denote real numbers) illustrate how the two operators work:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cup \begin{pmatrix} a & e & g \\ c & f & h \end{pmatrix} \equiv \begin{pmatrix} a & b & e & g \\ c & d & f & h \end{pmatrix} \tag{7}$$

and

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mathrm{V} \begin{pmatrix} a & e & g \\ c & f & h \end{pmatrix} \equiv \begin{pmatrix} a & b & a & e & g \\ c & d & c & f & h \end{pmatrix} \tag{8}$$

(d) We estimate Y by RR using the regressors in X_j with k set to the value prescribed by HKB, which is $k = \frac{ps^2}{b^T b}$. We denote this prescribed value of k by k^j_{HKB}. We use the value of $t^{(i)}_{RR}$ for each regressor i in X_j to determine whether it is statistically significant at an α level (i.e., type I error) of 15%. We denote the sub-matrix of X_j containing the statistically significant regressors found as $X_{j,\,HKB}$.

(e) We repeat step "(d)" by using the value of k prescribed by LW, which is $\frac{1}{F}$, where F is the usual F ratio in the analysis of variance (ANOVA) table resulting from estimating b by classical LS. We denote this prescribed value of k by k^j_{LW} and denote the resultant sub-matrix of X_j containing the significant regressors found as $X_{j,\,LW}$.

(f) We create $X_{j,\,RR} = X_{j,\,HKB} \cup X_{j,\,LW}$, which is the set of significant regressors in X_j identified by RR using either k^j_{HKB} or k^j_{LW}. We let $p_j > 0$ denote the number of regressors in $X_{j,\,RR}$.

(g) We repeat steps "(d)", "(e)" and "(f)" $\forall j$, where $\forall$ denotes the universal quantifier "for all", an operator borrowed from Whitehead and Russell [29].

(h) We define the statistically significant sub-matrix of regressors identified in X by RR as follows $X_{RR} \equiv \cup_{\forall j \ni p_j > 0} X_{j,\,RR}$, where $\ni$ denotes "such that". The idea of subscripting and superscripting alphabets, for creating notation, is borrowed from the idea of the "tensor" (see Ricci-Curbastro and Levi-Civita [30]).

(i) We re-estimate Y by GRR using the regressors in X_{RR} with the following vector of k values: $K_{HKB} \equiv \vee_{j=1}^{j=m} \vee_1^{p_j} k^j_{HKB}$. Under GRR (see Hoerl and Kennard [1]), the usual matrix diagonal increment kI will be replaced by $D(K)I$, where K is a row-vector of constants, rather than a single non-stochastic number as it is in the usual "familiar" form of RR; $D(K)$ is the corresponding diagonal matrix with the elements of K on the diagonal, and I is the usual identity matrix. We let the resultant GRR estimated coefficients be denoted by b_{HKB}.

(j) We repeat step "(i)" with the following vector of k values: $K_{LW} \equiv \vee_{j=1}^{j=m} \vee_1^{p_j} k^j_{LW}$. We let the resultant GRR estimated coefficients be denoted by b_{LW}.

(k) We will see that, for "low" to "moderate" levels of multicollinearity in X, b_{HKB} is a good solution, and, for "severe" levels of collinearity in X, b_{LW} is a good solution. The mathematical characterization of the adjectives in quotes will be clarified below. We define the subset (sub-vector) of b_{RR} containing only the true coefficients found by GRR as b^{true}_{RR}, where the subscript $RR \equiv HKB$ or LW as the case may be.

2.2. Simple Example to Illustrate Algorithm Use

To lessen the level of abstraction in the notation above, we illustrate it assuming we have a dataset X with 5 observations and 9 regressors. We let the 9 regressors be denoted by Z_1 through Z_9, respectively, and the regressand denoted, as usual, by Y. Because we have 5 observations, we can partition the first 8 regressors into subsets of two each and keep the ninth regressor alone. Thus, we have 5 partitions ($m = 5$), and the OPR for 4 partitions is $5 \div 2 = 2.5 > 2$, which meets the criterion of having an OPR of at least 2 to detect statistical

significance. If each Z_i is viewed as a 5×1 matrix (vector), we can define our partitions as follows:

$$X_1 \equiv Z_1 \cup Z_2, \;\; X_2 \equiv Z_3 \cup Z_4, \;\; X_3 \equiv Z_5 \cup Z_6, \;\; X_4 \equiv Z_7 \cup Z_8, \;\; X_5 \equiv Z_9 \,,$$

where $X = \cup_1^5 X_j$.

We can now perform 5 RRs to estimate Y and identify statistically significant regressors using the modified t-statistic. These RRs are: Y vs. X_1, Y vs. X_2, Y vs. X_3, Y vs. X_4 and Y vs. X_5. Each RR is used to identify the statistically significant regressors, given the value of k used (e.g., that of HKB or LW). Note, that on a standalone basis, each RR will be a misspecified model if more than two regressors generate Y in nature. However, this misspecification does not matter, because we ignore the estimated coefficients of the significant regressors yielded by each of the five RRs. At this stage, we are only interested in whether the regressors are statistically significant or not. This approach has a point of contact with the classical stepwise procedure of Efroymson [31], which starts with no regressors and adds them one at a time according to their partial F-statistics (see Hocking [32]) until either all regressors are included or until no excluded regressors' partial F-statistic is statistically significant. This procedure converges (see Miller [33]). Efroymson [31] selected one variable at a time to test its significance. We select predefined "blocks" of variables one at a time to test their significance. In Efroymson's [31] approach, we are also not interested in the intermediate coefficients of significant regressors. We are only interested in the coefficients yielded by the *jointly* estimated statistically significant regressors.

We let the values of k selected (e.g., per HKB or LW) to perform each of the 5 RRs be k_1, k_2, k_3, k_4 and k_5, respectively. Now we can suppose the statistically significant regressors determined by these values of k are Z_1, Z_2, Z_3, Z_5, Z_6 and Z_9. Then Y will be regressed against these 6 selected regressors using GRR with the following vectors of k values: $K \equiv (k_1 \quad k_1 \quad k_2 \; k_3 \quad k_3 \quad k_5)$. If $Z \equiv Z_1 \cup Z_2 \cup Z_3 \cup Z_5 \cup Z_6 \cup Z_9$, then the generalized RR estimate of the coefficients is:

$$\left(Z^T Z + D(K) \times I\right)^{-1} Z^T Y$$

where

$$D(K) = \begin{pmatrix} k_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & k_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & k_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & k_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & k_3 & 0 \\ 0 & 0 & 0 & 0 & 0 & k_5 \end{pmatrix} \text{ and } I \text{ is the } 6 \times 6 \text{ identity matrix}$$

The product of $D(K)$ and I produce the unequal diagonal increments for $Z^T Z$.

Analogous to Efroymson's [31] stepwise regressor selection procedure, intermediate RR estimated coefficients are ignored here. However, the statistically significant regressors and the corresponding values of k that identified them are retained. Then Y is re-estimated by RR using these retained values of k, as above, to determine the corresponding coefficients jointly.

2.3. Description of Simulation Design

An extensive simulation study is conducted by simulating several linear models by discreetly varying the following parameters: multicollinearity levels of X (which is varied from mild to extreme), the squared distance of b from the origin (which is varied from 10 to 600,000) and the *a priori* probability that an element of b is non-zero (which varied from 5% to 95%). Multicollinearity levels are varied by setting the trace (ω) of $Z^T Z$, where Z is a partition of X, to 100, 300, 1000, 4000 or an extreme value (e.g., several trillions). Statistical significance testing is done at an α of 15%. For each combination of simulation inputs (a.k.a. a scenario), 2000 outputs are generated, and there are 822 scenarios.

We borrow from the work of Gana [34], with a little simplification, the following notation:

$L_1^2(b_{EM} \mid X) \equiv$ the squared distance (length), from the true coefficients, of the estimated coefficients that are statistically significant (but not necessarily true) given X, under estimation method EM. This is a measure of coefficient precision. Note the connective "given" is symbolized by the stroke symbol (|).

$L_1^2(b_{EM}^{true} \mid X) \equiv$ the squared distance, from the true coefficients, of estimated coefficients that are true (and statistically significant) given X, under estimation method EM.

$EM_{found}(X_{EM}^{true} \mid X) \equiv$ the proportion of true regressors found by EM given X.

b_{EM}^{true} *is empty* $\equiv$ none of the coefficients in b_{EM} are true; that is all elements of b_{EM} are spurious but are statistically significant (i.e., they are "false positives", so to speak).

Following Hoerl, Schuenemeyer and Hoerl [35], we simulate linear models using an observable data matrix X_{base} as our "base" matrix. The matrix X_{base} was created as a modified combination of the 3 datasets presented in Hoerl, Schuenemeyer and Hoerl [35] and is presented as "Table A10" in Gana [34]. X_{base} has 33 regressors and 89 observations (and is shown in tab "X_base" in the Supplementary Materials Excel file "Supplementary_Material_file_mdpi_math.xlsx"). The X matrix used in this paper is the transpose of the base matrix, X_{base}^T. Thus, our chosen $X \equiv X_{base}^T$ matrix has 89 regressors and 33 observations.

Because X is not full-rank (it is an 89×33 matrix), we alter its eigenvalue spectrum by partitioning it so collinearity levels can be varied from low to high levels. A good measure of the multicollinearity level of a full-rank matrix Z, say, which is in correlation form, is the trace of $\left(Z^T Z\right)^{-1}$. We partition X into 3 mutually exclusive sub-matrices: $X_1^{1:30}$, $X_2^{31:60}$ and $X_3^{61:89}$, where a partition $X_r^{s:t}$ contains the columns s through t of X. Given a target multicollinearity level (trace) ω, the eigenvalue spectrum of a partition is altered as follows:

(i) We let λ_i be the ith eigenvalue of $\left(X_r^{s:t}\right)^T X_r^{s:t}$, where $\lambda_i \geq \lambda_{i+1}$ for all $i < t - s + 1 \equiv q$.

(ii) We let E be the $q \times q$ matrix whose columns are the eigenvectors corresponding to the eigenvalues.

(iii) Then $E^T\left(X_r^{s:t}\right)^T X_r^{s:t} E = D(\lambda)$, where $D(\lambda)$ is the diagonal matrix of the eigenvalues λ_i, stored in vector λ, and $E^T E = E E^T = I$, where I is a $q \times q$ identity matrix.

(iv) We choose a new vector of q arbitrary eigenvalues, θ, with θ_i denoting the ith eigenvalue entry in θ.

(v) We create $X_{r,1}^{s:t} \equiv X_r^{s:t}\, E\, D\left(\sqrt{\frac{\theta_i}{\lambda_i}}\right) E^T$ and transform $X_{r,1}^{s:t}$ to correlation form. We denote the transformed $X_{r,1}^{s:t}$ as $X_{r,2}^{s:t}$. We denote by $\varnothing_i$ the ith eigenvalue of $\left(X_{r,2}^{s:t}\right)^T X_{r,2}^{s:t}$. We calculate $\sum_1^q \frac{1}{\varnothing_i} = \Phi$.

(vi) We repeat steps "(iv)" and "(v)" (by trial and error) until $\Phi = \omega$.

The final selections of θ and the resultant values of $\varnothing_i$, by partitions and predetermined collinearity levels, are in the supplementary materials Excel file (see table "eigenvalues"). Each partition is altered to target the same value of ω, without loss of generality. That is, ω does not vary over partitions. However, there will be some variation in collinearity levels across the sub-matrices of X. Because the original eigenvalues (λ_i) indicate extremely high levels of collinearity by partition, this is considered to be the "extreme" multicollinearity case. We use the ratio of the trace of $\left(Z^T Z\right)^{-1}$ to the number of regressors (columns) in Z as an "invariant" measure of the collinearity of Z. For the sub-matrices of X to which RR is applied, these ratios are shown in the supplementary materials Excel file (see table "trace"). Generally, one can think of "mild" to "quite severe" levels of multicollinearity being present when this ratio varies from about 3 to 25, respectively.

The true coefficient vector, β (interchangeably b, as relevant), is simulated. Then ε is simulated from a normal distribution with a mean of zero and a variance of unity (without a loss of generality). Then the corresponding values of the regressand, Y (i.e., $X\beta + \varepsilon$), are

calculated. Such linear models are simulated by varying the multicollinearity levels of X, the squared distance of β from the origin (r^2) and the *a priori* probability (ρ) that an element of β is non-zero. The statistical significance (α) level used for testing the null hypothesis that an element of β is zero is set at 15%.

For each simulated Y, linear models are fit with X as the initial set of regressors using the EN and RR, respectively and fit-metrics compared. This is a comprehensive simulation covering a wide range of linear models and is executed via the following Steps:

(1) We pick a multicollinearity level (ω) from {100, 300, 1000, 4000, extreme}, where "extreme" denotes the original multicollinearity level of X.
(2) We use the multicollinearity level picked in Step "(1)" to create, as described previously (i.e., Steps "(i)" through "(vi)" above), the partitions: $X_{1,2}^{1:30}$, $X_{2,2}^{31:60}$ and $X_{3,2}^{61:89}$. We let $X_{\omega=\omega_0} \equiv X_{1,2}^{1:30} \cup X_{2,2}^{31:60} \cup X_{3,2}^{61:89}$, where ω_0 is the multicollinearity level picked in "(1)". Where the interpretation is clear, we will simply represent, for notational brevity, $X_{\omega=\omega_0}$ as X_ω.
(3) We pick a value of r^2 (i.e., $\beta^T\beta$) from {10, 25, 50, 100, 250, 500, 1000, 1500, 3000, 5000, 7500, 10,000, 15,000, 30,000, 60,000, 100,000, 150,000, 200,000, 300,000, 600,000}. That is, 20 possible values of r^2 are selected.
(4) We pick a value of ρ from {0.05, 0.20, 0.35, 0.50, 0.65, 0.80, 0.95}. For greater granularity, we include additional values of ρ from {0.10, 0.15, 0.25, 0.40, 0.45, 0.70} as necessary. These additional values of ρ are used sometimes, to approximately mark ρ-cutoffs where $L_1^2\left(b_{RR}^{true} \mid X_{RR}\right)$ starts being less than $L_1^2\left(b_{EN}^{true} \mid X\right)$.
(5) We generate a $p \times 1$ vector, v, of uniform random numbers in the interval [–1, 1].
(6) We generate a $p \times 1$ vector, c, of Bernoulli random variables (0 or 1) with $\Pr(c = 1) = \rho$.
(7) We generate an $n \times 1$ vector, ε, of independent, identically distributed normal (0, 1) random variables.
(8) We perform pairwise multiplication to create $a = v \times c$ until $a^T a \neq 0$.
(9) We generate $\beta = \left(\sqrt{\frac{r^2}{a^T a}}\right) a$.
(10) We generate $Y = X_\omega \beta + \varepsilon$.
(11) We partition X_ω into 9 mutually exclusive sub-matrices: $X_\omega^{1:10}$, $X_\omega^{11:20}$, $X_\omega^{21:30}$, $X_\omega^{31:40}$, $X_\omega^{41:50}$, $X_\omega^{51:60}$, $X_\omega^{61:70}$, $X_\omega^{71:80}$ and $X_\omega^{81:89}$. This ensures that, for our case, the OPR is at least 2.
(12) We estimate Y using RR with all of the regressors in $X_\omega^{s:t}$ under the HKB determined value of k. We use the modified t-statistic to select the RR-identified significant variables in $X_\omega^{s:t}$ and retain them in $X_{\omega,\ HKB}^{s:t}$, a non-null matrix. We let the value of k identifying the significant regressors be denoted by $k_{\omega,\ HKB}^{s:t}$ and the number of significant regressors identified be: $p_{\omega,\ HKB}^{s:t} > 0$.
(13) We collect the RR-identified significant regressors, over all partitions, under the HKB-determined value of k and concatenate them as $X_{\omega,HKB} \equiv \cup_{\forall\ s:t} X_{\omega,\ HKB}^{s:t}$. Here, "$\forall\ s:t$" includes only partitions for which at least one significant regressor is identified by RR. Where obvious, this interpretation of $\forall\ s:t$ is assumed in lieu of burdening the indexing by expanding the subscript to read as: $\forall\ s:t \ni p_{\omega,\ HKB}^{s:t} > 0$.
(14) We estimate Y using RR with all of the regressors in $X_\omega^{s:t}$ under the LW-determined value of k. We use the modified t-statistic to select the RR-identified significant variables in $X_\omega^{s:t}$ and retain them in $X_{\omega,\ LW}^{s:t}$, a non-null matrix. We let the value of k identifying the significant regressors be denoted by $k_{\omega,\ LW}^{s:t}$ and the number of significant regressors identified be: $p_{\omega,\ LW}^{s:t} > 0$.
(15) We collect the RR-identified significant regressors, over all partitions, under the LW value of k, and concatenate them as $X_{\omega,LW} \equiv \cup_{\forall\ s:t} X_{\omega,\ LW}^{s:t}$.
(16) We create $X_{RR} \equiv X_{\omega,\ HKB} \cup X_{\omega,\ LW}$.
(17) We create $K_{HKB} \equiv \vee_{\forall\ s:t} \vee_1^{p_{\omega,\ HKB}^{s:t}} k_{\omega,\ HKB}^{s:t}$ and $K_{LW} \equiv \vee_{\forall\ s:t} \vee_1^{p_{\omega,\ LW}^{s:t}} k_{\omega,\ LW}^{s:t}$.

(18) We estimate Y using generalized RR with the regressors in X_{RR} and $K = K_{HKB}$. We let the resultant estimated coefficient vector be denoted as: b_{HKB}.
(19) We calculate $L_1^2(b_{HKB} \mid X_{RR})$ and $L_1^2(b_{HKB}^{true} \mid X_{RR})$.
(20) We estimate Y using generalized RR with the regressors in X_{RR} and $K = K_{LW}$. We let the resultant estimated coefficient vector be denoted as: b_{LW}.
(21) We calculate $L_1^2(b_{LW} \mid X_{RR})$ and $L_1^2(b_{LW}^{true} \mid X_{RR})$.
(22) We calculate the proportion of true regressors in X_{RR} and denote it by $RR_{found}(X_{RR}^{true} \mid X_{RR})$.
(23) We estimate Y using the EN. We ese the Schwarz Bayesian criterion (SBC) [36] as the "stopping rule" with a maximum of 300 "steps" for EN calculations and $\delta_0 \equiv 10^{-4}$. The SBC is defined as $n \times log(SSE \div n) + p \times \log(n)$, where SSE denotes the sum of squared errors resulting from the fitted EN regression. The EN estimated β is stored in $p \times 1$ vector b_{EN}^{sig}, where the coefficients of all EN-selected regressors, assumed to be "significant", are in b_{EN}^{sig}.
(24) We calculate $L_1^2(b_{EN} \mid X)$, $L_1^2(b_{EN}^{true} \mid X)$ and $EN_{found}(X_{EN}^{true} \mid X)$.
(25) We repeat Steps 5 through 24, 2000 times.
(26) We calculate the hit rate: $\Pr\{L_1^2(b_{HKB} \mid X_{RR}) < L_1^2(b_{EN} \mid X)\}$. That is, we calculate the percentage of times this event occurs in 2000 trials (simulations) and interpret it as a probability.
(27) We calculate the hit rate: $\Pr\{L_1^2(b_{LW} \mid X_{RR}) < L_1^2(b_{EN} \mid X)\}$.
(28) We calculate the hit rate: $\Pr\{L_1^2(b_{HKB}^{true} \mid X_{RR}) < L_1^2(b_{EN}^{true} \mid X)\}$.
(29) We calculate the hit rate: $\Pr\{L_1^2(b_{LW}^{true} \mid X_{RR}) < L_1^2(b_{EN}^{true} \mid X)\}$.
(30) We calculate the miss rate: $\Pr\left\{RR_{found}(X_{RR}^{true} \mid X_{RR}) < EN_{found}(X_{EN}^{true} \mid X)\right\}$.
(31) We calculate the "empty" rates: $\Pr(X_{RR}^{true}\ is\ empty \mid X_{RR})$ and $\Pr(X_{EN}^{true}\ is\ empty \mid X)$.
(32) We redo Steps 5 through 31 for all of the combinations of the simulation starting conditions in Steps 1 through 4. That is, for each ω, the values of r^2 are crossed with the values of ρ.

The simulation done in this paper is executed in SAS [37]. The SAS code used for simulation is provided as supplementary materials and named "ElasticRR3F.sas".

3. Results

In real situations, the multicollinearity level of X would be knowable, *a priori*, but not necessarily the values of r^2 or ρ. So, a simple summarization of the simulation outputs would simply be to consider the summary statistics of these outputs across all of the simulation scenarios (i.e., r^2 and ρ). For example, several metrics can be summarized by considering the probabilities these metrics exceed 50%. Table 1 is a summary of the simulation results across *all* scenarios. The subscript RR for a metric in Table 1 means that the "best" value of that metric, yielded by either the HKB or LW value of k, is selected for computing it at the selected value of ω in the simulation. For example, $L_1^2(b_{RR} \mid X_{RR}) \equiv min\{\{L_1^2(b_{HKB} \mid X_{RR}),\ L_1^2(b_{LW} \mid X_{RR}) \mid \omega\}\}$.

Table 1. Metrics measured and summarized across all simulation scenarios.

Metric in Plain English (for Generalists) and Mathematized (for Specialists)	Measure *
Average probability that the squared length of the RR-estimated coefficient vector (from the true coefficient vector) is shorter than the corresponding one for the EN: $E\{Pr(0 < L_1^2(b_{RR} \mid X_{RR}) < L_1^2(b_{EN} \mid X))\}$	79%
Probability that the squared length of the RR-estimated coefficient vector is shorter than the corresponding one for the EN more frequently (viz. 50 + % of the time): $Pr\{Pr(0 < L_1^2(b_{RR} \mid X_{RR}) < L_1^2(b_{EN} \mid X)) > 50\%\}$	86%
Average probability that the squared length of the RR-estimated true coefficient vector is shorter than the corresponding one for the EN: $E\{Pr(0 < L_1^2(b_{RR}^{true} \mid X_{RR}) < L_1^2(b_{EN}^{true} \mid X))\}$	67%
Probability that the squared length of the RR-estimated true coefficient vector is shorter than the corresponding one for the EN more frequently: $Pr\{Pr(0 < L_1^2(b_{RR}^{true} \mid X_{RR}) < L_1^2(b_{EN}^{true} \mid X)) > 50\%\}$	74%
Average probability that RR finds more of the true regressors than does the EN: $E\left\{Pr\left(RR_{found}\left(X_{RR}^{true} \;\middle|\; X_{RR}\right) > EN_{found}(X_{EN}^{true} \mid X) > 0\right)\right\}$	89%
Probability that RR finds more of the true regressors than does the EN more frequently: $Pr\left\{Pr\left(RR_{found}\left(X_{RR}^{true} \;\middle|\; X_{RR}\right) > EN_{found}(X_{EN}^{true} \mid X) > 0\right) > 50\%\right\}$	99%
Average proportion of true regressors found by RR: $E\left\{RR_{found}(X_{RR}^{true} \mid X_{RR})\right\}$	65%
Average proportion of true regressors found by the EN: $E\left\{EN_{found}\left(X_{EN}^{true} \mid X\right)\right\}$	41%
Conditional probability that RR finds fewer of the true regressors than the EN more frequently, given that the squared length of the RR-estimated coefficient vector is shorter than the corresponding one for the EN less frequently (i.e., the RR downside to finding true regressors when RR coefficients are relatively imprecise is small): $\Pr[\Pr\left\{(RR_{found}\left(b_{RR}^{true} \;\middle|\; X_{RR}\right) < EN_{found}(b_{EN}^{true} \mid X)\right\} > 50\% \;\middle|\; \Pr(0 < L_1^2(b_{RR} \mid X_{RR}) < L_1^2(b_{EN} \mid X)) < 50\%]$	3%
Probability that RR finds none of the true regressors: $Pr(b_{RR}^{true}\ is\ empty)$	4%
Probability that the EN finds none of the true regressors: $Pr(b_{EN}^{true}\ is\ empty)$	25%
Probability that the proportion of times the squared length of the RR-estimated coefficient vector chosen with the HKB RR tuning parameter is shorter than the corresponding one chosen with the LW RR tuning parameter more frequently: $Pr\{Pr(0 < L_1^2(b_{HKB} \mid X_{RR}) < L_1^2(b_{LW} \mid X_{RR})) > 50\%\}$	11%
Probability that the proportion of times the squared length of the RR-estimated true coefficient vector using the HKB RR tuning parameter is shorter than the corresponding one using the LW RR tuning parameter more frequently: $Pr\{Pr(0 < L_1^2(b_{HKB}^{true} \mid X_{RR}) < L_1^2(b_{LW}^{true} \mid X_{RR})) > 50\%\}$	48%

* To the nearest integer. The operators Pr and *E* denote "probability" and "expectation" (a.k.a. "average"), respectively.

3.1. Results by Collinearity Level

In contrast to Table 1, wherein all scenarios are summarized, Table 2 below summarizes the results by ω.

The simulation results summarized in Table 2 lead us to the following conclusions:

1. RR is better than the EN at estimating true coefficients. Specifically, the double probability,

$$\Pr\left\{\Pr\left(L_1^2(b_{RR}^{true} \mid X_{RR}) < L_1^2(b_{EN}^{true} \mid X) \;\middle|\; r^2, \rho\right) > 50\%\right\}$$

wherein the subscript *RR* is either *HKB* or *LW*, is relatively high. Specifically, in rows 3 and 4 of Table 2, the average of this double probability across the 5 multicollinearity levels (ω) can be computed as (73.6 + 73.1 + 70.6 + 65.0 + 89.0) ÷ 5 or about 74%, wherein the first three numbers in the numerator correspond to the subscript setting of *b*: $RR \equiv HKB$ and the remaining two numbers correspond to the subscript setting (in the double probability statement) of *b*: $RR \equiv LW$. That is, when the multicollinearity level is low to severe, RR with the HKB values of *k* is better than the EN at estimating true coefficients, and, for very severe to extreme collinearity levels, RR with the LW values of *k* is better than the EN at estimating the true coefficients. For the sake of

notational simplicity, where obvious, we will leave out terms following " | " in double probability statements.

2. RR is better than the EN at jointly estimating both true and spurious, but "statistically significant", coefficients. In particular, the double probability

$$Pr\Big\{\ Pr\ \Big(L_1^2(b_{RR} \mid X_{RR}) < L_1^2(b_{EN} \mid X) \ \Big|\ r^2,\ \rho\Big) > 50\%\Big\}$$

is much higher than 50%. Specifically, the average probability across ω in rows 1 and 2 of Table 2 is (82.6 + 84.8 + 87.1 + 84.7 + 93.3) ÷ 5 or about 86%, wherein the first three numbers in the numerator correspond to the subscript setting of b: $RR \equiv HKB$ and the remaining two numbers correspond to the subscript setting of b: $RR \equiv LW$. That is, when the multicollinearity level is low to severe, RR with the HKB values of k is better than the EN at estimating significant coefficients, and, for very severe to extreme collinearity levels, RR with the LW values of k is better than the EN at estimating the significant coefficients.

3. RR is better than the EN at finding true regressors. In particular, the double probability

$$\Pr\Big\{Pr\Big(RR_{found}\big(X_{RR}^{true} \mid X_{RR}\big) > EN_{found}\big(X_{EN}^{true} \mid X\big) \ \Big|\ r^2,\ \rho\Big) > 50\%\Big\}$$

takes on values that are much higher that 50% in row 5 of Table 2.

4. RR is better than the EN at finding at least one true regressor. The probabilities, by ω, that b_{RR}^{true} and b_{EN}^{true} are empty, in rows 8 and 9 of Table 2, respectively, indicate this. As can be noted, the probability that b_{EN}^{true} is empty is generally much higher than the probability that b_{RR}^{true} is empty. In particular, the average probability that b_{RR}^{true} is empty across the entire simulation space is (2.58 + 1.75 + 2.94 + 4.91 + 7.98) ÷ 5 or about 4%, and the corresponding probability that b_{EN}^{true} is empty is about 25%, about six multiples of 4%.

5. The conditional probability that b_{RR}^{true} is empty, given r^2, ρ, and a trial (simulation) where X_{RR}^{true} turns out to be empty, is denoted by:

$$\Pr\Big\{\ b_{RR}^{true}\ is\ empty\ \Big|\ X_{RR},\ \ r^2,\ \ \rho\ and\ \exists(empty\ X_{RR}^{true})\ \Big\}$$

where the "existential quantifier" $(\exists\ a,\ b)$ means that "there exists at least one a and b", an operator borrowed from Whitehead and Russell [29]. The corresponding conditional probability for $b_{EN}^{sig\ \&\ true}$ is denoted similarly. The expected values (averages) of these two conditional probabilities are shown in rows 10 and 11 of Table 2, respectively. If we take the averages of these two expectations across ω, respectively, we get: (0.54 + 1.08 + 1.22 + 0.81 + 0.33) ÷ 5 or 3.98 and (1.60 + 1.53 + 1.26 + 2.39 + 9.03) ÷ 5 or 15.81. Furthermore, 15.81 is nearly four multiples of 3.98.

6. For given combinations of r^2 and ρ, the average value of the proportion of true regressors found is calculated in the simulation, by ω. For RR and the EN, these are $E\Big(RR_{found}\big(X_{RR}^{true} \mid X_{RR}\big) \ \Big|\ r^2,\ \rho\Big)$ and $E\Big(EN_{found}\big(X_{EN}^{true} \mid X\big) \ \Big|\ r^2,\ \rho\Big)$, respectively. The averages (and standard deviations) of these two expectations across r^2 and ρ, are shown, by ω, in rows 12 and 13 of Table 2, respectively. They also indicate that RR is better than the EN at finding true regressors and that RR does so with lower volatility.

7. Row 6 of Table 2 indicates that the LW value of k is better than the HKB value of k when estimating the coefficient vector that includes both the true and spurious coefficients. Row 7 of Table 2 indicates that, for low levels of collinearity, the HKB value of k is better than the LW value of k when estimating the true coefficients.

Table 2. RR vs. EN over all combinations of r^2 and ρ by ω, at $\alpha = 15\%$.

Row	Metric (Measured as Percent)	ω [1]				
		100	**300**	**1000**	**4000**	**Extreme**
1	$\Pr\{\Pr(\ L_1^2(b_{HKB} \mid X_{RR}) < L_1^2(b_{EN} \mid X) \mid r^2, \rho\ ) > 50\%\}$	82.6 (79.3)	84.8 (78.0)	87.1 (74.6)	73.6 (65.1)	0.0 (0.99)
2	$\Pr\{\Pr(\ L_1^2(b_{LW} \mid X_{RR}) < L_1^2(b_{EN} \mid X) \mid r^2, \rho\ ) > 50\%\}$	82.6 (79.5)	84.8 (78.8)	87.1 (78.4)	84.7 (76.4)	93.3 (83.7)
3	$\Pr\{\ \Pr\ (L_1^2(b_{HKB}^{true} \mid X_{RR}) < L_1^2(b_{EN}^{true} \mid X) \mid r^2,\ \rho) > 50\%\}$	73.6 (67.0)	73.1 (66.1)	70.6 (64.4)	65.0 (58.7)	0.0 (5.5)
4	$\Pr\{\ \Pr\ (L_1^2(b_{LW}^{true} \mid X_{RR}) < L_1^2(b_{EN}^{true} \mid X) \mid r^2,\ \rho) > 50\%\}$	71.0 (66.4)	68.4 (64.8)	67.1 (64.2)	65.0 (62.0)	89.0 (76.8)
5	$\Pr\{\Pr(RR_{found}(X_{RR}^{true} \mid X_{RR}) > EN_{found}(X_{EN}^{true} \mid X) \mid r^2,\ \rho) > 50\%\}$	96.8 (75.2)	99.4 (84.8)	100 (90.1)	100 (94.3)	100 (99.8)
6	$\Pr\{\Pr(\ L_1^2(b_{HKB} \mid X_{RR}) < L_1^2(b_{LW} \mid X_{RR}) \mid r^2, \rho\ ) > 50\%\}$	45.8 (43.1)	7.0 (36.8)	0.0 (26.8)	0.0 (17.7)	0.0 (0.29)
7	$\Pr\{\Pr(\ L_1^2(b_{HKB}^{true} \mid X_{RR}) < L_1^2(b_{LW}^{true} \mid X_{RR}) \mid r^2, \rho\ ) > 50\%\}$	78.7 (52.1)	69.0 (50.3)	51.2 (43.4)	41.7 (38.1)	0.0 (5.0)
8	$Pr(b_{RR}^{true}\ is\ empty \mid X_{RR},\ r^2,\ \rho)$	2.58	1.75	2.94	4.91	7.98
9	$Pr(b_{EN}^{true}\ is\ empty \mid X,\ r^2,\ \rho)$	4.52	8.77	24.11	28.83	58.28
10	$\mathrm{E}(Pr\ \{\ b_{RR}^{true}\ is\ empty \mid X_{RR},\ r^2,\ \rho\ and\ \exists(empty\ X_{RR}^{true})\ \})$ [2]	0.54 [0.9]	1.08 [1.3]	1.22 [1.6]	0.81 [1.3]	0.33 [0.7]
11	$\mathrm{E}(Pr\ \{\ b_{EN}^{true}\ is\ empty \mid X,\ r^2,\ \rho\ and\ \exists(empty\ X_{EN}^{true})\ \})$ [2]	1.60 [2.6]	1.53 [3.1]	1.26 [3.2]	2.39 [4.3]	9.03 [14.5]
12	$\mathrm{E}\{\ E(RR_{found}(X_{RR}^{true} \mid X_{RR}) \mid r^2,\ \rho)\ \}$ [2]	53.7 [12.1]	57.0 [9.9]	61.0 [8.7]	66.9 [6.6]	85.8 [5.8]
13	$\mathrm{E}\{\ E(EN_{found}(X_{EN}^{true} \mid X) \mid r^2,\ \rho)\ \}$ [2]	49.0 [16.8]	47.5 [15.1]	46.3 [13.8]	44.4 [12.6]	18.4 [11.4]
14	$\Pr\{Pr(\ L_1^2(b_{LW} \mid X_{RR}) < L_1^2(b_{EN} \mid X) \mid r^2, \rho\ )$ $\geq Pr(\ L_1^2(b_{HKB} \mid X_{RR}) < L_1^2(b_{EN} \mid X) \mid r^2, \rho\ )\ \}$	73.6	84.8	91.8	96.3	100
15	$\Pr\{Pr(\ L_1^2(b_{LW}^{true} \mid X_{RR}) < L_1^2(b_{EN}^{true} \mid X) \mid r^2, \rho\ )$ $\geq Pr(\ L_1^2(b_{HKB}^{true} \mid X_{RR}) < L_1^2(b_{EN}^{true} \mid X) \mid r^2, \rho\ )\}$	40.0	39.2	51.2	58.9	98.2
16	$\Pr[\ \Pr\{(RR_{found}(b_{RR}^{true} \mid X_{RR}) < EN_{found}(b_{EN}^{true} \mid X)\} > 50\%\ \mid \Pr(hit_{\Delta}) < 50\%\ ]$ [3]	7.41	3.85	0	0	0

[1] Average values of the probability within the curly braces, i.e., $\{Pr\ (\cdot)\}$, in the first column labeled "Metric" are within parentheses. [2] The number within square brackets ([·]) is the standard deviation (as %) of the corresponding metric. [3] $\Delta \equiv HKB\ for\ \omega \cong 100\ and\ \Delta \equiv LW\ for\ the\ remaining\ values\ of\ \omega$.

3.2. Recognizing Failure Scenarios

In order not to clutter the body of this paper, detailed simulation outputs are saved in the supplementary materials Excel file with tab-naming convention ω-Ti, wherein, for each value of ω, i takes the values 1, 2 and 3. We will refer to those tabs in the supplementary materials Excel file as necessary to highlight simulation scenarios wherein the EN does better than RR for particular values of ρ and r^2.

For example, when ω = 4000 (see table "4000-T1" in the supplementary materials Excel file):

$\Pr(L_1^2(b_{LW}^{true} \mid X_{RR}) < L_1^2(b_{EN}^{true} \mid X) \mid r^2 = 300,000,\ \rho \leq 55\%) < 50\%$ and
$\Pr(L_1^2(b_{LW} \mid X_{RR}) < L_1^2(b_{EN} \mid X) \mid r^2 = 300,000,\ \rho \leq 40\%) < 50\%$; but
$Pr\left(RR_{found}(X_{RR}^{true} \mid X_{RR}) > EN_{found}(X_{EN}^{true} \mid X) \mid r^2 = 300,000,\ \rho\right) > 50\%$

That is, EN does better than RR in terms of $L_1^2(\cdot)$ but not in terms of true regressors found.

More analysis of the above examples or "failure scenarios" is done next. To do so, we define, for simplicity, the following symbols:

$$I(\psi) \equiv \begin{cases} 1, & Proposition\ or\ Propositional\ function\ \psi\ is\ true \\ 0, & \psi\ is\ false \end{cases}$$

$$hit_{\Delta}^{\Omega} \equiv 0 < L_1^2\left(b_{\Delta}^{\Omega} \;\middle|\; X_{RR}\right) < L_1^2\left(b_{EN}^{\Omega} \;\middle|\; X\right)$$

We eyeball simulation outputs and note possible patterns. For example, if we observe the values of $\Pr\{L_1^2(b_{HKB} \mid X_{RR}) < L_1^2(b_{EN} \mid X)\}$ (see column labeled "a" in tab 100-T1 in the supplementary materials Excel file), we see that the corresponding hit rate (i.e., 0.307 < 50%) first fails when $r^2 = 500$ and $\rho = 0.05$. As we go down this column of probabilities, we notice that this hit rate tends to fail as r^2 is "large" and ρ is "small" (i.e., tends to be sparse). Furthermore, the tendency for hit rates to fail as r^2 increases are dampened as ρ increases.

3.3. Regression Modeling of Simulation Outputs to Understand Patterns

One way to conceptualize the abovementioned patterns observed in hit rate variations in the simulation is to model how simulation output is generated by its inputs, under the simplifying assumption that the inputs are fixed. We do this using LS linear regression and RR as shown in Table 3.

Based on eyeballing the output, we select simulation variables, with r^2 in logarithmic scale, as regressors to see how they impact the evolution of the simulation output (i.e., the regressand). First, a LS regression is fit to the chosen dependent variable shown in Table 3. Because variance inflation factors (VIFs) exceeding 10 may be problematic (see Marquardt and Snee [38]), the hypothesized regression equation is also re-estimated by RR, and the resultant coefficient estimates are shown following the slash (/) after the corresponding LS-estimated coefficients. In Table 3, the RR tuning parameter, k, is chosen graphically, and conservatively, by observing where the ridge traces "stabilize" (see Hoerl and Kennard [39]), indicating the near orthogonality of the regressors. This happens for the small value of k of 0.02 (bias in RR is proportional to k^2). Because observing the broad outlines of patterns among simulation inputs and outputs is the focus of this exercise, algorithmic selections of k are not pursued; nor are the RR ANOVA tables (see Hoerl and Kennard [40]) computed. The LS absolute t, F, R-squared and root mean squared error (RMSE) values are computed. White's [41] t values are also examined (but not reported) to confirm that no LS t values turn out to be statistically insignificant. Note that LS t values are relevant in RR as well (see Obenchain [42]). The RR RMSE value is also computed, but without correcting for the RR degrees of freedom in the RR ANOVA table.

Furthermore, for conceptualizing simulation patterns, RR is used to assess the sensitivity of the coefficients to small perturbations, because the collinearity among the simulation variables is considered "natural". That is, if this simulation is repeated, the expectation is that the present correlation structure among its inputs used as regressors will persist. Another way to think about the perturbations is that they can produce more accurate matrix inverses when the matrices to be inverted are close to being singular. In particular, for an ill-conditioned matrix A:

$$(\mathrm{A} + k\mathrm{I})^{-1} = \mathrm{A}^{-1} - k^2 \times \mathrm{A}^{-1}\left(\mathrm{I} + \mathrm{A}^{-1} \div k\right)^{-1}\mathrm{A}$$

which for "small" k approximates to A^{-1}, if we ignore second-order terms (see Piegorsch and Casella [10]). RR is a good way to invert matrices like A using "small" values of k while observing where coefficients stabilize as the matrix diagonal increment (k) starts increasing from zero. For our purpose, we will consider values of $k \leq 0.1$ to be "small". Furthermore, whether or not multicollinearity exists, some shrinkage of LS estimated coefficients may be desirable (see Brook and Moore [13]).

For our purpose, these regression approximations are quite reasonable in revealing broad patterns in simulation outputs and are better at revealing these patterns than simple two-dimensional graphs of simulation outputs vs. inputs. As can be noted from the second-to-last row of Table 3, the conditional probability of RR failing to find more of the true regressors than EN, given that the RR hit rate is less than 50%, is only about 7%.

Table 3. Approximating patterns in simulated hit rates, $\omega \cong 100$.

Regressand (Y)					
$Y \equiv \mathbf{Pr}\left(L_1^2(b_{HKB} \mid X_{\mathbf{RR}}) < L_1^2(b_{EN} \mid X)\right)$			$Y \equiv \mathbf{Pr}\left(L_1^2(b_{HKB}^{true} \mid X_{\mathbf{RR}}) < L_1^2(b_{EN}^{true} \mid X)\right)$		
Simulation variables used as regressors	LS/RR coefficients (LS VIF)	LS $\lvert t \rvert$ value	Simulation variables used as regressors	LS/RR coefficients (LS VIF)	LS $\lvert t \rvert$ value
Intercept	1.3924/1.32	38.39	Intercept	1.1454/1.12	48.56
$\log(r^2)$	$-0.0296/-0.03$ (8.7)	5.50	$\log(r^2)$	$-0.0394/-0.04$ (6.6)	8.40
ρ	$-0.3842/-0.25$ (5.3)	8.27	ρ	Not significant	
$I(r^2 > 250)$	$-0.4882/-0.42$ (8.9)	12.08	$I(r^2 > 50)$	$-0.3338/-0.35$ (5.0)	8.92
$\rho \times I(r^2 > 250)$	0.5043/0.37 (35.4)	4.72	$\rho \times I(r^2 > 50)$	0.2778/0.33 (21.6)	3.24
$\log(r^2) \times \rho \times I(r^2 > 250)$	0.0331/0.035 (32.7)	3.35	$\log(r^2) \times \rho \times I(r^2 > 50)$	0.0348/0.03 (21.5)	4.14
$I(\rho \leq 20\%)$	$-0.5394/-0.47$ (5.3)	17.79	$I(5\% \leq \rho \leq 35\%)$	$-0.4787/-0.44$ (5.6)	17.04
$\rho \times I(\rho \leq 20\%)$	1.3230/1.05 (3.2)	8.48	$\rho \times I(5\% \leq \rho \leq 35\%)$	1.0866/0.95 (3.0)	13.42
Sample size	155			155	
LS F value	309.90			474.59	
LS/RR RMSE	0.0727/0.0756			0.0727/0.0736	
LS R-squared	93.7%			95.1%	
Conditional probability that RR finds fewer of the true regressors than EN more frequently, given the RR-estimated coefficients are less precise than those of EN more frequently *	7.41%			NA (not applicable)	
Selected k using ridge traces	0.02			0.02	
Average Y (via double integration) in the rectangular region bounded by $\rho_L = 0.05$, $\rho_U = 0.20$, $r_L^2 = 10$ and $r_U^2 = 250$, and using the LS/RR regression coefficients	83.3%/84.7%			Not computed	
Average Y (via double integration) in the rectangular region bounded by $\rho_L = 0.20$, $\rho_U = 0.95$, $r_L^2 = 250$ and $r_U^2 = 600,000$, using the LS/RR regression coefficients	84.3%/81.0%			Not computed	

* $\Pr\left[\Pr\left\{(RR_{found}(b_{RR}^{true} \mid X_{RR}) < EN_{found}(b_{EN}^{true} \mid X)\right\} > 50\% \mid \Pr(hit_{HKB}) < 50\%\right]$.

One can use the equation in Table 3 to visualize movements in hit rates for different r^2 and ρ combinations. Furthermore, following Fubini [43,44], we can use an iterated integral approach to calculate the average hit rate in Table 3. For example, if the regression equation for the regressand $\Pr(hit_{HKB})$ is denoted by $f(\rho, r^2)$, say, then:

$$\begin{aligned} \textit{Average value of } \Pr(hit_{HKB}) &\equiv \overline{Pr}(hit_{HKB}) \\ &\cong \frac{1}{(\rho_U-\rho_L)\times\left(r_U^2-r_L^2\right)} \times \int_{r_L^2}^{r_U^2}\int_{\rho_L}^{\rho_U} f(\rho, r^2)\, d\rho\, dr^2 \end{aligned}$$

where ρ_L and ρ_U are the lower and upper bounds imposed on ρ, respectively; r_L^2 and r_U^2 are the lower and upper bounds imposed on r^2. For $\rho_L = 0.05$, $\rho_U = 0.20$, $r_L^2 = 10$ and $r_U^2 = 250$ and using the LS (RR)-estimated coefficients for $f(\rho, r^2)$, the value of $\overline{\Pr}(hit_{HKB})$ works out to about 83.3% (84.7%); and, for $\rho_L = 0.20$, $\rho_U = 0.95$, $r_L^2 = 250$ and $r_U^2 = 600,000$ and using the LS (RR)-estimated coefficients, the value of $\overline{Pr}(hit_{HKB})$ works out to about 84.3% (81.0%). These double integral values are consistent with the number in parentheses (79.3%) in row 1 of Table 2 for $\omega = 100$.

$\Pr(hit_{HKB})$ is a probability but has been modeled herein as a linear combination of simulation inputs. A more precise approach that constrains the predictions to naturally lie in the 0–1 range is to model the log-odds, $log\left(\frac{\Pr(hit_{HKB})}{1-\Pr(hit_{HKB})}\right)$, as a linear combination of simulation inputs and estimate $\overline{Pr}(hit_{HKB})$. As an example, the linear regression model for $\Pr(hit_{HKB})$ in log-odds form (for $\omega = 100$) was estimated. Because some LS VIFs are high (e.g., 198), we take the RR-estimated coefficients to do the double integration with $\rho_L = 0.20$, $\rho_U = 0.95$, $r_L^2 = 250$ and $r_U^2 = 600,000$. Because double integration for this case becomes a bit more computationally intensive, we use numerical integration per Shampine [45,46] in Matlab [47] and find that $\overline{Pr}(hit_{HKB})$, for this region, works out to about 81.6%. That is, the choice of regressand transformation has little impact on the value of the double integral. Similarly, the true regressor miss rate of RR relative to EN is modeled. The Matlab code used for double integration is in the supplementary materials Excel file (see table "matlab_code").

An example is shown in Table 4. The double integral computations for miss rates are shown in the last two rows of Table 4. For other values of ω, the regressions hold and are included in the supplementary materials (see table "output–input regressions" in the supplementary materials Excel file). When ω is 100 or 300, the HKB value of k is used for regressing hit rates associated with the coefficients that are both significant and true, because row 15 of Table 2 indicates that the HKB-prescribed k has an advantage in this regard.

3.4. Time-Series Modeling of Simulation Output to Understand Patterns

One can go deeper into the simulation output by artificially viewing it as "ordered" output. That is, we can think of each row of simulation output as "naturally" followed by the next row of simulation output. This makes the simulation output an artificial "time series", and one can then explore the memory and persistence in simulation outputs as they were generated.

As an example, we picked $\Pr\left\{RR_{found}\left(b_{RR}^{true} \,\middle|\, X_{RR}\right) < EN_{found}\left(b_{EN}^{true} \mid X\right)\right\}$ as the regressand and modeled it. For modeling these miss rates, we note that the application of Durbin's [48–50] and Vinod's [50] tests for serial correlation among the errors yields significant serial correlations for lags 1, 2, 7 and 8. We use the unconditional least squares (ULS) method of Spitzer [51] to re-estimate the model in order to account for these correlations by retaining the significant ones in the model. Lags 1, 2 and 8 are significant in the model (lag 7 drops off). No unit roots, per Elliott et al. [52], for the ULS-estimated equation are detected. The inputs used as regressors continue to be statistically significant, and, as expected, the regression R^2 value increases to about 96%. This example is included as supplementary materials (see the table labeled "Approximating patterns in simulated miss rates, $\omega \cong 300$" in the tab "output–input regressions" of the supplementary materials Excel file). Fitting such "time-series" models for the other regressions, on hit rates and miss rates, is not pursued at this time, but may be an interesting topic for future research on assessing the relationships between simulation inputs and outputs.

Table 4. Approximating patterns in simulated miss rates, $\omega \cong 100$.

$Y \equiv \Pr\{RR_{found}(b_{RR}^{true} \mid X_{RR}) < EN_{found}(b_{EN}^{true} \mid X)\}$				
	Regressand Is Y		**Regressand Is** $log\left(\frac{Y}{1-Y}\right)$	
Simulation variables used as regressors	LS/RR coefficients (LS VIF)	LS $\lvert t \rvert$ value	LS/RR coefficients (LS VIF)	LS $\lvert t \rvert$ value
Intercept	0.2332/0.2378	11.39	−1.061/−1.074	10.93
$\log(r^2)$	0.0068/0.0051 (2.3)	4.07	0.0279/0.021 (2.3)	3.51
ρ	−0.1981/−0.1719 (2.6)	10.23	−1.259/−1.097 (2.6)	13.71
$I(r^2 < 250)$	−0.1587/−0.1067 (4.8)	8.12	−0.8406/−0.57 (4.8)	9.07
$\rho \times I(r^2 < 250)$	1.2296/0.6209 (18.6)	18.72	5.8716/3.015 (18.6)	18.85
$\log(r^2) \times \rho \times I(r^2 < 250)$	−0.1918/−0.0532 (16.0)	11.41	−0.8453/−0.21 (16.0)	10.61
$I(\rho \leq 20\%)$	0.1205/0.1139 (2.4)	9.91	0.5044/0.49 (2.4)	8.75
$\rho \times I(\rho \leq 20\%)$	Not statistically significant at $\alpha = 5\%$			
Sample size	155		155	
LS F value	160.99		192.50	
LS/RR RMSE	0.0438/0.0551		0.2076/0.2606	
LS R-squared	86.7%		88.6%	
Selected k using ridge traces	0.04		0.04	
Average Y (via double integration) in the rectangular region bounded by $\rho_L = 0.05$, $\rho_U = 0.20$, $r_L^2 = 10$, and $r_U^2 = 250$, and using the LS/RR regression coefficients	Not computed		23.5%/28.2%	
Average Y (via double integration) in the rectangular region boundedby $\rho_L = 0.20$, $\rho_U = 0.95$, $r_L^2 = 250$, and $r_U^2 = 600,000$, and using the LS/RR regression coefficients	Not computed		19.5%/18.1%	

3.5. Examining Failure Scenarios for RR Miss Rates

Over all simulation scenarios, the observed probability that RR finds fewer of the true regressors than does the EN is greater than 50% is only 0.73%. This happens for six observations among the 822 miss rates output by the simulation. These six observations are shown in Table 5 and indicate that this event tends to occur when the number of true regressors (i.e., p^{true}) is "small" (i.e., $\rho = 5\%$) and r^2 is "large" or when ρ is "large" but r^2 is "small". Both sets of failures occur when multicollinearity is "mild" (i.e., $\omega \leq 300$). However, for the failures for which $\rho \geq 65\%$, the probabilities of all hit rates, hit_{RR} and hit_{RR}^{true}, where RR is HKB or LW, are nearly 100% each. When $\rho < \frac{n}{p} = \frac{33}{89} \cong 37.1\%$, p^{true} will tend to be less than n. In other words, the problem in nature reduces to a classical regression problem when the number of regressors is less than the number of observations; there is no need for the EN. However, we have to invoke the EN because p^{true} is unobservable, and, thus, we are forced to cast "the net" wide. Therefore, it is of interest to compute the probability (in the simulation) of the number of RR-selected regressors, p_{RR}, being less

than n and how miss rates conditioned on p_{RR} look. This will give us insight into how well RR recognizes (or fails to recognize) these scenarios. These probabilities are displayed, for these scenarios, in Table 5.

Table 5. Input combinations for which $Pr\{RR_{found}(b_{RR}^{true} \mid X_{RR}) < EN_{found}(b_{EN}^{true} \mid X)\} \geq 50\%$.

ρ	r^2	ω	$Pr(\, p_{\mathbf{RR}}<n \mid \rho, r^2, \omega)$	$Pr\left(\mathbf{RR}_{found}(b_{\mathbf{RR}}^{true} \mid X_{\mathbf{RR}}) \middle\langle EN_{found}(b_{EN}^{true} \mid X) \;\middle|\; \psi\right)$	
				$\psi \equiv \mathbf{Pr}(\, p_{\mathbf{RR}}<n\,)$	$\psi \equiv \mathbf{Pr}(\, p_{\mathbf{RR}}\geq n\,)$
5%	200,000	100	47.1%	54.7%	49.0%
5%	600,000	100	47.6%	52.8%	46.8%
5%	600,000	300	16.9%	55.8%	52.1%
65%	10	100	57.5%	76.7%	20.0%
80%	10	100	56.5%	82.0%	17.6%
95%	10	100	56.7%	90.5%	10.5%

Furthermore, RR miss rates (i.e., $\Pr\{RR_{found}(b_{RR}^{true} \mid X_{RR}) < EN_{found}(b_{EN}^{true} \mid X)\}$) for small ρ have "stabilized" at $r^2 = 600,000$, as shown in Table 6, wherein miss rates are calculated by fixing ρ and ω at 5% and 100, respectively. Table 6 indicates that, as r^2 increases, miss rates tend to be less than 55%.

Table 6. RR miss rate as r^2 increases with ρ fixed at 5% and ω fixed at 100.

r^2	Miss Rate	r^2	Miss Rate	r^2	Miss Rate	r^2	Miss Rate
6×10^6	0.5030	6×10^{18}	0.5330	6×10^{30}	0.5235	6×10^{60}	0.5445
6×10^9	0.5350	6×10^{21}	0.5415	6×10^{39}	0.5430	6×10^{69}	0.5100
6×10^{12}	0.5225	6×10^{24}	0.5260	6×10^{45}	0.5310	6×10^{75}	0.5400
6×10^{15}	0.5185	6×10^{27}	0.5265	6×10^{54}	0.5245	6×10^{99}	0.5175

The average miss rate in Table 6 is about 53%, with a standard deviation of about 1.2%, giving a signal-to-noise (SNR) ratio (i.e., $\frac{\mu}{\sigma}$) of about 44, which more than satisfies the Rose criterion [53,54] for large r^2 that " ... *to reduce the number of false alarms to below unity, we will need ... a signal whose amplitude is 4–5 times larger than the RMS noise ...* ".

3.6. Examining Failure Scenarios for RR Hit Rates

Over all simulation scenarios, $Pr\{Pr(hit_{LW}) \leq 50\% \mid r^2, \rho, \omega\}$ is 13.5%. That is, 111 of the 822 hit rates are less than or equal to 50%. Of these 111 observed probabilities, 105 (i.e., nearly 95%) are associated with $\rho \leq 20\%$, five are associated with $\rho = 35\%$ and the remaining one with $\rho = 40\%$.

Over all simulation scenarios, $Pr\{Pr(hit_{RR}^{true}) \leq 50\% \mid r^2, \rho, \omega\}$ is about 26%. That is, 212 of the 822 hit rates are less than or equal to 50%; here "RR" means that HKB k values were used when $\omega \leq 1000$ and LW k values otherwise, for computing the aforementioned double probability over the simulation inputs. Of these 212 observed probabilities, 201 (i.e., nearly 95%) are associated with $\rho \leq 40\%$; one is associated with $\rho = 45\%$; eight with $\rho = 50\%$ and the remaining two with $\rho = 55\%$. Furthermore, if $Pr(\, hit_{LW}) \leq 50\%$, then $\Pr(hit_{RR}^{true}) \leq 50\%$ as well. However, if $\Pr(hit_{LW}) > 50\%$, then $\Pr(hit_{RR}^{true}) \leq 50\%$ about 14.21% (i.e., $\frac{101}{822-111}$) of the time; about 89% of the time this happens, it happens for $\rho \leq 40\%$. These results are also shown as supplementary materials (see table "contingency tables" in the supplementary materials Excel file).

3.7. Examining Simulation Stability as r^2 Increases

Simulation outputs are stable as r^2 increases. This is illustrated for the case where $\omega = 100$, and the SNRs for the corresponding simulation outputs for 5 metrics, as r^2 becomes large, are shown in Table 7 (simulation outputs are in tab "rsq large for omega 100" in the Supplementary Materials Excel file). Ten values are assumed by r^2: 6×10^u, where $u \in \{6, 9, 12, 24, 36, 48, 60, 72, 84, 99\}$. There are only two instances in Table 7, as expected, wherein the Rose criterion is violated; in all other cases, the SNRs are much larger than 5. In these two instances, for $\rho = 5\%$, $\Pr(\ hit_{HKB})$ and $\Pr(hit^{true}_{HKB})$, each approach about 1% as r^2 approaches 6×10^{99}.

Table 7. SNRs by ρ over large r^2 values, where $\omega = 100$.

Metric	ρ							
	5%	**20%**	**35%**	**40%**	**50%**	**65%**	**80%**	**95%**
$\Pr(L_1^2(b_{HKB} \mid X_{RR}) < L_1^2(b_{EN} \mid X))$	1.22	42.67	60.67	78.13	140.88	297.41	188.19	318.21
$\Pr(L_1^2(b^{true}_{HKB} \mid X_{RR}) < L_1^2(b^{true}_{EN} \mid X))$	0.99	32.29	30.43	51.47	72.81	108.08	176.82	260.40
$E\left[EN_{found}(b^{true}_{EN} \mid X)\right]$	490.62	354.04	330.69	247.00	544.53	661.89	512.00	776.72
$E\left[RR_{found}(b^{true}_{RR} \mid X_{RR})\right]$	320.83	378.64	244.60	284.69	387.35	269.77	271.48	333.68
$Pr(\ p_{RR} < n \mid \rho,\ r^2,\ \omega = 100)$	56.44	23.37	22.19	20.75	28.01	12.83	23.43	17.14

Finally, to confirm that the simulation itself is stable, we re-ran the entire simulation for the 8 values of ρ, 20 values of r^2 and the 5 values of ω, at an α of 15%. The resultant $8 \times 20 \times 5$ or 800 rows of simulation outputs are shown in the supplementary materials Excel file (see table "alpha15pct"). The conclusions remain consistent, demonstrating the superiority of RR over the EN.

3.8. Examining Scenarios by Setting α Less than 15%

In our simulations, we have set α at 15%. There are two reasons we set α at 15%. First, it has been shown by Bendel and Afifi [55] that an α level of 15% to 25% is generally appropriate for variable subset selection when $p < n$, with an α level of 15% being "superior overall". Second, Gana's [34] recent work showed that setting α at 15% is more superior overall than setting it at 10%, when comparing RR and the Lasso. Notwithstanding these reasons, we redid some simulations by setting α at 5% and 10%. Because this is an illustration, we did these simulations for certain, but not all, choices of ω and r^2.

For α at 5%, when $\omega \in \{100, 1000, 4000\}$, all 20 values of r^2 are used; when ω is extreme, all values of r^2 except the "small" ones, 10, 25, 50 and 100, are used. For α at 10%, $\omega \in \{300, \text{extreme}\}$ and all 20 values of r^2 are used when ω is 300; when ω is extreme, only the smallest value of r^2, 10, is dropped. For both of these α settings, all eight values of $\rho \in \{0.05, 0.20, 0.35, 0.40, 0.50, 0.65, 0.80. 0.95\}$ are used. Simulation outputs are included as supplementary materials (see table "alpha_other_pct" in the supplementary materials Excel file).

Results from this simulation are shown in Table 8. Overall results, as well as results for $\rho \leq 20\%$, are shown in Table 8. A key difference between results at lower α values versus those at α = 15% is that RR finds fewer of the true regressors than does the EN for lower values of ω at an α setting of 5% (see row 5 of Table 8). Future research should explore how these metrics compare, on a trial-by-trial pairwise basis, for different α values relative to the chosen (baseline) α value of 15%.

Table 8. RR vs. EN by ω at α settings of 5% and 10%.

Row	Metric (Measured as Percent to the Nearest Integer)	α=5%				α=10%	
		ω				ω	
		100	**1000**	**4000**	**Extreme**	**300**	**Extreme**
1	$\Pr\{Pr(\ L_1^2(b_{LW} \mid X_{RR}) < L_1^2(b_{EN} \mid X) \mid r^2, \rho\ ) > 50\%\}$	82	79	74	91	84	93
2	$\Pr\{Pr(\ L_1^2(b_{LW} \mid X_{RR}) < L_1^2(b_{EN} \mid X) \mid r^2, \rho \leq 20\%\ ) > 50\%\}$	28	43	50	66	35	71
3	$\Pr\{Pr(L_1^2(b_{LW}^{true} \mid X_{RR}) < L_1^2(b_{EN}^{true} \mid X) \mid r^2, \rho) > 50\%\}$	54	55	54	85	64	89
4	$\Pr\{Pr(L_1^2(b_{LW}^{true} \mid X_{RR}) < L_1^2(b_{EN}^{true} \mid X) \mid r^2, \rho \leq 20\%) > 50\%\}$	13	25	30	41	20	55
5	$\Pr\{Pr(RR_{found}(X_{RR}^{true} \mid X_{RR}) > EN_{found}(X_{EN}^{true} \mid X) \mid r^2, \rho) > 50\%\}$	3	6	76	100	78	100
6	$\Pr\{Pr(RR_{found}(X_{RR}^{true} \mid X_{RR}) > EN_{found}(X_{EN}^{true} \mid X) \mid r^2, \rho \leq 20\%) > 50\%\}$	13	23	60	100	50	100
7	$Pr(b_{RR}^{true}$ *is empty* $\mid X_{RR},\ r^2,\ \rho)$	10	12	20	12	4	14
8	$Pr(b_{RR}^{true}$ *is empty* $\mid X_{RR},\ r^2,\ \rho \leq 20\%)$	18	28	55	47	15	47
9	$Pr(b_{EN}^{true}$ *is empty* $\mid X,\ r^2,\ \rho)$	6	22	31	48	11	57
10	$Pr(b_{EN}^{true}$ *is empty* $\mid X,\ r^2,\ \rho \leq 20\%)$	23	40	53	100	35	100

3.9. Results Summary

The proposed algorithm has unambiguously shown that RR has an advantage over the EN when searching for true regressors. Additionally, RR's computational simplicity makes it an important competitor to the EN.

Key high-level takeaways from the simulations done herein, in mostly plain English, are:

(1) RR finds more of the true regressors than does EN, with very high probability (≈99%). This is critically important for linear model discovery in the sciences, when $p > n$. Thus, it is advantageous to consider the set of significant regressors selected by RR using the hard constraint of 3 OPR built into the proposed algorithm. For example, the RR-selected regressors can be compared and contrasted with the EN-selected regressors in the context of the science of the process driving the regressand.

(2) When RR fails to find more of the true regressors than does the EN (0.73% of the time across all simulations), these failures occur when ρ is "small" ($\leq$20%) and r^2 is "large" ($\geq$200,000) or when ρ is "large" ($\geq$65%) and r^2 is "small" ($\leq$10). All of these failures occur under "mild" ($\omega \leq 300$) multicollinearity levels.

(3) The probability that the EN finds none of the true regressors is about six times higher than the corresponding probability (≈25% vs. 4%) that RR finds none of the true regressors. This re-emphasizes the fact that the RR-selected regressors should, at a minimum, be compared and contrasted with those selected by the EN.

(4) The squared length of the RR-estimated coefficient vector ($\hat{b}$, say) from the true coefficient vector (b) is less than the corresponding EN $\hat{b}$ with high probability (≈86%). Note that $\hat{b}$ includes the spurious coefficients that are either statistically significant (as in RR) or are the output of the optimization process (as in the EN). This indicates that the simpler RR-estimation process tends to produce more accurate estimates of b with high probability. Thus, comparing and contrasting both $\hat{b}$ vectors, in the context of the underlying science of the process generating the regressand would be quite advantageous.

(5) When the squared length of $\hat{b}$ from b, yielded by RR, fails to be less than the corresponding one yielded by the EN (13.5% of the time), 95% of these failures occur when $\rho \leq 20\%$ and 99% of them occur when $\rho \leq 35\%$. When $\rho \geq 35\%$, the failures occur for "large" r^2 ($\geq$100,000).

(6) The squared length of the true coefficients in the RR-estimated $\hat{b}$ (with the spurious ones set to zero), from b, is less than the corresponding one from the corresponding EN $\hat{b}$ with high probability (≈74%). This re-emphasizes, again, that comparing and contrasting the RR $\hat{b}$ with the EN $\hat{b}$ is quite important. We may not know *a priori* which coefficients in $\hat{b}$ are the true ones, but the science underlying the process generating the regressand may provide insights regarding wherein the truth lies. Conversely, the RR $\hat{b}$ would have an advantage over the EN $\hat{b}$ to alert scientists to the existence of possible causal variables hitherto unconsidered or undiscovered.

(7) When the squared length of the true coefficients in the RR-estimated $\hat{b}$ from b fails to be less than the corresponding one for the EN (25.79% of the time), about 95% of these failures occur when $\rho \leq 40\%$ and 99% of them occur when $\rho \leq 50\%$. These failure rates drop steeply under extreme multicollinearity.

(8) It is observed in the simulation that, whenever the squared length of $\hat{b}$ from b, yielded by RR, fails to be less than the corresponding one yielded by the EN, so does the corresponding squared length for the true coefficients in $\hat{b}$.

(9) On the other hand, if the squared length of $\hat{b}$ from b, yielded by RR, is less than the corresponding one yielded by the EN, there is about a 14% probability that the squared length of the true coefficients in the RR estimated $\hat{b}$ from b fails to be less than the corresponding one for the EN. Furthermore, when this event occurs, there is about an 89% probability that it happens when $\rho \leq 40\%$.

(10) For low to moderate levels of collinearity among the regressors, as measured by the traces of the respective matrices of regressors under consideration for RR estimation, the Hoerl, Kennard and Baldwin [26]-proposed values of the RR tuning parameters provide a good estimate of $\hat{b}$. For higher levels of collinearity, the corresponding values proposed by Lawless and Wang [27] are good. In practice, it would be best to compare and contrast the RR $\hat{b}$ solutions derived using the HKB and LW values of k, respectively.

(11) A wide range of input parameters are covered in the simulation. Specifically, the squared length of b from the origin (i.e., r^2), the *a priori* probability (i.e., ρ) that an element of b is zero and the multicollinearity level of the matrix of regressors (i.e., ω) are varied in the simulation. For a given set of data, ω is knowable, but, r^2 and ρ, in general, are not. However, the science underlying the data being examined may yield some insights into r^2 and ρ. If so, the tables with simulation output, in the body of this paper and those in the supplementary materials Excel file can indicate where RR is inferior to the EN in terms of metrics such as the accuracy of $\hat{b}$. Alternatively, the regression equations relating simulation inputs and outputs (e.g., as in Table 3) can indicate this. In such cases, the regressors found by RR and by EN can be pooled, and another EN or RR (or both) re-estimation can be done to see how it impacts the science underlying the data in terms of the causal relationships between the inputs believed to generate the output.

4. Discussion

In the following sub-sections, we discuss our work at a high-level and present some of its limitations.

4.1. This Work

Two questions of enormous importance in regression analysis (see Myers [56]), when used as a tool for model discovery in the sciences, is how many of the true regressors generating the regressand can be found and how accurate are their coefficients. If philosophical thinking were applied to science, a question of great importance is: what does $p > n$ mean in nature? In the context of our linear model, does it mean that data generation stopped before n could reach its "natural" state of being much larger than p and, thus, make the concept of the EN irrelevant? Or does it mean that $n < p$ is the "natural" state and that there will be no further data in nature (i.e., some type of "censoring" occurred in the fabric of nature)? Or does it simply mean that $n < p$ is, to borrow Penrose's [57] words, just "an arbitrary construction of the human mind" caused by the fact that, because we do not know what the true regressors are, we simply spread our net very wide in an attempt to find them?

Assuming that $n < p$ is a "natural" state and the underlying model generating the regressand is linear, the EN is a well-known procedure to estimate the regressors generating the regressand. Under these two assumptions, we have shown that RR improves upon the EN, with high probability, both in terms of finding the true regressors and the accuracy

of estimated coefficients. Furthermore, to accomplish this, we have bypassed "complex" optimization methods by using "elementary" mathematical operations within the RR framework. We use the word "elementary" in the sense that an operation like simply inverting a matrix in RR is much simpler than a set of operations needed to minimize a quadratic objective function subject to linear or nonlinear constraints.

The idea of using elementary, but not necessarily "simple", methods to tackle "complex" mathematical problems is well-known to mathematicians. For others, an introduction to this idea can be found in a fascinating paper by Levinson [58], which discusses elementary methods, not requiring more than the properties of the logarithm, to establish the prime number theorem. For example, Hardy's [59] keenness to find an elementary method establishing the prime number theorem is well known to mathematicians and reflected in the following statement of his: " ... *if Ramanujan really had proved (2.10.12), he would have found an elementary proof of the Prime Number Theorem, a proof involving no function-theory at all* ... ". Bohr [60] recognizes that despite Hardy's keenness to find an elementary method, Hardy concluded, in 1921, that: " ... *No elementary proof of the prime number theorem is known, and one may ask whether it is reasonable to expect one* ... *A proof of such a theorem, not fundamentally dependent upon the ideas of the theory of functions, seems to me extraordinarily unlikely* ... ". Our approach to tackling the EN has been inspired by mathematicians' keenness for pursuing that which is elementary. We have used two "old" and "elementary" ideas to find an alternative to the EN: the idea of "stepwise" variable selection, an idea which goes back to 1960, and the idea of RR, which goes back to 1970.

Furthermore, for our problem (i.e., the EN vs. RR), if a simpler method also produces better solutions to the problem, why not use it? This, we believe, is the value added to the practice of regression by our paper. To only rely on the EN to estimate b, when $p > n$, would not be as good, in a probabilistic sense, as also relying on RR to produce an estimate of b. Thus, using only RR to estimate b reasonably complements the EN estimate of b and, when compared, will likely produce more insight into the process generating Y.

Although the simulation results indicate that the probability of RR doing better than the EN is much higher than 50%, that probability is not 100%. Thus, there may be room to incorporate the Lasso into RR and further improve the search for significant regressors and their coefficients. This also indicates that RR can provide important feedback on the outputs of other fashionable competitors, such as machine learning (e.g., Breiman [61,62]), with its pervasive "black-box" focus on prediction, rather than on the process generating the data, as an end in itself.

There are several avenues for further research as outlined below:

(i) A few simulations were done using the following "ensemble" approach to selecting the RR tuning parameter k: when doing RR estimation by partitions, we use the LW value of k when the ratio of the trace associated with a partition to the number of regressors in that partition exceeds 10; otherwise, we use the HKB value of k. However, the few simulations done in this regard did not improve upon the selections of k considered herein. It may be of interest to pursue some variant of this idea further by increasing the cutoff value of 10 to a higher number and see if the ensemble approach to selecting k is beneficial.

(ii) The hard constraint of keeping 3 OPR when partitioning X can be relaxed by increasing OPR to a higher number like 10. This can be done by running simulations on a "wider" dataset, for example one with p = 5000 and n = 100. For the dataset used herein, p = 89 and n = 33, which yields a p-to-n ratio of about 2.7. For the "wider" dataset, this ratio would be 50.

(iii) Alternative selections of k can be considered, for example, those of Hoerl and Kennard [63] and Inoue [64]. For RR "failure scenarios", the conditional probability that RR finds fewer of the true regressors than does the EN is small (see row 16 of Table 2). However, for these failure scenarios, the RR $\hat{b}$ is less accurate (in terms of squared distance, $L_1^2(\cdot)$) than the corresponding EN $\hat{b}$. It may be worth examining if other se-

lections of k, such as those proposed by Hemmerle [65], can improve RR performance in these failure scenarios.

(iv) Following up on "(iii)" above, another consideration worth pursuing is iterative RR estimation for regressor selection. In this paper, significant regressors are selected by identifying significant values of the RR-estimated t-ratio (i.e., $t_{RR}^{(i)}$). However, this is done only once for each partition following Hoerl, Schuenemeyer and Hoerl [35]—i.e., all regressors associated with insignificant RR t-ratios are dropped, and the remaining regressors are retained as significant. Two versions of iterative RR estimation were proposed by Gana [66]. The first version, called "backward stepwise eliminating" (BSE), is the following: (a) we fit a LS regression to the partition under consideration; (b) we calculate the LW k; (c) we re-estimate the LS regression with RR using the LW k and calculate the RR t-ratio; (d) we drop the regressor having an RR t-ratio with the highest p-value above 20%; and (e) we redo steps "a" through "d" until the p values of the RR t-ratios of the remaining regressors are below 20% or until no regressors meet this criterion. The second version, called "backward group eliminating" (BGE), is the following: we follow all of the steps laid out for a BSE RR after modifying only step, "d", to drop all regressors with RR t-ratios whose p-values are greater than 20%. That is, in BGE RR, groups of insignificant regressors are dropped, in contrast to BSE RR, wherein the "most" insignificant regressors are dropped one at a time iteratively. For our problem, BSE RR may have some advantages. For example, some initial simulation done indicate that there is a greater than 50% probability that RR $\hat{b}$ will improve in terms of accuracy.

(v) Deeper explorations linking simulation outputs and inputs can be pursued. As mentioned before, multicollinearity levels (ω) would be known a priori but not the squared length of the true coefficient vector from the origin (r^2) or the probabilities (ρ) generating the true coefficients. Linking simulation inputs to outputs may shed light on the nature of r^2 and ρ by looking at observable outputs. For example, in Table 5, we note that, when RR fails to find more of the true regressors than EN, the probability of RR missing true regressors drops when RR finds more significant regressors than the number of observations (i.e., when $p_{RR} \geq n$). Because p_{RR} is observable, the question is whether such results hold over the entire simulation space with high probability for p_{RR} or for other observable simulation outputs.

(vi) Exploring connections between simulation outputs and inputs can also be pursued by researching whether connections exist between EN outputs and RR outputs.

4.2. Limitations

A key limitation of this study is that it is based on simulation. This limitation can be seen from two perspectives as explained next.

This paper was motivated by the work of Gana and Vasudevan [67]. In that paper, a search was attempted to find the true regressors (causality) driving a fundamental, and very complex, biological process called O-glycosylation. In that paper, the problem of finding true regressors was tackled by keeping $p < n$. Examining the problem under $p > n$ was not considered, because it was unclear how to go about that without resorting to other fashionable, but "black-box", methods, such as machine learning, that focus on prediction, rather than on the process generating the data, as an end in itself. In the sciences, and to a certain extent in economics, finding the variables that generate the data of interest (e.g., the target variable) is of enormous importance. That is, in science, the focus of model discovery is always on causality. This brings us to an important question: prior to the availability of experimental evidence, what are the limits on the discovery of true regressors imposed by the use of RR when $p > n$ in nature? Simulation is one way to get a sense for how probable it is to find causality. Then there arises the difficult question of how simulation results can be used efficiently to understand how nature generates the target variable. We will touch upon these next.

Given the mathematical intractability of studying how one can find true regressors in the context of the linear model, there is no obvious path to be computer-independent with regard to this exercise. That is, there is no obvious path to analytically prove things like the following: there exists a positive vector of GRR tuning parameters such that the mean squared error (MSE) of the coefficients of the true regressors found by GRR is less than the MSE of the corresponding coefficients found by the EN or there exists a positive vector of GRR tuning parameters such that the number of true regressors found by GRR is greater than those found by the EN. If the linear model is assumed to be true, as is (i.e., in non-model-discovery mode), there are several theoretical studies that have explored the properties of GRR (e.g., Ishwaran and Rao [68], Lawless [69] and Hemmerle [65]). Thus, extensive simulation is done in this present paper to provide answers to such critically important questions.

Because this study is simulation-based, another key limitation is that it is somewhat unclear how the methodology can be applied to empirical data. For us, in this paper, the matrix of potential regressor values, X, is synthetic, and we know exactly how the corresponding vector Y is generated. The advantage of a simulation is that the truth is known. For empirical data, especially in the sciences, experimental verification is of enormous importance. As a thought experiment, we can suppose that X and Y are empirical matrices and that the regression model linking them is linear. The multicollinearity spectrum of X (by partition) would be easy to estimate. That is, ω can be easily calculated. How RR estimates will behave in regions of high or low collinearity is well-known. For example, it would be fairly easy to select appropriate alternative algorithms for calculating k by partition. Two key unobservable parameters are ρ and r^2. Without some prior information about these two parameters, we will not be able to estimate whether or not we are applying RR to empirical data in $(\rho,\ r^2)$ "regions" where RR fails to be better than the EN.

If we can estimate ρ and r^2 given X and Y, we will have better insight for recognizing "failure regions". When RR fails in a particular region, EN can be used to estimate b; otherwise, RR can be used to estimate b. Although our research in regard to this is truly in its infancy, it appears, from several trial-and-error simulations that we have done, that there may be some conditional probability statements we could make, such as the following one, for the case of an X matrix with $p < n$ (i.e., a partition):

$$Pr\left\{ abs\left(\frac{\hat{b}^T\hat{b} - s^2 \times trace\left(\left(X^TX\right)^{-1}\right)}{r^2} - 1 \right) \leq 2 \;\middle|\; \hat{b}^T\hat{b} - s^2 \times trace\left(\left(X^TX\right)^{-1}\right) > 0 \right\} > 50\% \tag{9}$$

where "abs" denotes the "absolute value" operator, the X matrix in (9) has more rows than columns, and $\hat{b}$ is the LS estimate of b such that only the statistically significant coefficients, at α of 15%, are retained. Our experiments also indicate that the probability of the statement following the conditional operator (" | ") in (9) is reasonably high. When $p < n$, it is easy to see that (9) has a point of contact with the following corresponding result under LS for the *true* model:

$$E\left(\hat{b}^T\hat{b}\right) = b^Tb + \sigma^2 \times trace\left(\left(X^TX\right)^{-1}\right) \tag{10}$$

where σ^2 is the usual true residual variance, b is the true coefficient vector associated with X and $\hat{b}$ is the corresponding LS estimate of b. That is, in (10), no variable selection is involved, because it represents the textbook case of a fully known linear model. In other words, for this X, coefficients for *all* regressors are estimated. In contrast, the X in (9), is a subset of regressors drawn from a larger set of potential regressors such that all regressors in this X are statistically significant per LS. Because we are discussing this mostly in plain English for now, we have not developed notation to distinguish X in (9) and (10).

A challenge posed by (9) is how to generalize it to the case of $p > n$. The thought process behind (9) is only presented here as a crude example for thinking about future research along similar lines. For example, there may be probability statements in the spirit of (9) that hold, under GRR, when $p > n$ with probabilities materially higher than

50%. Similarly, there may be statements estimating probabilistic bounds on ρ. We are pursuing research along these lines and hope to share the results, if promising, in the future. However, because the research in support of (9) is in its infancy, it would be rash of us to discuss it more and unintentionally create the perception of its legitimacy.

Another limitation worth mentioning, in empirical work, is that we have to assume that p is known. That is, we have to assume that the number of potential regressors from which to select the causal regressors is known. This will rarely be the case in practice.

Notwithstanding these limitations, if data science is defined as the search for the causal or true drivers of a target variable, and the resultant discovery of a theory generating the target, both primarily depending on the data, then this paper has a point of contact with data science. For mathematically intractable problems, such as the one to which this paper is dedicated, computation by simulation is a powerful way to understand the merits of new algorithms, and as stressed by Summermann et al. [70], for imparting that knowledge to students and researchers.

It is hoped that the results herein are useful for data scientists engaged in model discovery via regression analysis in the sciences, wherein causality is germane to understanding the process generating the data.

5. Conclusions

In this paper, we have studied how well elementary methods, more than half a century old, can successfully compete with the EN. The two key elementary methods we have used are RR and the concept of stepwise variable selection. A computational advantage in doing what we did is that mathematical operations more complex than matrix inversion are not necessary for estimation. This is so because the use of these elementary methods keeps us well within the LS framework. In contrast, because the EN uses both the Lasso and RR to estimate the coefficients of the linear model, the EN is computationally more "complex". In particular, the EN requires operations outside of the LS framework for estimation and optimization. Studying how competitive these old methods are with the EN is a mathematically intractable problem. Thus, we had to resort to doing a comprehensive simulation for studying this. Using elementary methods, and computer methods, for understanding mathematical problems is well-known to mathematicians. Furthermore, such approaches have a long and rich tradition in mathematics (e.g., see Diamond [71]).

Our simulation starts by defining the linear model using three key metrics: the squared distance of the true coefficients from the origin (r^2), the *a priori* probability (ρ) that a true coefficient is zero and the correlation structure (ω) binding the regressors together. Twenty discrete values of r^2 in the closed interval $[10, 6 \times 10^5]$, and eight discrete values of ρ in the closed interval $[0.05, 0.95]$, were chosen to simulate linear models. Four values of ω were chosen, representing low to extreme levels of multicollinearity among the regressors. We applied these metrics to a synthetic dataset having 33 observations and 89 regressors and created a large number of true linear models. Starting with ordinary RR, we developed a simple algorithm for selecting statistically significant regressors. We selected significant regressors by partitioning the matrix of potential regressors into a number of partitions, each of which has three observations per regressor. Then, we used GRR to successfully estimate the coefficients of all of the significant regressors jointly.

We observed that, given empirical data under the assumption of a linear model, a formidable challenge is to estimate r^2 and ρ. We conjecture that there are probability-based inequalities lying deeper below the surface that place bounds on r^2 and ρ, based on the empirical data. However, our research regarding such bounds is in its infancy, and, thus, we cannot, as yet, make concrete conjectures as to what such bounds may be. Thus, in practice, a key limitation of our work is that we have offered no "optimal" estimates of r^2 and ρ that can be derived from an empirical matrix of regressors and the corresponding empirical regressand.

Our simulation results clearly show that our simple algorithm is "superior" to the EN with high probability, in theory. Superiority is measured using four metrics: the precision

of the estimated statistically significant coefficients (which may be either true or spurious), the precision of the estimates of the true coefficients found, the chance that our algorithm finds none of the true coefficients and the chance that our algorithm finds more of the true coefficients than does the EN. Our algorithm is superior to the EN with high probability in regard to all of these metrics. This means that, at the very least, our algorithm should be used in conjunction with the EN for linear model discovery.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/math10173057/s1. Supplementary_Material_file_mdpi_math.xlsx and ElasticRR3F.sas.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used for this study is synthetic data and is included as supplementary material.

Acknowledgments: I thank the two anonymous reviewers for their valuable comments on the earlier version of my manuscript, which enhanced the presentation of my research. I thank the assistant editor for keeping me informed on manuscript progression at all times. I thank the academic editor for comments that helped enhance my presentation structure. The prior version of my manuscript was posted on SSRN's preprint server in September 2020, and I thank SSRN for offering this option. This paper is dedicated to the memory of Arthur Edwin Hoerl, who encouraged my pursuit of Regression and showed me what it is about.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

BGE	Backward group eliminating
BSE	Backward stepwise eliminating
EN	Elastic Net
GRR	Generalized ridge regression
HKB	Hoerl, Kennard and Baldwin
LS	Least squares
LW	Lawless and Wang
MSE	Mean squared error
NLP	Nonlinear programming
RR	Ridge regression
SM	Supplementary material
SNR	Signal-to-noise ratio
VIF	Variance inflation factor

References

1. Hoerl, A.E.; Kennard, R.W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
2. Petkovsek, M.; Wilf, H.S.; Zeilberger, D. *A = B*; CRC Press, Taylor & Francis Group: Boca Raton, FL, USA, 1996.
3. Schott, J.R. *Matrix Analysis for Statistics*; John Wiley, Inc.: Hoboken, NJ, USA, 2016.
4. Seber, G.A.F. *A Matrix Handbook for Statisticians*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2007.
5. Vinod, H.D. Equivariance of ridge estimators through standardization—A note. *Commun. Stat.-Theory Methods* **1978**, *7*, 1157–1161. [CrossRef]
6. Frank, I.E.; Friedman, J.H. A Statistical View of Some Chemometrics Regression Tools. *Technometrics* **1993**, *35*, 109–135. [CrossRef]
7. Halawa, A.M.; El Bassiouni, M.Y. Tests of regression coefficients under ridge regression models. *J. Stat. Comput. Simul.* **2000**, *65*, 341–356. [CrossRef]
8. Gokpinar, E.; Ebegil, M. A study on tests of hypothesis based on ridge estimator. *Gazi Univ. J. Sci.* **2016**, *29*, 769–781.
9. Muniz, G.; Kibria, G.B.M.; Shukur, G. On Developing Ridge Regression Parameters: A Graphical Investigation. *Stat. Oper. Res. Trans.* **2012**, *36*, 115–138.

10. Piegorsch, W.W.; Casella, G. The Early Use of Matrix Diagonal Increments in Statistical Problems. *SIAM Rev.* **1989**, *31*, 428–434. [CrossRef]
11. Hoerl, R.W. Ridge Analysis 25 Years Later. *Am. Stat.* **1985**, *39*, 186–192.
12. Hoerl, R.W. Ridge Regression: A Historical Context. *Technometrics* **2020**, *62*, 420–425. [CrossRef]
13. Brook, R.J.; Moore, T. On the expected length of the least squares coefficient vector. *J. Econom.* **1980**, *12*, 245–246. [CrossRef]
14. Smith, G.; Campbell, F. A Critique of Some Ridge Regression Methods. *J. Am. Stat. Assoc.* **1980**, *75*, 74–81. [CrossRef]
15. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
16. Osborne, M.; Presnell, B.; Turlach, B. A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.* **2000**, *20*, 389–403. [CrossRef]
17. Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. Least angle regression. *Ann. Stat.* **2004**, *32*, 407–499. [CrossRef]
18. Delbos, F.; Gilbert, J.C. Global Linear Convergence of an Augmented Lagrangian Algorithm to Solve Convex Quadratic Optimization Problems. *J. Convex Anal.* **2005**, *12*, 45–69.
19. Taylor, H.L.; Banks, S.C.; McCoy, J.F. Deconvolution with the ℓ_1 norm. *Geophysics* **1979**, *44*, 39–52. [CrossRef]
20. Santosa, F.; Symes, W.W. Linear Inversion of Band-Limited Reflection Seismograms. *SIAM J. Sci. Stat. Comput.* **1986**, *7*, 1307–1330. [CrossRef]
21. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* **2005**, *67*, 301–320. [CrossRef]
22. Boissonnade, A.; Lagrange, J.L.; Vagliente, V.N. *Analytical Mechanics*; Springer: Berlin/Heidelberg, Germany, 1997.
23. Bussotti, P. On the Genesis of the Lagrange Multipliers. *J. Optim. Theory Appl.* **2003**, *117*, 453–459. [CrossRef]
24. Plackett, R.L. Studies in the History of Probability and Statistics. XXIX. *Biometrika* **1972**, *59*, 239–251. [CrossRef]
25. Stigler, S.M. Gauss and the Invention of Least Squares. *Ann. Stat.* **1981**, *9*, 465–474. [CrossRef]
26. Hoerl, A.E.; Kannard, R.W.; Baldwin, K.F. Ridge regression: Some simulations. *Commun. Stat.* **1975**, *4*, 105–123. [CrossRef]
27. Lawless, J.F.; Wang, P. A simulation study of ridge and other regression estimators. *Commun. Stat.-Theory Methods* **2010**, *5*, 307–323.
28. Austin, P.C.; Steyerberg, E.W. The number of subjects per variable required in linear regression analyses. *J. Clin. Epidemiol.* **2015**, *68*, 627–636. [CrossRef]
29. Whitehead, A.N.; Russell, B.A.W. *Principia Mathematica to *56*; Cambridge University Press: Cambridge, UK, 1962.
30. Ricci, M.M.G.; Levi-Civita, T. Methodes de calcul diffrentiel absolu et leurs applications. *Math. Ann.* **1900**, *54*, 125–201. [CrossRef]
31. Efroymson, M.A. Multiple Regression Analysis. In *Mathematical Methods for Digital Computers*; Ralston, A., Wilf, H.S., Eds.; John Wiley: New York, NY, USA, 1960.
32. Hocking, R.R. A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression. *Biometrics* **1976**, *32*, 1–49. [CrossRef]
33. Miller, A.J. The Convergence of Efroymson's Stepwise Regression Algorithm. *Am. Stat.* **1996**, *50*, 180–181.
34. Gana, R. Ridge regression and the Lasso: How do they do as finders of significant regressors and their multipliers? *Commun. Stat.-Simul. Comput.* **2020**, 1–35. [CrossRef]
35. Hoerl, R.W.; Schuenemeyer, J.H.; Hoerl, A.E. A Simulation of Biased Estimation and Subset Selection Regression Techniques. *Technometrics* **1986**, *28*, 369–380. [CrossRef]
36. Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
37. SAS. *SAS Enterprise Guide 7.15 HF9 (64-bit)*; SAS Institute Inc.: Cary, NC, USA, 2017.
38. Marquardt, D.W.; Snee, R.D. Ridge Regression in Practice. *Am. Stat.* **1975**, *29*, 3–20.
39. Hoerl, A.E.; Kennard, R.W. Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics* **1970**, *12*, 69–82. [CrossRef]
40. Hoerl, A.E.; Kennard, R.W. Ridge Regression: Degrees of Freedom in the Analysis of Variance. *Commun. Stat.-Simul. Comput.* **1990**, *19*, 1485–1495. [CrossRef]
41. White, H. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica* **1980**, *48*, 817–838. [CrossRef]
42. Obenchain, R.L. Classical F-Tests and Confidence Regions for Ridge Regression. *Technometrics* **1977**, *19*, 429–439. [CrossRef]
43. Fubini, G. Opere scelte. *Cremonese* **1958**, *2*, 243–249.
44. Fubini, G. Sugli integrali multipli. *Rom. Acc. L. Rend.* **1907**, *16*, 608–614.
45. Shampine, L.F. Matlab program for quadrature in 2D. *Appl. Math. Comput.* **2008**, *202*, 266–274. [CrossRef]
46. Shampine, L.F. Vectorized adaptive quadrature in MATLAB. *J. Comput. Appl. Math.* **2008**, *211*, 131–140. [CrossRef]
47. Matlab. *R2021b Update 2*; The MathWorks Inc.: Natick, MA, USA, 2022.
48. Durbin, J. Testing for Serial Correlation in Least-Squares Regression When Some of the Regressors are Lagged Dependent Variables. *Econometrica* **1970**, *38*, 410–421. [CrossRef]
49. Durbin, J. Tests for Serial Correlation in Regression Analysis Based on the Periodogram of Least-Squares Residuals. *Biometrika* **1969**, *56*, 1–15. [CrossRef]
50. Vinod, H.D. Generalization of the durbin-watson statistic for higher order autoregressive processes. *Commun. Stat.* **1973**, *2*, 115–144. [CrossRef]
51. Spitzer, J.J. Small-Sample Properties of Nonlinear Least Squares and Maximum Likelihood Estimators in the Context of Autocorrelated Errors. *J. Am. Stat. Assoc.* **1979**, *74*, 41–47. [CrossRef]
52. Elliott, G.; Rothenberg, T.J.; Stock, J.H. Efficient Tests for an Autoregressive Unit Root. *Econometrica* **1996**, *64*, 813–836. [CrossRef]

53. Rose, A. *Vision: Human and Electronic*; Plenum Press: New York, NY, USA, 1973.
54. Burgess, A.E. The Rose Model, Revisited. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* **1999**, *16*, 633–646. [CrossRef]
55. Bendel, R.B.; Afifi, A.A. Comparison of Stopping Rules in Forward "Stepwise" Regression. *J. Am. Stat. Assoc.* **1977**, *72*, 46–53.
56. Myers, R.H. *Classical and Modern Regression with Applications*; International Thomson Publishing: London, UK, 1990.
57. Penrose, R. *The Emperor's New Mind*; Oxford University Press: Oxford, UK, 2016.
58. Levinson, N. A Motivated Account of an Elementary Proof of the Prime Number Theorem. *Am. Math. Mon.* **1969**, *76*, 225–245. [CrossRef]
59. Hardy, G.H. *Ramanujan: Twelve Lectures on Subjects Suggested by His Life and Work*; Chelsea Publishing Company: New York, NY, USA, 1978.
60. Bohr, H. Address of Professor Harald Bohr. In Proceedings of the International Congress of Mathematicians, Cambridge, MA, USA, 3 August–6 September 1950; American Mathematical Society: Cambridge, MA, USA, 1950; p. 129.
61. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
62. Breiman, L. Statistical Modeling: The Two Cultures (With Comments and a Rejoinder by the Author). *Stat. Sci.* **2001**, *16*, 199–231. [CrossRef]
63. Hoerl, A.E.; Kennard, R.W. Ridge regression iterative estimation of the biasing parameter. *Commun. Stat.-Theory Methods* **1976**, *5*, 77–88. [CrossRef]
64. Inoue, T. Improving the 'Hkb' Ordinary Type Ridge Estimator. *J. Jpn. Stat. Soc.* **2001**, *31*, 67–83. [CrossRef]
65. Hemmerle, W.J. An Explicit Solution for Generalized Ridge Regression. *Technometrics* **1975**, *17*, 309–314. [CrossRef]
66. Gana, R. Did COVID-19 Force a President to Face the St. Petersburg Paradox and lose the White House? *Soc. Sci. Res. Netw.* **2021**, *1*, 7. Available online: www.ssrn.com/abstract=3898613 (accessed on 6 August 2021). [CrossRef]
67. Gana, R.; Vasudevan, S. Ridge regression estimated linear probability model predictions of O-glycosylation in proteins with structural and sequence data. *BMC Mol. Cell Biol.* **2019**, *20*, 21. [CrossRef]
68. Ishwaran, H.; Rao, J.S. Geometry and properties of generalized ridge regression in high dimensions. *Contemp. Math.* **2014**, *622*, 81–93.
69. Lawless, J.F. Mean Squared Error Properties of Generalized Ridge Estimators. *J. Am. Stat. Assoc.* **1981**, *76*, 462–466. [CrossRef]
70. Sümmermann, M.L.; Sommerhoff, D.; Rott, B. Mathematics in the Digital Age: The Case of Simulation-Based Proofs. *Int. J. Res. Undergrad. Math. Educ.* **2021**, *7*, 438–465. [CrossRef]
71. Diamond, H.G. Elementary methods in the study of the distribution of prime numbers. *Bull. Am. Math. Soc.* **1982**, *7*, 553–589. [CrossRef]