Graphical Comparison of TRACEs from CRAN **R**-packages for Ridge Regression

Robert L. Obenchain
Risk Benefit Statistics, Clayton, CA 94517

December 2024

Abstract

We discuss ten **TRACE** graphics generated using functions from four CRAN **R**-packages [genridge, lmridge, ridge or RXshrink] that implement diverse forms of Ridge Regression. The "10-Factor" data of Gorman and Toman (1966) are used to illustrate alternative "shrinkage" paths. "Efficient" shrinkage **TRACE**s, Obenchain (2022), are shown not only for all 10 X-predictors but also for the "Best Subset" of 7 X-predictors proposed by Gorman and Toman.

1 Introduction

The Ordinary Least Squares (OLS) estimator of the $\beta-$coefficient vector in linear regression is $\hat{\beta}^o = (X'X)^+X'y$, where the superscript + denotes the Moore–Penrose pseudo-inverse of $X'X$. This OLS estimator is unquestionably the most "well-known" estimator that achieves "maximum likelihood" under normal distribution-theory. In fact, $\hat{\beta}^o$ is the "BLUE" (Best Linear Unbiased Estimator) for linear models with independent and homoscedastic error-terms because it achieves minimum MSE risk under these assumptions.

A potential "difficulty" with OLS in practical applications of linear regression is its unbiasedness: $E(\hat{\beta}^o) = \beta$. After all, when $p \geq 2$ X-predictors are **highly inter-correlated** (i.e. **ill-conditioned**), unbiased estimation implies that the variances of (and covariances between) $\hat{\beta}^o-$estimates can then be quite large. Ridge "shrinkage" (i.e. **biased estimation**) can typically reduce these problems.

The ridge estimators proposed in Hoerl and Kennard (1970a) use a single (scalar) parameter, k, commonly characterized as being a "small positive numerical constant". This k is added to each diagonal element of the $X'X-$matrix before it is inverted: $\hat{\beta}^*(k) = [X'X + k{\cdot}I]^{-1}X'y$. These $\hat{\beta}^*(k)$ estimators are *biased* when $k > 0$, but the **normal distribution-theory "likelihood"**

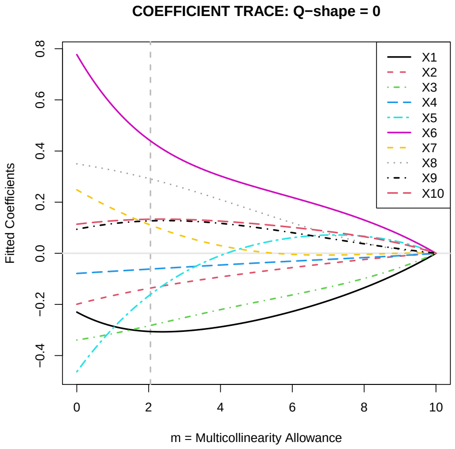

Figure 1: This Q=0 TRACE from the **RXshrink R**-package corresponds to Hoerl-Kennard (1970b) shrinkage. All 10 $\hat{\beta}-$coefficient estimates are "shrunken" to **0** at **m=p=10**; See equation (4) in Section 3. Note that none of these 10 "curves" is truly straight.

of these estimators to be **optimally biased** had not been quantified before the publication of Obenchain(1975).

Every "generalized" ridge regression shrinkage "Path" starts at $\hat{\beta}^o$, the OLS Best estimator. Since the *terminus* of the shrinkage path is (usually) taken to be $\hat{\beta}^* = 0$, and the overall length of the shrinkage path is always finite, it is somewhat unfortunate that Hoerl and Kennard's $k-$parameter must approach $+\infty$ to actually approach this terminus.

Version 2.3 (2023) of my **RXshrink R**-package provides three main **R**-functions.

[1]. **YonX()** treats the special cases where a **single X-variable** is used to predict the Y-variable.

[2]. **qm.ridge()** uses two strictly **finite parameters**, q and m, to determine not only the most likely $q-$Shape or "curvature" for its "shrinkage path" $[-5 \leq q \leq +5]$ but also the most likely $m-$Extent of shrinkage $[0 \leq m \leq p]$ along that path.

[3]. **eff.ridge()** treats models where the "centered" $X-$matrix has rank ≥ 2. This **Efficient Shrinkage Path** typically applies a different $\delta-$factor to each "uncorrelated component" of the OLS estimator, $\hat{\beta}^o$. Its path consists of **two-piece linear functions** that first connect each (unbiased) OLS coefficient estimate to it's optimally biased estimate, and then heads directly for the shrinkage terminus at $\hat{\beta}^* = 0$.

1.1 Multiple Regression Notation and Standardization

The usual model for multiple linear regression is written as

$$E(y|X) = 1\mu + X\beta \text{ and } Var(y|X) = \sigma^2 I \ , \tag{1}$$

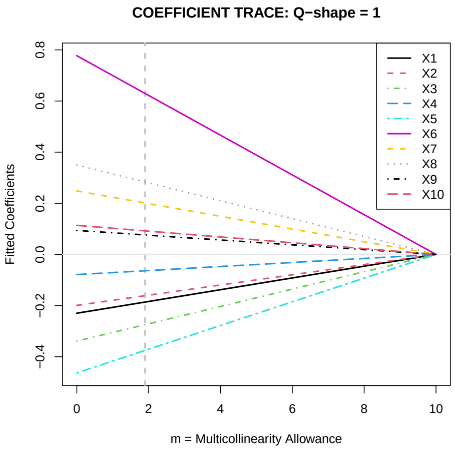

Figure 2: This Q=1 TRACE from the **RXshrink R**-package depicts Mayer-Willke (1973) "Uniform" shrinkage. Note that all 10 coefficient estimates now form **perfectly straight lines** and thus have **stable relative magnitudes**!

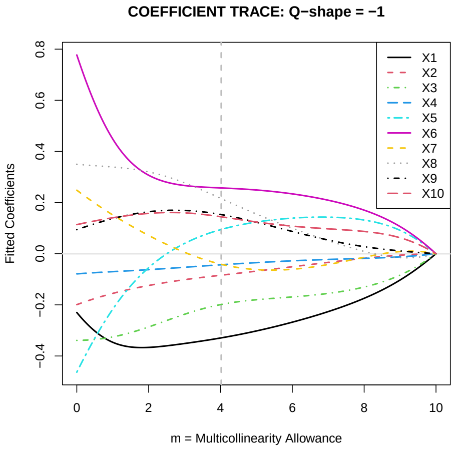

Figure 3: This TRACE for $Q = -1$ is the $Q-$shaped path most likely to lead to minimum MSE risk. It maximizes the "CRLQ" and minimizes the "CHISQ" criteria used by the **RXshrink R**-package.

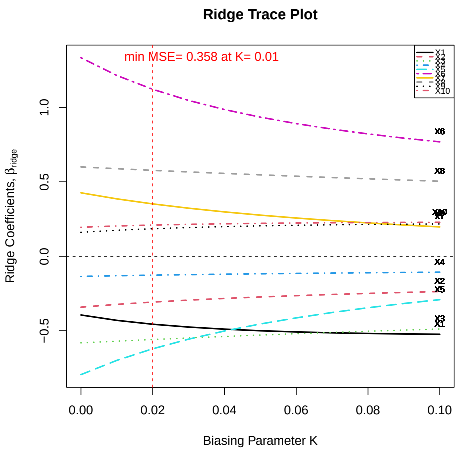

Figure 4: This TRACE from the **lmridge R**-package ends abruptly at $K = 0.10$. While the red text states that minimum MSE risk occurs at $K = 0.01$, the vertical red dashed line at $K = 0.02$ corresponds, instead, to the Generalized Cross Validation criterion, Golub et al. (1979).

where y is a $n{\times}1$ vector of observed response values, X is a $n{\times}p$ matrix of non-constant coordinates for p predictor variables, 1 is a $n{\times}1$ vector of ones, μ is an unknown intercept, β is a $p{\times}1$ vector of unknown regression coefficients, and $\sigma^2 I$ is an unknown and positive scalar multiple of a $p{\times}p$ identity matrix. Note that the $X-$matrix is assumed given, while the $y-$vector of response outcomes is assumed to be (conditionally) stochastic in the sense that it consists of uncorrelated observations with constant unknown variance, σ^2.

Generalized Ridge Regression may possibly best be viewed as the "shrinkage" version of **Principal Components Regression**, Massy(1965), where each of p individual δ_i-factors is either exactly 0 or exactly 1.

1.2 Optimal Shrinkage Δ Factors

To apply normal-theory maximum likelihood to shrinkage estimation, we will need a *definition* for the Δ factors that make $\hat{\beta}(\Delta) = G\Delta c$ the minimum MSE, linear estimator of β. We start by computing the risk of δ_i times c_i as an estimator of the $i-$th true component γ_i:

$$MSE(\delta_i c_i) = E[(\delta_i c_i - \gamma_i)^2] = E\{[\delta_i(c_i - \gamma_i) - (1 - \delta_i)\gamma_i]^2\}. \tag{2}$$

Obenchain(1978) considered alternative definitions for MSE optimal shrinkage but ultimately concluded that minimizing (2) is the most reasonable definition overall.

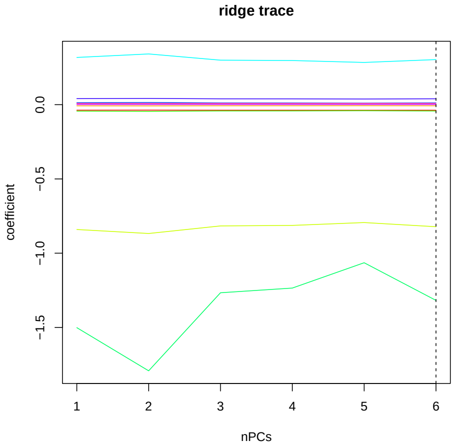

Figure 5: This TRACE from the **ridge** **R**-package strikes me as being confusing and minimally informative. While it suggests that only three $\beta-$coefficient estimates are clearly different from 0, which 3 out of 10 are they? Also, in what sense does the vertical dashed line at $nPC = 6$ signify a "good" or "best" choice?

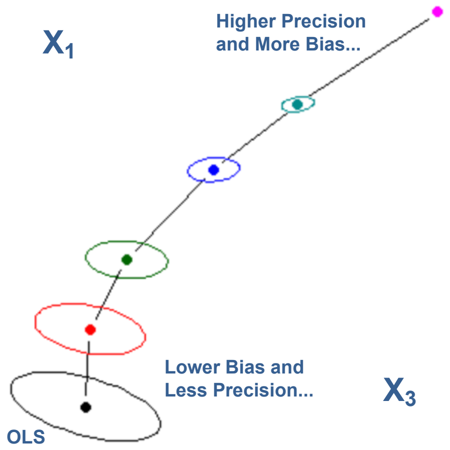

Figure 6: This is the portion of the "genridge" **pairs** graphic depicting the $p = 2$ model $logY \sim X1 + X3$. While a "trade-off" between Bias and Precision clearly occurs here, this graphic strikes me as providing no useful information on an optimal amount of shrinkage.

2 Two-Parameter Shrinkage Paths

Obenchain (1975) proposed restricted GRR Shrinkage Paths of the general form:

$$\delta_j = \lambda_j / (\lambda_j + k \times \lambda_j^q) = 1/(1 + k \times \lambda_j^{(q-1)}) \ . \tag{3}$$

for $j = 1, ..., p$, where k and q are scalars such that $0 < k < +\infty$ and $-5 \leq q \leq +5$.

The qm.ridge() function in the *RXshrink* R-package of Obenchain(2022) searches, by default, over only integer and half-integer $q-$Shapes between qmin = -5 and qmax = $+5$. The limit as q approaches $+\infty$ is optimal for the "unfavorable" case where the true β vector is parallel to the eigen-vector with the smallest eigen-value, λ_p. Shrinkage to $m = p - 1$ ($\delta_1 = \cdots = \delta_{p-1} = 0$) then reduces all components of $\hat{\beta}^o$ orthogonal to the true β to zero! This is essentially an extreme form of Massy(1965) "type (b)" principal components regression.

3 Likelihood in Shrinkage Estimation

Equation (2) can be inverted to yield

$$\gamma_i^2/\sigma^2 = \delta_i^{MSE}/[\lambda_i(1 - \delta_i^{MSE})] \ . \tag{4}$$

Equation (4) quantifies the KEY relationships used in Obenchain (1975) to define the normal distribution-theory likelihood that any given set of p shrinkage $\delta-$factors achieves overall minimum MSE-risk.

The "efficient path" of Obenchain (2022) consists of p **two-piece linear functions**, each having a **single interior knot** at the $\hat{\beta}-$estimator with Maximum Likelihood of achieving minimum MSE risk under normal distribution-theory. This **efficient path** is the shortest path and, at least when $p > 2$, essentially the only known shrinkage path that always contains the $\hat{\beta}-$vector that is most likely to be **optimally biased**. Functions in R-packages freely distributed via **CRAN** perform the calculations and produce graphics that illustrate optimal shrinkage. These basic concepts and visualization tools provide invaluable data-analytic insights and improved self-confidence to applied researchers and data scientists fitting linear models to data.

Each GRR TRACE typically displays estimates of p quantities that change as *shrinkage* occurs. The "coef" TRACE shows how fitted linear-model $\hat{\beta}-$estimates change with shrinkage. The "rmse" TRACE displays corresponding estimates of **relative mean-squared-error** given by dividing each diagonal element of the MSE-matrix by the OLS-estimate of σ^2. Three other types of *RXshrink* TRACE displays ("exev", "infd" and "spat") are not illustrated here.

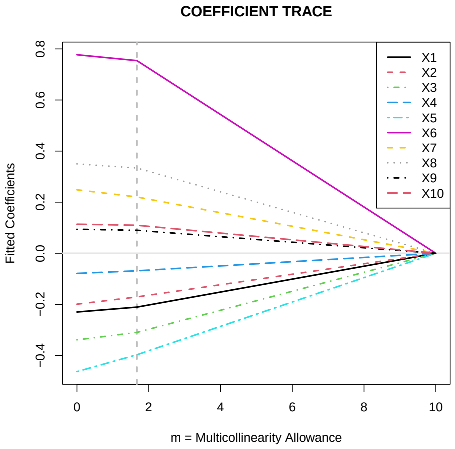

Figure 7: This is the Efficient 10−parameter TRACE display for shrunken $\beta-$coefficients. The vertical dashed-line indicates that Minimal MSE Risk occurs at $m = 1.85$.

4 Quantifying the m-Extent of Shrinkage

A measure of the *extent of shrinkage* applied by equation (2) is given by

$$m \equiv p - \delta_1 - \cdots - \delta_p = rank(X) - trace(\Delta). \tag{5}$$

This scalar, called the *multicollinearity allowance*, Obenchain (1977), is always ≥ 0 and $\leq p =$ Rank of the $X-$matrix.

In the TRACE plot proposed by Hoerl and Kennard (1970), their $k-$factor starts at 0 but must end abruptly at some finite $k-$max value specified by the user. Their shrinkage terminus is the vector of p zeros, and $k-$max is apparently chosen via trial-and-error. In Figure 4, $k-$max= 0.10.

5 Summary

Freely available CRAN R-packages provide computational and visual "fast-tracks" into the strengths and weaknesses of alternative formulations for Generalized Ridge Regression "Shrinkage". Here, we have focused primarily on TRACE plots of Regression $\hat{\beta}-$coefficient estimates. Besides TRACE-plots of Relative MSE Risk, functions in the **RXshrink R**-package can also display TRACE-plots of "Excess Eigenvalues", "Inferior Direction" cosines as well as the "Shrinkage Pattern" of generalized ridge $\delta-$factors.

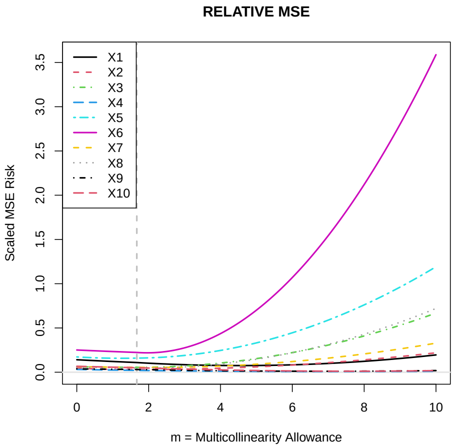

Figure 8: This is the Efficient 10−parameter TRACE display of Relative MSE Risk estimates. The vertical dashed-line again indicates that Minimal MSE Risk occurs at $m = 1.85$.

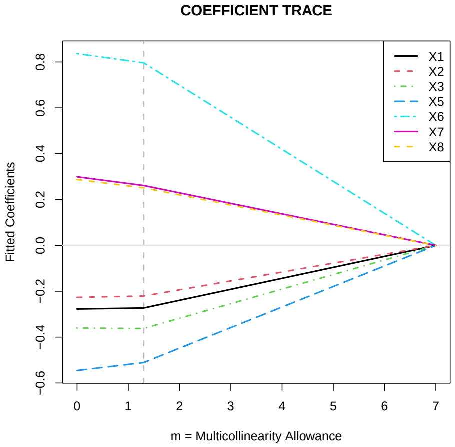

Figure 9: This is the Efficient 7−parameter TRACE for the subset of 7 predictors selected by Gorman and Toman(1966) that results from dropping $X4$, $X9$ and $X10$. Minimal Relative MSE Risk occurs here at $m = 1.30$; see Figure 10.

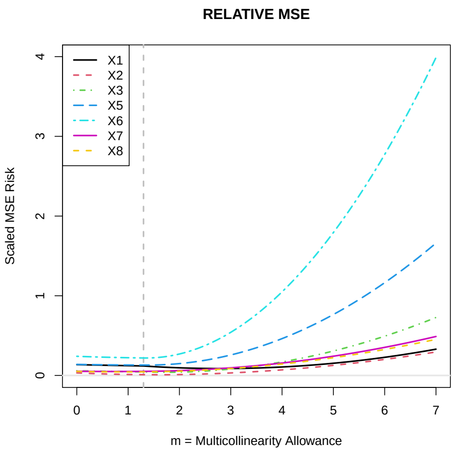

Figure 10: This is the Efficient "rmse" TRACE resulting from dropping $X4$, $X9$ and $X10$. The vertical dashed-line indicates that Minimal MSE Risk occurs here at m = 1.30.

References

[1] Friendly, M. (2024), "*genridge*: Generalized Ridge Trace Plots for Ridge Regression", ver 0.8.0, `https://CRAN.R-project.org/package=genridge`

[2] Golub, G. H., Heath, M. and Wahba. G. (1979), "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter". *Technometrics* 21, 215—223.

[3] Gorman, J. W. and Toman, R. J. (1966), "Selection of Variables for Fitting Equations to Data." *Technometrics* 8, 27—51.

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

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

[6] Imdad Ullah Muhammad (2023), "*lmridge*: Linear Ridge Regression with Ridge Penalty and Ridge Statistics", ver 1.2.2, `https://CRAN.R-project.org/package=lmridge`

[7] Massy, W. F. (1965), "Principal components regression in exploratory statistical research." *Journal of the American Statistical Association* 60, 234—246.

[8] Mayer, L. S. and Willke, T. A. (1973), "On biased estimation in linear models." *Technometrics* 15, 497—508.

[9] Moritz, S., Cule, E. and Frankowski, D. (2022), "*ridge*: Ridge Regression with Automatic Selection of the Penalty Parameter", ver 3.3, `https://CRAN.R-project.org/package=ridge`

[10] Obenchain, R. L. (1975), "Ridge analysis following a preliminary test of the shrunken hypothesis." *Technometrics* 17, 431−441.

[11] Obenchain, R. L. (1978), "Good and optimal ridge estimators." *Ann. Statist.* 6, 1111−1121.

[12] Obenchain, R. L. (2022a), "Efficient Generalized Ridge Regression", *Open Statistics* 3(1), 1−18. https://www.degruyter.com/document/doi/10.1515/stat-2022-0108/html

[13] Obenchain, R. L. (2022b), "Maximum Likelihood Ridge Regression", https://arxiv.org/abs/2207.11864v1

[14] Obenchain, R. L. (2023), "*RXshrink*: Maximum Likelihood Shrinkage using Generalized Ridge or Least Angle Regression Methods", ver 2.3, https://CRAN.R-project.org/package=RXshrink