m-Extent of Shrinkage: The Horizontal Axis on Ridge TRACE Diagnostic Plots

Robert L. Obenchain, Risk Benefit Statistics, http://localcontrolstatistics.org

December 2022

Abstract

We discuss five main reasons for plotting Generalized Ridge Regression *TRACE diagnostics* using an index, **m**, on its horizontal axis that measures the current "extent of shrinkage". In particular, we illustrate why **m** can be meaningful called the "multicollinearity allowance" (rank deficiency) when analyzing "ill-conditioned" linear models fit using *confounded* (correlated) $X-$variables.

1 Introduction

Generalized Ridge Regression (GRR) estimators apply various forms of "regularization" to the process of fitting a linear model to data containing ill-conditioned (possibly confounded) $X-$variables. Different forms of GRR typically use rather different "shrinkage-paths". All paths start at the ordinary least squares (OLS) estimator of regression $\hat{\beta}_0$ coefficients and end at the shrinkage "terminus", which is usually $\hat{\beta} \equiv 0$, the vector of all zeros. The fundamental objective of shrinkage in GRR is to exploit *variance-bias trade-offs* that can reduce the Mean Squared Error (MSE) Risk associated with an optimal form and extent of *shrinkage* applied to $\hat{\beta}_0$, the vector of OLS regression coefficient estimates.

The "ordinary" ridge path of Hoerl and Kennard (1970) possesses the dual properties that it contains not only [i] the shortest $\hat{\beta}-$vectors with any given Residual Sum-of-Squares (RSS) but also [ii] the smallest RSS for any given length of the $\hat{\beta}-$vector of estimates. Unfortunately, the overall "length" of the $\hat{\beta}-$vector is really not a key determinant of the MSE risk matrix of an estimator. What is important is applying just the "right" amount of shrinkage to each of the "uncorrelated components" of the OLS estimator; see equation (2) on page 3.

Three Main Types of Ridge Shrinkage Path

- A highly-flexible new "Efficient" Maximum Likelihood GRR path, Obenchain(2021), was recently implemented via the **eff.ridge()** function in Versions 2.1 and 2.2 of the *RXshrink* **R-**package, Obenchain(2022).

- The $q-$shape paths of Obenchain(1975) and Goldstein and Smith(1974) include the Hoerl and Kennard(1970) "ordinary" ridge path as the special case, $q = 0$. The **qm.ridge()** function in *RXshrink* can search a specified lattice of plausible values to determine a "most-likely" $q-$shape. Unfortunately, these 2$-$parameter paths (i.e. $q-$shape and $m-$extent) typically fail to pass through the overall Maximum Likelihood GRR estimate of regression $\beta-$coefficients in $p-$dimensional $X-$space whenever $p > 2$.

- The aug.lars() and uc.lars() functions within the *RXshrink* library focus on the GRR paths within the "least angle regression" family of lars() estimators, Hastie and Efron (2013). The "TRACE-like" plots produced by lars() **R-**functions are literally "backwards"; their left-hand starting point is $\hat{\beta} \equiv 0$, and they finish equaling OLS estimates at their right-hand end. These lars() paths can feature *selection* of subsets of $X-$variables (with non-zero coefficient estimates) when all other coefficient estimates are zeros. In other words, $X-$variables with NULL coefficient estimates can (correctly) be said to be of "no real value" in predicting observed $y-$outcomes!

In order to be able to **compare** and **contrast** this rather wide variety of different types of GRR shrinkage-paths, it is important (if not essential) for GRR software to produce "ridge TRACE diagnostic" plots with *standardized characteristics*. For example, it is quite helpful to display **full paths** that start at the OLS $\hat{\beta}_0$ estimate and extend all of the way to $\hat{\beta} \equiv 0$. In particular, to expedite comparisons among different paths, it really does help to scale the horizontal axis of all TRACE diagnostic plots the *same way*. This common scaling is provided by the *multicollinearity allowance* $m-$extent of shrinkage defined below in equation (6).

2 GRR Estimation Basics

Linear models and the OLS estimator, $\hat{\beta}_0$, of $\beta-$coefficients can be placed in a canonical form that is easy to generalize when defining GRR estimators. We assume that the $y-$outcome vector has been both centered and re-scaled to have an observed mean of zero and variance 1, that each column of the $X-$matrix $(n \times p)$ has been standardized in this same way, and that the resulting $X-$matrix has full (column) *rank* p that is ≥ 2 and $\leq (n-1)$. Recalling that the OLS fit corresponds to an orthogonal projection in the $n-$dimensional space of

individual observations onto the column-space of X, we can write the following well-know matrix-expressions:

$$\hat{y} = HH'y = X\hat{\beta}_0 \ . \tag{1}$$

These results follow by writing the singular-value decomposition (SVD) of X as $X = H\Lambda^{1/2}G'$. HH' then denotes an orthogonal projection; it is a $n \times n$ symmetric and idempotent matrix of rank p known as the **Hat-matrix** for OLS. In particular, $\hat{\beta}_0 = Gc$ where G represents an orthogonal "rotation" within the column-space of X, and $c = \Lambda^{-1/2}H'y$ is the $p \times 1$ column vector containing the *uncorrelated components* of $\hat{\beta}_0$, Obenchain (1975).

Our interest here is in generalized ridge regression (GRR) estimators that apply a given scalar-valued *shrinkage-factor*, δ_j, to each of the p uncorrelated components, c_j, of $\hat{\beta}_0 = Gc$. A key restriction here is that $0 \leq \delta_j \leq 1$ for $j = 1, 2, ..., p$. The resulting GRR estimators are then of the form:

$$\text{shrunken } \hat{\beta} = G\Delta c = \sum_{j=1}^{p} g_j \delta_j c_j \ , \tag{2}$$

where Δ denotes the diagonal matrix containing the p given shrinkage δ_j-factors and g_j denotes the j^{th} column of G.

In practical applications of GRR estimation, a important issue is whether the δ_j-shrinkage factors in equation (2) can be realistically viewed as *given constants* when, in reality, they actually are either informal user "choices" that depend upon the observed $y-$outcome vector or else specific (non-linear) functions of y. Janson, Fithian and Hastie (2013) comment on issues in determining the *degrees of freedom* of GRR estimates and support the arguments given here in Section §3.

2.1 The Hoerl-Kennard "ordinary" Ridge Path

This is a 1-parameter Path where the j^{th} shrinkage $\delta-$factor is of the form:

$$\delta_j = \lambda_j/(\lambda_j + k) \ , \tag{3}$$

where k denotes a non-negative scalar *constant*. I called k the "additive eigenvalue inflation factor" in Obenchain (1977). While several early authors remarked that k should be "small," it is clear that $k = +\infty$ is needed to shrink the $\hat{\beta}-$vector to 0 when at least one the $X'X$ eigenvalues, (λ_j), is strictly positive.

To display a Hoerl-Kennard "Ordinary" ridge TRACE plot like Figure (1), a finite upper $k-$limit, k^{max}, must be specified by the regression practitioner. I recently read a guideline suggesting use of $k^{max} = 1$, but we will see below that optimal GRR on the "Portland cement" data requires a $k-$value of at least ≈ 2.5.

Empirical guidelines from numerous early simulation studies (1970-1980) were somewhat ambiguous about appropriate choices for k^{max} in TRACEs like Figure (1). After all, the

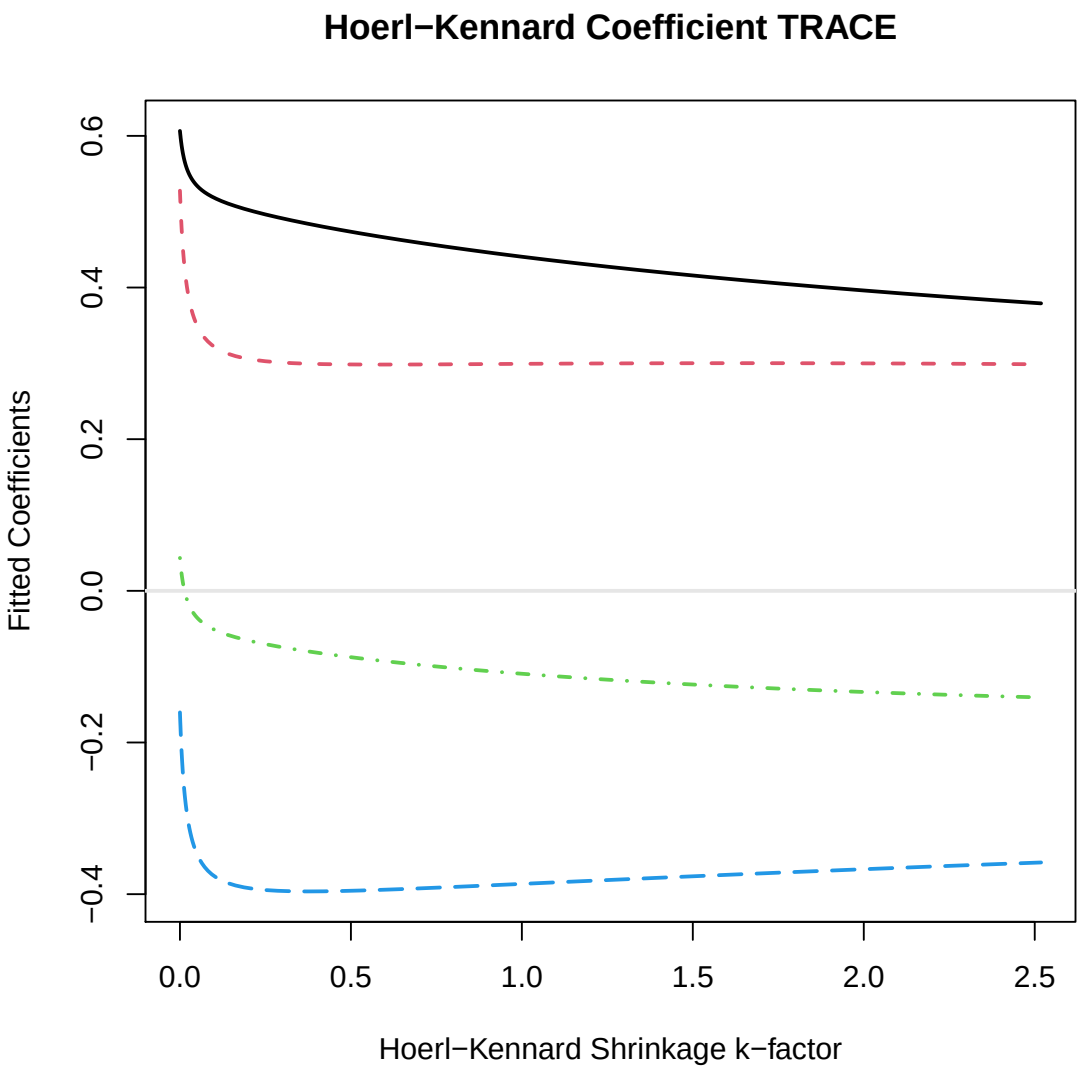

Figure 1: GRR Coefficient estimates for the Portland cement data with $p = 4$ $X-$variables. This *TRACE* displays the Hoerl-Kennard (1970) "ordinary" ridge path ($q = 0$) as a function of their $k-$parameter over the restricted range $0 \leq k \leq 2.5$. Note that this type of display always ends **abruptly** ...falling well short of $\hat{\beta} \equiv 0$ at the true GRR shrinkage terminus.

algorithms for choice of k that they studied used at least two different sources of information: [1] properties of the given ill-conditioned $X-$matrix, and [2] (non-linear) δ_j-estimates derived from an observed (stochastic) $y-$outcome vector.

2.2 The 2-parameter Goldstein-Smith Paths

Obenchain (1975) used the 2-parameter family of Paths initially proposed by Goldstein and Smith (1974) to illustrate ML estimation methods for GRR. The j^{th} shrinkage $\delta-$factor for these Paths is of the form:

$$\delta_j = \lambda_j/(\lambda_j + k\lambda_j^q) = 1/[1 + k\lambda_j^{(q-1)}] \ , \tag{4}$$

where the choice $q = 0$ corresponds to the Hoerl-Kennard "ordinary" Path, while $q = 1$ yields **uniform-shrinkage** ($\delta_1 = \delta_2 = \ldots = \delta_p$). Since the computed eigenvalues of the $X'X-$matrix are deliberately ordered ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$), it follows that $q-$shaped shrinkage factors are monotone-decreasing ($\delta_1 \geq \delta_2 \geq \ldots \geq \delta_p > 0$) when q is less than $+1$, but monotone-increasing when q exceeds $+1$. The shapes of traditional $1-$ or $2-$parameter paths are almost "predetermined" by the eigenvalues of the X$-$matrix, a disadvantage pointed out by Hoerl and Kennard (1975). However, "optimal" shrinkage $\delta-$factors along principal axes are rarely strictly monotone (increasing or decreasing) in their ordered eigenvalues!

2.3 The p-parameter "Efficient" Path

Obenchain (2021) proposed a new GRR Path that heads directly towards and always passes directly through the $\beta-$coefficient estimate that has Maximum Likelihood, under normal distribution-theory, of achieving minimum MSE Risk. The j^{th} shrunken estimate is

$$\hat{\delta}_j^{ML} \cdot c_j = \frac{n \cdot \hat{\rho}_j^3}{n \cdot \hat{\rho}_j^2 + (1 - R^2)} \cdot \sqrt{\frac{y'y}{\lambda_j}} \ . \tag{5}$$

Thompson (1968) studied a non-linear estimator somewhat like this using numerical integration. Note that the $\delta_j \times \hat{c}_j$ numerator terms in equation (5) are "cubic" in $\hat{\rho}_j$ while the denominator can be shown to be "quadratic" in all p of its $\hat{\rho}$ terms!

3 Quantifying Extent of Shrinkage

To illustrate *TRACE* diagnostic plots of GRR estimates, we use the "Portland cement" data of Woods, Steinour and Starke (1932) that were featured in Hald (1952). Our linear regression model uses $p = 4$ confounded $X-$ingredient percentages to predict the $y-$Outcome: "heat (cals/gm) evolved during hardening" for $n = 13$ cement mixtures. Note that the

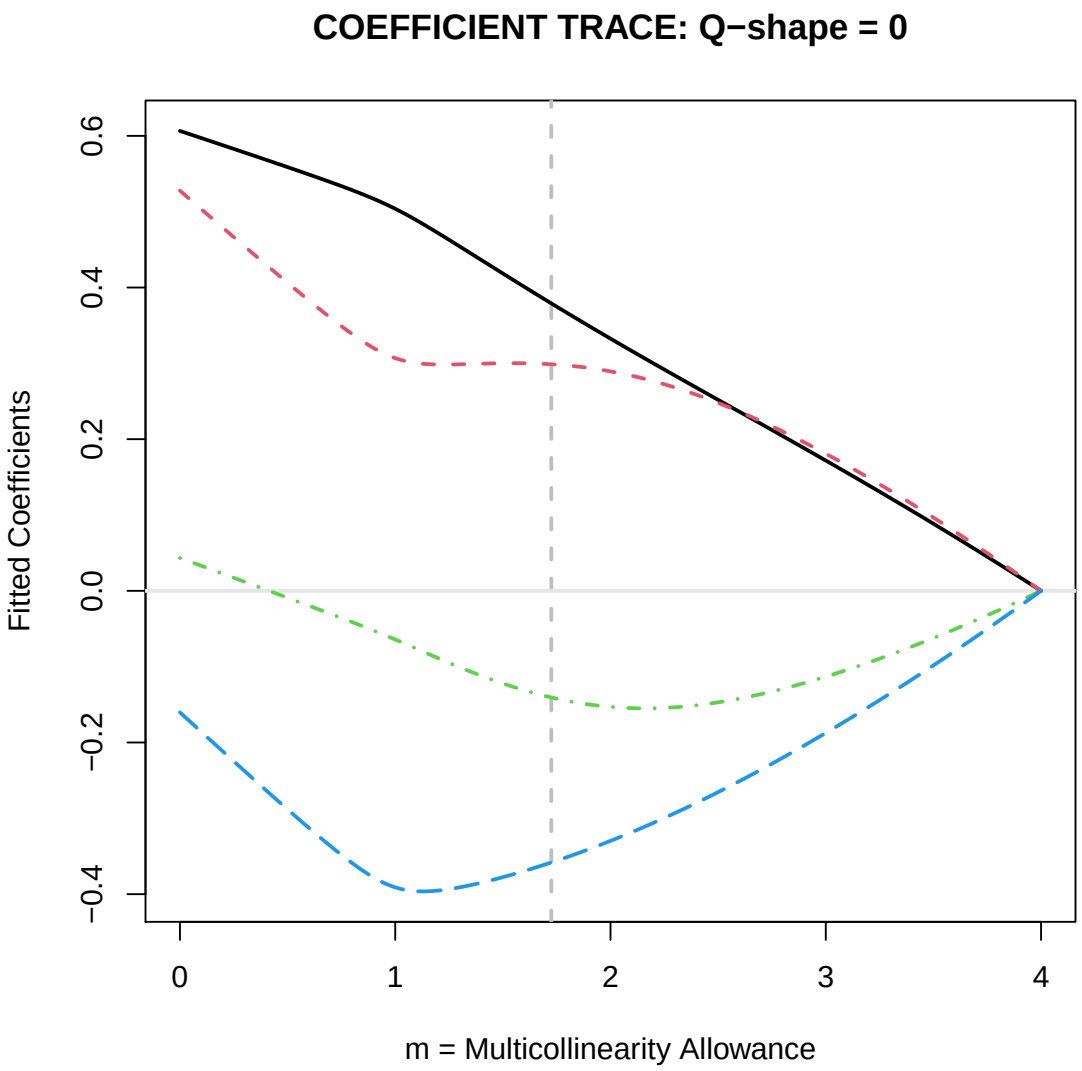

Figure 2: GRR Coefficient estimates for the Portland cement data utilizing $m-$extent of Shrinkage on its horizontal axis. Although hardly recognizable as such, this *TRACE* actually begins with the same $\hat{\beta}-$estimates as in Figure (1) for $m \leq 1.73$. But this TRACE continues from that point onward to $\hat{\beta} = 0$ at $m = p = 4$ ...the upper $m-$limit as k approaches $+\infty$.

4 Maximum Likelidood $\delta-$factor estimates are $(0.9986, 0.07430, 0.9266,$ and $0.1528)$ in this example. Only one of the eigenvalues of $X'X$ $(26.8, 18.9, 2.24, 0.0195)$ is "small". But the 2^{nd} "uncorrelated component" of the OLS solution is not only numerically "small" but also statistically "insignificant" in this example.

3.1 Interpreting m as a Measure of "Rank Deficiency"

The *multicollinearity allowance* parameter, m, defines the "extent" of shrinkage applied by equation (2) as being

$$m = p - \delta_1 - \cdots - \delta_p = rank(X) - trace(\Delta), \tag{6}$$

where $0 \leq m \leq p$, Obenchain (1977).

In practical applications of linear models, the given $X-$matrix typically has more rows, n, than columns, p, and to have "full" column-rank p. These conditions (or assumptions) assure that all p elements of the OLS estimator, $\hat{\beta}_0$, are uniquely determined.

In these cases, the *trace* of the GRR Hat-matrix, $H\Delta H'$ in (2) is $trace(\Delta)$ because $H'H = I$. Thus, m can definitely be interpreted as a measure of *inferred* "rank deficiency" in the given $X-$matrix whenever the set of $\Delta-$factors used in estimation are maximum likelihood (ML) estimates under normal-theory of the *unknown true* minimum MSE risk $\delta-$factors, Obenchain (1975, 1978).

Use of this $m-$scale for displaying *TRACE diagnostics* also suggests using the short-hand notation, $\hat{\beta}_m$, to denote individual $\hat{\beta}$ GRR point-estimates in equation (2). Note that the OLS estimate, $\hat{\beta}_0$, always occurs at the beginning of a TRACE plot $(m = 0)$, while the shrinkage terminus, $\hat{\beta}_p \equiv 0$, always occurs at the very end $(m = p)$.

3.2 The "Finite Range" Restriction

Since the width (and height) of a plot must be finite, Figure (2) is more complete (and meaningful) than Figure (1) simply because it displays a "full" shrinkage Path rather than just the very beginning of one. Since the (euclidean) distance between the OLS estimate, $\hat{\beta}_0$, and $\hat{\beta}_p \equiv 0$ is finite, it strikes me as extremely counter-productive to use an unbounded parameter, like k, to index a TRACE plot of finite width. Unfortunately, this is precisely what Hoerl and Kennard (1970) originally proposed.

3.3 Stable Relative Magnitudes

Shrunken regression coefficients with *perfectly stable relative magnitudes* form (perfectly) "straight lines" in TRACE plots using $m-$scaling.

As noted above for the Portland cement example where m ranges from 0 to 4, there are multiple reasons to expect the "effective rank" of the ill-conditioned $X-$matrix to be roughly $m = 2$. Thus one expects appropriately shrunken GRR estimates of coefficients to display "stable" relative magnitudes starting at about $m = 2$. This expectation is supported rather "weakly" by Figure (2) and much more fully by Figure (3).

3.4 Bayesian Posterior Precision

For any given value of m, the average value of all p shrinkage $\delta-$factors is $(p - m)/p$, which is the Theil(1963) proportion of Bayesian posterior precision due to *sample information* ...rather than due to *prior information*. This observation supports an empirical Bayes interpretation for the effects of shrinkage. [This was first called to my attention by Rick Vinod, an Economist and colleague at AT&T Bell Laboratories.]

In particular, note that this proportion of posterior precision *decreases linearly* as m increases. The OLS solution, $\hat{\beta}_0$, at $m = 0$ gets ALL of its posterior precision from sample information. Furthermore, $\hat{\beta}_p \equiv 0$ at the shrinkage terminus uses NO sample information.

3.5 Generality

One final reason for plotting a TRACE against m is perhaps the *most obvious reason*. Hoerl and Kennard (1970), bottom of page 63, stated that "There is no graphical equivalent to the RIDGE TRACE" ...when the additive eigenvalue inflation-factors are not all equal. However, there certainly is a graphical **improvement!** Any GRR family or shrinkage Path can be displayed in a TRACE diagnostic that uses $m-$scaling along its horizontal axis.

4 Summary

Since the range of the $m-$index of equation (6) is always finite, this $m-$scale is "ideal" for use as the horizontal axis on all types of *TRACE diagnostic* plots [coefficients, relative MSEs, excess risk eigenvalues, inferior direction cosines and $\delta-$factors].

When linear models are fit to ill-conditioned or confounded *narrow-data*, *TRACE* diagnostics are useful in demonstrating and justifying deliberately *biased* estimation. This makes *TRACE diagnostics* powerful "visual" displays for use in training of advanced students and persuasion of all people capable of elementary *statistical thinking*.

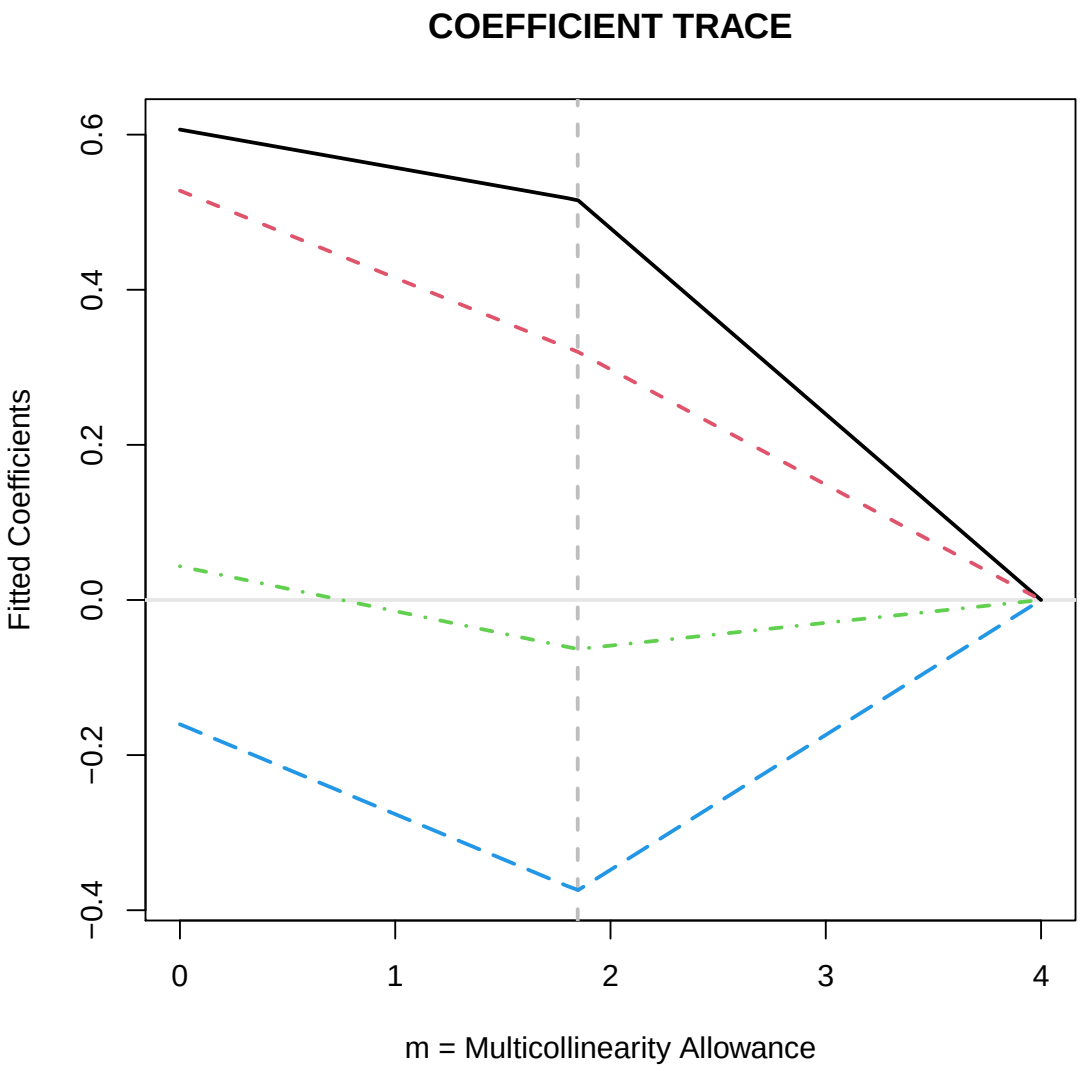

Figure 3: GRR Coefficient estimates for the 4-parameter "Efficient" Path on the Portland cement data. Thanks to its "Standardized" $m-$scaling, this *TRACE* looks somewhat like that of Figure (2). But there are some relatively "small" but *important* differences. For example, the vertical dashed-line marking the estimates with minimum MSE risk occurs here at $m = 1.85$ rather than $m = 1.73$. Furthermore, all coefficient estimates fall on a *two-piece linear function* with a single "interior" Knot also at $m = 1.85$. Finally, coefficient relative-magnitudes are *perfectly stable* over the "final" range: $1.85 < \mathrm{m} < 4$.

5 References

Hald A. (1952). *Statistical Theory with Engineering Applications*. (Page 647.) New York; Wiley

Hastie, T. and Efron, B. (2013). "Least Angle Regression, Lasso and Forward Stagewise." ver 1.2, `https://CRAN.R-project.org/package=lars`

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

Hoerl, A. E. and Kennard, R. W. (1975). "A Note on a Power Generalization of Ridge Regression." *Technometrics* 17, 269.

Goldstein, M. and Smith, A. F. M. (1974). "Ridge-type estimators for regression analysis." *J. R. Statist. Soc.*, B, 36, 284−291.

Lucas Janson, Will Fithian and Trevor Hastie. (2013). "Effective Degrees of Freedom: A Flawed Metaphor." Stanford, Statistics Department.

Obenchain, R. L. (1975). "Ridge analysis following a preliminary test of the shrunken hypothesis." *Technometrics* 17, 431−441. DOI: 10.1080/00401706.1975.10489369

Obenchain, R. L. (1977). "Classical F-tests and confidence regions for ridge regression." *Technometrics* 19, 429−439. DOI: 10.1080/00401706.1977.10489582

Obenchain, R. L. (1978). "Good and optimal ridge estimators." *Annals of Statistics* 6, 1111−1121. DOI: 10.1214/aos/1176344314

Obenchain, R. L. (2016). "Intro to Regression Shrinkage Concepts." Select Main-Menu item: **Shrinkage**... `http://localcontrolstatistics.org`

Obenchain, R. L. (2021). "The Efficient Shrinkage Path: Maximum Likelihood of Minimum MSE Risk." `http://arxiv.org/abs/2103.05161`

Obenchain, R. L. (2022). "Efficient Generalized Ridge Regression." *Open Statistics* 3, 1−18.

Obenchain, R. L. (2022). "*RXshrink*: Maximum Likelihood Shrinkage using Generalized

Ridge or Least Angle Regression." Version 2.2. https://CRAN.R-project.org/package=RXshrink

McDonald, G. C. (1975). Discussion of "Ridge analysis following a preliminary test of the shrunken hypothesis." *Technometrics* 17, 443−445.

Theil, H. (1963). "On the use of incomplete prior information in regression analysis." *Journal of the American Statistical Association* 58, 401−414.

Vinod, H. D. (1976). "Applications of new ridge regression methods to a study of Bell System scale economies." *Journal of the American Statistical Association* 71, 835−841.

Woods, H., Steinour, H. H., and Starke. H. R. (1932). "Effect of composition of Portland cement on heat evolved during hardening." *Industrial Engineering and Chemistry* 24, 1207−1214.