shrink: a vignette for the RXshrink R-package…

Generalized Ridge and Least Angle Regression

Version 1.0-7, December 2011 … License: GPL (>= 2)

"Personal computing treatments for your data analysis infirmities …since 1983"

Bob Obenchain
Principal Consultant, Risk Benefit Statistics LLC
13212 Griffin Run, Carmel, IN 46033-9935
317-580-0144, wizbob@att.net
website: http://members.iquest.net/~softrx

Table of Contents

1. A Personal Summary of 50+ Years of "Shrinkage in Regression"

As someone who has been fascinated with the possibility that shrunken regression coefficient estimates can reduce MSE risk via variance-bias trade-offs and who has conducted and published research in this area starting in the 1970s, I must say that I am absolutely delighted by the recent wide-spread tolerance for (if not outright acceptance of) shrinkage methods. Anyway, I wish to summarize here some personal perspectives on why and how professional statisticians may have become somewhat enlightened about shrinkage over the last 50+ years …since ~1955.

Early optimism about a theoretical basis for and the practical advantages of shrinkage almost surely started with the work of Stein(1955) and James and Stein(1961). Unfortunately this shrinkage was always "uniform," thus really doing nothing to adjust the relative magnitudes of correlated regression coefficient estimates for ill-conditioning. Furthermore, although an overall improvement in the scalar value of "summed MSE risk" was guaranteed, there was no way to know "where," in an **X**-space of 3 or more dimensions, risk was actually being reduced. In fact,

researchers on normal-theory minimax estimation in regression [such as Strawderman(1978) and Casella(1980,1985)] found that, when a desired "location" for improved risk was specified, their estimates succeeded only by concentrating shrinkage somewhere else! Actually, the earlier work of Brown (1975) and Bunke(1975a, 1975b), was really the beginning of the end for minimax research. After all, only OLS estimation can be minimax when one's risk measures are truly **multivariate** (matrix rather than scalar valued.) I personally would like to think that modern researchers and regression practitioners view shrinkage estimators as attractive, practical alternatives to OLS estimation in ill-conditioned models even though there cannot be any truly meaningful way to uniformly "dominate" OLS on MSE risk.

On the other hand, the real gold-rush of interest in (non-uniform) shrinkage in regression is undoubtedly due to the pioneering "ridge" work of Hoerl (1962) and Hoerl and Kennard (1970a, 1970b.) Some of their terminology was misleading (e.g. their "too longness" argument was actually based upon a simple measure of coefficient variability), and their conjectures that it should be "easy" to pick shrunken estimators from a graphical trace display that would have lower MSE risk than OLS were, in fact, unquestionably naïve.

Meanwhile, a major frustration for me, personally, was that my shrinkage work at AT&T Bell Labs lead to open conflict with John Tukey. This unfortunate turn of events started when my management learned that Tukey had been consistently disparaging shrinkage methods at professional meetings in the 1970s and culminated when we were asked to formally comment on each other's 1975 papers and on my Bell Labs internal "regression training" materials. Luckily for me, Colin Mallows ultimately intervened, and cooler heads prevailed.

The most widely accepted forms of shrinkage in regression today are undoubtedly the random coefficient BLUP estimates from Henderson's mixed model equations, as implemented in SAS proc mixed and the lme( ) and nlme( ) R functions. See Robinson (1991), Littel, Milliken, Stroup and Wolfinger(1996) and Pinheiro and Bates(1996).

Looking back upon my personal contributions to the literature on shrinkage in regression, I can only lament that my writings lacked focus and simplicity. I clearly love details, myself, and my papers have always been chuck-full of many-too-many alternative concepts. For example, my 1975 invited paper in **Technometrics** might have had more positive impact if I had only picked a better title! With some minor changes in emphasis, that paper could have easily been, say, "Maximum Likelihood Shrinkage in Regression." Instead, this work became identified with both "ridge analysis" (as averse the ridge regression) and "preliminary-test estimation" …and rightfully remains obscure today. I guess practitioners do not really want (or need) an extremely powerful statistical test for ill-conditioning! After all, in practical applications of regression, the presence of at least some ill-conditioning tends to be more of a rule than an exception.

Next, I became sufficiently frustrated by the **Technometrics** refereeing process on a second shrinkage paper (publication delayed until 1977) that I decided to submit a third manuscript (with important implications for practical applications of shrinkage) to **Annals of Statistics**. Some agonizing delays again occurred, and that publication was delayed until 1978. This annals paper presented the "ridge function theorem," the "excess mean squared error matrix," the

"inferior direction," and the "2/P-ths rule of thumb" for limiting shrinkage …plus their individual Maximum Likelihood (ML) estimators for display in TRACE plots.

Unfortunately, while my basic shrinkage equations and theorems were published, I had much less success publishing descriptions of practical shrinkage applications, including "how-to" information about interpretations for my five types of TRACE plots. In fact, only three of my papers on shrinkage applicants or software, Obenchain(1984, 1991, 1995), were ultimately accepted for publication. As illustrated in Section §3 of this vignette, shrinkage TRACE displays reveal "where" MSE risk can be reduced by shrinkage.

Similarly, I also developed a closed form expression, Obenchain(1981), for the normal-theory ML estimator within the 2-parameter Goldstein and Smith (1974) shrinkage family. Unfortunately, none of my attempts to present that material in a peer-reviewed publication have succeeded. Closed form expressions speed shrinkage estimation and are particularly helpful when simulating MSE risk profiles.

My "bottom-line" on the topic of normal-theory ML shrinkage is simply this: The linear estimator identified as being most likely to be optimal is, in reality, a **nonlinear** estimator. The true MSE risk of this ML shrinkage estimator can be computed exactly in certain special cases and can always be accurately simulated. While having a MSE risk profile that is clearly not "dominant" like that of the unknown, optimal linear estimator, achievable ML shrinkage profiles can nevertheless be fairly impressive:

> In simple rank-one cases, ML shrinkage can reduce MSE risk by about 50% in favorable cases (with low signal and/or high uncertainty) while increasing risk by at most 20% in unfavorable cases.
>
> In high-dimensional situations, a savings of more than 50% is possible, and worst case situations result in an increase of less than 5% in MSE risk.

As Burr and Fry(2005) have noted, the key strategy and/or tactic in shrinkage estimation is definitely to be "cautious" rather than "greedy."

Frank and Freidman(1993), Breiman (1995), Tibshirani (1996), LeBlanc and Tibshirani (1998) and Efron et al. (2004) are currently keeping the shrinkage regression "home fires" burning for exploratory analyses of gigantic datasets.

2. Introduction to Shrinkage Regression Concepts and Notation

The following formulas define the ***Q*** "shape" and the ***k*** "extent" of shrinkage yielding 2-parameter generalized ridge regression estimators.

$$\beta^{*} = [\, X'X + k \times (X'X)^{Q} \,]^{-1} X'y$$

Our first formula, above, represents the 2-parameter family using notation like that of Goldstein and Smith(1974). Here we have assumed that the response vector, **y**, and all p columns of the (nonconstant) regressors matrix, **X**, have been "centered" by subtracting off the observed mean value from each of the n observations. Thus Rank(**X**) = r can exceed neither p nor (n−1).

Insight into the form of the shrinkage path that results as k increases (from zero to infinity) for a fixed value of ***Q*** is provided by the "singular value decomposition" of the regressor **X** matrix and the corresponding "eigenvalue decomposition" of **X'X**.

$$X = H\Lambda^{+1/2}G'$$

$$(X'X)^{Q} = G\Lambda^{Q}G'$$

The **H** matrix above of "regressor principal coordinates" is (n by r) and semi-orthogonal (**H'H** = **I**.) And the **G** matrix of "principal axis direction cosines" is (p by r) and semi-orthogonal (**G'G** = **I**.) In the full-column-rank case (r = p), **G** is orthogonal; i.e. **GG'** is then also an identity matrix.

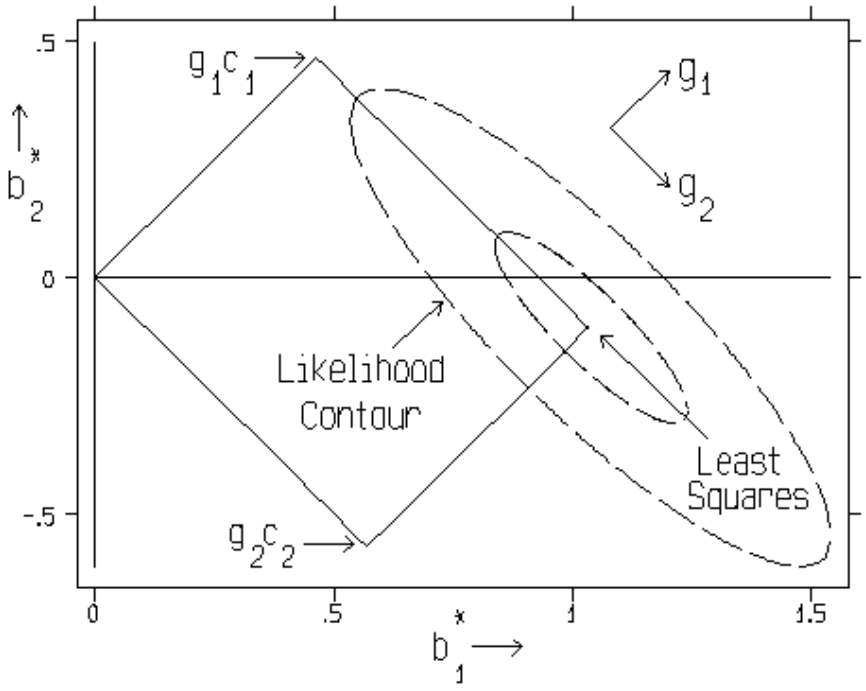
The (r by r) diagonal "Lambda" matrix above contains the **ordered** and **strictly positive** eigenvalues of **X'X**; $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$. Thus our operational rule for determining the ***Q***-th power of **X'X** (where ***Q*** may not be an integer) will simply be to raise all of the positive eigenvalues of **X'X** to the ***Q***-th power, pre-multiply by **G**, and post-multiply by **G'**.

Taken together, these decompositions allow us to recognize the above 2-parameter (**k** and ***Q***) family of shrinkage estimators, β^* (beta-star), as being a special case of r-dimensional generalized ridge regression...

$$\beta^* = G\Delta\Lambda^{-1/2}H'\,y = G\Delta c$$

where the (r by r) diagonal **Δ** matrix contains the multiplicative **shrinkage factors** along the **r** principal axes of **X**. Each of these Delta(i) factors range from 0 to 1 (i = 1, 2, ..., r.)

Note that the (r by 1) column vector, **c**, contains the **uncorrelated components** of the ordinary least squares estimate, beta-hat = $\mathbf{G\,c} = \mathbf{g}_1 c_1 + \mathbf{g}_2 c_2 + \ldots + \mathbf{g}_r c_r$ of the unknown, true regression coefficient β vector. The variance matrix of **c** is the diagonal $\mathbf{\Lambda^{-1}}$ matrix times the scalar value of the error sigma-square. The P = r = 2 dimensional case is depicted below.



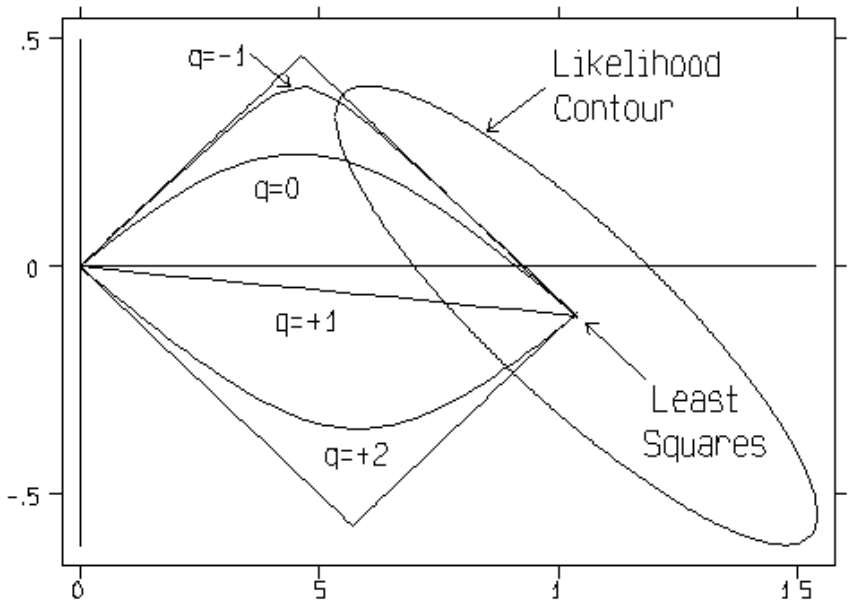
In fact, we now see that the 2-parameter family of shrinkage estimators from our **first equation**, above, is the special case of the **last equation** in which...

$$\delta_i = \frac{\lambda_i}{\left(\lambda_i + k \cdot \lambda_i^{Q}\right)} = \frac{1}{\left(1 + k \cdot \lambda_i^{Q-1}\right)}$$

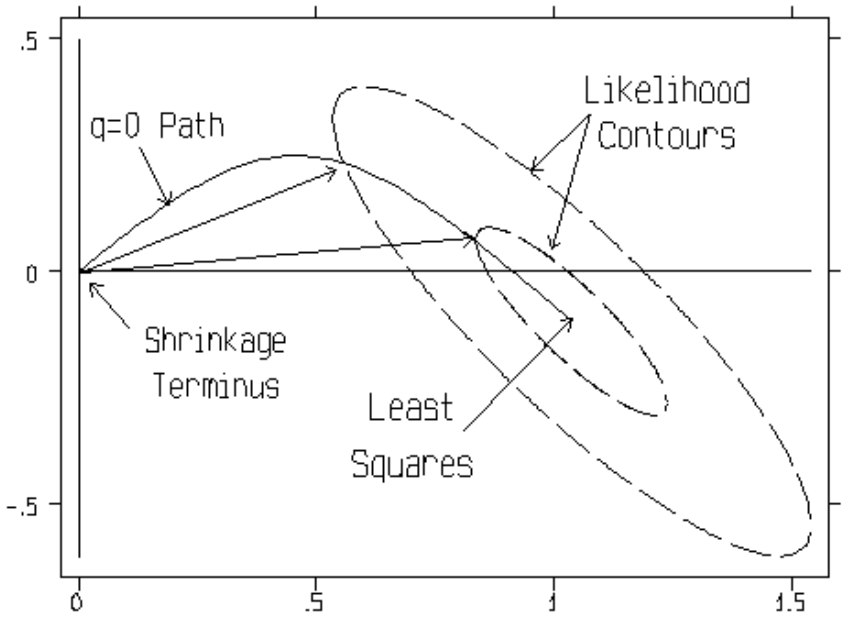
Q = the ridge parameter that controls the "shape" (or "curvature") of the ridge path through regression coefficient likelihood space.

Q = +1 ...yields uniform shrinkage (all Shrinkage Factors equal.)
Q = 0 ...yields Hoerl-Kennard "ordinary" ridge regression.
Q = –5 ...is usually very close, numerically, to "Principal Components Regression," with exact agreement in the limit as ***Q*** approaches minus infinity.

The display below shows a variety of shrinkage path ***Q*-shapes** for the rank(**X**) = p = 2 case.



The best known special case of a ***Q***-shaped path is probably ***Q*** = 0 for Hoerl-Kennard(1970) **"ordinary" ridge regression**. This path has a dual "characteristic property," illustrated in the figure below. Namely, the ***Q*** = 0 path contains not only the shortest beta estimate vector of any given likelihood but also the most likely beta estimate of any given length.



Another well known special case of a ***Q***-shaped path is ***Q*** = +1 for **uniform** shrinkage. The coefficient trace and shrinkage factor trace for this path are both rather "dull," but the estimated risk and inferior direction TRACES can still be interesting even when ***Q*** = +1.

Again, an extremely important limiting case is ***Q*** = minus infinity for **principal components regression**. [Marquardt(1970) called this limit "assigned rank" regression.] My experience is that the ***Q*** = –5 path is frequently quite close, numerically, to this limiting case. Note in the figure at the top of page 6 that the path with ***Q*** = –1 shape is already near the limit in the p = 2 dimensional case depicted there.

2.1 The m = *MCAL* = "multicollinearity allowance" parameter

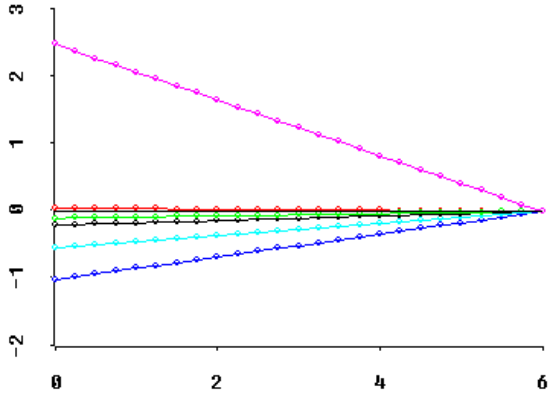
Unfortunately, the "**k**" parameter is really not a very good measure of the **extent** of shrinkage. After all, the sizes of the **r** shrinkage factors, **Δ**, can depend more on one's choice of ***Q*** than on one's choice of **k**. Specifically, the **k**–values corresponding to two rather different choices of ***Q*** are usually **not** comparable.

Thus my shrinkage regression algorithms use the **m** = ***MCAL*** = "multicollinearity allowance" parameter of Obenchain and Vinod(1974) to index the **M-extent of Shrinkage** along paths. This parameter is defined as follows:

$$MCAL = r - \delta_1 - \delta_2 - \ldots - \delta_r = Rank(X) - Trace(\Delta)$$

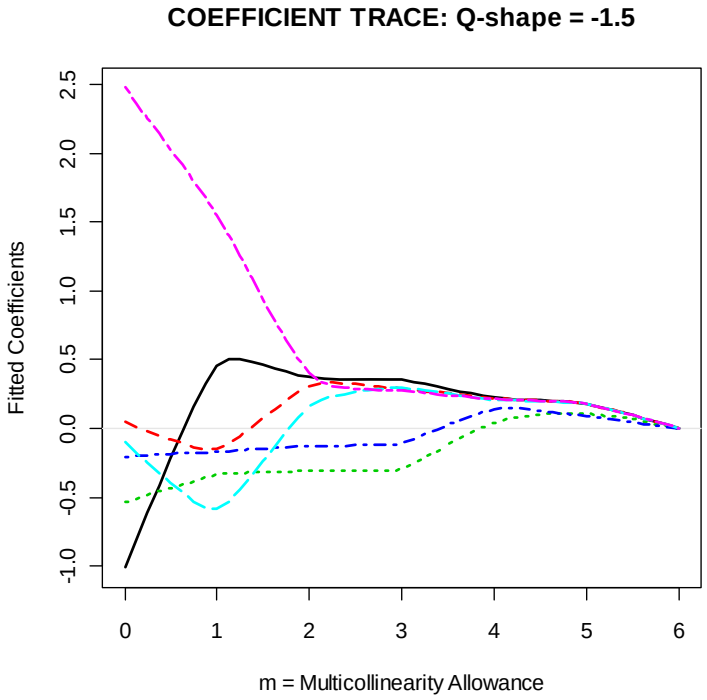
Note that the range of ***MCAL*** is finite; ***MCAL*** ranges from **0** to **r** = Rank(**X**), inclusive. Whatever may be your choice of ***Q***-shape, the OLS solution always occurs at the beginning of the shrinkage path at ***MCAL*** = 0 (**k** = 0 and **Δ** = **I**) and the terminus of the shrinkage path, where the fitted regression hyperplane becomes "horizontal" (slope=0 in all p-directions of **X**-space) and ***y***-hat = ***y***-bar, always occurs at ***MCAL*** = **r** (k = +∞ and **Δ** = **0**). `RXridge()` uses Newtonian descent methods to compute the numerical value of **k** corresponding to given values of ***MCAL*** and ***Q***-shape.

In addition to having finite (rather than infinite) range, ***MCAL*** has a large number of other advantages over **k** when used as the scaling for the horizontal axis of ridge trace displays. For example, shrunken regression coefficients with stable relative magnitudes form **straight lines** when plotted versus ***MCAL***.



Similarly, the average value of all **r** shrinkage factors is **(r – MCAL)/r**, which is the Theil(1963) proportion of Bayesian posterior precision due to sample information (rather than to prior information.) Note that this proportion decreases linearly as ***MCAL*** increases.

Perhaps most importantly, ***MCAL*** can frequently be interpreted as the ***approximate deficiency in the rank of X.*** For example, if a regressor **X'X** matrix has only two relatively small eigenvalues, then the coefficient ridge trace for best ***Q***-shape typically "stabilizes" at about ***MCAL*** = 2. This situation is illustrated below using the ridge coefficient trace for the path of shape ***Q*** = –1.5 for the original Longley(1967) dataset where the response is ***y*** = Employed. Compared with the major initial shifts in relative magnitudes and numerical signs of coefficients between ***MCAL*** = 0 and ***MCAL*** = 2, note that the trace below becomes relatively much more stable (somewhat "straight") between ***MCAL*** = 2 and ***MCAL*** = r = 6.



As a general rule-of-thumb, paths with ***Q***-shapes in the [–1,+2] range generally tend to be fairly **smooth** ...i.e. have "rounded" corners. Paths with ***Q***-shapes greater that +2 or less than –1 can display quite "sharp" corners. In fact, the paths with limiting shapes of ±∞ are actually linear splines with join points at integer ***MCAL*** values!

My computing algorithms provide strong, objective guidance on the choice of the ***Q***-shape that is best for your data. Specifically, they implement the methods of Obenchain(1975, 1978, 1981) to identify the path Q-shape (and the ***MCAL***-extent of shrinkage along that path) which have **maximum likelihood** (under a classical, fixed coefficient, normal-theory model) of achieving overall minimum MSE risk in estimation of regression coefficients.

2.2 Shrinkage δ–factors for Least Angle and Lasso Estimators

The **`RXlarlso()`** and **`RXuclars()`** functions in the RXshrink R-package re-interpret lar and lasso regression estimators as generalized ridge estimators simply by solving equations such as

$$\beta^{lar} = G\,\Delta^{lar}\,c$$

for the implied **Δ**-factors. With the **i**th column of **G** again denoted by $\mathbf{g_i}$ (as in the figure on page 5), the solutions of the above **r** equations are

$$\delta_i^{lar} = \mathbf{g_i}'\beta^{lar} / \mathbf{c_i} \quad \text{for } i = 1, 2, \ldots, r.$$

Because these equations clearly do not constrain the resulting lar or lasso "delta-factors" to be non-negative and less than +1, the resulting estimates may have neither of these properties. In other words, lar and lasso estimators can correspond to "non-standard" generalized ridge estimators and, thus, can correspond to higher MSE risk than would be possible with a true "shrinkage" estimator.

On the other hand, the **`RXuclars()`** function applies lar estimation directly to the uncorrelated components vector, **c**, and this restriction yields a true generalized ridge (shrinkage) estimator. In fact, the delta-factors from **`RXuclars()`** will always then be of the following form:

$$\delta_i^{uclars} = \max[\,0,\,(\,1 - k / |\rho_i|\,)\,],$$

where ρ_i is the **i**th "principal correlation" …i.e. the correlation between the response **y**-vector and the **i**th column of the **H** matrix of "principal coordinates" of **X** (page 4.) Note that the k-factor in this shrinkage formulation is limited to a subset of [0, 1]. ***MCAL*** = 0 occurs at k = 0, while ***MCAL*** = r results when k is the maximum absolute principal correlation.

3. Interpretation of ridge TRACE Displays

We will use the **`longley2`** numerical example here in Section §3 to illustrate interpretation of ridge TRACE displays. These data, compiled by Art Hoerl using the 1976 "Employment and Training Report of the President," are an updated version of the infamous Longley(1967) dataset for benchmarking accuracy of regression computations. The **`longley2`** data.frame contains some slightly different numerical values from those used by Longley(1967) within the original 16 years (1947 through 1962) and also adds data for 13 subsequent years (1963 through 1975.)

Start by loading the **`RXshrink`** package, then execute the following R-code:

```
data(longley2)
form <- GNP~GNP.deflator+Unemployed+Armed.Forces+Population+Year+Employed
rxrobj <- RXridge(form, data=longley2)
rxrobj
```

Because **rxrobj** is an R-object of class **RXridge**, the fourth line of code prints the default **RXridge( )** output. This output is rather detailed and extensive, so it is abbreviated below.

Principal Axis Summary Statistics of Ill-Conditioning...

	LAMBDA	SV	COMP	RHO	TRAT
1	124.55432117	11.1603907	0.466590166	0.98409260	179.451944
2	34.04395492	5.8347198	-0.009779055	-0.01078296	-1.966301
3	7.97601572	2.8241841	0.228918857	0.12217872	22.279619
4	1.31429584	1.1464274	-0.557948473	-0.12088200	-22.043160
5	0.06505309	0.2550551	0.613987118	0.02959472	5.396677
6	0.04635925	0.2153120	-0.471410409	-0.01918176	-3.497845

COMP = 6 × 1 vector of Uncorrelated Components of the OLS estimator, $\mathbf{c} = \mathbf{G}'\beta^{o}$.
RHO = 6 × 1 vector of Principal Correlations between the response **y** and the columns of **H**. In this example, the first RHO is huge, and the other 5 are all relatively small.

Note that the ill-conditioning in this example is quite extreme. **The last three uncorrelated components are (numerically) the three largest.** This is the case only because the corresponding singular values, SV = sqrt(LAMBDA), are small. Again, the last three principal correlations are all quite small relative to the only large one, the first.

```
Residual Mean Square for Error = 0.0008420418
Estimate of Residual Std. Error = 0.02901796
```

Classical Maximum Likelihood choice of SHAPE(Q) and EXTENT(M) of shrinkage in the 2-parameter generalized ridge family...

	Q	CRLQ	M	K	CHISQ
1	5.0	0.03065132	5.973237	9.992836e+06	212.2772
...					
9	1.0	0.52547213	2.111210	5.428963e-01	202.9410
10	0.5	0.79341430	1.816359	4.358166e-01	183.5424
11	0.0	0.89070908	2.678418	1.513692e+00	166.6511
12	-0.5	0.93599740	3.140371	7.907552e+00	151.8817
13	-1.0	0.95935445	3.453422	5.035840e+01	139.1481
...					
20	-4.5	0.98439456	4.586356	3.768549e+08	112.1289
21	-5.0	0.98446554	4.729924	4.185069e+09	112.0005

Q = -5 is the path shape most likely to lead to minimum MSE risk because this shape maximizes CRLQ and minimizes CHISQ.

RXridge: Shrinkage PATH Shape = -5 ← **RXridge( ) choice of Q.**

The extent of shrinkage (M value) most likely to be optimal in the Q-shape = -5 two-parameter ridge family can depend upon whether one uses the Classical, Empirical Bayes, or Random Coefficient criterion. In each case, the objective is to minimize the minus-two-log-likelihood statistics listed below:

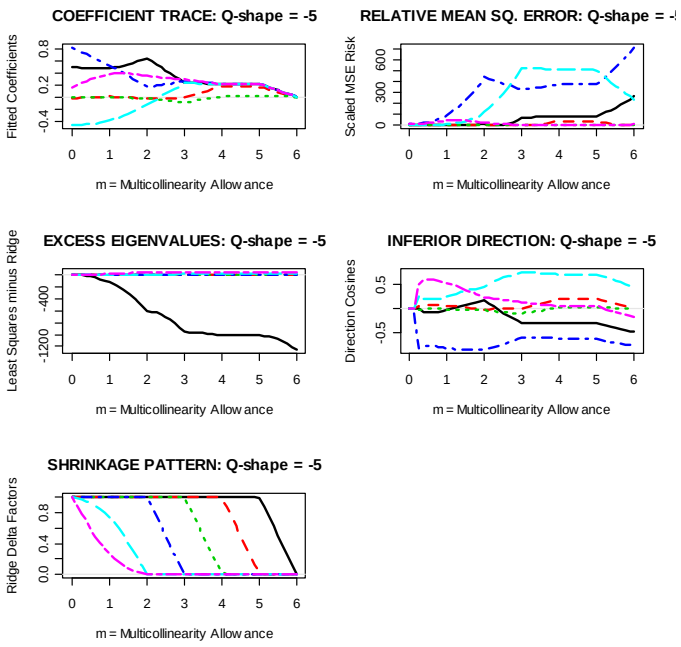
```
        M             K          CLIK         EBAY      RCOF
0   0.000 0.000000e+00           Inf          Inf       Inf
1   0.125 1.216886e-09 1.756397e+12     113.2484 113.7283
2   0.250 2.723817e-09 1.759946e+12     112.8258 113.6267
3   0.375 4.619196e-09 1.761921e+12     113.3184 114.2927
..
37  4.625 2.588153e+09 1.157462e+02    1056.0587 120.4012
38  4.750 4.641076e+09 1.121409e+02    1073.4679 120.1243
39  4.875 1.062368e+10 1.206503e+02    1124.3692 120.4956
..
47  5.875 2.615094e+13 2.083713e+02   29207.1835 208.6979
48  6.000          Inf 2.123044e+02   33230.5079 212.3044
```

Before abbreviation, the above listing described 49 choices for the M-extent of shrinkage (m = 0.0 to m = 6.0 in steps of 0.125.) The search over this lattice suggests that m = 4.750 minimizes the CLIK criterion; the earlier output using the normal-theory closed form expression suggested m = 4.7299, which is not on the lattice. No closed form expressions exist for the EBAY or RCOF criteria, but the lattice search suggests that m = 0.250 is best for these criteria, which is MUCH less shrinkage than suggested by the CLIK criterion!

Applying the "(2/P)ths Rule-of-Thumb" of Obenchain (1978) with P = 6, it follows that the most shrinkage likely to produce a "good" ridge estimator (better than OLS in every MSE sense) along the ***Q*** = –5 path for the **`longley2`** data is m = 1.58.

With all of the above background information in mind, it is now high time to examine and interpret ridge trace displays!

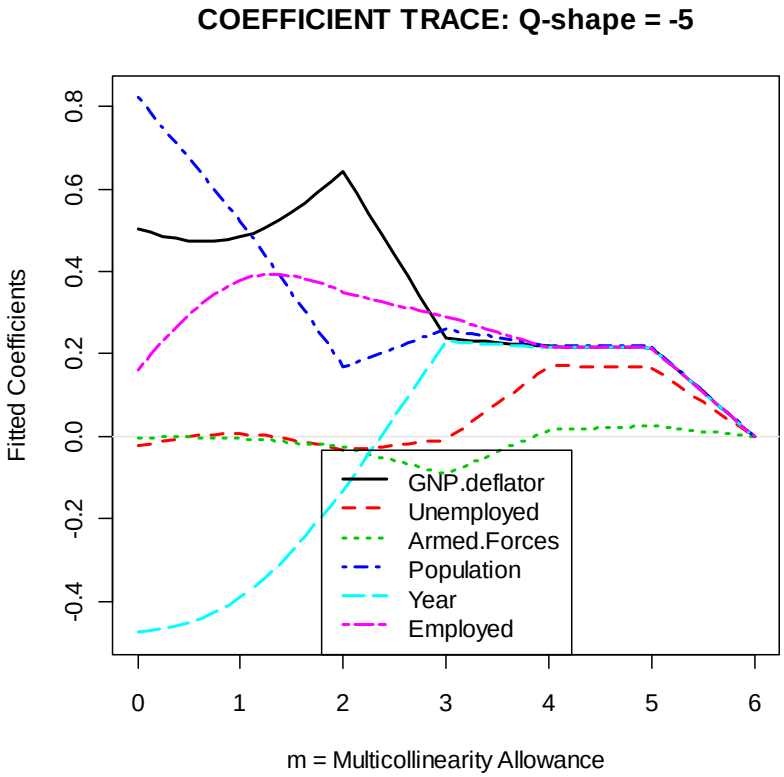
`plot(rxrobj)` ← **Default display of all 5 TRACEs.**



3.1: Shrinkage Coefficient Trace

The COEFFICIENT trace display shows how point estimates of β-coefficients change as shrinkage progresses along a path of shape ***Q***. Coefficient estimates that are numerically "stable" will tend to plot close to the straight line from their (left-hand end) **least-squares estimates** at MCAL=0 to **zero** at MCAL=P (right-hand end.) Relatively unstable coefficient estimates will change non-linearly, possibly switching numerical sign, as MCAL increases. Super-stable estimates will display traces that initially change very little (remaining almost horizontal), finally approaching zero only as MCAL approaches P.

```
plot(rxrobj, trace = "coef", trkey = TRUE)
```



Note that most of the clearly undesirable features of the OLS estimates in this **`longley2`** example have been mitigated once the shrinkage extent reaches at least m = 3. From that point on, four of the six estimates have become essentially equal.

"Wrong Sign" Problem(s):

A theoretical basis for detecting "sign problems" by comparing the numerical signs of fitted coefficients with their marginal correlations is provided by Remark (d) on page 1118 of Obenchain (1978). When the α vector in my Theorem 2 is parallel to the unknown, true β, the corresponding optimal generalized ridge estimator is KNOWN to be proportional to **X'y**, a vector that clearly has elements with the same numerical signs as the vector of marginal correlations of **y** with **X**.

Because the vector of OLS estimates is of the form $\mathbf{X}^{+}\mathbf{y}$, its elements can have different signs from those of **X'y** when the data are ill-conditioned. When this does occur, it's relatively bad news!

```
rxrobj$coef[1,]   ← OLS regression coefficient estimates

      0.50356   -0.02370     -0.00258    0.82122   -0.47560    0.16119
```

```
(cor(longley2))[1:6,7]   ← marginal correlations between y = GNP
                                                    and 6 Xs.

GNP.deflator Unemployed Armed.Forces Population      Year  Employed
      0.9936     0.6967       0.0735     0.9838    0.9479    0.9841
```

Note that the OLS coefficient for Year is large and negative here. This signals a major "wrong sign" problem because the marginal correlation between GNP and Year is quite strongly positive (+0.9479.) The problem clearly disappears once ***Q*** = –5 shrinkage reaches m = 3.

Similar (but minor) problems exist due to negative OLS estimates for Unemployed and Armed.Forces. However, these OLS coefficients are already relatively small numerically, and the corresponding marginal correlations with GNP are much less positive.

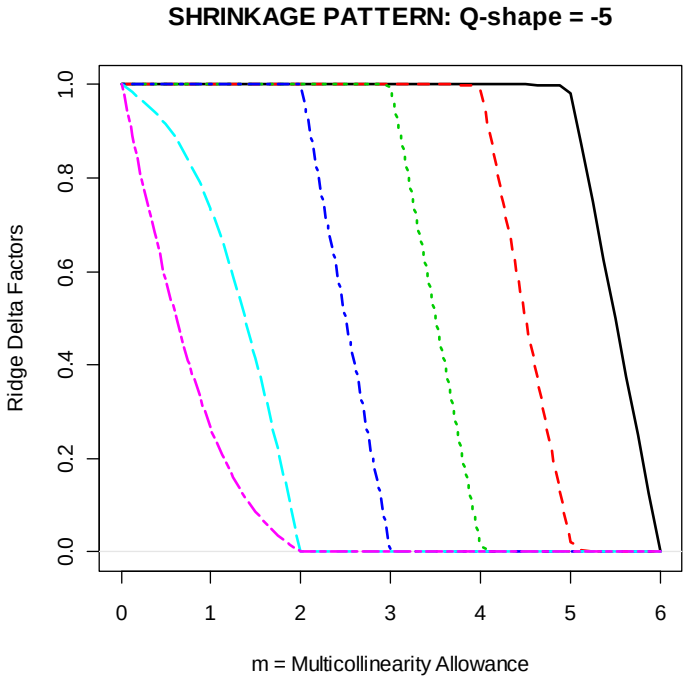
3.2. Shrinkage Pattern Trace

The SHRINKAGE PATTERN trace shows how the generalized ridge "Delta Shrinkage-Factors" applied to the ordered "uncorrelated components" vector, ***c***, decrease as shrinkage of shape ***Q*** occurs. All such delta factors start out as 1 at M=0 (the OLS solution.) As M increases, all deltas remain equal when ***Q*** = 1; the trailing deltas are smallest when ***Q*** < 1; and the leading deltas are smallest when ***Q*** > 1.

Colors have somewhat different interpretations in SHRINKAGE PATTERN traces than in the COEFFICIENT trace. In both cases, colors are ordered: **FIRST, SECOND, THIRD, FOUTRH, FIFTH, SIXTH,** etc. In a COEFFICIENT trace, colors represent the X-variables in the order that they were specified in the regression formula: **Y** ~ **X1** + **X2** + **X3** + **X4** + **X5** + X6. But in a SHRINKAGE PATTERN trace, these same colors represent the regressor principal axes in the decreasing order of the eigenvalues of **X'X**: $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_r > 0$.

Since we are following an extreme shrinkage path shape of ***Q*** = –5 for the **`longley2`** dataset, we see in the SHRINKAGE PATTERN trace displayed below that essentially only the last two out of six shrinkage factors, δ_5 and δ_6, change between M=0 and M=2. After all, the last two singular values (square roots of eigenvalues of **X'X**) are nearly equal and are much smaller than the other four singular values. In fact, the last two shrinkage factors have essentially been reduced to zero at M=2.

```
plot(rxrobj, trace = "spat")
```



As shrinkage then continues from M=2 to M=3, the fourth shrinkage factor, δ_4 , essentially decreases from 1 to 0 …while δ_1, δ_2 and δ_3 all remain near 1. As was clear from the COEFFICIENT trace displayed above, the majority of the severe ill-conditioning in the **`longley2`** dataset (i.e. switches in β-coefficient signs and drastic changes in their relative magnitudes) is confined to the last three out of six total principal components of **X**-space.

3.3. Relative (or "Scaled") MSE Risk Trace

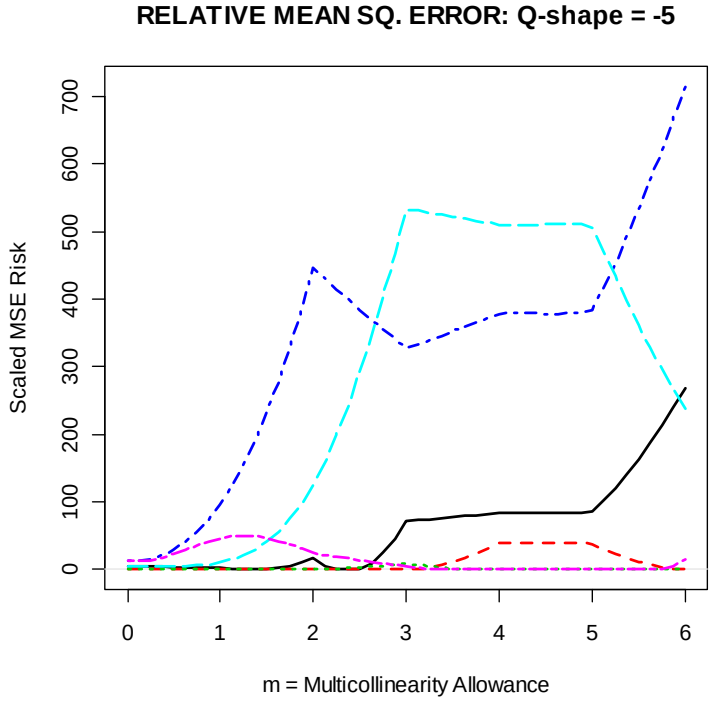
The RELATIVE MSE trace displays normal distribution theory, "modified" maximum likelihood estimates of "scaled" MSE risk in individual–coefficient estimates as shrinkage of shape ***Q*** occurs.

> Risks are "scaled" by being divided by the usual estimate of the error (disturbance term) variance. In other words, scaled risk expresses imprecision in fitted coefficients as a multiple of the variance of a single observation. Furthermore, when regression disturbance terms are assumed to be uncorrelated and homoskedastic, the "scaled" MSE risks of the unbiased OLS estimates (at the extreme left of the trace where Δ = I) are **known quantities**, being the diagonal elements of the $(\mathbf{X'X})^{-1}$ matrix.
>
> When shrinkage Δ factors are less than 1, maximum likelihood scaled risk estimates are "modified," first of all, so as to be unbiased under normal theory. Then they are adjusted upward, if necessary, to have correct range relative to a known lower bound on scaled risk, which may re-introduce some bias.

As in the COEFFICIENT trace, colors in the RELATIVE MSE trace represent the X-variables in the order that they were specified in the regression formula: **Y** ~ **X1** + **X2** + **X3** + **X4** + **X5** + X6.

In the Relative MSE trace (below) for the **`longley2`** data, shrinkage appears to be injecting considerable bias into the **4th (Population)** and **5th (Year)** β-coefficient estimates.

```
plot(rxrobj, trace = "rmse")
```



Changes in the **6th (Employment)** β-coefficient estimate between M=0 and M=3 first increase but then decrease MSE risk. Initial increases in the **1st (GNP.deflator)** β-coefficient estimate between M=0 and M=2 are relatively unimportant, but subsequent shrinkage increases MSE risk at M=3 and beyond. Increases in the **2nd (Unemployment)** β-coefficient estimate between M=3.5 and M=4.5 also increase MSE risk somewhat.

3.4. Excess Eigenvalues Trace

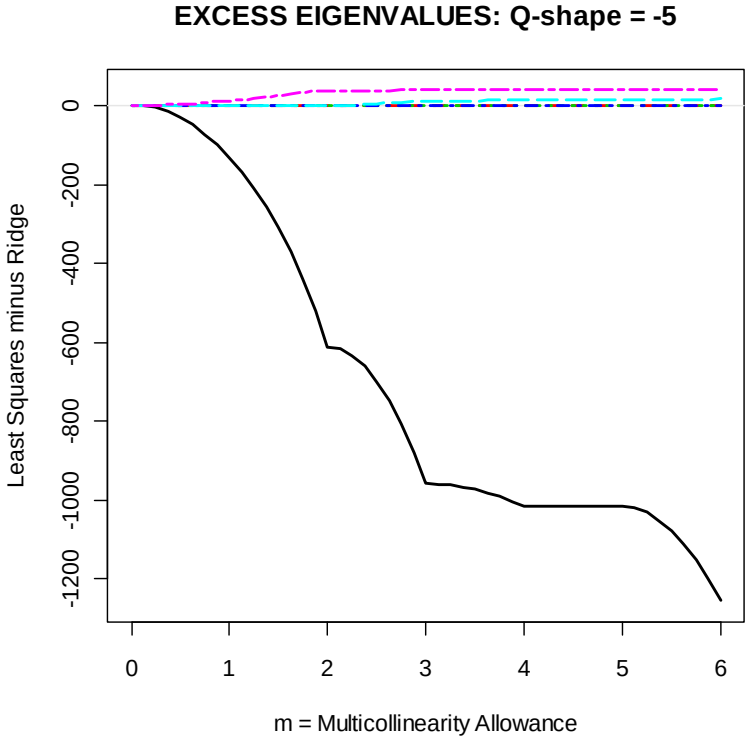
The EXCESS EIGENVALUES trace plots the eigenvalues of the estimated difference in Mean Squared Error matrices, ordinary least squares (OLS) minus ridge. As long as all eigenvalues are non-negative, there is reason to hope that the corresponding shrunken estimators yield smaller MSE risk than OLS in all directions of the r-dimensional space spanned by X-predictors (i.e. all possible linear combinations.) As shrinkage continues, **at most one negative eigenvalue will appear**.

The colors in the EXCESS EIGENVALUE trace represent only the observed order (smallest to largest) of these eigenvalues. Specifically, the **SMALLEST** (possibly negative) is drawn in **black**, while the **SECOND SMALLEST** (never negative) is **red**. At the top end when the **X**

matrix has rank 6, the **LARGEST** eigenvalue is **magenta**, while the **SECOND LARGEST** is shown in **cyan**.

In the EXCESS EIGENVALUE trace (below) for the **`longley2`** data, the smallest eigenvalue becomes negative at the 3rd computational step of M = 0.250, which also happens to be the approximate extent of shrinkage suggested by the EBAY and RCOF likelihood criteria. The negative eigenvalue at M = 0.250 is –3.26 while the corresponding largest eigenvalue is only +2.00. In other words, more MSE "harm" is already being done in the "inferior direction," Obenchain(1978), corresponding to M = 0.250 along the path of shape ***Q*** = –5 that in the (unspecified) direction of greatest MSE decrease due to shrinkage.

```
plot(rxrobj, trace = "exev")
```



The negative eigenvalue at the M = 4.750 extent of shrinkage suggested by the CLIK criterion is **–1017** while the corresponding two largest eigenvalues are **+13.9** and **+39.1**. In other words, the **`longley2`** dataset is rather clearly very highly ill-conditioned. In fact, ill-conditioning is sufficiently bad that the amount of shrinkage needed to stabilize coefficient relative magnitudes (including correction of a "wrong sign" problem in the **Year** coefficient) cannot be justified from a MSE reduction perspective.

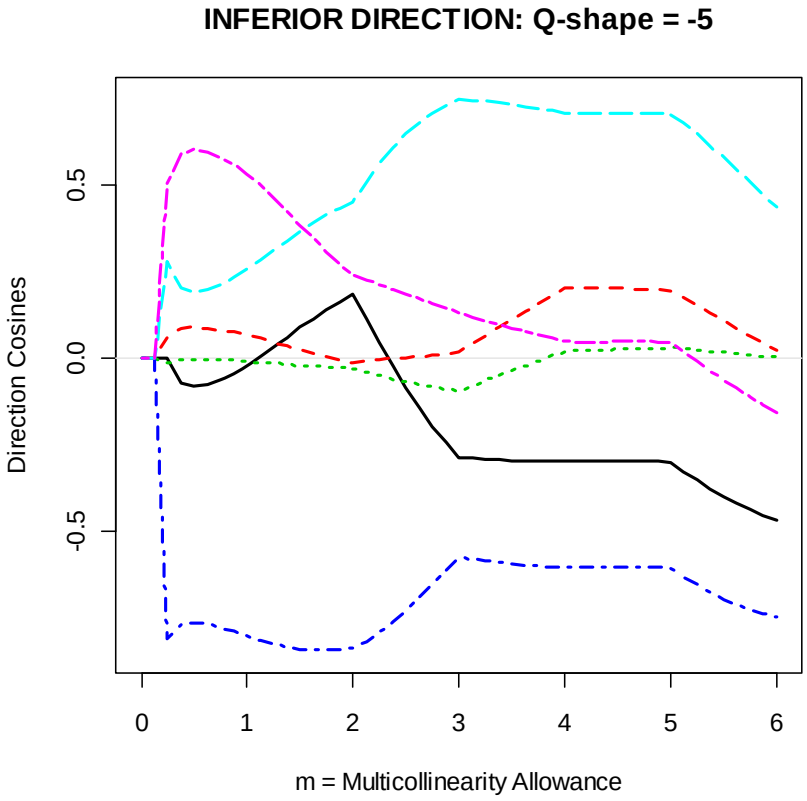
3.5. Inferior Direction-Cosine Trace

The INFERIOR DIRECTION trace displays the **direction cosines** (elements of the normalized eigenvector) corresponding to any negative eigenvalue of the difference in MSE matrices, OLS – ridge. This direction gives that single linear combination of ridge regression coefficients that not

only fails to benefit from ridge shrinkage of shape ***Q*** but probably actually suffers increased risk due to shrinkage.

Because the rows and columns of these MSE matrices are in the order specified on the right-hand-side of the regression formula **Y** ~ **X1** + **X2** + **X3** + **X4** + **X5** + X6, the direction cosines relative to these given **X** axes are colored in this same order.

```
plot(rxrobj, trace = "infd")
```



Interpretation of direction cosines in 6-dimensions can be problematic, to say the least. Thus we will focus here on only relatively simple things that can be seen in an INFERIOR DIRECTION trace. Note that all values in the plot could be multiplied my **–1** (turning it upside-down) without changing its basic interpretation.

First of all, all fitted regression coefficients have been shrunken to (0, 0, …, 0) at the right-hand extreme of all TRACE displays, M = rank(**X**). This is usually much-too-much shrinkage, so the inferior direction typically points backwards from (0, 0, …, 0) essentially towards the original ±OLS coefficient vector at M = 0. In the above plot for the **`longley2`** dataset, the displayed direction cosines at M = 6 clearly point to the negative of the original OLS vector.

When two curves on an INFERIOR DIRECTION trace cross, their direction cosines are clearly equal at that value of M. This happens with the cosines for the **1st (GNP.deflator)** and **2nd (Unemployed)** regressors at M = 1.295, where the common cosine value is +0.041. Thus, at M = 1.295, the shrunken estimate of the SUM of the **1st** and **2nd** β-coefficients (0.512) can have

higher MSE risk than its OLS estimate(0.480); after all, the vector (1,1,0,0,0,0) is clearly NOT orthogonal to the inferior direction at M = 1.295. In sharp contrast, the vector (+1,–1,0,0,0,0) IS orthogonal to the inferior direction at M = 1.295, and thus the DIFFERENCE between the shrunken estimates of the **1st** and **2nd** β-coefficients (0.513) should have the same or lower MSE risk than the corresponding difference in OLS estimates (0.527.)

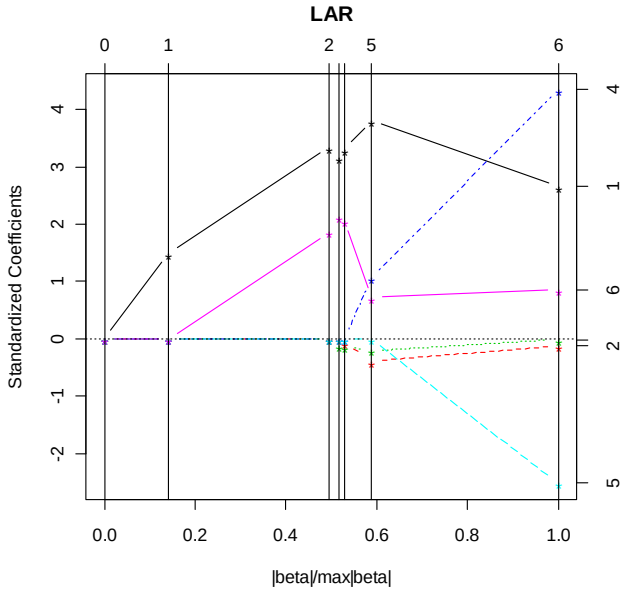
A similar crossing of cosines for the **5th (Year)** and **6th (Employed)** regressors occurs at M = 1.535, where the common cosine value is +0.373. Thus, at M = 1.535, the shrunken estimate of the SUM of the **5th** and **6th** β-coefficients (–0.119) can have higher MSE risk than its OLS estimate(–0.314.) Meanwhile, the DIFFERENCE between the shrunken estimates of the **5th** and **6th** β-coefficients (–0.656) should have the same or lower MSE risk than the corresponding difference in OLS estimates (–0.637.)

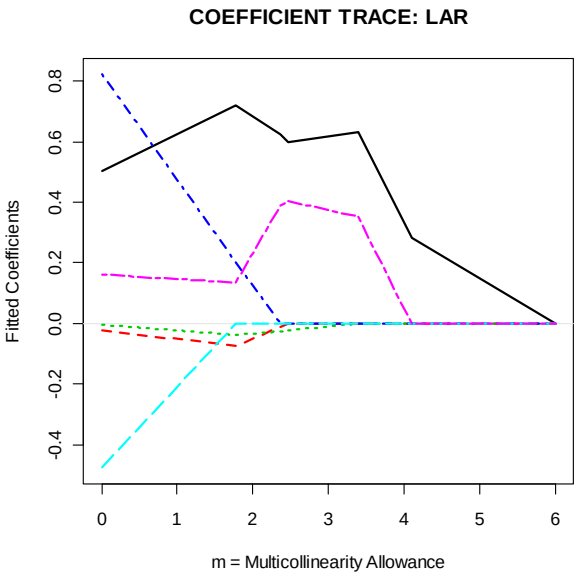
M-extents of shrinkage such that two regressors have inferior direction cosines with equal magnitudes but opposite numerical signs have the opposite effects on the MSE risks of sums and differences. The SUM of the corresponding shrunken coefficients then has the same or reduced MSE risk, while the corresponding DIFFERENCE has increased MSE risk. This happens for the **1st (GNP.deflator)** and **6th (Employed)** regressors at M = 2.67, where the direction cosines are ±0.164. Unfortunately, shrinkage to M = 2.67 has inappropriately reduced the difference between coefficient estimates (from 0.34 to 0.06) while leaving the sum mostly unchanged (0.68 rather than 0.66.)

4. Interpretation of least angle regression TRACE displays

```
xlong2 <- as.matrix(longley2[,1:6])
ylong2 <- as.matrix(longley2[,7])
larsobj <- lars(xlong2,ylong2,type="lar")
plot(larsobj)
```

```
rxlobj <- RXlarlso(form,data=longley2)
plot(rxlobj, trace = "coef")
```

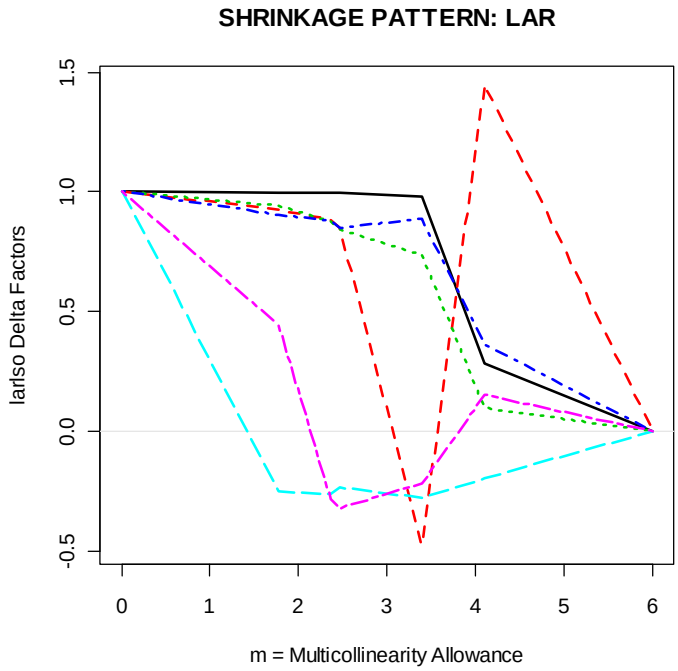




The first thing to note about the coefficient TRACE displays from the **`RXlarlso()`** and **`RXuclars()`** functions within the **RXshrink** package is that they are essentially "backwards" relative to the default coefficient displays from the **lars** R-package. This point is illustrated above. Close examination of this pair of graphs also shows that, besides being backwards relative to each other, there are some additional, rather minor differences between the **|beta|/max|beta|** scaling used along the horizontal axis by **lars** and the **m = Multicollinearity Allowance** scaling used by **RXshrink**.

Least angle regression (lar) may yield an initial solution vector that is longer that the OLS vector. As explained at the end of Section §2, this means that one or more of the shrinkage "delta" factors implied by the **`lars()`** estimate starts out being greater than one. Similarly, as lar shrinkage occurs, one or more of these implied delta-factors may eventually become negative. These points are illustrated in the graph below.

```
plot(rxlobj, trace = "spat")
```

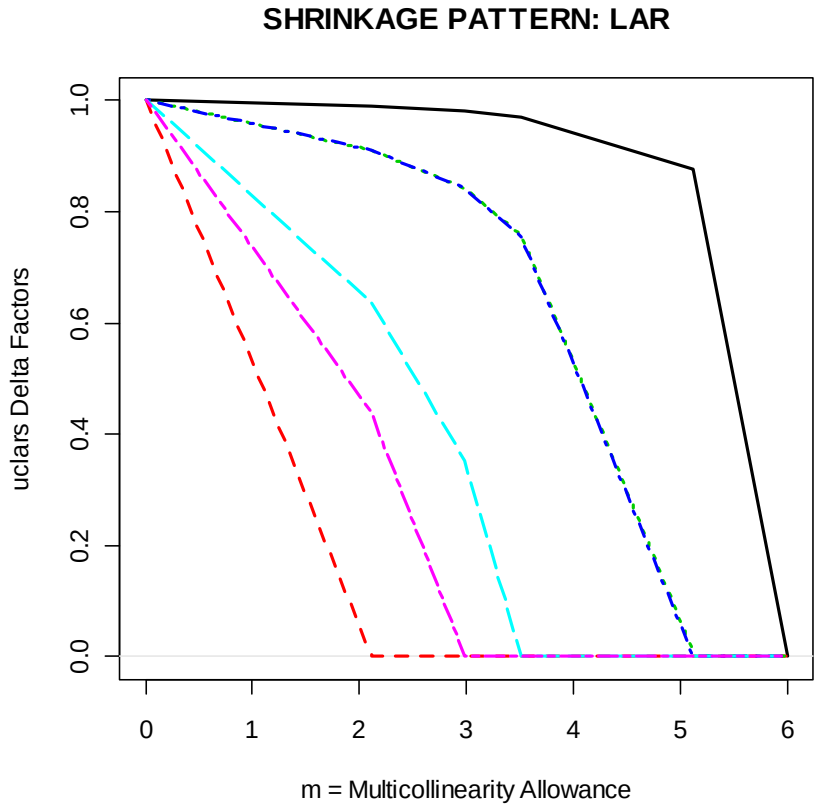


Reductions in MSE risk relative to OLS usually occur only when all of the delta-factors implied by **`lars()`** estimates are non-negative and strictly less than +1. Exceptions can occur when the unknown true "gamma" component corresponding to an "out of range" delta-factor is nearly zero.

The ***Q***-shape shrinkage paths typically used in "generalized ridge" regression depend upon the **eigenvalue spectrum** of the centered **X'X** matrix as well as upon the **principal correlations** with the centered response vector, ***y***. In sharp contrast, the shrinkage paths implied by "least angle" regression methods typically depend only upon **correlations** (marginal or principal) with the response ***y***-vector. As a direct result, the relative sizes of the shrinkage delta-factors implied by **`lars()`** estimates are not ordered in a predetermined way.

Use of implied delta-shrinkage factors outside of the usual range of [0, 1) can be avoided by use of the **`RXuclars()`** function rather than the **`RXlarlso()`** function illustrated above. In this special case, the **principal correlations** with the response ***y***-vector determine the implied delta-shrinkage factors. Specifically, the general expression $\delta_i^{uclars} = \max[\ 0, (\ 1 - k\ /\ |\rho_i|\ )\ ]$ then shows that the smallest delta-factor will always correspond to the smallest principal correlation. The **`RXridge()`** output listed at the top of page 10 shows that the **2nd principal coordinates** of X-predictors have the smallest absolute correlation (`0.01078`) with the response ***y***-vector for the **`longley2`** dataset. This is also clear in the graph below.

```
rxuobj <- RXuclars(form,data=longley2)
     plot(rxuobj, trace = "spat")
```



Note that, because the **3rd** and **4th** **principal coordinates** of **X**-predictors have nearly equal absolute correlations (`0.1222` and `0.1209`) with the response ***y***-vector, the **3rd** and **4th** shrinkage delta-factors in the above graph are essentially equal.

5. Summary

The **RXshrink** package for **R** is fully documented with *.rd* and *.html files. The additional information provided in this vignette [1] comments on the history of shrinkage in regression, [2] discusses the 2-parameter family of generalized ridge estimators and interpretation of TRACE

displays, and [3] orients the shrinkage implied by lars and lasso estimates relative to the principal axes of the given **X**-variables and the uncorrelated components, **c**, of the OLS β estimator.

Visualization of shrinkage regression results requires examination and interpretation of the TRACE plots produced by **RXshrink** functions. In a trace, **r = Rank(X)** quantities (several estimated coefficients, risks, shrinkage factors, etc.) are plotted vertically against a horizontal indicator of the extent of shrinkage. Traditional TRACES display the OLS solution at their left-hand extreme and cover the full range of shrinkage that culminates in "total" shrinkage at their right-hand extreme (where all "centered" regression coefficient estimates become zero.) **RXshrink** functions require **r** to be at least 2.

Measures of MSE risk (expected loss) are defined for all forms of statistical distributions, but the **RXshrink** functions focus on Likelihoods implied by assuming that the OLS estimator has a multivariate normal distribution with mean vector β and variance σ^2 **I**. The classical, empirical Bayes and random coefficient perspectives thus suggest using the extent of shrinkage that minimizes the CLIK, EBAY or RCOF –2log(likelihood ratio) criterion, respectively.

A "good" shrinkage estimator, Obenchain(1979), achieves equal or lower matrix-valued MSE risk than OLS for the true values of the β and σ parameters. Brown (1975) and Bunke(1975a, 1975b), showed that no single, realizable estimator can be "good" under normal distribution theory for all possible values of β and σ. Thus, users of **RXshrink** functions need to focus attention on the question: "Are the most likely values of the β and σ parameters for a given regression model either **highly favorable to shrinkage** or else **possibly unfavorable to shrinkage**?" Shrinkage TRACES display sample information that goes a long way towards "answering" this question, especially the Excess Eigenvalue and Inferior Direction TRACES.

> For example, this vignette uses the **`longley2`** dataset to illustrate interpretation of TRACE displays, and we have seen that this particular regression problem appears to be quite **unfavorable to shrinkage**. The original Longley(1967) data and model with ***y*** = Employed is more favorable to shrinkage. To see the TRACES for a setup **quite favorable to shrinkage**, the reader can run: **`demo(haldport)`**

Unfortunately, this vignette has not illustrated interpretation of the output from the **`RXtrisk()`** and **`RXtsimu()`** functions. Obenchain(1984, 1995) discussed uses for these types of TRACES based upon early implementations in SAS/IML and Stata, respectively.

REFERENCES

Breiman L. Better subset regression using the non-negative garrote. **Technometrics** 1995; 37: 373-384.

Brown L. Estimation with incompletely specified loss functions (the case of several location parameters.) **Journal of the American Statistical Association** 1975; 70: 417-427.

Bunke 0. Least squares estimators as robust and minimax estimators. **Math. Operations forsch u. Statist.** 1975(a); 6: 687-688.

Bunke 0. Improved inference in linear models with additional information. **Math. Operations forsch u. Statist**. 1975(b); 6: 817-829.

Burr TL, Fry HA. Biased Regression: The Case for Cautious Application. **Technometrics** 2005; 47: 284-296.

Casella G. Minimax ridge regression estimation. **Annals of Statistics** 1980; 8: 1036-1056.

Casella G. Condition numbers and minimax ridge-regression estimators. **Journal American Statistical Association** 1985; 80: 753-758.

Efron B, Morris CN. "Discussion" (of Dempster, Schatzoff and Wermuth.) **Journal American Statistical Association** 1976; 72: 91-93. (empirical Bayes.)

Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. **Annals of Statistics** 2004; 32: 407-499 (including discussion.)

Frank IE, Freidman JH. A statistical view of some chemometrics regression tools. **Technometrics** 1993; 35: 109-148 (including discussion.)

Goldstein M, Smith AFM. Ridge-type estimators for regression analysis. **Journal of the Royal Statistical Society B** 1974; 36: 284-291. (2-parameter shrinkage family.)

Golub GH, Heath M, Wahba G. Generalized cross-validation as a method for choosing a good ridge parameter. **Technometrics** 1979; 21: 215-223.

Hoerl AE. Application of Ridge Analysis to Regression Problems. **Chemical Engineering Progress** 1962; 58: 54-59.

Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. **Technometrics** 1970(a); 12: 55-67.

Hoerl AE, Kennard RW. Ridge Regression: Applications to Nonorthogonal Problems. **Technometrics** 1970(b); 12: 69-82.

James W, Stein C. Estimation with quadratic loss. **Proceedings of the Fourth Berkeley Symposium** 1961; 1: 361-379. University of California Press.

LeBlanc M, Tibshirani R. Monotone shrinkage of trees. **Journal of Computational and Graphical Statistics** 1998; 7: 417-433.

Littel RC, Milliken GA, Stroup WW, Wolfinger RD. **SAS System for Mixed Models.** 1996. Cary, NC: SAS Institute.

Longley JW. An appraisal of least-squares programs from the point of view of the user. **J. Amer. Statist. Assoc.** 1967; 62: 819-841.

Marquardt DW. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. **Technometrics** 1970; 12: 591-612.

Pinheiro JC, Bates DM. Unconstrained Parametrizations for Variance-Covariance Matrices. **Statistics and Computing** 1996; 6: 289-296.

Obenchain RL. Ridge Analysis Following a Preliminary Test of the Shrunken Hypothesis. **Technometrics** 1975; 17, 431-441. (Discussion: McDonald GC, 443-445.)

Obenchain RL. Classical F-tests and confidence regions for ridge regression. **Technometrics** 1977; 19: 429-439.

Obenchain RL. Good and optimal ridge estimators. **Annals of Statistics** 1978; 6: 1111-1121.

Obenchain RL. Maximum likelihood ridge regression and the shrinkage pattern alternatives. **I.M.S. Bulletin** 1981; 10: 37 [Absract 81t-23.]

Obenchain RL. Maximum likelihood ridge displays. **Communications in Statistics - A** 1984; 13; 227-240.

Obenchain RL. Ridge regression systems for MS-DOS personal computers. **The American Statistician** 1991; 45: 245-246.

Obenchain RL. Maximum likelihood ridge regression. **Stata Technical Bulletin** 1995; 28: 22-35.

Obenchain RL. "Shrinkage Regression: ridge, BLUP, Bayes, spline and Stein." **www.iquest.net/~softrx** eBook-in-Progress 1992--2005. (200+ pages.)

Robinson GK. That BLUP is a good thing: the estimation of random effects. **Statistical Science** 1991; 6: 15-51 (including discussion.)

Stein C. Inadmissibility of the usual estimate of the mean of a multivariate normal distribution. **Proceedings of the Third Berkeley Symposium** 1955; 1: 197-206. University of California Press.

Strawderman WE. Minimax adaptive generalized ridge regression estimators. **Journal of the American Statistical Association** 1978; 73: 623-627.

Tibshirani, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society B** 1996; 58: 267-288.

Tukey JW. "Instead of Gauss-Markov Least Squares; What?" **Applied Statistics,** ed. R. P. Gupta. 1975. Amsterdam-New York: North Holland Publishing Company.