<div style="border:1px solid black; padding:10px;">

RXshrink: a softRX freeware R-package
implementing shrinkage methods…

Generalized Ridge and Least Angle Regression

Version 1.0-5, September 2010 … License: GPL (>= 2)

</div>

softRX is a trademark of The Software Prescriptions Company.
"Personal computing treatments for your data analysis infirmities …since 1983"

Bob Obenchain, PhD FASA
Principal Consultant, Risk Benefit Statistics LLC
13212 Griffin Run, Carmel, IN 46033-9935
317-580-0144, wizbob@att.net

Table of Contents

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

As some one who has been fascinated with the po ssibility 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 m ust say that I am absolutely delighted by the recen t wide-spread tolerance for (if not outright accep tance of) shrinkage m ethods. Anyway, I wish to summarize here some personal perspectives on why and how professional statisticians m ay have become somewhat enlightened about shrinkage over the last 50+ years …since ~1955.

Early optimism about a theoretical basis for and the practical a dvantages of shrinkage alm ost surely started with the work of Stein(1955) and Jam es and Stei n(1961). Unfortunately this shrinkage was always "uniform ," thus really doing nothing to adjust the relative m agnitudes of correlated regression coefficient estimates for ill-conditioning. Furthermore, although an overall improvement in the scalar va lue of "summ ed MSE risk" was guaranteed, there was no way to know "where," in an X-space of 3 o r 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 "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 dominate OLS in MSE risk were, in fact, unquestionably naïve.

Meanwhile, a major frustration for me, personally, was that my shrinkage work at 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 papers and on my Bell Labs internal "regression training" materials.

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 (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 m y basic shrinkage equati ons and theorem s were published, I had m uch less success publishing de scriptions of practical shrinkage applications, including "how-to" information about interpretations for my five types of TRACE plot s. In fact, only three of my papers on s hrinkage ap plicants o r softwa re, Obenchain (1984, 1991, 1 995), were ultim ately 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 for m e xpression, Obenchain(1981), for the norm al-theory ML estim ator within the 2-param eter Gold stein and Sm ith (1974) shrinkage fam ily. Unfortunately, none of my attem pts to present th at material in a peer-reviewed publication have succeeded. Closed form expressio ns speed sh rinkage es timation and are p articularly h elpful when simulating MSE risk profiles.

My "botto m-line" on the topic of norm al-theory ML shrinkage is s imply this: The linear estimator identified as being most likely to be o ptimal is, in reality, a **nonlinear** estimator. The true MSE risk of this ML shrinkage estim ator ca n be com puted exactly in certain s pecial cases and can alw ays be accu rately s imulated. W hile ha ving a MSE risk p rofile that is clearly not "dominant" like that of the unknow n, 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 inc reasing risk by at m ost 20% in unfavorable cases.

> In high-dimensional situations, a savings of <u>more than 50%</u> is possible, and worst case situations result in an increase of <u>less than 5%</u> in MSE risk.

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

Frank and F reidman(1993), Breim an (1995), Tibshirani (1996), Le Blanc and Tibshirani (1998) and Efron et al. (2004) are curren tly keeping the shrinkage regre ssion "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 <u>generalized ridge regression</u> 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 = **G c = g₁** c_1 + **g₂** c_2 + … + **gᵣ** c_r of the unknown, true regression coefficient β vector. The variance matrix of **c** is the diagonal Λ^{-1} matrix times the scalar value of the error sigma-square. The P = r = 2 dimensional case is depicted below.
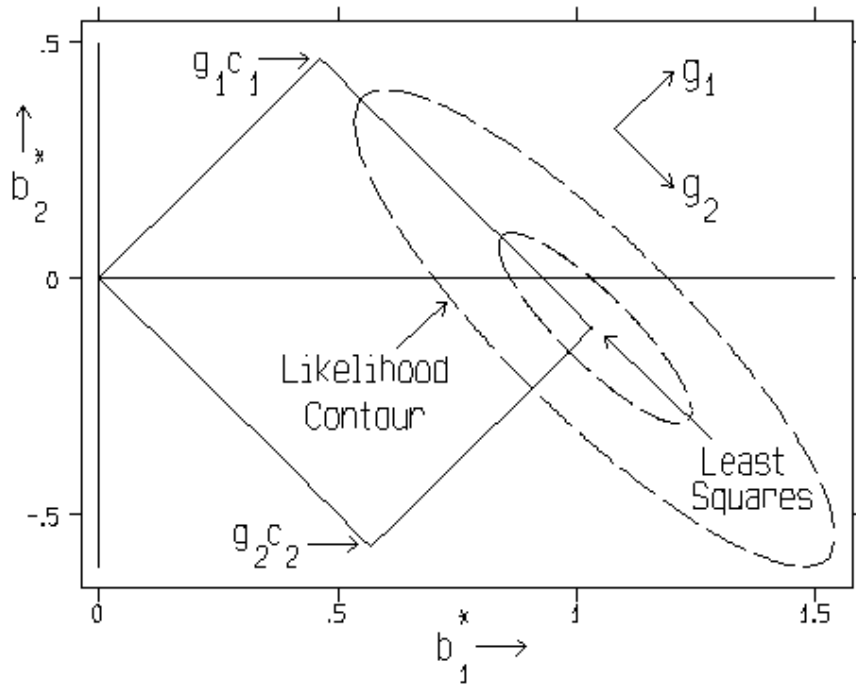
In fact, we now see that the <u>2-parameter family</u> of shrinkage estimators from our **first equation**, above, is the special case of the **last equation** in which...

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

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($\mathbf{X}$) = p = 2 case.
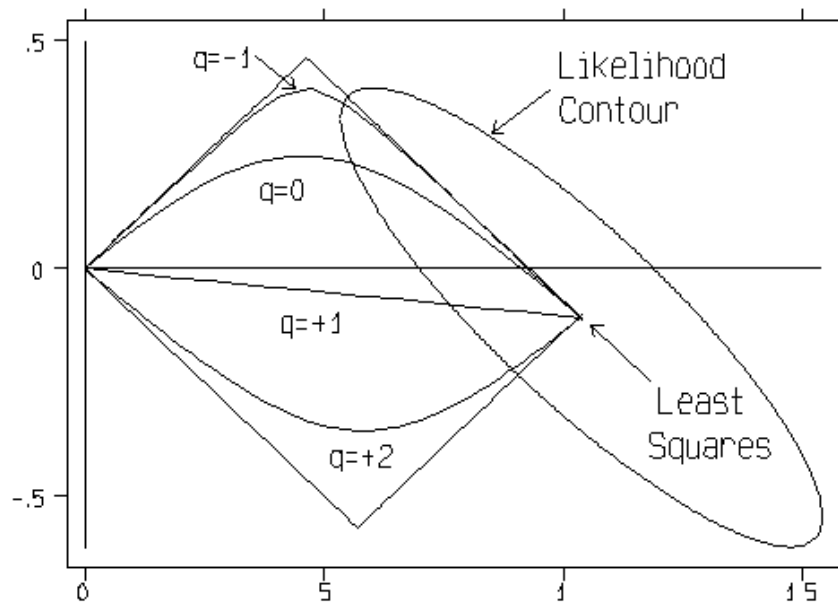
The best known special case of a **Q**-shaped path is probably **Q** = 0 for Hoerl-Kennard(1970) **"ordinary" ridge regression**. This path has a dual "characteristic property," illustrated in the figure below. Namely, the **Q** = 0 path contains not only the shortest beta estimate vector of any given likelihood but also the most likely beta estimate of any given length.
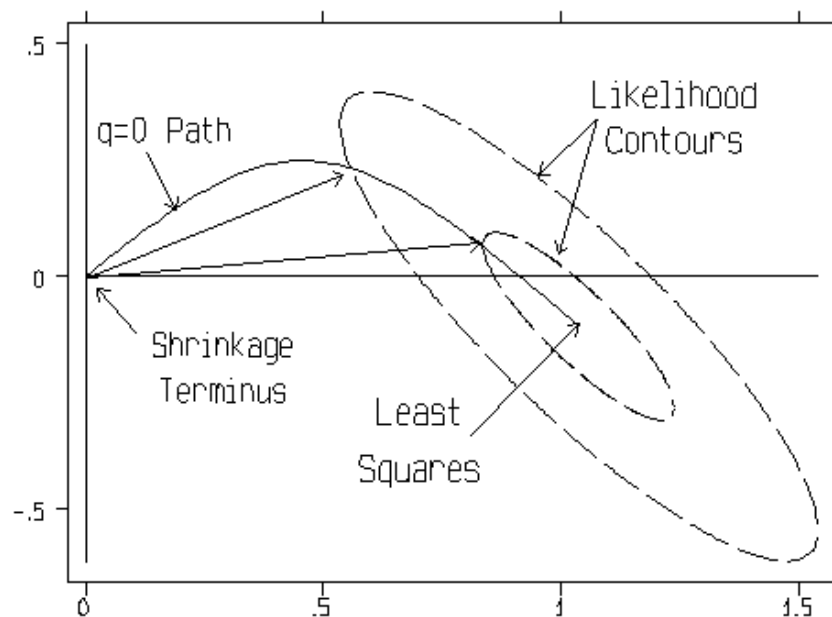
Another well known special case of a **Q**-shaped path is **Q** = +1 for **uniform** shrinkage. The coefficient trace and shrinkage factor trace for this path are both rather "dull," but the estimated risk and inferior direction TRACES can still be interesting even when **Q** = +1.

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.

Unfortunately, the "**k**" parameter is really not a very good measure of the **extent** of shrinkage. After all, the sizes of **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 <u>finite (rather than infinite) range</u>, *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 <u>coefficients with stable relative magnitudes</u> form **straight lines** when plotted versus *MCAL*.
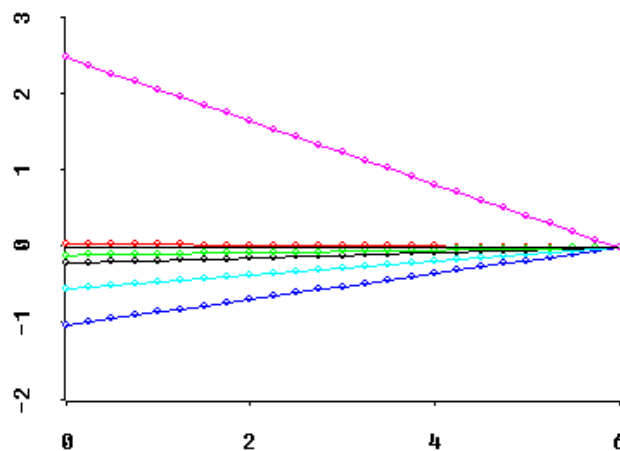
Similarly, the average value of all **r** shrinkage factors is **(r − MCAL)/r**, which is the Theil(1963) proportion of Bayesian posterior   precision due to sam   ple inform ation (rather than to prior information.) Note that this proportion decreases linearly as *MCAL* increases.

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 b est *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) data set when the response is y = Em   ployed. Com pared with the major initial shifts in relative magnitudes and numerical signs of coefficients between *MCAL* = 0 and *MCAL* = 2, note th at the trace becom es relatively much m ore stable (som ewhat "straight") between *MCAL* = 2 and *MCAL* = r = 6.

COEFFICIENT TRACE: Q-shape = -1.5
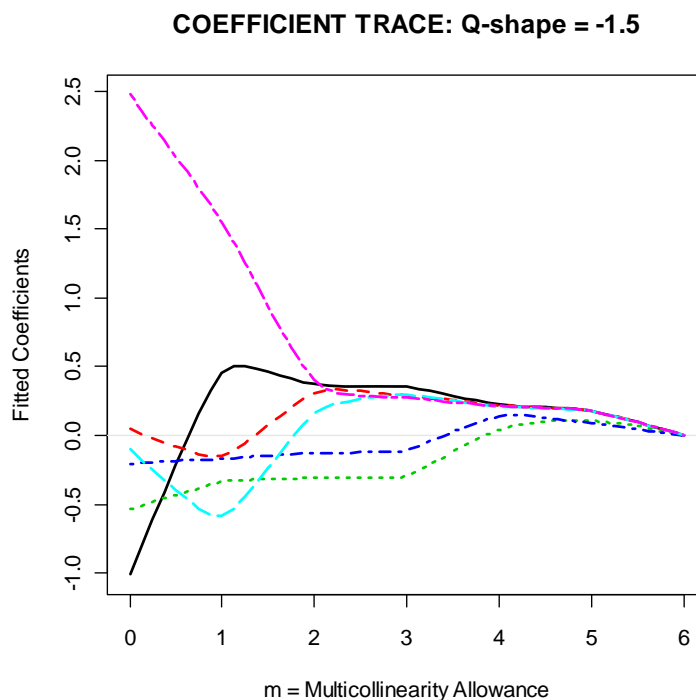
As a general rule-of-thumb, paths with *Q*-shapes in the [−1,+2] range generally tend to be fairly **smooth** ...i.e. have "rounded" corners. Paths with *Q*-shapes greater that +2 or less than −1 can display quite "sharp" corners. In fact, the paths with limiting shapes of ±∞ are actually linear splines 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.

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 ith 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 th e resulting lar or las so "delta-factors" to be positive and less than +1, the resulting estimates may have neither of these properties. In oth er words, lar and lasso estim ators can correspond to "non-standard" generalized ridge estim ators 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 es timation directly to the <u>uncorre lated components vector</u>, **c**, and this restriction yields a true ge neralized ridge (shrinkage) estim ator. 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 ith "principal correlation" …i.e. the correlation between the response **y**-vector and the ith column of the **H** matrix of "principal coordinates" of **X** (page 4.) Note that the k-factor in this shrink age formulation is lim ited 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, **c** = **G'β°**.
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.**

softRX Shrinkage in Regression
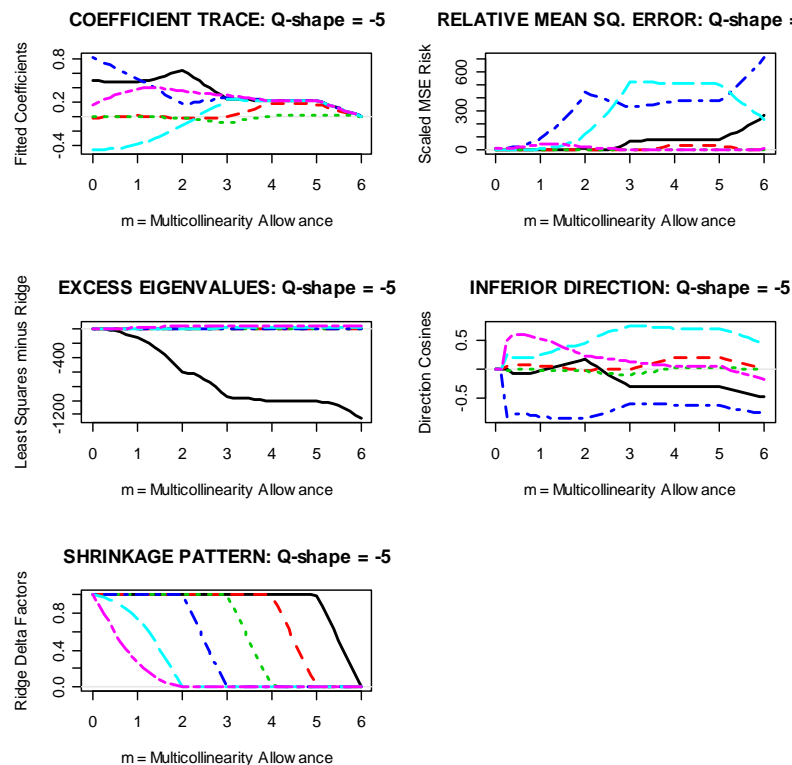
3.1: Shrinkage Coefficient Trace

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) <u>**least-squares estimates**</u> at MCAL=0 to <u>**zero**</u> 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)
```

COEFFICIENT TRACE: Q-shape = -5
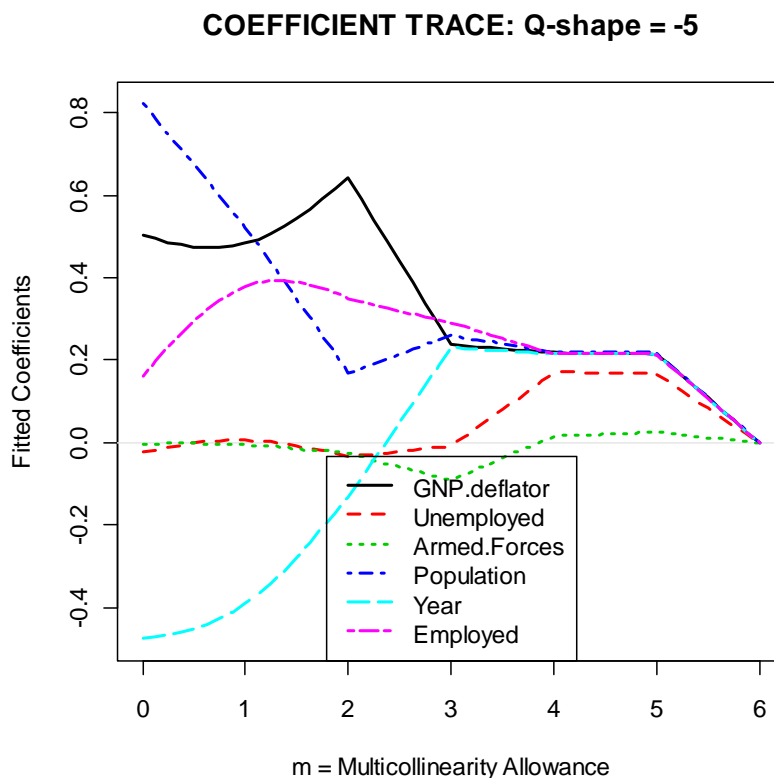
Note tha t most of the clea rly un desirable f eatures of th e OLS estim ates in th is `longley2` example have been m itigated once the shrink age ex tent reaches at least m = 3. From that point on, four of the six estimates have become essentially equal.

<mark>"Wrong Sign" Problem(s):</mark>

A theoretical basis for detecting "sign probl ems" by com paring the num erical signs of fitted coef ficients with their m arginal co rrelations is p rovided by Rem ark (d) on p age 1118 of Obenchain (1978). W   hen the α vector in m y 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 h as elements with the sam e numerical signs as the vector of marginal correlations of **y** with **X**.

Because the vecto r of OLS estim ates is of the f orm $\mathbf{X}^+\mathbf{y}$, its elem ents can have different signs from t hose of $\mathbf{X'y}$ when the data ar e ill-conditioned. W hen 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 th e OLS coefficien t for Y ear is la rge and negative here. This signa ls a major "wrong sign" problem because the m arginal correlation between GNP and Year is quite strongly pos itive (+0.94 79.) The p roblem clearly disappears once $Q = -5$ shrinkage reaches m = 3.

Similar (but m inor) pro blems exist due to negative OLS estim ates for Une mployed and Armed.Forces. However, these OLS coe fficients are already relatively s mall numerically, and the corresponding m arginal correlations with GNP are m uch less positive.

3.2. Shrinkage Pattern Trace

The SHRINKAGE PATTERN trace shows how the gene ralized ridge "Delta Shrinkage-Factors" applied to the ordered "uncor related com ponents" 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 increas es, a ll deltas rem ain equal when $Q = 1$; th e tr ailing d eltas ar e sm allest when $Q < 1$; and the leading deltas are smallest when $Q > 1$.

Colors have som ewhat different interpretations in SHRINKAGE P ATTERN traces than in the COEFFICIENT trace. In both cases, co lors are ord ered: **FIRST, SECOND, THIRD, FOUTRH, FIFTH, SIXTH,** etc. In a COEFFICIENT trace, colo rs represent the X-variables in the order that they were specified in the regression formula: $\mathbf{Y} \sim \mathbf{X1} + \mathbf{X2} + \mathbf{X3} + \mathbf{X4} + \mathbf{X5} + \mathbf{X6}$. But in a SHRINKAGE PATTERN trace, these s ame colors represent the regressor principal axes in the decreasing order of their eigenvalues of $\mathbf{X'X}$: $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_r > 0$.

Since we are following an extreme shrinkage path shape of $Q = -5$ for the **longley2** dataset, we see in the S HRINKAGE PATTERN trace d isplayed below that essentially on ly the last two o ut of six shrin kage factors , δ_5 and δ_6, change between M=0 and M=2. Af ter all, th e las t two singular values (square roots of eigenvalues of $\mathbf{X'X}$) are nearly equ al and are m uch smaller than the other four singular values. In fact, the las t two shrin kage factors have essentially been reduced to zero at M=2.

SHRINKAGE PATTERN: Q-shape = -5
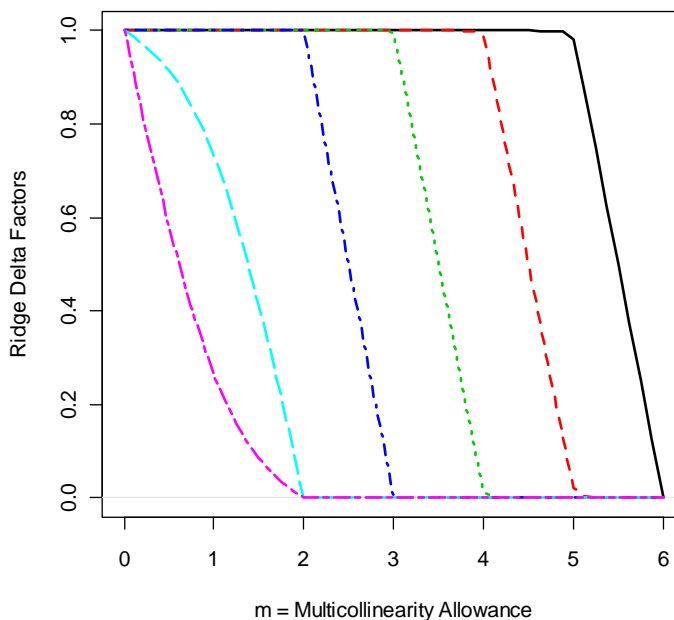
As shrinkage then continues from M=2 to M=3, the fourth shrinkage factor, δ_4 , essentially decreases from 1 to 0 …while δ_1, δ_2 and δ_3 all remain near 1. As was clear from the COEFFICIENT trace displayed above, the majority of the severe ill-conditioning in the longley2 dataset (i.e. switches in β-coefficient signs and drastic changes in their relative magnitudes) is confined to the last three out six total principal components of X-space.

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

The RELATIVE MSE trace displays norm   al   distribution theory, "m   odified" m aximum likelihood estim ates of "scaled" MSE risk in indi   vidual–coefficient estim ates as shrinkage of shape Q occurs.

Risks are "s caled" by b eing divid ed by the usual estim ate of the erro r (dis turbance term) variance. In other words, scaled risk expresses imprecision in fitted coefficients as a multiple of the variance of a single observation. Furthe rmore, when regression disturbance term s are assumed to be uncorrelated and homoskedastic , the "scaled" MSE risks of the unbiased OLS estimates (at the ex treme left of the trace where $\Delta = I$) a re **known quantities**, being the diagonal elements of the $(\mathbf{X'X})^{-1}$ matrix.

When shrinkage Δ fa ctors are less than 1, maxim um likelihood scaled risk estim ates are "modified," first of all, so as to be unbiased   under norm al theory. Then they are adjusted upward, if necessary, to have correct range re lative 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")

RELATIVE MEAN SQ. ERROR: Q-shape = -5
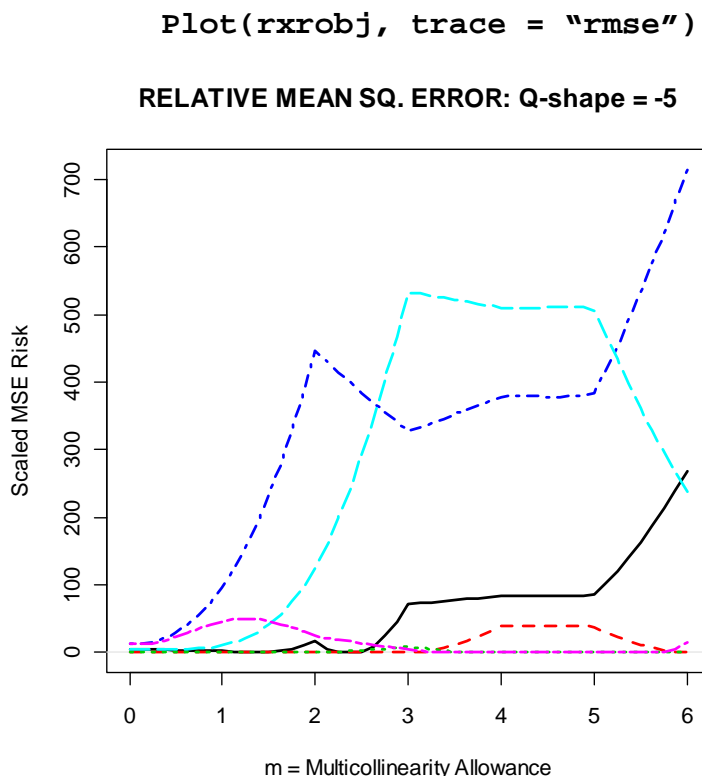
Changes in the **6th (Employment)** β-coefficient estimate between M=0 and M=3 first increase but then decrease MSE risk. Initial increases in the **1st (GNP.deflator)** β-coefficient estimate between M=0 and M=2 are relatively unimportant, but subsequent shrinkage increases MSE risk at M=3 and beyond. 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 eigenva lues of the estim ated 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 th e corresponding shrunken estim ators yield sm aller MSE risk than OLS in a ll directions of the r-dimensional space spanned by X-predictors (i.e. all possible linear com binations.) 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 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")

EXCESS EIGENVALUES: Q-shape = -5

m = Multicollinearity Allowance
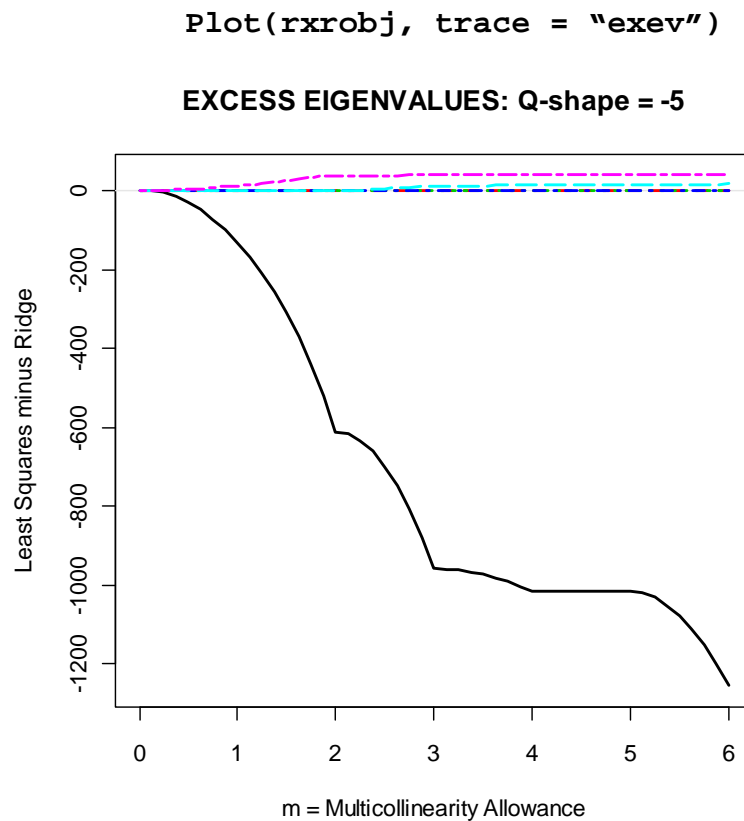
The negative eigenvalue at the M = 4.750 extent of shrinkage suggested by the CLIK criterion is **−1017** while the corresponding two largest eigenvalues are **+13.9** and **+39.1**. In other words, the **longley2** datase t is r ather cle arly very h ighly ill-conditioned. In f act, ill- conditioning is sufficiently bad that the am ount 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.
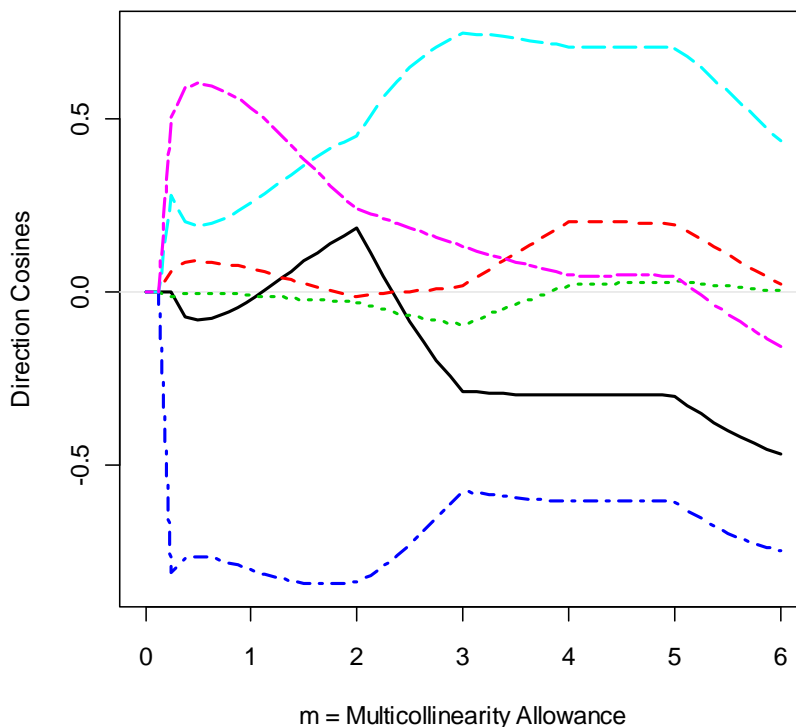
3.5. Inferior Direction-Cosine Trace

The INFERIOR DIRECTION trace displays the **direction cosines** (elem ents of the norm alized 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 colum ns of t hese MSE m atrices are in the order specified on the right-hand-side of the regression for mula **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")

INFERIOR DIRECTION: Q-shape = -5

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 I NFERIOR 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 reg ression coefficients have been shrunken to (0, 0, …, 0) at the right-hand extreme of all TRACE displays, M = rank(**X**). This is usually m uch-too-much shrinkage, so the inferior direction typically points backwards from (0, 0, …, 0) essentiall y towards the or iginal ±OLS coefficient vecto r at M = 0. In the abov e plot for th e **longley2** data set, the displayed direction cosines at M = 6 clearly point to the negative of the original OLS vector.

When two curves on an INFERIOR DIRECTION tr ace cross, their direction cos ines are clearly equal at that value of M. This happens with the cosines for the **1**st **(GNP.deflator)** and **2**nd **(Unemployed)** regressors at M = 1.295, where the comm on cosine value is +0.041. Thus, at M

= 1.295, the shrunken estim ate of t he SUM of the **1st** and **2nd** β-coefficients (0.512) can have higher MSE risk than its OLS esti mate(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 contra st, the vector (+1,−1,0,0,0,0) IS orthogonal to the inferior direction at M = 1.295, and thus the DIF FERENCE between the shrunken estimates of the **1st** and **2nd** β-coefficients (0.513) should have the sam e or lower MSE risk than the corresponding difference in OLS estimates (0.527.)

A sim ilar crossing of cosines for th e **5th (Year)** and **6th (Employed)** reg ressors occu rs at M = 1.535, where the common cosine value is +0.373. Thus, at M = 1.535, the shrunken estim ate 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 D IFFERENCE between the shrunken estim ates of the **5th** and **6th** β-coefficients (−0.656) should have the sam e or lowe r MSE risk than the corresponding difference in OLS estimates (−0.637.)

M-extents o f shrinkage such that two regresso rs have inf erior direction cosines with equal magnitudes <u>but opposite numerical signs</u> have the opposite effects on the MSE risks of sums and differences. The SUM of the corresponding shrunken coefficients then has the sam e 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, wher e the direction cosines are ±0.164. Unfortunately, shrinkag e to M = 2.67 has inappropria tely 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])              rxlobj <- RXlarlso(form,data=longley2)
larsobj <- lars(xlong2,ylong2,type="lar")      plot(rxlobj, trace = "coef")
plot(larsobj)
```
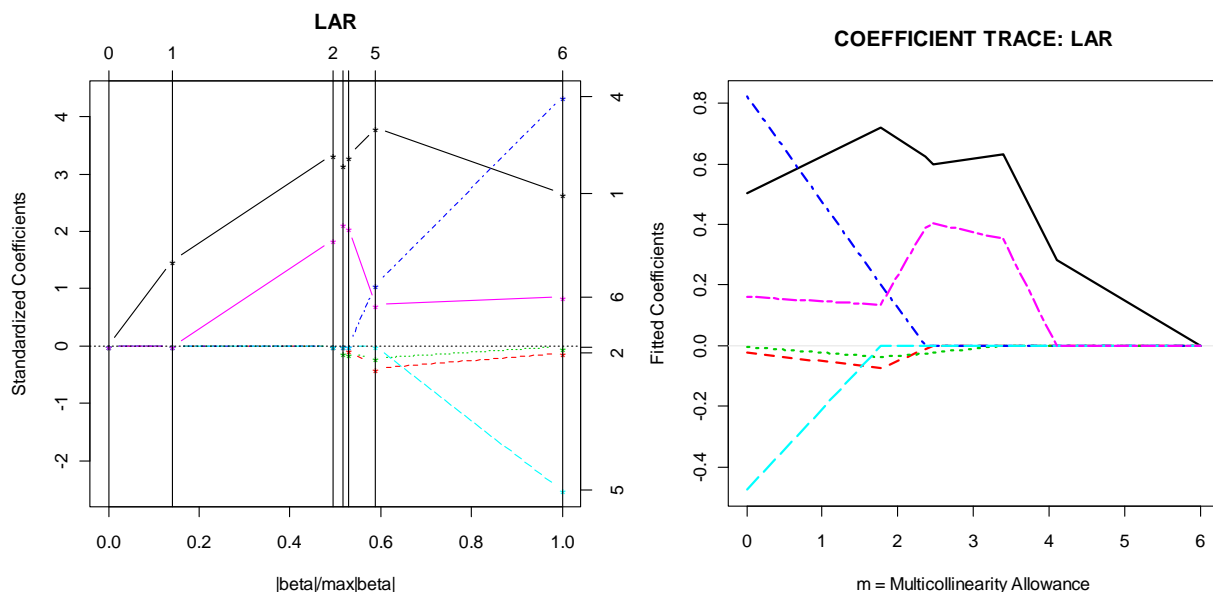
The first thing to note about the coefficient TR ACE displays from the RXlarlso() and Rxuclars() functions within the **RXshrink** pack age is that they are essen tially "bac kwards" relative to th e default coef ficient d isplays from the **lars** R-package. This point is illustrated above. Close examination of this pair of graphs also shows that, besides being backwards relative to each other, th ere are som e addition al, r ather m inor dif ferences b etween 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 S ection §2, this m eans the one or more of the shrinkage "delta" factors im plied by the `lars()` estimate starts out being g reater than one. Sim ilarly, as lar shrinkage occurs, one or m ore of these im plied delta-factors may eventually becom e negative. These points are illustrated in the graph below.

**Plot(rxlobj, trace = "spat")**

SHRINKAGE PATTERN: LAR
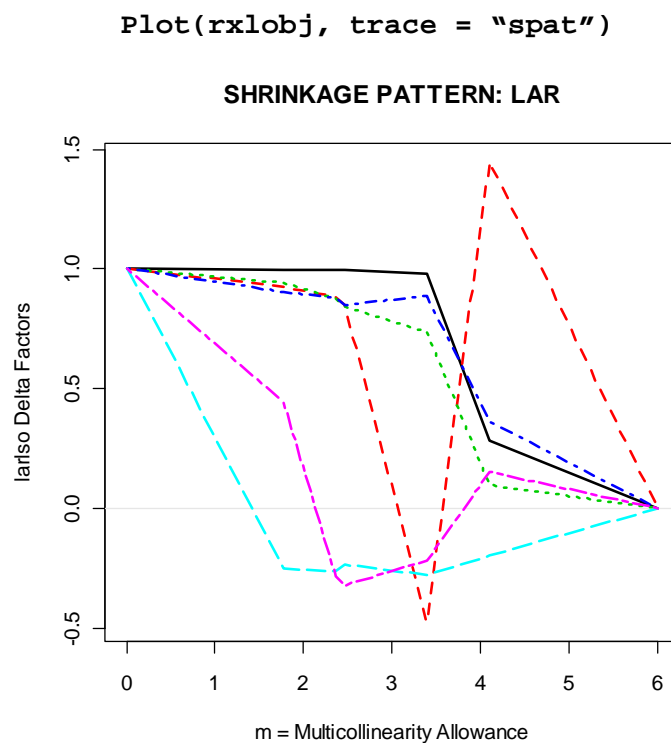
Reductions in MSE risk rela tive to OLS usually o ccur only when all of the delta-factors im plied by `lars()` estimates are non-negative and strictly less than +1. Exceptions can occur when the unknown true "gamm a" com ponent co rresponding to an "out of ra nge" delta-factor is nearly zero.

The Q-shape shrinkage paths typically used in "generalized ridge" regression depend upon the **eigenvalue spectrum** o f the centered **X**'**X** m atrix as well as upon the **principal correlations** with the centered respo nse vecto r, **y**. In sharp contrast, the shrinkage paths im plied by "least angle" regression methods typically depend only upon **correlations** (marginal or pr incipal) with the response **y**-vector. As a direc t 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 **<u>principal correlations</u>** 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")
```

SHRINKAGE PATTERN: LAR

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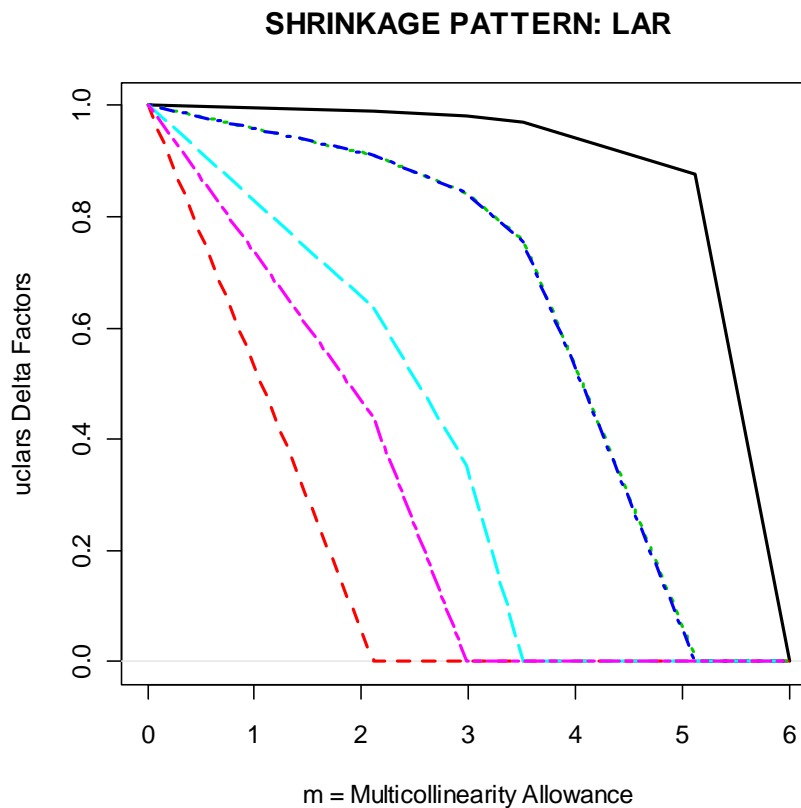
Note that, because the **3rd** and **4th principal coordinates** of **X**-predictors have nearly equal absolute correlations (`0.1222` and `0.1209`) with the response y-vector, the **3rd** and **4th** shrinkage delta-factors in the above graph are essentially equal.

5. Final Remarks

The **RXshrink** package for **R** is fully documented with *.Rd, *.tex, *.html and *.chm files. The additional information provided here is supplemental.

Traditional visualizations of shrinkage regression computations use "trace" plots. In a trace, P quantities (several estimated coefficients, risks, shrinkage factors, etc.) are plotted vertically

against a ho rizontal indicator of the extent of shrinkage. Trad itional "ridge" traces display th e Ordinary-Least-Squares (OLS) solution at the ir le ft-hand extrem e and cove r the full range of shrinkage that culm inates in "total" shrinkage at their right-hand extrem e (where all "centered" regression coefficient es timates becom e zero.) Here, P denotes the num ber of non-constant predictor variables in the regression model. **RXshrink** functions require P to be at least 2.

RXshrink functions attempt to iden tify shrunken coefficient estim ates that are either "good" in the sense that they dominate least squares estimates in every (multivariate) Mean Squared Er ror sense or are "optim al" in one well-defined (univa riate) MSE sense. Definition s for "good" o r "optimal" ridge shrinkage factors are based upon ri sk (expected loss) calcu lations that app ly to all forms of statistical distributions. But the ML inferences for the P-parameter and 2-parameter shrinkage paths explored by RXshrink functions are based upon standard norm al-distribution-theory.

REFERENCES

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

Brown L. " Estimation with in completely sp ecified lo ss f unctions (the case of sev eral loc ation parameters.)" **Journal of the American Statistical Association** 1975; 70: 417-427.

Bunke 0. "Least squares estim ators as robust and m inimax estim ators." **Math. Operations forsch u. Statist.** 1975(a); 6: 687-688.

Bunke 0. "Im proved inference in linear models with addition al inform ation." **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 num bers and m inimax ridge-regression estim ators." **Journal American Statistical Association** 1985; 80: 753-758.

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

Efron B, Ha stie T, Johnstone I, Tibshi rani R. "Least angle regress ion." **Annals of Statistics** 2004; 32: 407-499 (including discussion.)

Frank IE, Freidm an JH. "A statis tical v iew of som e chem ometrics regression tools." **Technometrics** 1993; 35: 109-148 (including discussion.)

Goldstein M, Sm ith 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. "Applica tion of Ridge Analysis to Regression Problems." **Chemical Engineering Progress** 1962; 58: 54-59.

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

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

James W , Stein C. "Estim ation with quadra tic loss." **Proceedings of the Fourth Berkeley Symposium** 1961; 1: 361-379. University of California Press.

LeBlanc M, Tibshirani R. "M onotone 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 Parametri zations for Varian ce-Covariance Matrices." **Statistics and Computing** 1996; 6: 289-296.

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

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

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

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

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

Obenchain RL. Ridge regression syst ems for MS-DOS personal com puters. **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, Baye   s, spline and Stein." **www.iquest.net/~softrx** eBook-in-Progress 1992--2005. (200+ pages.)

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

Stein C. "Inadm   issibility of the usual esti   mate of the m ean of a m   ultivariate norm al 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 shrinkag   e and selection via the lasso."   **Journal of the Royal Statistical Society B** 1996; 58: 267-288.

Tukey J W. "Instead of Gauss-Ma rkov Least Squares; W hat?" **Applied Statistics,** ed. R. P. Gupta. 1975. Amsterdam-New York: North Holland Publishing Company.