RXshrink_in_R

Generalized Ridge and Least Angle Regression estimates most likely to have minimum MSE risk

"vignette-like" documentation for R-package RXshrink

version 2.3 - August 2023

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

Robert L. (Bob) Obenchain, PhD
Principal Consultant, Risk Benefit Statistics
1006 Pebble Beach Dr, Clayton, CA 94517-2211
wizbob@att.net http://localcontrolstatistics.org

Table of Contents

1. A Personal Summary of 65+ 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 am delighted with today's acceptance of regularization (ridge/shrinkage) methods for fitting linear models. Anyway, I will summarize some personal perspectives on why and how researchers, statisticians and data scientists appear to have become somewhat enlightened about shrinkage over the last 65+ 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" and thus did nothing to adjust the ***relative magnitudes*** of correlated estimates for their 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 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 ...rather than "length" of the OLS beta-vector), 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.

My early ridge-shrinkage papers (1975, 1977, 1978, 1981) lacked focus and simplicity and were certainly less impactful than the regression publications of AT&T Bell Labs "giants" like Collin Mallows (1973, 1995) and John Tukey (1975). I truly love details, and my papers have always been chuck-full of many-too-many alternative concepts.

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

Unfortunately, only three of my papers on shrinkage **applicants** and **software**, Obenchain (1984, 1991, 1995), have been accepted for publication. As illustrated in Section §3 of this vignette, shrinkage TRACE displays suggest "where" MSE risk can be reduced by shrinkage. My 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: Any “linear” estimator identified as being most likely to be optimal is actually 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 tend to result in increases of less than 5% in MSE risk.

As Burr and Fry(2005) noted, sound strategy / tactics in **shrinkage estimation** require analysts fitting linear models to be “cautious” rather than “greedy” (overly optimistic about risk reduction.)

Frank and Freidman (1993), Breiman (1995), Tibshirani (1996), LeBlanc and Tibshirani (1998) Efron et al. (2004) and Hastie (2020) are currently keeping the shrinkage / GRR “home fires” burning for analyses of Linear Models using truly “Big Data” (very large **n**) and/or “Wide Data” (**p** much greater than **n**) in genomics or “document classification” studies.

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.

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

This first formula defines the (**p** by 1) vector of regression coefficient estimates within 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.

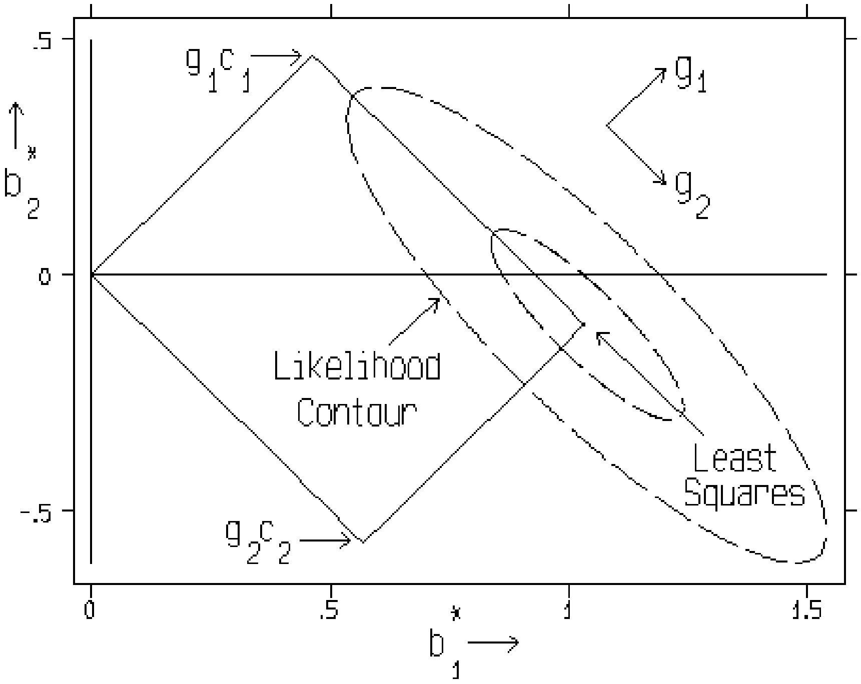
The (**r** by **r**) diagonal "Lambda" matrix above contains the **ordered** and **strictly positive eigenvalues** of **X'X**; $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$. Thus our operational rule for determining the ***Q***-th power of **X'X** (where ***Q*** may not be an integer) will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

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

where the (**r** by **r**) diagonal **Δ** matrix contains multiplicative **shrinkage factors** along the **r** principal axes of **X**. Each of these **Δ** shrinkage-factors is within **[0,1]** …i.e. $\mathbf{0} \leq \delta_j \leq \mathbf{1}$ (j = 1, 2, ..., **r**.)

Note that the ($\mathbf{r}$ by 1) column vector, **c**, contains the **uncorrelated components** of the ordinary least squares estimate, beta-hat = **Gc** = $\mathbf{g}_1$ c_1 + $\mathbf{g}_2$ c_2 + … + $\mathbf{g}_r$ c_r , of the **unknown, true OLS** regression coefficient β^o vector. The variance matrix of **c** is the diagonal Λ^{-1} matrix multiplied by the scalar value of the **sigma-square for error, σ^2**. A **p = r = 2**-dimensional case is depicted in the plot below…



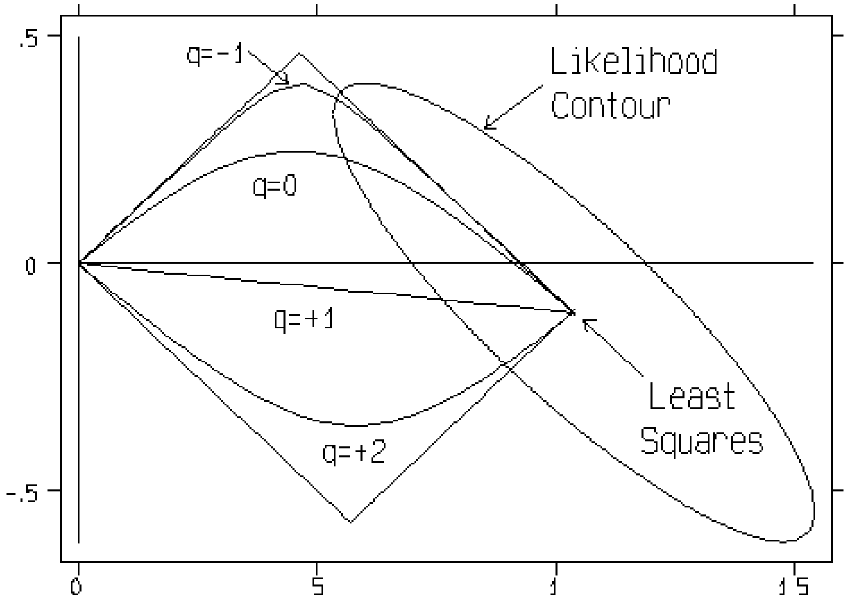
In fact, we can now literally "see" that the 2-parameter family of shrinkage estimators from our **first (Q-shape) equation**, Page 3, is the special case of the **last (GRR) equation** above when **Q = q** and...

$$\delta_j = \frac{\lambda_j}{\lambda_j + k \cdot \lambda_j^q} = \frac{1}{1 + k \cdot \lambda_j^{(q-1)}}$$

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

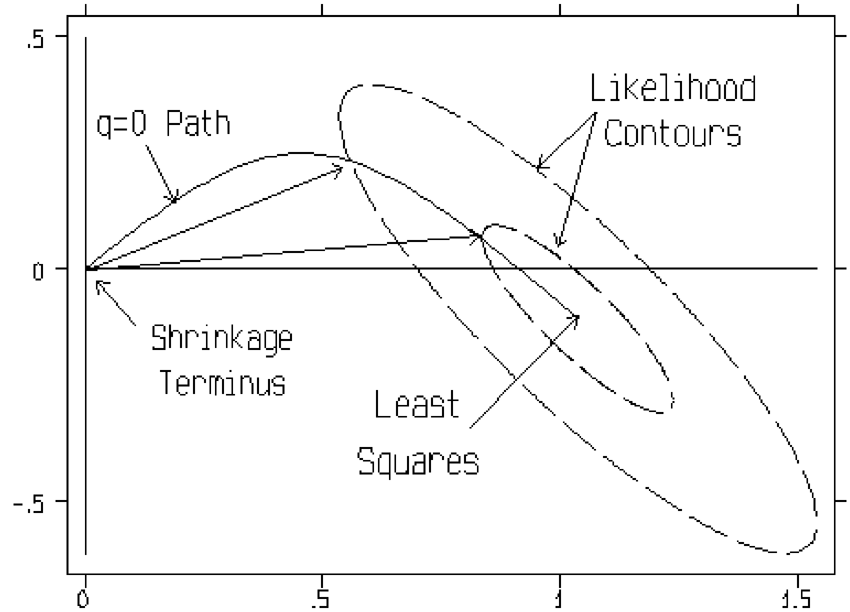
q = +1 ...yields uniform shrinkage (all Shrinkage Factors equal.)
q = 0 ...yields Hoerl-Kennard "ordinary" ridge regression.
q = −5 ...is usually very close, numerically, to "Principal Components Regression," with exact agreement in the limit as ***Q*** approaches minus infinity.

Here, then, is a graphic showing a <u>range of shrinkage path **q-shapes**</u> for the rank(**X**) = p = 2 case previously depicted above (Page 5).



The most widely recognized <u>special case</u> of a ***q***-shaped path is unquestionably ***q*** = 0 for Hoerl-Kennard(1970) **"ordinary" ridge regression**. This path has a dual "characteristic property," illustrated in the figure on Page 6. Namely, the ***q*** = 0 path contains not only

1) the <u>shortest</u> beta estimate vector of any given <u>likelihood</u> …but also
2) the <u>most likely</u> beta estimate of any given <u>length</u>.



Unfortunately, the length of a beta estimate vector has relatively little to say in determining its MSE risk characteristics! After all, the **statistical concept** that **$E(x^2) \geq [E(x)]^2$** is called "**variance**" …not "too bigness" or "too longness".

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.

An important limiting case is ***q* = -∞** 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. For example, note in the graphic near the top of Page 6 that the path with shape ***q* = −1** is already somewhat near this limit in the **p = 2 dimensional** case depicted there.

2.1 The Shrinkage "Extent" parameter:

m = *MCAL* = "multicollinearity allowance"

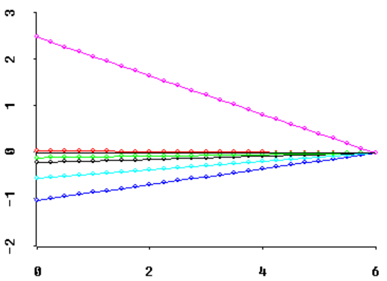
Unfortunately, the scalar **k** parameter (of pages 4 & 5) is not a particularly meaningful measure of **extent** of shrinkage. After all, the numerical sizes of the **r** shrinkage-factors, **(δ_1, δ_2, …, δ_r)**, can depend more on one's choice of ***Q* = *q*** than on one's choice of **k**. Specifically, the **k**–values corresponding to two rather different choices of ***Q*** are usually **not** comparable (i.e. they use two different scales.)

RXshrink 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:

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

Note that the range of ***MCAL*** is finite: **0 ≤ *MCAL* ≤ p** = Rank(**X**). Whatever may be your choice of *Shrinkage PATH* (unrestricted or ***Q***-shaped), 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*** = **p** (**k** = +∞ or **k^***=0 and **Δ** = **0**). For example, `qm.ridge()` uses Newtonian descent methods to compute the numerical value of **k** corresponding to given values of ***MCAL*** and ***Q***-shape.

In addition to always being finite, ***MCAL*** has a number of other distinct advantages over **k** when used as the scaling for the horizontal axis of **ridge TRACE diagnostic** plots. For example, shrunken regression coefficients with **stable relative magnitudes** form **straight lines** when plotted versus ***MCAL***.

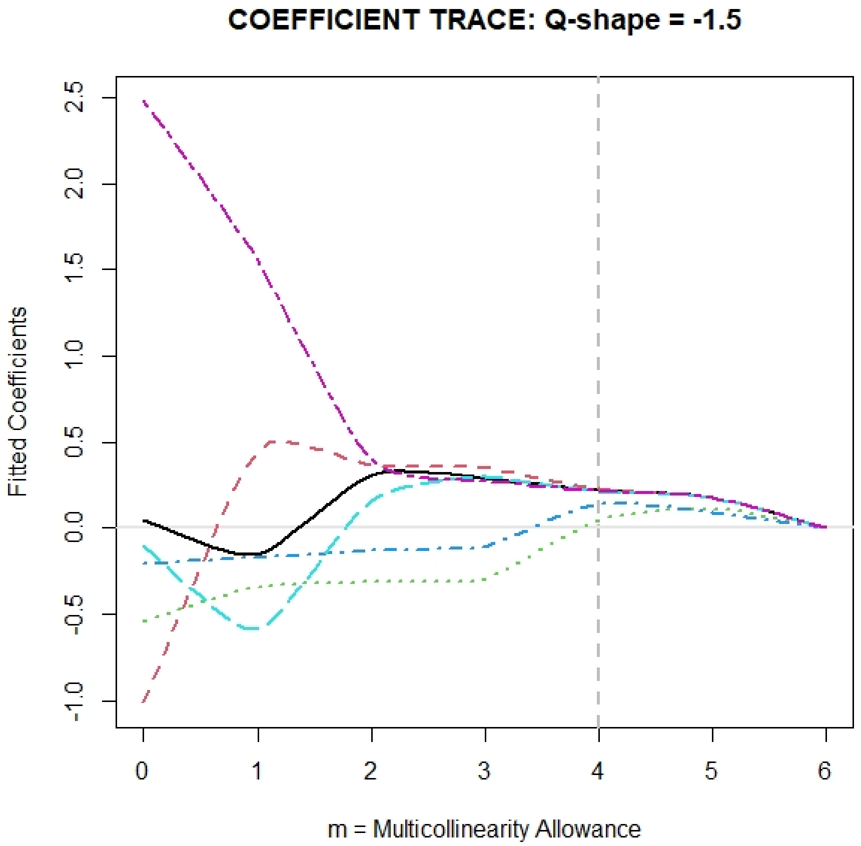


Similarly, the average value of all **r** shrinkage factors is **(r − MCAL)/r**, which is the Theil(1963) proportion of Bayesian posterior precision due to sample information (rather than to prior information.) Note that this proportion decreases linearly as ***MCAL*** increases.

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 [Page 9] for the ridge coefficient TRACE with path of shape ***Q*** = −1.5 using the **R** data.frame containing the data analyzed in Longley (1967), where the response is ***y*** = **Employed**. Compared with the major initial shifts in relative magnitudes and numerical signs of coefficients within **0 ≤ *MCAL* ≤ 2**, note that the trace [Page 9] becomes relatively much more stable (somewhat more "straight" curves) within **2 ≤ *MCAL* ≤ r = p = 6**. The vertical gray dashed-line at **m** = 4 marks the location of the shrunken (biased) coefficients most likely to have minimum MSE risk when restricted to the 1-dimensional ***Q*** = −1.5 Path through 6-dimensional likelihood space.

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

RXshrink computing algorithms provide strong, objective guidance on the choices between, say, **"efficient"** (**p**-parameter) and **[*Q* , m]** (2-parameter) Paths that are best for your data when **p ≥ 2**. Specifically, RXshrink implements the methods of Obenchain(1975, 1978, 1981, 2021) to identify the Path (and the **m**-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 the "Efficient" p-parameter Path

The shrinkage **δ–factors** used by the **eff.ridge( )** function assure that its p-parameter PATH is **as short as possible**. The shortest distance between two points (in Euclidean space) is along a straight line. Thus, the efficient path…

[1] starts at the vector of unbiased OLS β-estimates ($\delta_i \equiv 1$),

[2] follows a straight line until it reaches the point where all "p" shrinkage factors are most likely to yield optimally biased β-estimates ($0 < \delta_i = \delta_i^{MSE} < 1$), then

[3] follows a (different) straight line until it reaches the shrinkage "terminus" at ($\delta_i \equiv 0$).

Clearly, any other PATH through **these same three points** would be longer!
Obenchain (1975) showed that

$$\delta_j^{MSE} = \frac{\gamma_j^2}{\gamma_j^2 + (\sigma^2/\lambda_j)} = \frac{\lambda_j}{\lambda_j + (\sigma^2/\gamma_j^2)} = \frac{\varphi_j^2}{\varphi_j^2 + 1}$$

where $\varphi_j^2 = \gamma_j^2\lambda_j/\sigma^2$ is the unknown non-centrality of the F-test for $\gamma_j = 0$. Thus, the **Normal-theory Maximum Likelihood estimate** of ($\delta_j^{MSE} \times c_j$) is of the "cubic" form…

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

2.3 Shrinkage δ–factors for Least Angle and Lasso (Selection) Estimators

The `aug.lars()` and `uc.lars()` functions in the **RXshrink** R-package re-interpret "lar" and "lasso" regression estimators as generalized ridge estimators simply by solving the **p**-equations…

$$\boldsymbol{\beta}^{lar} = \mathbf{G}\boldsymbol{\Delta}^{lar}\mathbf{c}$$

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

$$\boldsymbol{\delta_j^{lar} = (g_j'\beta^{lar}) \,/\, c_j} \quad \textbf{for } \mathbf{j = 1, 2, \ldots, r.}$$

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 `uc.lars()` **function applies lar estimation directly to the uncorrelated components vector**, **c**, and this restriction does yield a true generalized ridge (shrinkage) estimator. In fact, the delta-factors for `uc.lars()` are of the form:

$$\boldsymbol{\delta_j^{uc.lars} = \max[\ 0,\ 1 - k^{uc} \,/\, |\rho_j|\ ],}$$

where ρ_j is the $\mathbf{j}^{th}$ "principal correlation" estimate …i.e. the observed correlation between the response **y**-vector and the $\mathbf{j}^{th}$ column of the **H** matrix of "principal coordinates" of **X** (Page 4.) Note that the k^{uc}-factor in this shrinkage formulation is limited to a subset of [0, 1]. Note that $\mathbf{m = 0}$ (and $\boldsymbol{\Delta}^{uc.lars} = \mathbf{I}$) occurs at $k^{uc} = 0$. The **m-Extent** of shrinkage then increases as k^{uc} increases, and $\mathbf{m = p = r}$ results when k^{uc} = the **maximum absolute principal correlation**.

2.4 Minimizing MSE Risk in unknown directions: Either Orthogonal to or Parallel to the true Beta.

A theoretical basis for detecting "wrong-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).** Because the vector of OLS estimates is of the form $\mathbf{X}^+\mathbf{y}$, its elements can have different signs from those of $\mathbf{X}'\mathbf{y}$ when the data are ill-conditioned. After all, when the "centered" columns of $\mathbf{X}$ are mutually orthogonal and of equal length, $\mathbf{X}^+\mathbf{y} \propto \mathbf{X}'\mathbf{y}$. So, when these sorts of "wrong-signs" do occur, it has traditionally been considered relatively bad news!

When the α vector (of length 1) in my Theorem 2 is **parallel** to the unknown, true β (i.e. when $\alpha = \beta / \sqrt{\beta'\beta}$), the corresponding MSE optimal ridge shrinkage factors are $\Delta \propto \Lambda$, yielding a $\beta^{(=)}$ estimate of the form $\mathbf{k}^{(=)} \times \mathbf{X}'\mathbf{y}$, where $\mathbf{k}^{(=)} \geq 0$ is the unknown scalar $(\gamma'\gamma) / [(\sigma^2 + \gamma'\Lambda\gamma) \times \mathbf{ssy}]$. The Normal-theory ML estimate of the true $\mathbf{k}^{(=)}$ scalar is…

$$\widehat{k}^{(=)} = (\Sigma r_{yj}^2/\lambda_j) \times \mathrm{n} /\{ \mathrm{ssy} \times [1+(\mathrm{n}-1)\mathrm{R}^2] \}.$$

[When the α vector lies within the p–1 dimensional space **orthogonal** to the unknown, true β, MSE risk is clearly reduced to **zero** by shrinking one's estimate of β to **zero.** While "interesting," this (secondary) observation does not appear to have clear practical implications.]

Noting that [1] the p-coordinates of the $\mathbf{k}^{(=)} \times \mathbf{X}'\mathbf{y}$ estimate with $\mathbf{k}^{(=)} \geq 0$ have same numerical signs as the vector of marginal correlations of **y** with **X** (...which is $\mathbf{X}'\mathbf{y}$) and [2] that the columns of **X** and **y** have been "centered" and rescaled to be of the same length (which defaults to $\sqrt{n-1}$ in **RXshrink** functions)

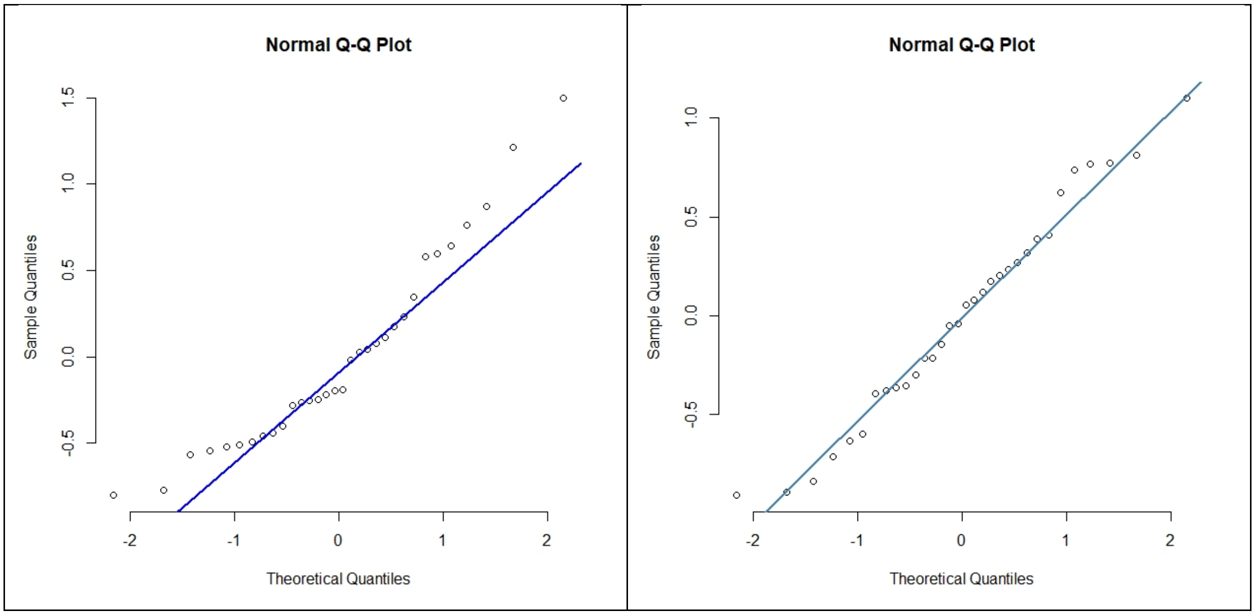
The maximum likelihood estimate, $\widehat{k}^{(=)}$, is computed by the **correct.signs()** function, where **ssy** defaults to (n-1). Here is an example invocation...

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

# NOTE: a "bottom portion" of the full correct.signs() output is...
#
# Comparison of Beta Coefficient Statistics...
#
#            OLS        X'y        Delta        B(=)       Bfit
# 1  0.503561780 0.99363526 1.877013e-01 0.041926973 0.22361771
# 2 -0.023697455 0.69671799 5.130368e-02 0.029398389 0.15679645
# 3 -0.002581592 0.07349887 1.201972e-02 0.003101324 0.01654093
# 4  0.821217919 0.98377362 1.980622e-03 0.041510856 0.22139835
# 5 -0.475600380 0.94794839 9.803393e-05 0.039999191 0.21333587
# 6  0.161192483 0.98412737 6.986262e-05 0.041525782 0.22147796
#
#     OLS Beta estimate uses No Shrinkage: Delta = I.
#     X'y is expressed here as Correlations.
#     B(=) Delta 'Shrinkage' Factors are proportional to LAMBDAs
#     B(=) also uses a Common Rescaling k-Factor = 0.001506983 here.
#     Bfit estimate parallel to B(=) has minimum Residual Sum-of-Squares.
#     Multiplicative B(=) Factor yielding Bfit estimate = 5.333505
#
# Residual Sum-of-Squares: Lack-of-Fit...
#            OLS    B(=)      Bfit
# RSS 0.01852492 18.7639 0.8217201 ...Since B(=) estimates are drastically
#       shrunken here, their predictions display considerable lack-of-fit.
#
# Squared Correlation between the response and its predictions...
#           OLS      Bfit
# Rsq 0.9993384 0.9706529
```

2.5 MLtrue( ) : Generate Data.Frames for Linear Models with Known Parameters and i.i.d. Normal "Errors"...

Since RXshrink functions stress calculation of Normal-theory maximum-likelihood estimators for GRR applications to linear models, users interested in simulating the true distributions of such estimates should find the **MLtrue()** function [as well as the **MLboot(), MLcalc() and MLhist()** functions of Section §5] particularly helpful. Here are some examples of **MLtrue()** usage:

```
library(RXshrink)
data(mpg)
# Suppose we try a simple Model with only p=2 X-variables...

form2 <- mpg~cubins+weight

# Since OLS provides "optimal" residuals [Obenchain(1975),JASA]
# while linear model residuals from biased estimates can be much
# larger, we start by examining a "q-q" plot of OLS residuals
# for the above "form2" model...

lm2mpg <- lm(form2, mpg)
qqnorm(lm2mpg$residuals, pch = 1, frame = FALSE)
qqline(lm2mpg$residuals, col = "blue", lwd = 2)

# This "q-q" plot, displayed at the top-left of the next page
# (13), suggests that either the above linear model is "wrong"
# or else that any errors-in-measurement of mpg are not very
# close to being Normal with mean=0 !
```



```
# The MLtrue() function can generate a Y-outcome vector, Yvec,
# such that the above "form2" model is a "correct" model and
# all error-terms in Yvec are generated [using rnorm()] to be
# more like i.i.d. Normal variates. The "q-q" plot above (on the
# right) shows that the MLtrue() OLS residuals (generated as
# shown below are indeed more like "i.i.d. Normal" errors.
```

The **"go"** argument to the **MLtrue()** function implements a logical **(TRUE,FALSE)** switch between its 2 alternative modes of operation:

go = TRUE generates a (new) y-Outcome variable for specified parameter values, while
go = FALSE estimates parameter settings for a current (existing) y-Outcome variable.

We now demonstrate how this “feature” can be helpful…

```
mpgF <- MLtrue(form2, mpg, go=FALSE)

mpgF    # Print current estimates... (greatly abbreviated below)

# OLS Residual Mean Square for Error: truv = 0.2341777
# OLS Beta Coefficients    = -0.3644932 -0.5439967
# Uncorrelated Components = -0.6423994 0.1269281

# Because these values look reasonable to use as TRUE values, we
# simply use most estimates as new “defaults”, but call MLtrue()
# with "truv = 0.4" (almost doubling the error variance) and
# "go = TRUE"...

mpgT <- MLtrue(form2, mpg, truv=0.4, go=TRUE)
```

The `mpgT` output list from `MLtrue()` now contains a data.frame named `new` …with 2 new columns, named `Yvec` and `Yhat` as well as copies of x-variables `cubins` and `weight` and the original y-Outcome variable, `mpg`. Specifically, `Yvec` = `Yhat` + errors, where `Yhat` is the (new) generated vector of (true) `expected values`.

```
names(mpgT$new)
# "Yvec"    "Yhat"    "cubins" "weight" "mpg"
```

We now illustrate fitting of a linear model with optimally biased estimates from `eff.ridge()` using the new `Yvec` of simulated (higher variance) `mpg` values...

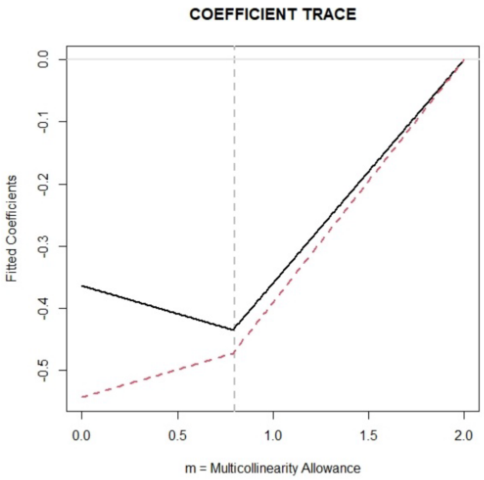
```
formT <- Yvec~cubins+weight   # New model formula...

# In fact, let us now compare...
eff.fit2 <- eff.ridge(formT, mpgT$new)
# with...
eff.fit1 <- eff.ridge(form2, mpg)
# using just their respective "coef" and "rmse" plots...
```

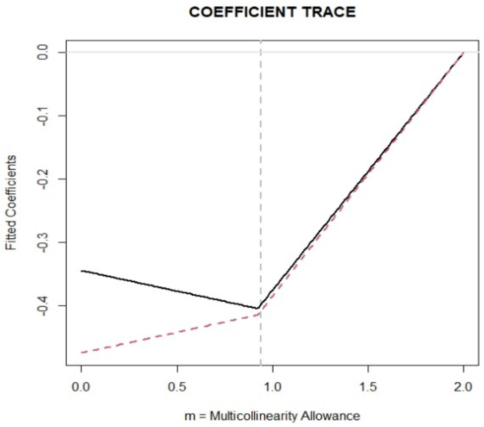
At least two types of things are going on that could make the below pairs of **TRACE Diagnostics** “look” somewhat different…

- Unlike the simulated `MLtrue()` `Yvec` Outcome, the linear model for predicting the original `mpg` y-Outcome from `cubins` and `weight` may be a **WRONG model** or (as noted above) may have **non-Normal** errors-in-measurement.

- `Yvec` is known to have **true** variance = `0.4000` while the “errors” in `mpg` were estimated by `MLtrue()` to have variance `truv = 0.2342.`

```
plot(eff.fit1, trace="coef")        plot(eff.fit2, trace="coef")
```

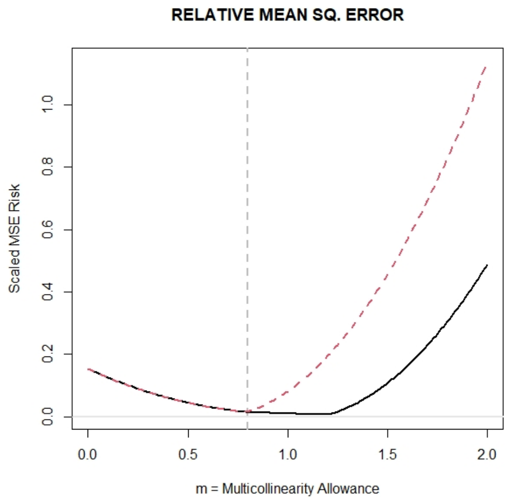




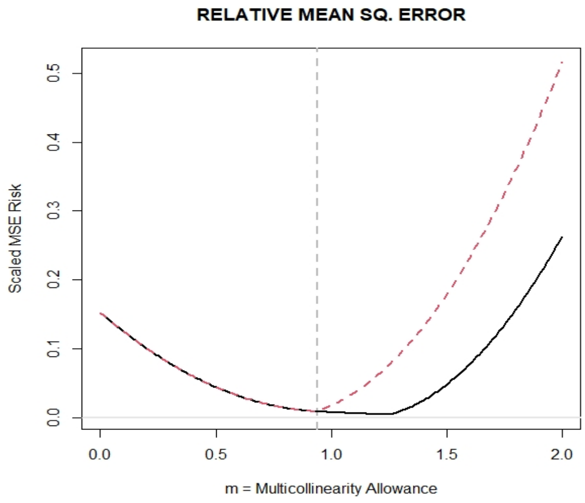
The above "**`coef`**" TRACEs show that ML estimates of "optimal" **m-Extents** for shrinkage are somewhat similar (**`m = 0.8001`** for **`unr.fit1`** and **`m=0.9416`** for **`unr.fit2`**), but the β-coefficient estimate for **`weight`** is slightly less negative in **`unr.fit2`**, where the ML **`truv`** estimate is **`0.3919`** (rather than **`0.4000`**).

There are almost surely **"some" TRUE Differences** in "optimal" **Variance-Bias Trade-Offs** between the two different Linear Model setups that we are trying to compare here!

```
plot(eff.fit1, trace="rmse")
```



```
plot(eff.fit2, trace="rmse")
```



While the above pair of "Relative MSE" plots appear rather "similar," note their considerable difference in **Vertical Scaling!** Although **`unr.fit2`** suffers a near doubling of the Error Variance within **`unr.fit1`**, this lead to a modeest (17.7%) increase in overall **m-**Extent of Shrinkage (from **`0.8001`** to **`0.9416`**) in **`eff.fit2`**. In summary, while the **Relative RISK** is generally less than half that of **`eff.fit1`** in **`unr.fit2`**, the corresponding **full MSE Risk** (**σ^2 × rMSE**) is roughly the same in these 2 example situations.

3. Interpretation of Ridge TRACE Diagnostic plots

In addition to the original `longley` data, we will use the `longley2` numerical example here in Section §3 to illustrate interpretation of Ridge TRACE Diagnostics. The `longley2` 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 in ill-conditioned regression computations. Note that 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.)

The "motivating" feature of the **Efficient Shrinkage** Path used by **eff.ridge( )** is that it deliberately heads directly towards (and always passes directly through) the Minimum MSE Risk point-estimate of optimally biased β-coefficients; see Section §2.2 (page 9).

The Deprecated **unr.ridge( )** function implemented a slightly **longer** and **more complicated** Path with (p – 1) "interior" Knots. The **eff.ridge( )** Path always has only one "interior" Knot, and the optimally biased β-coefficient estimates occur at this m-Extent of Shrinkage.

Together, the `longley` and `longley2` data.frames represented a pair of uniquely challenging examples of fitting simple linear models to ill-conditioned (confounded) data. One clear difficulty is that unusually high Coefficients of Determination, $\mathbf{R}^2 > 0.999$, result. Unusually low Error Mean-Square estimates, $\sigma^2 < 0.001$, thus apply.

We start by loading the **`RXshrink`** package, then executing the following **R**-code:

```
library(RXshrink)
data(longley2)
form <- Employed~GNP.deflator+Unemployed+Armed.Forces+Population+Year+GNP
effobj <- eff.ridge(form, data=longley2)
effobj
```

Because `effobj` is an R-object of class `eff.ridge`, the last line of code prints the (default) `eff.ridge()` output. [The command `str(effobj)` would display the full "structure" of this output object.] Since the default output is rather detailed, only a small portion of it is displayed below:

```
    Number of Regressor Variables, p = 6
    Number of Observations, n = 29

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

COMP = 6 × 1 vector of Uncorrelated Components of the OLS estimator, **c = G'β**^{o}.
RHO = 6 × 1 vector of Principal Correlations between the response **y** and the columns of **H**.
Note that the 1st RHO is huge, while the 2nd , 5th and 6th are small.

The ill-conditioning in this example is extreme. Note that the (in-significant) **Sixth uncorrelated COMP is larger (numerically) than the First Three** (highly significant) components**!** This happens primarily because the LAST singular value, SV_6 = sqrt($LAMBDA_6$), is **relatively small**.

$$c_j = \sqrt{y'y}\ (\frac{\hat{\rho}_j}{\sqrt{\lambda_j}})$$

[Technical Note: Under the conditional distribution-theory of interest here, the X-vectors (and their corresponding H-vectors of Principal-Coordinates) are considered GIVEN (rather than RANDOM). Thus, while ρ-estimates certainly have the "functional form" of correlation coefficients, their true distribution is simply that of linear combinations of random y-outcomes. Such distributions are typically assumed to be (at least approximately) Normal when fitting linear models.]

The final 5-lines of default OUTPUT that are displayed for each `eff.ridge()` object provide key information about optimal m-Extents of Shrinkage for the specified linear model. Here is that information for the `Longley(1967) model` on the `longley2` data.frame (which has n=29 observations for 1947 through 1975)…

```
    Most Likely Unrestricted Shrinkage Extent, mStar = 0.2509
    Corresponding Expected -2*log(Likelihood Ratio)  = 0.0

    Most Likely m-Value on Observed Lattice,     mClk = 0.25
    Smallest Observed -2*log(Likelihood Ratio), minC = 36.29

    dMSE Estimates = 0.9999764 0.8359725 0.9984740
                     0.9984412 0.9746134 0.9416156

                   M          CLIK         EBAY       RCOF
           0   0.000           Inf          Inf        Inf
           1   0.125  9.031336e+00     38.28041   38.83900
           2   0.250  1.910701e-04     36.29434   37.17625
           3   0.375  2.441361e+02    740.22821  120.77747
           4   0.500  2.192520e+02   1459.92565  137.73989
```

Meanwhile, here are corresponding results when fitting the model of `Longley (1967)` using the (original) `longley` data.frame (n=16 observations for years 1947-1962)…

```
    Most Likely UNRestricted Shrinkage Extent, mStar = 0.2108
    Corresponding Expected -2*log(Likelihood Ratio)  = 0.0
```

```
Most Likely m-Value on Observed Lattice,    mClk = 0.20
Smallest Observed -2*log(Likelihood Ratio), minC = 32.17
dMSE Estimates = 0.9999686 0.9827091 0.9966629
                 0.9507096 0.9664612 0.8926613
```

SUMMARY COMMENTS on “Lattice Density”

Like the three “other” primary **RXshrink** package functions for generating GRR **TRACE Diagnostic** plots, the `eff.ridge()` function uses a uniform lattice of **m-extents** for shrinkage that now **defaults** to **steps = 20** per unit increase in **m**. Each **m**-value on the default lattice is some unit-multiple of 1/20 = 0.05.

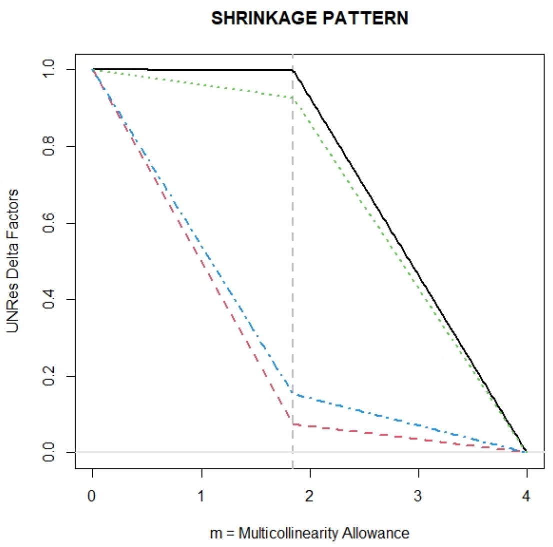
While users are free to specify (integer) values for `steps` that are larger or smaller than `20`, it’s difficult to fully anticipate which integer value for `steps` will yield a lattice that either “contains” or else “will come close to” an observed MSE optimal m-Extent (say, m = 1.83368.)

…Location of “Knots” in Piecewise Linear Functions

The **eff.ridge()** shrinkage Path is most “efficient” (shortest) and has only a single “interior” knot. Its Path “ends” at **m** = **p** (the Shrinkage Terminus where $\boldsymbol{\beta}^* = \mathbf{0}$).

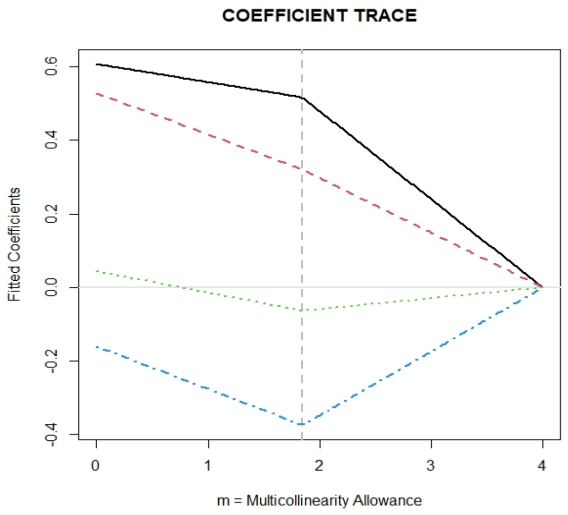
Next, we display four **`Trace`** plots [**`“spat”`**, **`“coef”`**, **`“rmse”`** and **`“exev”`**] for the **`haldport`** data.frame using **`p = 4`** “carrier” X-variables to explain the **`y = heat`** “catcher” variable. The first two **`Traces`** display “two-piece linear functions” that confirm that changes in shrinkage δ-factors and β-coefficient estimates are (purely) linear as **m-Extent** increases…

`eff.ridge “spat” plot`



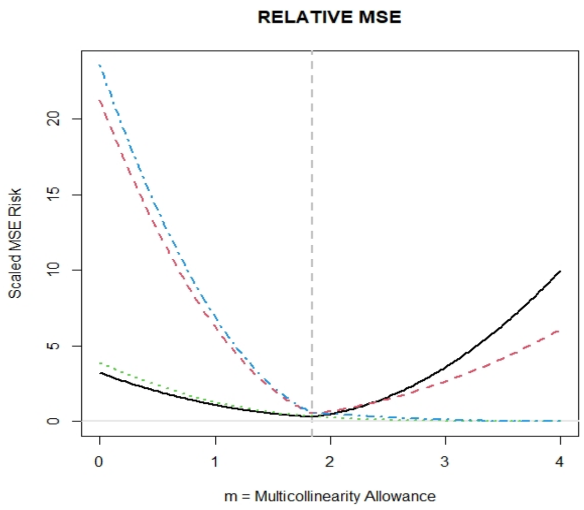
single “interior” knot at m = mStar

`eff.ridge “coef” plot`

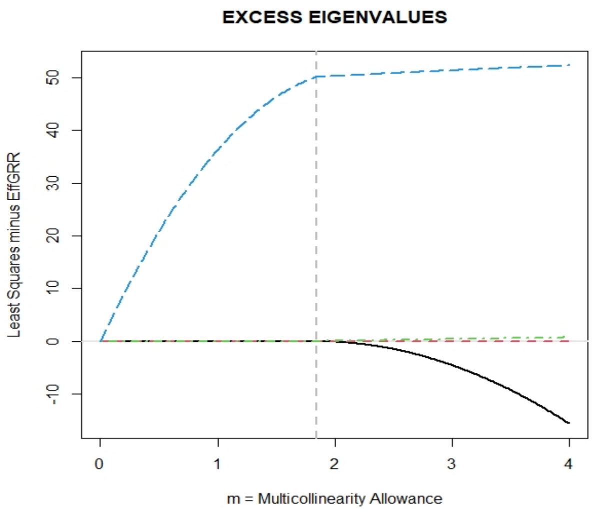


single “interior” knot at m = mStar

eff.ridge "rmse" plot



eff.ridge "exev" plot



In sharp contrast, the two final Traces shown here (page 19) illustrate that the MSE RISK characteristics of **eff.ridge( )** estimates change in distinctly non-linear ways as "linear" shrinkage occurs!

Finally, note that each **eff.ridge( ) Trace** plot displays a vertical (gray) dashed line at the Optimal (minimum MSE risk) extent, **m = mStar = 1.847759,** of shrinkage.

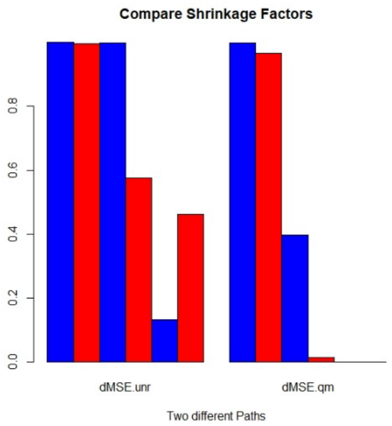
While the **eff.ridge( ) Path** is rather "new" (2020-21), the **2-parameter [Q, m] Paths** were originally proposed more than 45 years ago [Obenchain (1975)] and were discussed in our **Introduction (§2)** …see Pages 4 through 10. Note that **[Q, m] Paths** are sufficiently flexible to pass through the overall minimal MSE risk point-estimate only when **p = 2**. Furthermore, locating the truly **Optimal Q-shape** (and/or **m-Extent)** of Shrinkage may require using many decimal places!

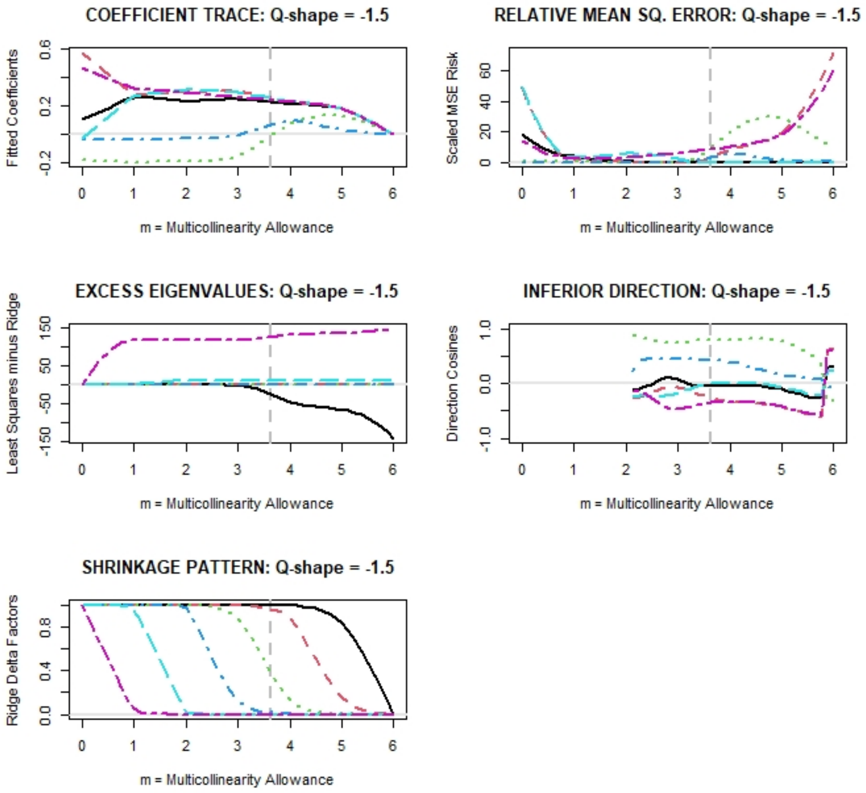
By default, **qm.ridge( )** searches for the ML choice of Q-shape only on a lattice of `nq = 21` numerical values between `qmax = +5` and `qmin = -5`. While users of **qm.ridge( )** may, of course, use whatever parameter choices they wish, this default is to search for the "best" integer or half-integer **Q-shape** within `[-5,+5]`.

Let us now examine Trace displays for another dataset (`longley2`) included in the ***RXshrink*** **R**-package:

```
library(RXshrink)
data(longley2)
form <- Employed~GNP.deflator+GNP+Unemployed+Armed.Forces+Population+Year
qmL2obj <- qm.ridge(form, data=longley2)
qmL2obj # print (detailed) results . . .
plot(qmL2obj)  # 5 Traces on 1 plot . . .
```

The `barplot()` display at right compares the optimal δ^{MSE} shrinkage factor estimates for the `eff.ridge()` Path with those from the "best" `qm.ridge()` Path. A fundamental property of [**Q**, **m**] Shrinkage is that all **p** of these factors are monotone **increasing** when **Q** > 1, all are **equal** when **Q** = 1, and all are monotone **decreasing** when **Q** < 1. Since "unrestricted" paths have no monotonicity properties, the 3rd δ-factor can be (slightly) larger than the 2nd, and the 6th can be (much) larger than the 5th. But the ML **Q**-shape is negative (-1.5) here, so **monotone decreasing** δ-factors result. In any case, it is rather clear that too-much shrinkage results from using a [**Q**,**m**] Path on the `longley2` data.frame.





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

See (abbreviated) listing below...

```
          Q         CRLQ        M            K      CHISQ
1       5.0 0.008436440 5.997945 9.746344e+10 175.29167
--
5       3.0 0.009074721 5.997623 6.002364e+06 175.29135
```

```
--
9   1.0 0.542865117 1.991512 4.968238e-01 165.19751
--
11  0.0 0.944642286 2.297756 7.142932e-01 111.22110
--
13 -1.0 0.984808745 3.153400 1.949507e+01  75.88091
14 -1.5 0.985920395 3.571474 1.958865e+02  73.85887
15 -2.0 0.984203423 3.908793 2.400662e+03  76.92452
--
21 -5.0 0.973332426 4.845362 7.706452e+09  91.11278
```

Q = -1.5 is the path shape most likely to lead to minimum MSE risk because this shape maximizes CRLQ and minimizes CHISQ.

qm.ridge: Shrinkage PATH Shape = -1.5 **← qm.ridge( ) choice of Q.**

The extent of shrinkage (M value) most likely to be optimal in the Q-shape = -1.5 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  0.0          0.0          Inf        Inf       Inf
1  0.1    7.216e-07    7.691e+09   94.12064  94.14125
2  0.2    1.610e-06    3.421e+09   89.44364  89.48460
3  0.3    2.756e-06    1.999e+09   86.41638  86.47742
---
22 2.2     0.843162    6,422.124   30.90636  31.59084
23 2.3     1.418777    3,786.496   30.32946  31.20015
24 2.4     2.168025    2,456.944   30.52813  31.52124
---
34 3.4     109.1686      79.6369  161.83109  71.94440
35 3.5     153.6432      74.7373  192.89231  76.02665
36 3.6     215.9928      73.9848  225.31073  79.58418
37 3.7     306.6918      76.2110  259.10920  82.72066
38 3.8     444.1371      80.8060  294.68446  85.54040
---
58 5.8      742,299     170.1663 7590.09809 169.76727
59 5.9    1,651,650     171.8291 8420.52713 172.65441
60 6.0          Inf     175.2937 9258.34433 175.29373
     M            K         CLIK       EBAY      RCOF
```

Before abbreviation, the above listing described 61 choices for the M-extent of shrinkage (m = 0.0 to m = 6.0 in steps of 0.1). The search over this lattice suggests that m = 3.6 minimizes the CLIK criterion; the qm.ridge( ) output using the Normal-theory closed form expression also suggested m = 3.6 ! No closed form expressions exist for the EBAY or RCOF criteria, but the lattice search suggests that m = 2.6 is best for these criteria, which is certainly less shrinkage than suggested by the CLIK criterion!

Applying the **"(2/p)$^{\text{ths}}$ Rule-of-Thumb"** of **Obenchain (1978) [Remark (b), page 1115]** with **p = 6**, it follows that the **most shrinkage** likely to produce a "good" ridge estimator (i.e. better than OLS in every MSE sense) along any Path for the `longley2` data would be approximately $m^{MAX} = 3.6 \times (2 / 6) = 1.2$. {When making that same calculation using **eff.ridge( ),** where $m^{EFF} = 1.834$ along a more "efficient" Path, the corresponding upper bound is only $m^{MAX} = 0.61$ …a much more conservative requirement.}

With all of the above background information in mind, let us now "see" and interpret the 5 distinct "individual" types of GRR TRACE Diagnostic plot!

```
     plot(efrobj) ← Default display of all 5 TRACEs
                                       in ONE Plot.

     plot(efrobj, trace= "seq")
                  Display 5 TRACEs in sequence (5 Plots.)
     plot(efrobj, trace= "coef", LP=7)
                  Display only the COEF TRACE plus a
                  Legend in Position # 7 (upper-right).
```

3.1: Shrinkage Coefficient Trace

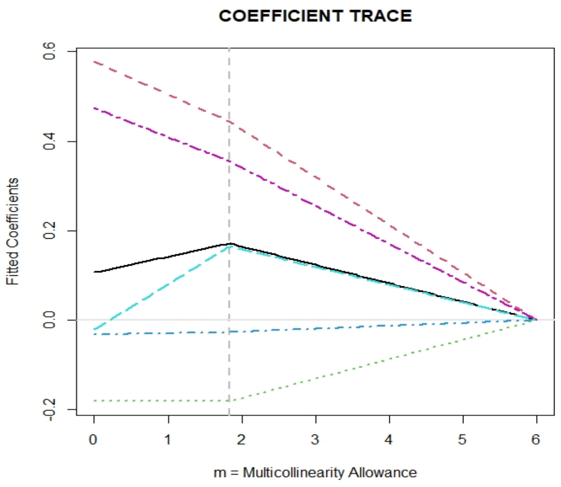
- Each "Coefficient TRACE" shows how individual point-estimates within a vector of β-coefficient estimates with $\mathbf{p} \geq \mathbf{2}$ coordinates **change** as "GRR Shrinkage" progresses from $\mathbf{m} = \mathbf{0}$ (where $\mathbf{\Delta} = \mathbf{I}$) to $\mathbf{m} = \mathbf{p}$ (where $\mathbf{\Delta} = \mathbf{0}$) along a specified Shrinkage Path.

- Coefficient estimates are said to be "numerically stable" over a given **RANGE of Shrinkage m-Extents** when the fitted coefficient estimates form smooth curves that appear to be (approximately) linear over the given Range.

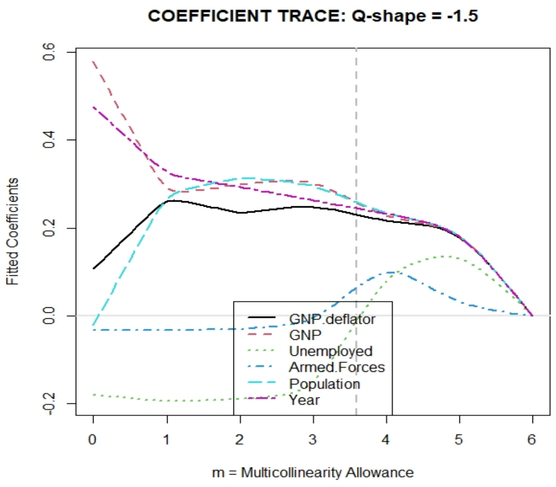
```
unL2obj$coef[1,]  ← OLS regression coefficient estimates

GNP.deflator Unemployed  Armed.Forces  Population     Year      GNP
      0.1066     0.5776       -0.1800    -0.03215 -0.02134   0.4741

(cor(longley2))[c(1:5,7), 6]   ← marginal correlation with Employed.

GNP.deflator Unemployed  Armed.Forces  Population     Year      GNP
      0.9936     0.6968        0.1769      0.9924   0.9752   0.9841
```

The pair of **Coefficient TRACE** displays (below) for the **Longley (1967) model** on the `longley2` data.frame illustrate use the **eff.ridge() function** [LEFT] and the **qm.ridge() function** with its "ML" **Q-shape (q = -1.5)** [RIGHT]. The **qm.ridge** TRACE was displayed earlier (on page 21) in the compact "5 Traces on 1 Plot" format, but these two Coefficient TRACEs look "somewhat different" here simply because their "Aspect Ratio" [Height / Width] is now closer to 1.

`plot(efL2obj, trace="coef")`



`plot(qmL2obj, trace="coef", LP = 2)`



Note that the (left) **`eff.ridge()` `COEF`** trace contains only **two** "sections" where its β-estimates form straight line segments; the final segment extends from **m=1.833** to **m = p = 6** (right-hand shrink-age terminus.) Relatively unstable coefficient estimates (e.g. `Population` here) change markedly (note switch in numerical sign near **m = 0.8** here). Super-stable estimates may display traces that initially change very little (remaining almost horizontal), finally approaching the zero-vector only as **m** becomes nearly **p**.

While most of the clearly undesirable features of the **OLS estimates** in this `longley2` example have been mitigated once `qm.ridge()` shrinkage extent reaches at least **m** = **3**, that's **much too late** to provide a fully satisfactory **Variance-Bias trade-off** on the `longley2` data. Specifically, the "optimal" `qm` Path made the `Population` coefficient much too positive near **m** ~ **1**, and ultimately merged it into a group of **four essentially equal β-estimates** (**`GNP.deflator`**, **`GNP`**, `Population` **&** `Year`). Indeed, all **p** = **6** **β**-coefficient `qm`-estimates became (desirably?) positive for **`m > 3.8.`**

The `qm`-**TRACE** displays featured in the **next four sections (§3.2 - §3.5)** use a **`formula (i.e. model)`** for the `longley2` data.frame that differs from that used by **`Longley(1967)`**. Instead of **predicting Employed**, Art Hoerl was (quite understandably, I think) interested in **predicting GNP** from the other 6 variables. Unfortunately, this particular model is even more highly ill-conditioned and, also, more unfavorable to "optimal" shrinkage via any `q-Shaped` Path. Here, the **ML shape** is `Q=-5.0` and an appropriate revised model formula is:

```
formAH <- GNP ~ GNP.deflator+Unemployed+Armed.Forces+Population+Year+Employed
```

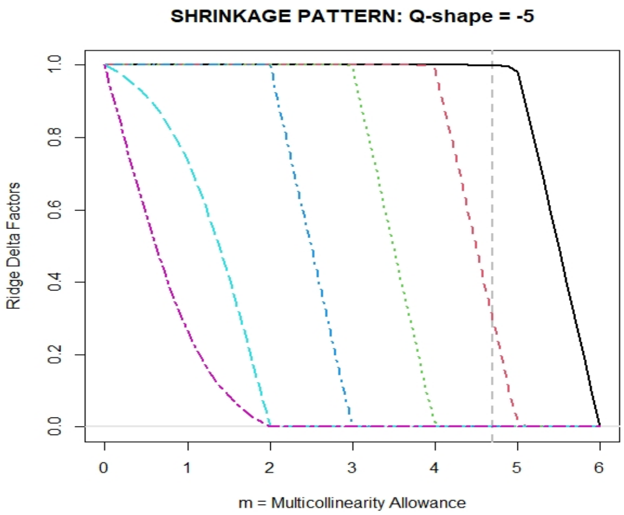
3.2. Shrinkage Pattern Trace

The SHRINKAGE PATTERN trace shows how the generalized ridge "Delta Shrinkage-Factors" applied to the ordered "uncorrelated components" vector, ***c***, decrease as `qm`-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 will be 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 here and are much smaller than the other four singular values. In fact, the last two shrinkage factors have both essentially been reduced to zero at **m=2**.

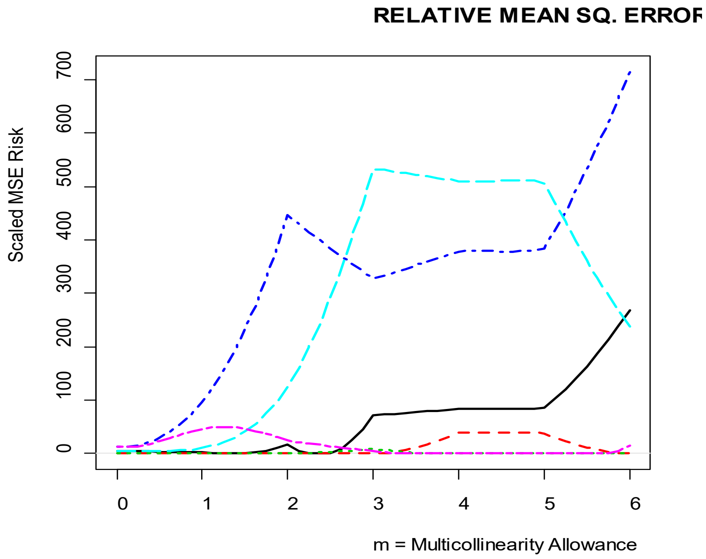
```
plot(qmAHobj, trace = "spat")
```



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 four out of six total principal components of **X**-space.

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

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



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 $\mathbf{\Delta} = \mathbf{I}$) are **known quantities**, being the diagonal elements of the $(\mathbf{X'X})^{-1}$ matrix.

As in the COEFFICIENT trace (and unlike the "SPAT" 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.

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.

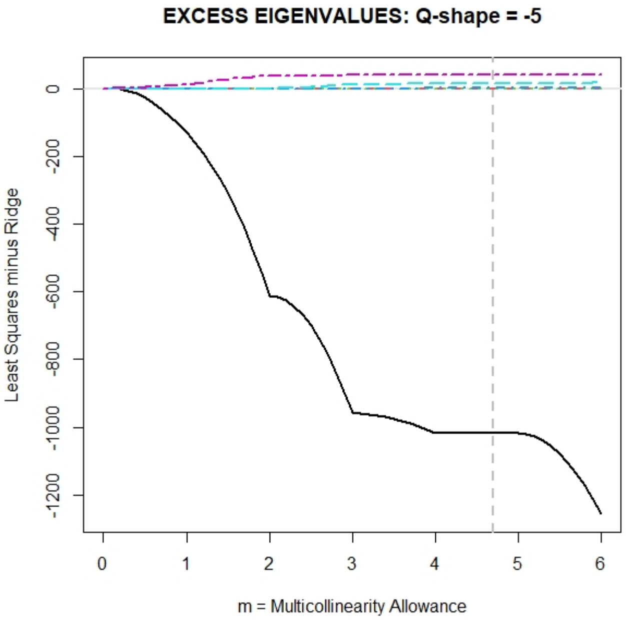
3.4. Excess Eigenvalues Trace

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 can 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 for the **`longley2`** data shown below (page 27), the smallest eigenvalue becomes negative at the 3rd computational step of **m = 0.250** (when using the **`steps=8`** default value), which also happens to be the MOST shrinkage favored by the EBAY and RCOF criteria. The (undesirable) negative eigenvalue at **m = 0.250** is already –3.26 while the corresponding (positive and desirable) largest eigenvalue is only +2.00. In other words, more MSE "harm" has already been done in the single **"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(qmAHobj, trace = "exev")`



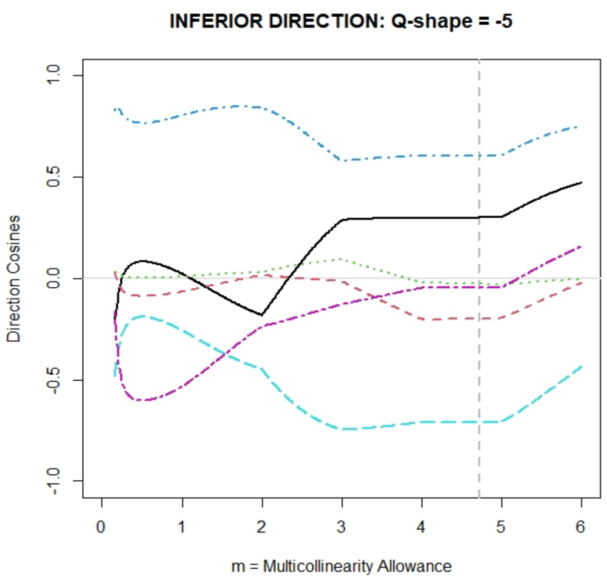
The negative eigenvalue at the **m = 4.750** extent of shrinkage "optimal" for the CLIK criterion is a whopping **–1017** while the corresponding two largest (positive and desirable) eigenvalues are only **+13.9** and **+39.1**. In other words, while the **`longley2`** dataset is "ill-conditioned", it is also highly unfavorable to all **q-Shaped shrinkage Paths**. As outlined at the **top of Page 23**, a strictly "monotone" Shrinkage Path can easily do "too much" shrinkage in some "least appropriate places". Again, the **eff.ridge( )** Path side-steps issues concerning monotonicity, use of only truly "linear" information from "given" X-variables, and avoidance of "non-linear" information from the observed y-vector. As a direct result, the **eff.ridge( )** can make much better use of the CLIK criterion to **favor stopping shrinkage** at **m = 0.251**, where the estimated **δ^{MSE}-vector** is (**0.99998**, **0.83597**, **0.99847**, **0.99844**, **0.97461**, **0.94162**) and thus extra shrinkage is applied primarily to δ_2.

3.5. Inferior Direction-Cosine Trace

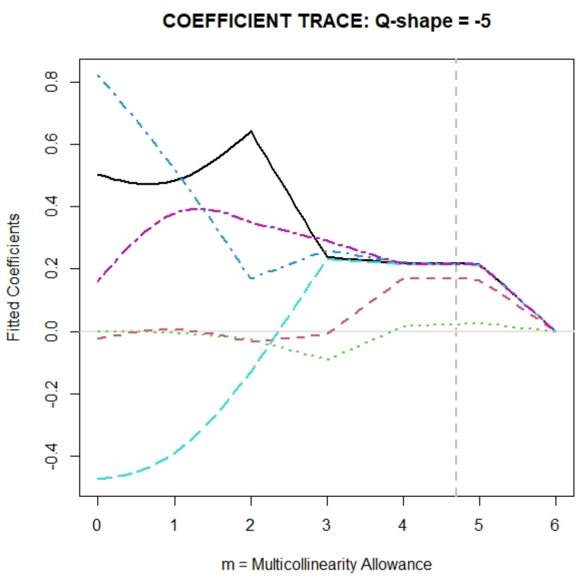
The INFERIOR DIRECTION trace displays the **direction cosines** (elements of the normalized eigenvector) corresponding to any negative eigenvalue of the difference in MSE matrices, OLS – ridge. This direction gives that single linear combination of ridge regression coefficients that not only fails to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



```
plot(qmAHobj, trace = "coef")
```



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** = **p** = 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**. The + sign is the correct choice in the above side-by-side pair of **TRACE**s for the `longley2` dataset; the `infd` direction cosines at **m = 6** on the left-hand side do “point” (approximately) towards the +OLS `coef` estimates at **m = 0** on the right-hand side!

Next, 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 their 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 second example of **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.)

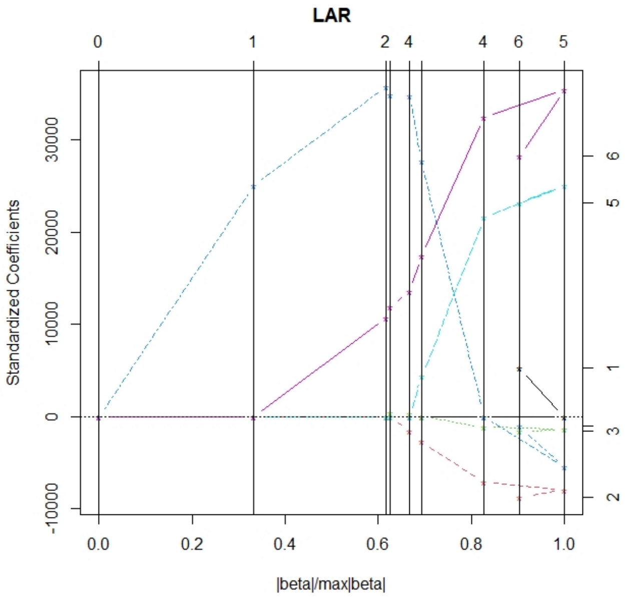
Note that **m-extents** of shrinkage such that **two regressors have inferior direction cosines with equal magnitudes but opposite numerical signs** have the opposite effects on the MSE risks of sums and differences. The SUM of the corresponding shrunken coefficients then has the same or reduced MSE 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, `Q=-5` shrinkage to **m = 2.67** has inappropriately reduced the difference between coefficient estimates (from 0.34 to 0.06) while leaving their sum mostly unchanged (0.68 rather than 0.66.)

Summary of Information from Inferior Direction Cosine TRACEs: These plots contain highly relevant **DETAILS** about the good and bad effects of Shrinkage on the MSE Risk of Linear Model estimates.

4. **Interpretation of least angle regression TRACE displays**

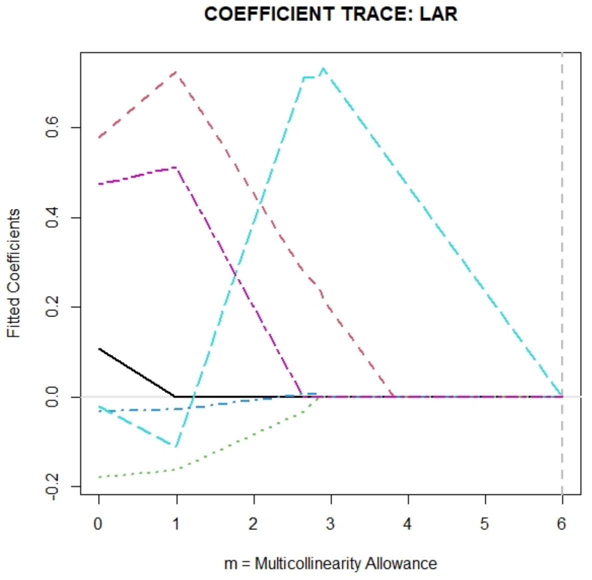
`lars()calls`…

```
xlong2 <- as.matrix(subset(longley2,
                    select=-6))
ylong2 <- as.vector(longley2[,6])
larobj <- lars(xlong2,ylong2,type=“lar”)
plot(larobj)
```



`aug.lars() calls`…

```
form <- Employed~GNP.deflator+GNP+
        Unemployed+Armed.Forces+
        Population+Year
alrobj <- aug.lars(form,data=longley2)
plot(alrobj, trace = “coef”)
```



The first thing to note about the coefficient TRACE displays from the **`aug.lars()`** and **`uc.lars()`** functions in the **`RXshrink`** package is that they are essentially "backwards" relative to the coefficient displays from the **`lars`** R-package. This point is illustrated visually above.

Close examination of this pair of graphs also shows that, besides being backwards relative to each other, there are some "real" differences between the **|beta|/max|beta|** scaling used along the horizontal axis by **lars** and the **m = Multicollinearity Allowance** scaling used by **`RXshrink`** functions: **qm.ridge()**, **eff.ridge()**, **aug.lars()** and **uc.lars()**. However, the coefficients in both types of TRACEs certainly "look like" piecewise linear spline functions!

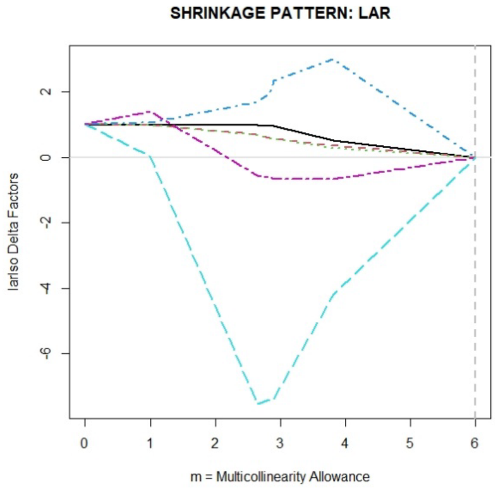
The output.list object from the lars( ) function contains a [1:7, 1:6] matrix of "beta" estimates with the property that ith row contains exactly "i" non-zero beta coefficient estimates…

```
larobj$beta
   GNP.deflator Unemployed Armed.Forces  Population    Year      GNP
  0      0.00000  0.0000000   0.00000000  0.00000000  0.0000  0.000000
  1      0.00000  0.0000000   0.00000000  0.31106374  0.0000  0.000000
  2      0.00000  0.0000000   0.00000000  0.44402264  0.0000  5.570831
  3      0.00000  0.0000000   0.12736366  0.43309352  0.0000  6.194481
  4      0.00000 -0.2442717   0.09223042  0.43250955  0.0000  7.045301
```

```
5      0.00000 -1.2118593  -0.42904523 -0.06832497 555.1066 18.443953
6     30.10233 -1.3431426  -0.48712630 -0.01297783 514.5042 14.719297
attr(,"scaled:scale")
[1]   173.09448  6553.53990  3227.08467 80382.32218    45.05552  1918.48296
```

Least angle regression (type = "lasso", "lar", "forward.stagewise" or "stepwise") **may** yield a final "row" vector that is longer that the OLS vector. As explained at the end of Section §2, this means that one or more of the implied δ-factors used by **lars()** is greater than 1 …implying "expansions" rather than "shrinkage." Similarly, as lar shrinkage occurs, one or more of these implied delta-factors may be negative. These points are illustrated in the graph below.

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



[Above: red dashes overstrike green dots.]

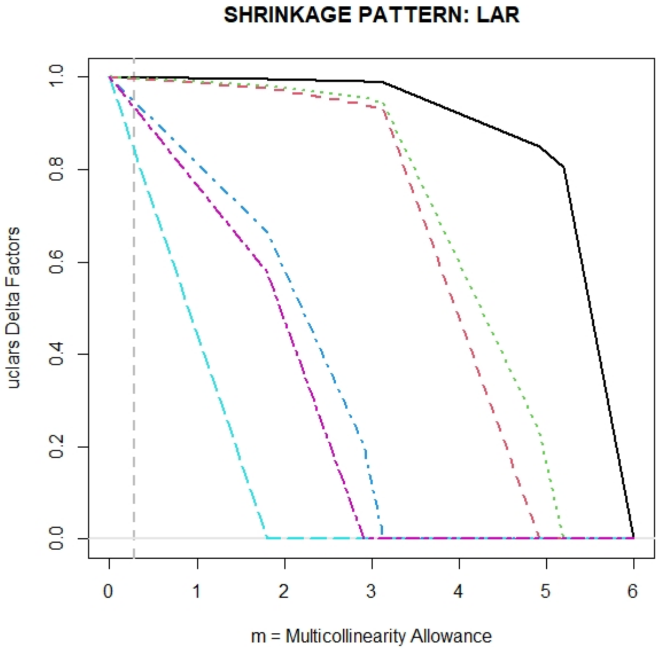
Reductions in MSE risk relative to OLS usually occur only when all of the delta-factors implied by **lars()** estimates are non-negative and strictly less than +1. Exceptions can occur when the unknown true "gamma" component corresponding to an "out of range" delta-factor is nearly zero.

The ***Q***-shape shrinkage paths typically chosen for "generalized ridge" shrinkage 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 monotone ...neither always decreasing nor increasing with their i-index. This can inject desirable flexibility into **lars()** shrinkage patterns.

Use of implied delta-shrinkage factors outside of the usual range of [0, 1) can be avoided by use of the **uc.lars()** function rather than the **aug.lars()** 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 δ_i^{uclars} = **max[0, 1-(k/|ρ_i|)]** then shows that the smallest delta-factor will always correspond to the smallest principal correlation. The **qm.ridge()** 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.

```
uclobj <- uc.lars(form,data=longley2)
    plot(uclobj, 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 (i.e. **blue** over-striking **green**).

5 Behavior of Maximum Likelihood Estimates under Normal distribution-theory: Simulation results using RXshrink functions.

Here, we illustrate some simulation findings for the (most simple) special case of Linear Models: models with only **p=2** predictor **X**-variables, where the distributions of interest are either univariate or bivariate.

The **MLboot(), MLcalc() and MLhist()** functions focus attention on only two forms of **Maximum Likelihood** estimation:

OLS estimates that are **Best Linear Unbiased Estimates (BLUE)**
and
Optimally Biased GRR (non-linear) estimates with **minimum MSE Risk**

Our simulation will use some of the variables from the **mpg** data.frame distributed with the **RXshrink R**-package. To be able to make comparisons that are as valid as possible, we will do some preliminary adjustments to the mpg data …the characteristics of only 32 different autos featured in **Motor Trends** magazine in the mid-1970's. (NOTE: data.frame **mtcars** contains the same numbers, but it's variable names tend to be shorter.)

```
library(RXshrink)
data(mpg)

# Create a data.frame with 5 Replicates of 32 auto "designs"…

 mpg5 <- data.frame(rbind(mpg,mpg,mpg,mpg,mpg))     # 160 autos…

# Focus on a Model with only p=2 X-variables...

form2 <- mpg ~ cubins + weight

# Calculate "preliminary" estimates of Linear Model parameters…

mlt <- MLtrue(form2, mpg5, go=FALSE)    # see §2.5, page 13.
mlt    # Print out not shown here...

# We now use these "estimates" as TRUE Values...

mlt2 <- MLtrue(form2, mpg5, seed=987, go=TRUE)

mlt2    # Output greatly abbreviated...
# OLS Residual Mean Square for Error = 0.2218601
# OLS Beta Coefficients   = -0.3644932 -0.5439967
# Uncorrelated Components = -0.6423994 0.1269281
# Random Number SEED value  = 987    (for REPRODUCABILITY)
# True error Variance, tvar = 0.2218601
# True Uncorrelated Components, tcomp = -0.6423994 0.1269281
```

NEXT Steps…

The "preliminary" steps detailed above have placed us in an ideal position to perform highly relevant Bootstrap Resampling (with replacement) via the `MLboot()` function of `RXshrink`. We can now simulate the properties of key `eff.ridge()` estimates using [1] data with **known** (user supplied) parameters, [2] a linear model **known** to be of "correct" form, and [3] with error (disturbance) terms **known** to come from `rnorm()`, a state-of-the-art generator of i.i.d. Normal random-numbers.

While there are "only" `n = 160` numerical values of `Yvec` and only **`32`** distinct auto "designs" that will be resampled literally thousands of times, note also that the total number of distinct "resampled" data.frames that could result is truly astronomical (> 3e+300)!

```
summary(mlt2$new$Yvec)
#     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
# -3.02547 -0.53560  0.01532  0.05450  0.85987  2.12640

summary(mlt2$new$Yhat)
#     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
#  -1.9256  -0.4660  -0.0087   0.0000   0.6181   1.3640
```

NOTE: The above difference in **y**-means vanishes in analyses when variables are "centered".

```
formBT <- Yvec ~ cubins + weight

system.time( mlBT <- MLboot(formBT, mlt2$new, reps=10000,
                     seed=768) )
#    user  system elapsed
#   17.36    0.16   17.51    << Less than 18 seconds...

mlBT
# MLboot Object: Resmpling WITH Replacement...
# Data Frame: mlt2$new
# Regression Equation:
# Yvec ~ cubins + weight
#
#     Number of Replications,       reps = 10000
#     Random Number Seed Value,     seed = 768
#     Number of Observations,          n = 160
#
# OLS Beta Coefficient matrix            = ols.beta
# ML Optimally Biased Coefficient matrix = opt.beta
# OLS Relative MSE Risk matrix             = ols.rmse
# ML Optimally Biased Coefficient matrix = opt.rmse
# ML Shrinkage Delta-factor matrix        = opt.dmse
```

Subsections §5.1 through §5.4 will now discuss and illustrate many of the findings from the above **MLboot()** analyses.

5.1 Joint Distribution of δ^{MSE} Estimates

Something that strikes me as particularly interesting is that, although the **MLboot()** function clearly introduced considerable variation into the resulting estimates of **MSE-optimal** shrinkage δ-factors, the observed **average values** of both δ-factors tend to agree almost **exactly** (4 decimal places) with their **true values** from the **MLtrue()** function. And both distributions (on [0, 1]) are also clearly skewed, especially that of the 2nd (smaller) **δ-factor**!

The Four displays on Page 39 include two histograms for the marginal distribution of each δ^{MSE} **estimate** using **MLhist().** Here is some code for generating the histogram for the 2nd **dmse** component:

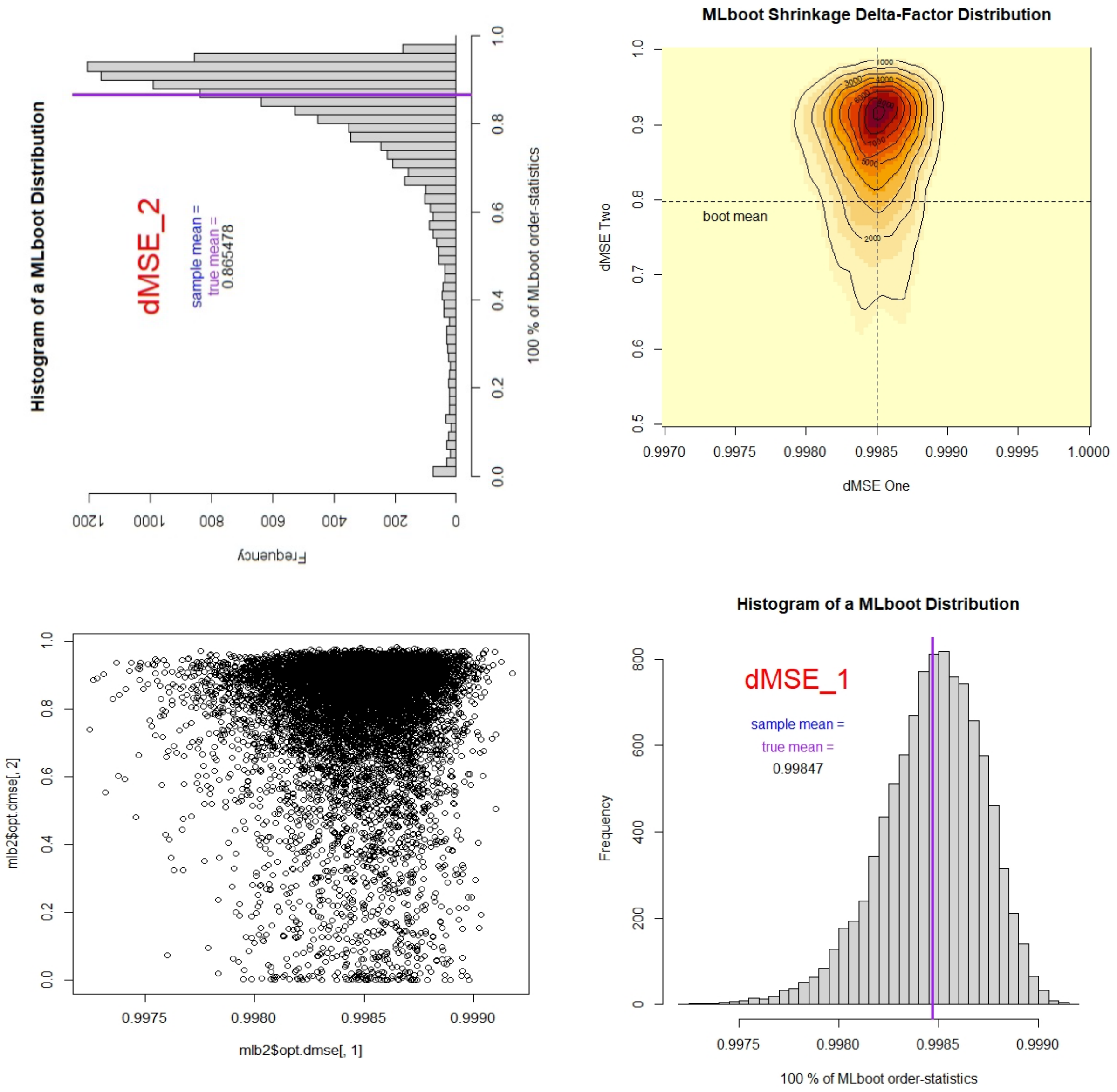
```
MLhist(mlBT, comp="opt.dmse", xvar=2, npct=100)
text(0.5, 1000, "dMSE_2", col='red', cex = 2)
abline(v=0.86548, col="purple", lwd=3, lty=1)
text(0.5, 850, "sample mean =", col='blue', cex = 1)
text(0.5, 800, "true mean =", col='purple', cex = 1)
text(0.5, 750, "0.865478", col='black', cex = 1)
#
summary(mlBT$opt.dmse[,2])
#      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
# 0.0000031 0.7627881 0.8659767 0.7978522 0.9155832 0.9814450
```

And here is code for generating Kernel Density estimates, contours, titles, etc. for the joint (bivariate) distribution of **p=2** optimal **δ-factor estimates…**

```
biv.kde3 <- kde2d(mlBT$opt.dmse[,1], mlBT$opt.dmse[,2], n=100,
                  lims=c(0.997,1.0,0.5,1.0))
image(biv.kde3)
contour(biv.kde3, add = TRUE)
abline(v=0.9985, lty=2)
abline(h=0.7979, lty=2)
text(0.9975, 0.78, "boot mean")
title(xlab="dMSE One", ylab="dMSE Two")
title(main="MLboot Shrinkage Delta-Factor Distribution")
```

Note that this joint distribution is somewhat skewed …with a somewhat longer downward tail than its leftward tail…

Optimal δ-Shrinkage Factor Visualizations

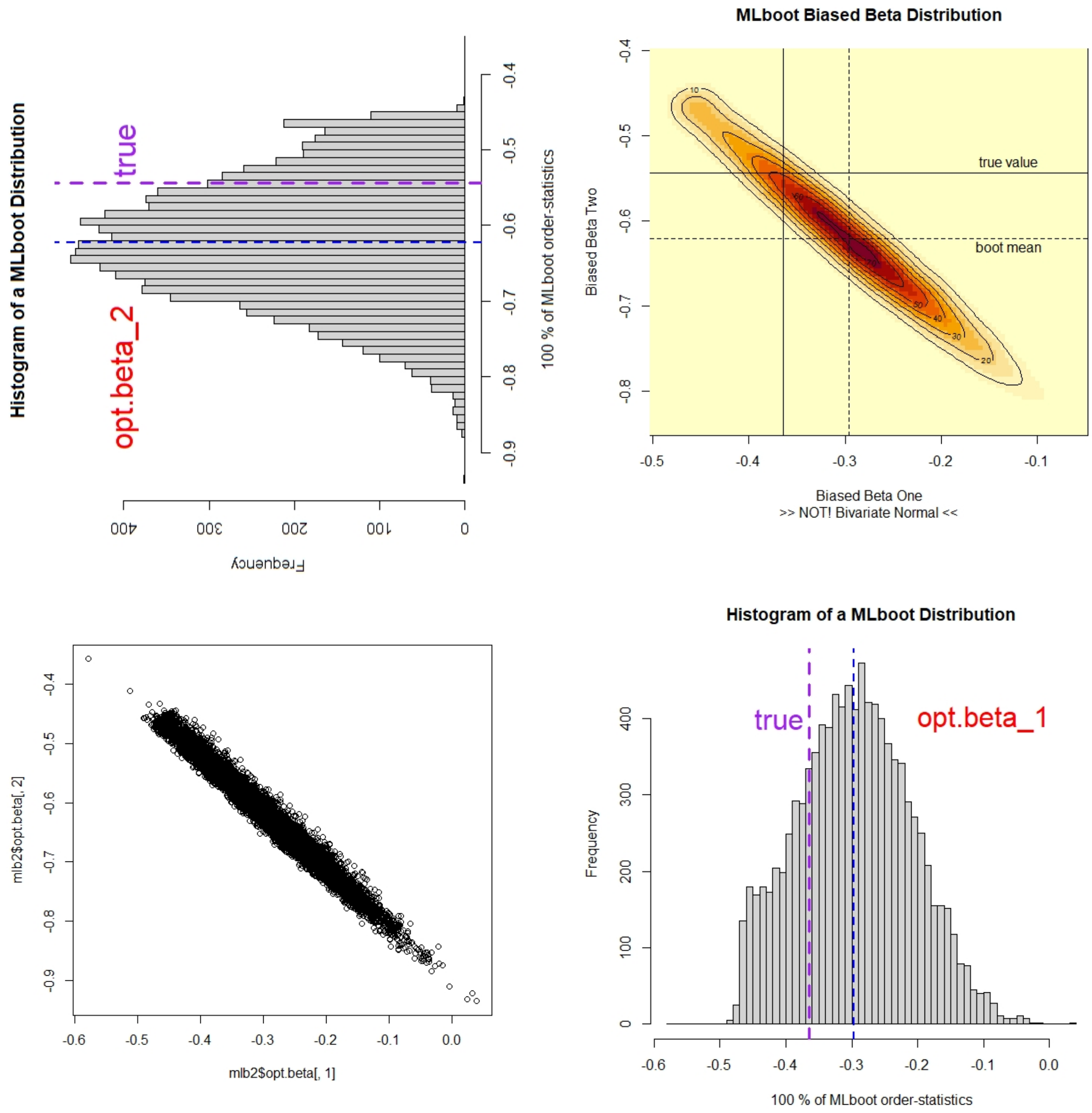


5.2 Joint Distribution of Optimally Biased β-Coefficient Estimates

The key observations here are:

- These estimates are negatively correlated and biased in opposite directions; β_1 is larger (less negative) while β_2 is smaller (more negative) than their true model values.
- And their joint distribution deviates considerably from **Bivariate Normal** …at least in its (somewhat "rounded") upper-left tail.

Optimally Biased β-Coefficient Visualizations



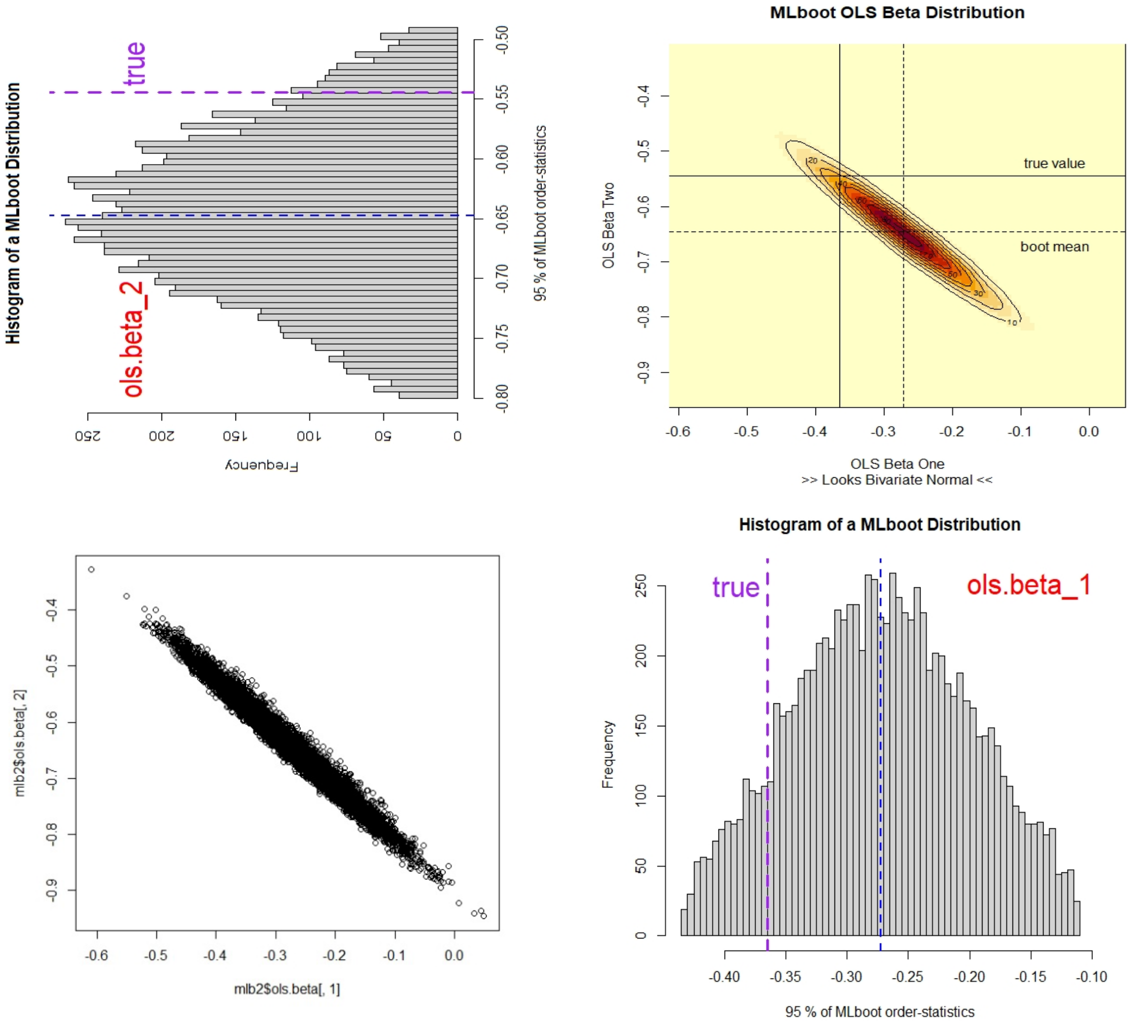
5.3 Joint Distribution of OLS β-Coefficient Estimates

The key observations here are:

- The joint distribution of **OLS β-Coefficient Estimates** does appear to be **Bivariate Normal**.

- However, these estimates are **NOT Unbiased**. They are negatively correlated and appear to have **measurable bias in opposite directions**; β_1 is again larger (less negative) while β_2 is smaller (more negative) than their known true model values.

Bootstrap OLS β-Coefficient Visualizations



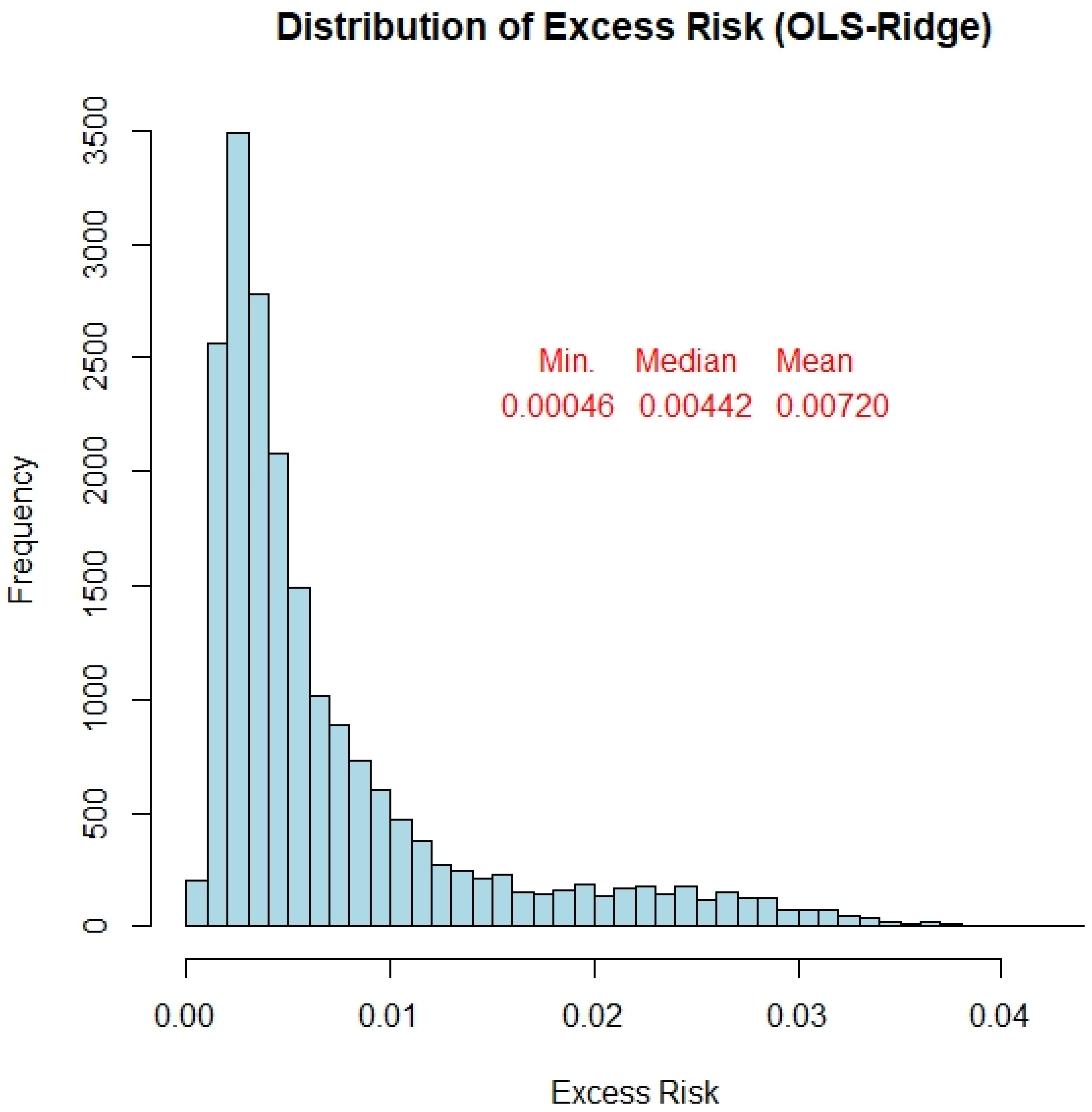
5.4 Distribution of OLS and Optimaly-Biased Estimates of Relative MSE

The key take-aways here are:

- Estimates of Relative Risk from optimally Shrunken linear model estimates are not only smaller on-average than the corresponding OLS estimates …but also smaller in every single case!

- The general forms of individual **"rmse"**-estimate distributions are different. OLS estimates are slightly skewed with a somewhat longer right-hand tail; optimally biased estimates are clearly skewed with a heavy left-hand tail towards lower risk.

Visualizations of Estimated Relative-MSE Risk



The above histogram shows all 20,000 estimated risk differences (**OLS** minus **eff.ridge**) for both β-coefficients ("cubins" & "weight") in our bivariate **MLboot()** simulation. Although none are particularly large, numerically, the important take-away is that **all 20,000 are strictly positive**.

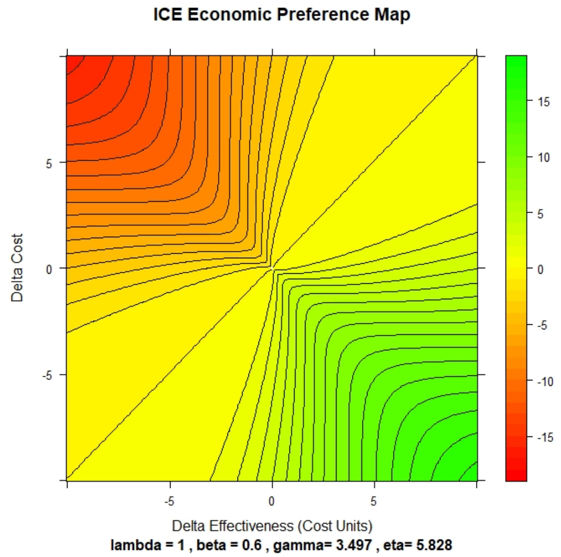
6. Shrinkage in Analysis of Net Benefits

The **ICEinfer R**-package [https://CRAN.R-project.org/package=ICEinfer] provides several functions for Incremental Cost-Effectiveness analyses. It also contains a small data.frame (**`sepsis`**) that we will use here to illustrate use of new **RXshrink** functions, **`eff.aug()`** and **`eff.biv()`**, in applications of GRR to any linear model with **`p >= 2`** predictor x-Variables.

In the special case illustrated here, the y-Outcome to be predicted is **Net Benefit** measured in “cost” units (Net Monetary Benefit) when comparing two Intensive Care Units (**`icu`** is **`0`** or **`1`**) on treatment of **`sepsis`** patients …while also adjusting for 3 “continuous” patient baseline confounder characteristics: **`age`**, Apache II score, and number of organ failures.

```
library(ICEinfer)   # load “ICEinfer” [version 1.3] functions…

data(sepsis)        # load the “sepsis” data.frame

names(sepsis)       # 7 variables observed on 94 patients…
[1] "patid" "icu" "qalypres" "totcost" "age" "apache" "orgfails"

# Create a variable “nldr” to quantify non-linear diminishing returns...

nldr <- ICEpref(sepsis$icu, sepsis$qalypres, sepsis$totcost,
                   lambda=50000, beta=0.6)
```

Since the ICU treatment “effectiveness” measure (**`qalypres`**) is expressed here in **Quality Adjusted Life Years, $50,000** (US) is a widely accepted value for lambda (λ). While the value of beta (**$\beta = 0.6$**) is positive here, returns-to-scale are diminishing because beta=0.6 is less than 1.0. In fact, here is the “ICE Preference Map” used in the above call to **`ICEpref()`**.



The above mapping is clearly **Non-Linear** …particularly in the highly "desirable" **South-East quadrant** and in the highly "undesirable" **North-West quadrant**. The subtitle for this illustration {**lambda =1, beta = 0.6, gamma =3.497, eta = 5.828**}is quite esoteric and raises at least two key questions: [1] Why does the sub-title say "**lambda=1**" rather than 50,000? [2] What do the odd-looking numerical values for **gamma** (γ) and **eta** (η = γ/β) imply?

> The **`ICEpref()`** function assumes that all economic preferences are to be expressed in **`"cost"`** units called Monetary Benefits. While the **`totcost`** variable in the **`sepsis`** data is already expressed in "cost" units, the **`"qalypres"`** variable is not! Thus, the "effectiveness" of ICU treatment for sepsis is measured by multiplying **`"qalypres"`** by **λ** = $50,000 per QALY**.** Thus, the answer to question [1] above is that **λ** becomes "standardized" to **1** whenever the **cost** and **effectiveness** measures are both re-expressed in the **same units** …here, both are cost measures, but some measure of effectiveness units (QALYs, "survival" or "cure" rates, time delays in disease progression, etc.) could be used instead.
>
> The answer to question [2] above is that the maximum value for **eta** (η = γ/β) in an ICE Map that satisfies the Monotonicity Axium is 5.828 = 3 + 2*sqrt(2) = **Ω** , Obenchain (2008), and these are the "most non-linear" of all **realistic** ICE-maps. In fact, the above ICE Map shows a **Green Mountain** of positive preferences in the **South-East ICE** quadrant (less costly & more effective) and a **Red Crater** of negative preferences in the **North-West ICE** quadrant (more costly and less effective). Furthermore, diminishing returns (β > 0 but less than 1) causes the constant-preference contours in the graphic to steadily become further and further apart in both the **South-East** and **North-West** quadrants. In terms of ICE "radius" {distance from (0,0)}, the **Green Mountain** of positive preferences gradually becomes less-and-less steep, while the **Red Crater** of negative preferences is bottoming out.

A key aspect of **ICE** methods is that they are always "**Incremental.**" They focus on "Differences" in Cost-Effectiveness between a pair of alternative treatment regimens. In the **`sepsis`** example, we focus on differences in costs and clinical outcomes between two Intensive Care Units. When reading any "ICE map," the point at the Origin (0,0) is the "anchor" representing the location of the "standard" treatment regimen (i.e. the regimen practiced at ICU = 0). Relative to this "central" location, the location of the "new" treatment regimen (used at ICU = 1) can be anywhere …including also at the Origin (which would then imply no differences between the two ICUs.)

Next, we append the **`nldr`** variable created above to the **`sepsis`** data.frame and define the GRR model of interest…

```
sndr <- data.frame(cbind(nldr$pref, sepsis))  # create new data.frame...

library(RXshrink)  # load "RXshrink" [version 2.1] functions...

form4 <- nldr.pref ~ icu + age + orgfails + apache    # p=4 x-vars...
esndr <- eff.aug(eff.ridge(form4, sndr))   # to compare ICUs...

str(esndr)         # print the "structure" of this "eff.aug" object...
plot( erobj <- eff.biv(esndr, 2, 4) )      # First Plot below...
erobj                                      # print "eff.biv" object...
```

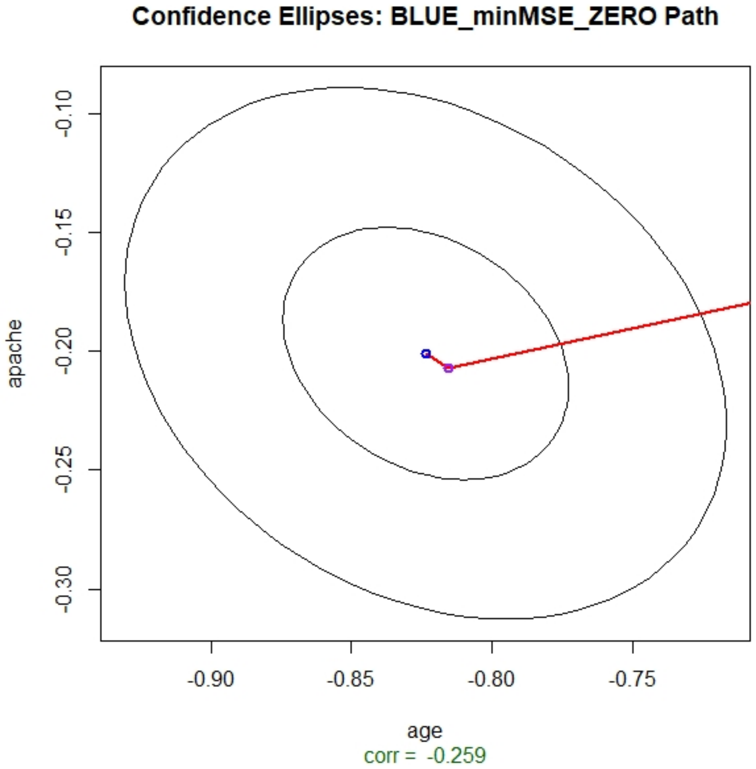
```
plot(erobj, type = "tr")                       # Second Plot below...
```

First Plot…

The two ellipses shown here represent 95% and 50% Confidence Regions for the GRR fitted coefficients of the **"age"** and **"apache"** scores of ICU patients.

The **BLUE** point at the Centroid of both Ellipses represents the **OLS Coefficient Estimates**.

The **Purple** point represents the **Minimum MSE Risk Estimates** that clearly are well within the 50% Confidence Region.

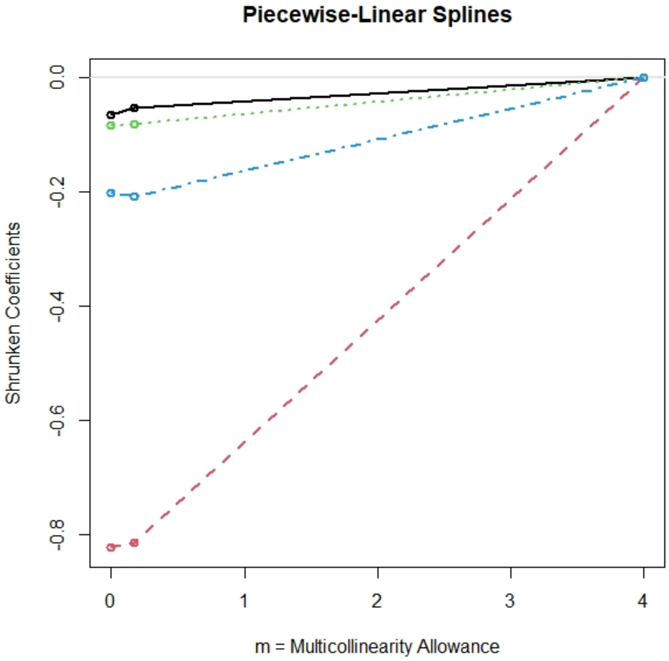


Printed Output… `(erobj)`

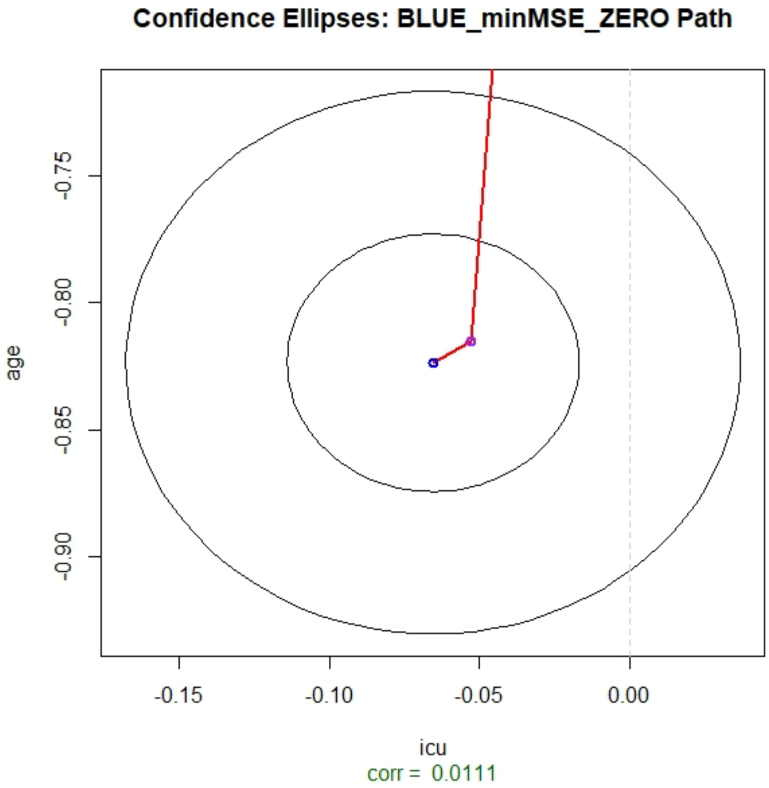
```
eff.biv Object: Bivariate displays of Efficient Shrinkage

    Current Horizontal Coefficient Number = 2    ...is patient age
    Current Vertical   Coefficient Number = 4    ...is Apache II score

    Matrix of Fitted Coefficients and their mcal-Extents:

        icu          age    orgfails       apache      mcal  (extent)
                     ===                   ======
-0.06549163 -0.8235724 -0.08371927 -0.2009351 0.0000000 ...OLS(BLUE)
-0.05267014 -0.8153471 -0.08003150 -0.2069178 0.1704141 ...minMSErisk
 0.00000000  0.0000000  0.00000000  0.0000000 4.0000000 ...Shrinkage
                                                            Terminus
```

The Second Plot {from: `plot(erobj, type = "tr")` } is shown at the top of the next page (# 42).

A **new type of GRR coefficient TRACE** displays only three sets of coefficient estimates connected via straight lines. The initial "knot" for OLS estimates always occurs at `m=0`. The "second" (or "middle") knot corresponds to the Minimum MSE Risk estimates (here at `m=0.17`). [Needless to say (I hope), but it would be quite misleading if the corresponding TRACEs of (non-linear) **RMSE**, **EXEV** or **INFD** estimates were also depicted as being Linear Splines! In reality, these TRACEs are distinctly non-linear functions of the multicollinearity allowance index, **m**.]



Finally, here is the **"plot(erobj <- eff.biv(usndra, 1, 2))"** display that shows [1] that the **"icu"** and **"age"** coefficients are nearly uncorrelated, and [2] that ICU=1 has lower expected Net Monetary Benefit than ICU=0 …but this difference is NOT statistically significant.

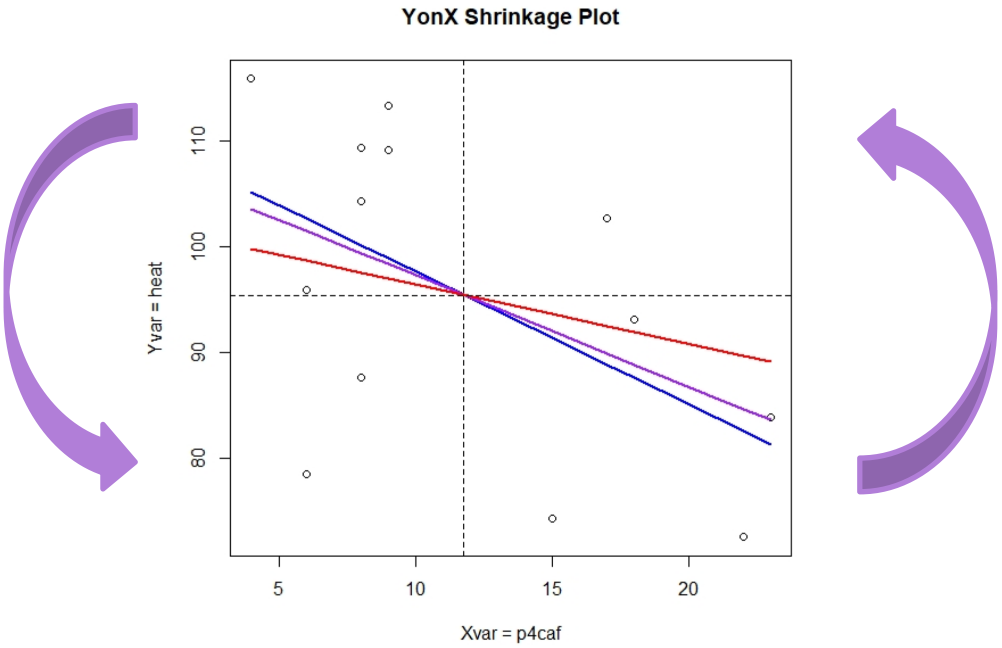


7. YonX( ) - Shrinkage in "Simple" Regression Models

Since **YonX( )** treats only "Simple" (p = 1) regression models of the form: **`y ~ x`** that contain an implicit intercept term, these models cannot be "ill-conditioned" in the "usual sense". [When p > 1, two or more (right-hand-side) variables can be (highly) inter-correlated.] These p = 1 models are easy to "visualize" and, thus, can turn out to be particularly informative! Specifically, the (**`y`**, **`x`**) numerical values then constitute a simple scatter of n points in just two-dimensions, and a regression "fit" is then a straight line that passes through the (geometric) **centroid** of that "scatter": {mean(**`y`**), mean(**`x`**)}.

The specific example of a **YonX()** model explored here uses the **`haldport`** data.frame with model formula **`heat ~ p4caf`** . In the plot below, note that this example has n = 13 observations, where **`cor(heat, p4caf)=-0.5347`** is negative and somewhat smaller than the correlations between **`heat`** and the 3 other potential **`x`**-Variables (**`p3ca, p3cs`** and **`p2cs`**). Also, **R²** (i.e. the proportion of the variance of **`y`** that is predictable from this **`x)`** is only **0.286** for the **`heat ~ p4caf`** model.

The **Optimally Biased Fit (β* = -1.053)** corresponds here to a (counter-clockwise) "rotation" of the **OLS fitted line (Best Linear Uunbiased, β° = -1.256)** towards becoming horizontal, i.e. towards having slope = 0. Note that all three fitted lines pass through **y** = mean(**`heat`**) = **95.42** at **x** = mean(**`p4caf`**) = **11.77**. Finally, the **red line** represents **the "most" shrinkage** consistent with "keeping estimated **MSE** relative risk numerically less than or equal to that of **OLS**."

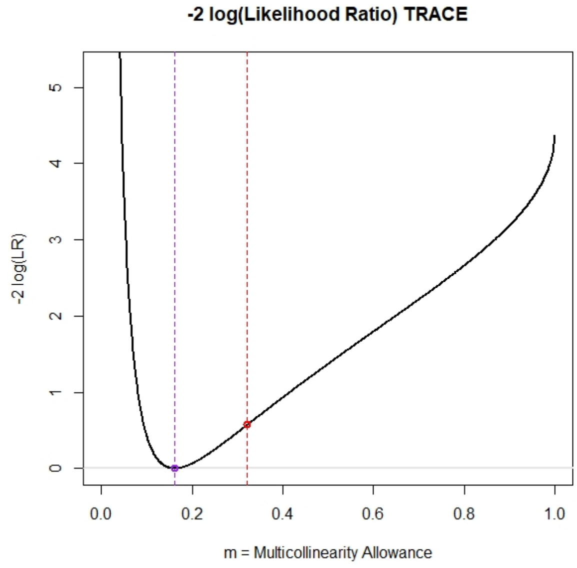


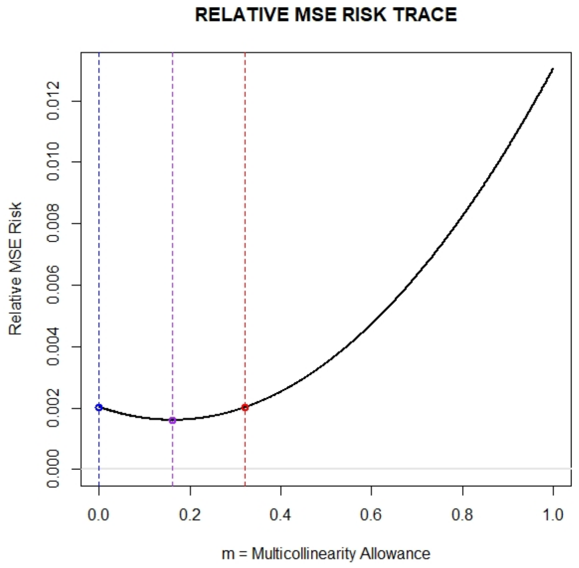
Note: When the **OLS β^o** estimate is **positive**, the **shrinkage rotation** would then be **clockwise!**

Potential inconsistencies between alternative estimates (Maximum Likelihood, or Unbiased or "Correct Range") become **visually obvious** in "Simple" (p = 1) regression models because all TRACE displays then consist of a single line or curve. For example, while an unbiased estimate of the Noncentrality (φ^2) of the F-test for $\beta = 0$ does exist, this estimate can be negative! The MSE risk optimal δ - shrinkage factor is "dMSE" = $\varphi^2 / (1 + \varphi^2)$, but the unbiased estimate of φ^2 cannot be routinely used in shrinkage computations.

Calculating "Angles of Rotation" (θ) of Fitted Lines in the Above Plot: Once ($\mathbf{y}$, $\mathbf{x}$) data are "centered" at their observed mean values, a fitted regression line, $\mathbf{y} = \mu + \beta\mathbf{x}$, passes through the points **(0, 0)** and **(β, 1)**. It follows that **tan(θ) = β**, and thus **θ = atan(β)**.

The main improvement to the **YonX( )** function implemented in version **1.6** (Jan 2021) is its use of a new quadratic approximation for MSE risk. The "rmse" vector of traditional risk estimates is still contained in the YonX output object, but the new estimates in the "qrsk" vector are plotted in the "rmse" TRACE on the right below. The new "lglk" TRACE of -2*log(Likelihood Ratio), shown on the left below, starts at +∞ for OLS (m = 0) but dives to essentially 0 at m = 0.161.





Note that **(1)** the **Maximum Likelihood** solution occurs at **m = (1 - dMSE) = 0.161** and **(2)** the **Maximum m-Extent** of Shrinkage yielding approximate MSE risk no larger than that of **OLS (m = 0)** occurs in the right plot at **m = 2*(1 - dMSE) = 0.322** (upper limit for "Good" estimates).

In the current example, the θ^o for **OLS** is atan(-1.255781) * 180 / pi = **-51.47 degrees**, while θ^* for **minMSE** is atan(-1.053368) * 180 / pi = **-46.49 degrees**. The angle between these two fits is thus (-46.49) - (-51.47) = **+ 4.98 degrees** (counter-clockwise rotation) in the plot on page 44.

Similarly, the maximum angle of rotation, θ^{max} = atan(**be**) * 180 / pi = **-29.47 degrees,** where **be** = x$coef[1] * (1 - x$mReql) = -0.5651016. The angle between the **OLS fit** and the limiting **Red Fit** on page 46 is thus (-29.47) - (-51.47) = + **22 degrees** (counter-clockwise rotation).

Excess MSE and **Inferior Direction** TRACEs are not very "meaningful" when **p = 1**. [Technically, the Inferior Direction is either almost always Upwards (+1) or almost always Downwards (-1).] And the **Coefficient** and **Shrinkage (δ-factor)** TRACEs are both simple **straight lines**!

8. Final Remarks…

The **RXshrink R-**package [https://CRAN.R-project.org/package=RXshrink] and its PDF manual can be downloaded / installed via CRAN. The additional information provided in this vignette: [1] comments on the history of shrinkage in regression, [2] discusses and illustrates the **2-parameter** **`qm.ridge()`** function and the **p-parameter** **`eff.ridge()`** family of generalized ridge regression (GRR) 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 Diagnostic** plots produced by **RXshrink** functions. In a trace, **p = Rank(X)** quantities (several estimated coefficients, risks, shrinkage factors, etc.) are plotted vertically against a horizontal indicator of the **m-**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 terminus (where all "centered" regression coefficient estimates become zero.) **RXshrink** functions (other than **YonX**) require $\mathbf{p \geq 2}$.

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 estimated OLS coefficients have a multivariate normal distribution with mean vector $\boldsymbol{\beta}$ and variance $\boldsymbol{\sigma}^2$ **I**. The "empirical Bayes" and "random coefficient" perspectives on fitting linear models suggest using the extent of shrinkage that minimizes the EBAY or RCOF criterion. The CLIK criterion quantifies the "−2log(Likelihood Ratio) under Normal distribution-theory, and this measure can be driven **all-of-the-way-to-ZERO** via the new "Unrestricted" **p**-parameter Path of **eff.ridge()** …or at least small enough not to be "significantly" greater than ZERO via a (traditional) **Q-shaped Path** using **qm.ridge()**.

A "good" shrinkage estimator, Obenchain(1979), achieves equal or lower matrix-valued MSE risk than **OLS** for the true values of the $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$ 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 $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$. Thus, users of **RXshrink** functions need to focus attention on the question: "Are the most likely values of the $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$ 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 Relative MSE, Excess Eigenvalue and Inferior Direction **TRACES**.

> Much of this "vignette" has used the **`longley2`** data.frame to illustrate interpretation of TRACE displays, and we have seen that this particular data.frame appears to be rather **unfavorable to shrinkage**. The original Longley(1967) data (16 years, 1947-1962) and model with $\boldsymbol{y}$ = Employed are slightly more favorable to shrinkage.

In sharp contrast, we now will display some **TRACE**s that are **quite favorable to shrinkage** by using the **`haldport`** data.frame that is also distributed with the **`RXshrink`** R-package.

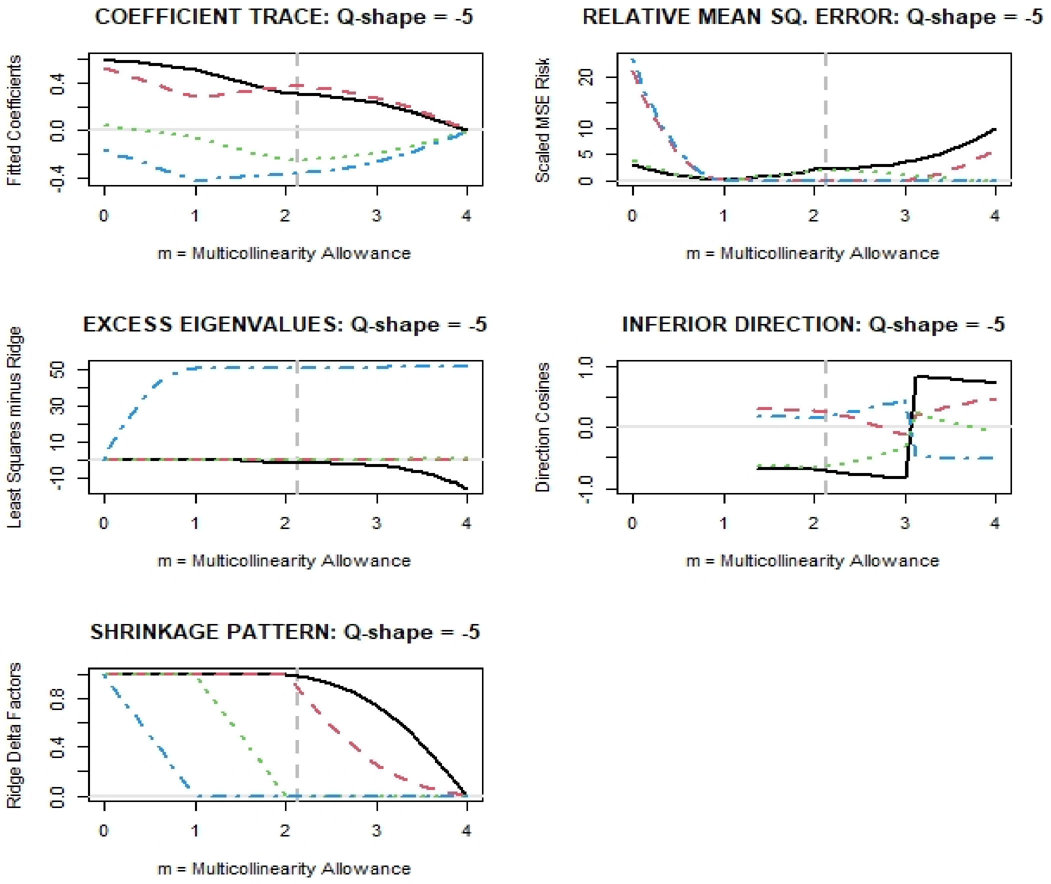
`haldport` is an **R** data.frame containing 13 observations on the following 5 variables:

p3ca is an Integer percentage of 3CaO.Al2O3 in each cement mixture.
p3cs is an Integer percentage of 3CaO.SiO2 in the mixture.
p4caf is an Integer percentage of 4CaO.Al2O3.Fe2O3 in the mixture.
p2cs is an Integer percentage of 2CaO.SiO2 in the mixture.
heat is a measure of the heat in cals/gm that evolved in the "setting" of a sample of concrete (nearest tenth).

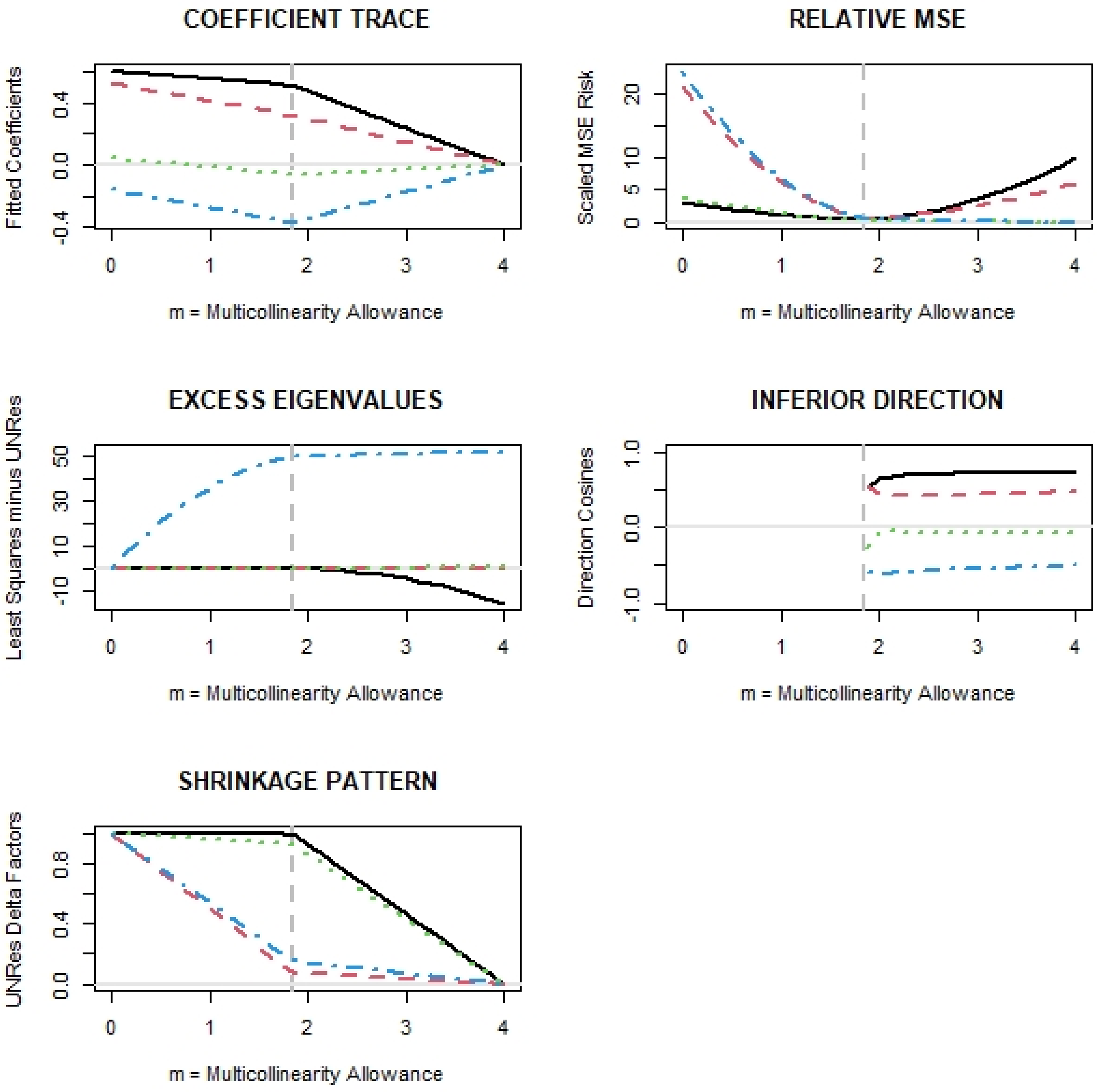
All 13 different cement mixtures contain the same four basic ingredients, but each ingredient percentage **appears to be rounded down** to a full integer. The sum of the four mixture percentages varies from a maximum of 99% to a minimum of 95%. If all four regressor X-variables always summed to exactly 100%, the centered X-matrix **would then be of rank only 3**. Thus, the regression of heat on the four X-percentages is clearly **ill-conditioned** …**with a rank deficiency of at least MCAL = 1.**

The graphics below display all 5 types of TRACEs for [i] the 2-parameter shrinkage PATH of MSE optimal shape, Q = -5, from the **`qm.ridge()`** function and [ii] the new 4-parameter path that passes through the **Maximum Likelihood** estimate of the true β-vector from the **`eff.ridge()`** function, Obenchain (2021). The δ-shrinkage factors corresponding to the ordered eigen-values (and singular values) are usually neither monotone increasing nor monotone decreasing along this "efficient" PATH.

qm.ridge( ) PATH **(haldport data, ML shape, q = -5)**



eff.ridge( ) PATH **(haldport data)**



Optimal m-Extent of Shrinkage: mStar = 1.848

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. (**EBAY: the "empirical Bayes" criterion**.)

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

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

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.

Hastie T. (2020). Ridge Regularizaton: an Essential Concept in Data Science. [Invited contribution to ***Technometrics*** (62, 426–433) celebrating the 50th Anniversary of the original Hoerl-Kennard papers.] https://doi.org/10.1080/00401706.2020.1791959

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. ***Journal American Statistical Association*** 1967; 62: 819-841.

Mallows CL. Some Comments on C_p. ***Technometrics*** 1973; 15: 661-675.

Mallows CL. More Comments on C_p. ***Technometrics*** 1995; 37: 362-372.

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

Obenchain RL. Residual Optimality: Ordinary vs. Weighted vs. Biased Least Squares. ***Journal of the American Statistical Association*** 1975; 70, 375-379. **http://doi.org/10.1080/01621459.1975.10479876**

Obenchain RL. Ridge Analysis Following a Preliminary Test of the Shrunken Hypothesis. ***Technometrics*** 1975; 17, 431-441. (Discussion: McDonald GC, 443-445.) **http://doi.org/10.1080/00401706.1975.10489369**

Obenchain RL. Classical F-tests and confidence regions for ridge regression. ***Technometrics*** 1977; 19: 429-439. **http://doi.org/10.1080/00401706.1977.10489582**

Obenchain RL. Good and optimal ridge estimators. ***Annals of Statistics*** 1978; 6: 1111-1121. **http://doi.org/10.1214/aos/1176344314**

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.** Unfinished eBook (185 pages), 1992--2005. Free Download… **http://localcontrolstatistics.org**

Obenchain RL. ICE Preference Maps: Nonlinear Generalizations of Net Benefit and Acceptability. ***Health Services and Outcomes Research Methodology*** 2008, 8: 31-56. **https://doi.org/10.1007/s10742-007-0027-2** Open Access.

Obenchain RL. Ridge TRACE Diagnostics**.** ***arXiv*** 2020. **https://arxiv.org/abs/2005.14291**

Obenchain RL. The Efficient Shrinkage Path: Maximum Likelihood of Minimum MSE Risk. ***arXiv*** 2021. **https://arxiv.org/abs/2103.05161**

Obenchain RL. Efficient Generalized Ridge Regression. ***Open Statistics***, 2022, 3(1), 1-18. **https://doi.org/10.1515/stat-2022-0108**

Obenchain RL. Maximum Likelihood Ridge Regression. ***arXiv,*** 2022, **https://arxiv.org/abs/2207.11864**

Obenchain RL. (2023) **RXshrink_in_R.PDF -- RXshrink R-package, ver. 2.3** Vignette-like documentation. **<This Document and its earlier versions.>**

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

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.