RXshrink: a softRX freeware implementation of Shrinkage Regression in R…

Generalized Ridge and Least Angle Methods

Version 1.0-4 February 2009

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

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

A Personal Summary of 50 Years of "Shrinkage in Regression"

As someone who has been fascinated with the possibility that shrunken regression coefficient estimates might reduce MSE risk via variance-bias trade-offs and who has conducted and published research in this area, I must say that I am absolutely delighted by the recent wide-spread tolerance for (if not outright acceptance of) shrinkage methods. Anyway, I wish to summarize here some personal perspectives on why and how professional statisticians may have become somewhat enlightened about shrinkage over the last 50+ years, 1955--2009.

Early optimism about a theoretical basis for and the practical advantages of shrinkage almost surely started with the work of Stein(1955) and James and Stein(1961). Unfortunately this shrinkage was always "uniform," thus really doing nothing to adjust the relative magnitudes of correlated regression coefficient estimates for ill-conditioning. Furthermore, although an overall improvement in the scalar value of "summed MSE risk" was guaranteed, there was no way to know "where," in an X-space of 3 or more dimensions, risk was actually being reduced. In fact, researchers on normal-theory minimax estimation in regression [such as Strawderman(1978) and Casella(1980,1985)] found that, when a specific "location" for improved risk was specified, their estimates succeeded only by concentrating shrinkage somewhere else! Besides, the earlier work of Brown (1975) and Bunke(1975a, 1975b), was really the beginning of the end for minimax research. After all, only OLS estimation can be minimax when one's risk measures are truly **multivariate** (matrix rather than scalar valued.) I personally would like to think that modern researchers and regression practitioners view shrinkage estimators as attractive, practical alternatives to OLS estimation in ill-conditioned models even though there cannot be any truly meaningful way to "dominate" OLS on MSE risk.

On the other hand, the real gold-rush of interest in (non-uniform) shrinkage in regression is undoubtedly due to the pioneering "ridge" work of Hoerl (1962) and Hoerl and Kennard (1970a, 1970b.) Some of their terminology was misleading (e.g. their "too longness" argument was actually based upon a simple measure of coefficient variability), and their conjectures that it should be "easy" to pick shrunken estimators from a graphical trace display that would dominate OLS in MSE risk were, in fact, unquestionably naïve. Meanwhile, a major frustration for me, personally, was that my research on shrinkage at Bell Labs in the 1970s lead to open conflict with John Tukey. This started when my management was informed that Tukey had been consistently disparaging shrinkage methods at professional meetings in the 1970s and continued to the point where, ultimately, we formally commented on each other's papers and regression training materials. There were many authoritative critics of shrinkage "optimism" back then, and I hope that their unyielding skepticism will someday be discounted and forgotten.

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

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

Next, I became sufficiently frustrated by the process of getting a second shrinkage paper published in **Technometrics** (delayed until 1977) that I decided to submit an important applications manuscript on shrinkage to **Annals of Statistics**. Some agonizing delays again occurred, and that paper was delayed until 1978. This paper derived the "ridge function theorem," the "excess mean squared error matrix," the "inferior direction," and the "2/P-ths rule of thumb" for limiting shrinkage …plus their individual Maximum Likelihood (ML) estimators for display in TRACE plots. These visual aids have clear practical implications; they show exactly "where and how" MSE risk might be reduced by shrinkage. These diverse TRACE visualization tools are implemented in my freeware algorithms for regression shrinkage in XlispStat, R, S-plus, Stata, GAUSS and SAS/IML.

Finally, I developed a closed form expression, Obenchain(1981), for the normal-theory ML estimator within the 2-parameter Goldstein and Smith (1974) shrinkage family. Unfortunately, none of my attempts to present this material in a peer-reviewed publication have yet succeeded.

My “bottom-line” on the topic of ML shrinkage is simply that the linear estimator that is Most Likely to be “optimal” under normal-distribution-theory is actually a **nonlinear** estimator. The MSE risk profile of an ML shrinkage estimator can always be simulated, if not computed exactly. While being nowhere close to the “dominant” risk profile of the unknown optimal linear estimator, achievable ML profiles can nevertheless be impressively “conservative.” In simple one-dimensional cases, ML shrinkage can reduce MSE risk by about 50% in favorable cases (with low signal and/or high uncertainty) while increasing risk by at most 20% in unfavorable cases. In high-dimensional situations, a savings of more than 50% is possible, and worst case situations result in a loss of less than 5% in MSE risk. However, as Burr and Fry(2005) have noted, the key tactic in shrinkage estimation is definitely to be “cautious” rather than “greedy.”

Frank and Freidman(1993), Breiman (1995), Tibshirani (1996), LeBlanc and Tibshirani (1998) and Efron et al. (2004) are currently keeping the shrinkage regression “home fires” burning for exploratory analyses of gigantic datasets. Least Angle (LA) regression usually starts with an initial solution vector longer that the OLS vector. No reduction in MSE risk relative to OLS is then possible until the LA solution ultimately becomes a genuine shrinkage estimator.

The RXshrink Package

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

Traditional visualizations of shrinkage regression computations use "trace" plots. In a trace, P quantities (several estimated coefficients, risks, shrinkage factors, etc.) are plotted vertically against a horizontal indicator of the extent of shrinkage. Traditional “ridge” traces display the Ordinary-Least-Squares (OLS) solution at their left-hand extreme and cover the full range of shrinkage that culminates in "total" shrinkage at their right-hand extreme (where all “centered” regression coefficient estimates become zero.) Here, P denotes the number of non-constant predictor variables in the regression model. RXshrink functions require P to be at least 2.

> RXridge, RXtrisk and RXtsimu use the "Multicollinearity Allowance," denoted by MCAL (or simply M), as its measure of the EXTENT of shrinkage along generalized ridge paths whose SHAPE (or curvature) is controlled by a parameter denoted by QPAR (or simply Q.) See the TECHNICAL APPENDIX at the end of this documentation for definitions of both MCAL and QPAR.
>
> RXlarlso and RXuclars use a horizontal trace scaling equivalent to "Multicollinearity Allowance" but display a P-parameter path with shrinkage factors determined by the strengths of observed correlations with Y instead of by the relative spreads (inverse precision) of the given X-regressor coordinates. In fact, RXuclars uses a closed form

expression for its LA shrinkage delta-factors that exists because X-regressor principal coordinates are uncorrelated.

RXshrink functions attempt to identify shrunken coefficient estimates that are either "good" in the sense that they dominate least squares estimates in every (multivariate) Mean Squared Error sense or are "optimal" in one well-defined (univariate) MSE sense. Definitions for "good" or "optimal" ridge shrinkage factors are based upon risk (expected loss) calculations that apply to all forms of statistical distributions. But the ML inferences for the P-parameter and 2-parameter shrinkage paths explored by RXshrink functions are based upon standard normal-distribution-theory.

GUIDELINES for Interpretation of Shrinkage Trace Plots

```
+-----------------------------+
| Shrinkage Coefficient Trace |
+----+------------------------+----------------------------------------+
     | This trace shows how regression coefficient point estimates     |
     | change as shrinkage (along a path of shape Q) occurs.  Any      |
     | coefficient estimate that is numerically "stable" will plot     |
     | close to the straight line from its least-squares estimate at   |
     | MCAL=0 to zero at MCAL=P.  Unstable coefficient estimates       |
     | will change more quickly, possibly switching numerical sign, as |
     | soon as MCAL starts increasing from zero.  Super-stable         |
     | estimates will change only very little initially, finally       |
     | approaching zero only as MCAL approaches P.                     |
     +-----------------------------------------------------------------+
```

```
+----------------------------------------------------+
| Estimated "Scaled" Risk (Mean-Squared-Error) Trace |
+----+-----------------------------------------------+-----------------+
     | This trace gives normal distribution theory, "modified" maximum |
     | likelihood estimates of "scaled" risk (mean-squared-error       |
     | loss) as shrinkage of shape Q occurs.                           |
     |                                                                 |
     |     The risk is "scaled" by dividing it by an 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.           |
     |                                                                 |
     |     Maximum likelihood scaled risk estimates are "modified,"    |
     |     first of all, so as to be unbiased.  Then they are adjusted |
     |     upward, if necessary, to have correct range relative to a   |
     |     known lower bound on scaled risk, which re-introduces bias. |
     +-----------------------------------------------------------------+
```

```
+---------------------------+
| Excess EigenValues Trace  |
+----+----------------------+------------------------------------------+
     | This trace plots the EigenValues of the estimated difference in |
     | Mean Squared Error matrices, ordinary least squares minus ridge.|
     | As long as all EigenValues are zero or positive, there is good  |
     | reason to hope that the corresponding ridge estimators yield    |
     | smaller MSE risk than Least Squares for all directions in       |
     | P-space (i.e. all possible linear combinations.)  As shrinkage  |
     | continues, at most one negative EigenValue will appear.         |
     +-----------------------------------------------------------------+

+----------------------------------+
| Inferior Direction-Cosine Trace  |
+----+-----------------------------+-----------------------------------+
     | This trace plots the Direction Cosines (normalized EigenVector) |
     | corresponding to any negative EigenValue of the difference in   |
     | MeanSquaredError 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.  |
     +-----------------------------------------------------------------+

+--------------------------------+
| Shrinkage Factor Pattern Trace |
+----+---------------------------+-------------------------------------+
     | This trace plots the Delta Shrinkage-Factor Pattern as shrinkage|
     | of shape Q occurs.  All deltas are equal when Q=1; the trailing |
     | deltas are small when Q < 1; and the leading deltas are small   |
     | when Q > 1.                                                     |
     +-----------------------------------------------------------------+
```

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

```
+--------------------------------------------------------------------+
|  TECHNICAL APPENDIX......"Extent" and "Shape" of Shrinkage in the  |
|                          Two-Parameter Generalized Ridge Family.   |
+--------------------------------------------------------------------+
```

```
MCAL = the "Multicollinearity Allowance" parameter that indexes
           the "extent" of ridge shrinkage along any ridge path.
     = R - trace( R x R diagonal matrix of Delta Shrinkage Factors ).

MCAL = 0 ...yields zero shrinkage.  This is the "starting point" of the
            ridge path, where the ridge estimator coincides with the
            Ordinary Least Squares estimator at the left-hand extreme
            (because all R of the Delta factors are equal to 1.)

MCAL = R ...yields "total" shrinkage.  This is the right-hand "end
            point" of the path, where the ridge estimator is all ZEROS.
```

Again, MCAL = R - Delta[1] -...- Delta[R], where Delta[j] is the ridge "shrinkage factor" applied to the j-th uncorrelated component, c[j], of Bzero . The average value of Delta[1],...,Delta[R] is (R-M)/R, which is Theil's "proportion of posterior precision in Bstar due to sample information." More importantly, MCAL can be interpreted as the approximate deficiency in the rank of (I - 11'/ N) X. For example, if the regressor matrix has only two relatively small singular values, then the coefficient ridge trace is expected to "stabilize" at about MCAL = 2. Perfectly stable relative magnitudes plot on the MCAL-scale as straight lines all intersecting at MCAL = R and Bstar = 0.

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.

Shrinkage Factor Formulas...

P = Number of Predictor Variables (non-constant Regressors),
R = Rank of the Centered Predictor Variable X-matrix,
N = Number of Observations (or Regressor Combinations), and

generalized ridge regression "Shrinkage Factors" are of the form...

$$\text{Delta} = \frac{\text{EigenValue}}{\text{EigenValue} + \text{Konstant}*\text{EigenValue}^Q}$$

or, equivalently,...

$$= \frac{1}{1 + \text{Konstant}*\text{EigenValue}^{(Q-1)}}$$

Empirical evidence that choice of "shape" as well as "extent" of shrinkage can be rewarding is given in the following table...

Data Set Name	Number of Observations	Number of Predictors	Min.MeanSqErr Extent of Shrinkage	Min.MeanSqErr Shrinkage Shape
FACE data	N = 21,	R = 10,	MCAL = 2.3,	Qshape = +.77
Air Pollution and Mortality	N = 60,	R = 15,	MCAL = 5.4,	Qshape = +.07
Acetylene	N = 16,	R = 9,	MCAL = 5.2,	Qshape = -.35
Ten-Factor	N = 36,	R = 10,	MCAL = 3.6,	Qshape = -.78
Stack Loss	N = 15,	R = 3,	MCAL = 0.24,	Qshape = -.95
Mantell, Bell Productivity	N = 25,	R = 3,	MCAL = 0.95,	Qshape = -1.1
Wood Beam	N = 10,	R = 2,	MCAL = 0.26,	Qshape = -1.4
Longley	N = 16,	R = 6,	MCAL = 4.0,	Qshape = -1.4
Hocking MPG	N = 32,	R = 10,	MCAL = 8.8,	Qshape = -7.6
Diesel data	N = 44,	R = 9,	MCAL = 4.9,	Qshape = -20
Portland Cement, Hald.	N = 13,	R = 4,	MCAL = 3.0,	Qshape = -INF.
Data Set Name	Number of Observations	Number of Predictors	Min.MeanSqErr Extent of Shrinkage	Min.MeanSqErr Shrinkage Shape