Maximum Likelihood Shrinkage in Regression

Robert L. Obenchain[1]

We use standard normal theory methods to derive a closed form expression for the (nonlinear) estimator of the regression coefficient vector that is most likely to achieve minimum Mean-Squared-Error (MSE) risk within a 2-parameter shrinkage family. This family contains Hoerl and Kennard(1970) "ordinary" ridge regression and uniform shrinkage as special cases and principal components regression as a limiting case. The closed form solution provides insights into MSE optimal shrinkage, enables rapid computation in practical applications, and allows the entire MSE risk profile for certain special cases of the estimator to be easily simulated.

KEY WORDS: normal theory maximum likelihood; extent and shape of shrinkage; minimum mean-squared-error estimation.

1 Introduction

Under normal distribution theory for multiple linear regression models with independent and homoscedastic observations, it is well known that the Ordinary Least Squares (OLS) estimator is not only most likely to be the true β vector but also is Best Linear Unbiased (BLU) and minimax. Here, we consider shrinkage estimators that reduce variance as bias is introduced and, thus, have the potential to reduce overall Mean-Squared-Error (MSE) risk. The linear estimator, Ly, achieving minimum MSE risk is not operational; the optimal choice for the L matrix depends upon both the unknown β coefficients and the unknown error variance σ^2. Thus we use standard methods to identify the operational estimator within a 2-parameter shrinkage family that is most likely to be the optimal Ly non-estimator. This restricted maximum likelihood estimator is expressed in closed form, but its mean and variance remain unknown because it turns out to be nonlinear in y.

The remainder of this introduction presents multiple regression notation and fundamental shrinkage concepts. Section 2 outlines the derivation of the restricted max-

[1] Robert L. (Bob) Obenchain PhD (UNC-CH, Mathematical Statistics, 1969) is a retiree from both AT&T Bell Laboratories and Eli Lilly & Company. His professional web site is http://localcontrolstatistics.org

imum likelihood estimator, and Section 3 displays some simulated MSE risk profiles that illustrate why the maximum likelihood estimator has advantages in practical regression applications.

1.1 Multiple Regression Notation and Standardization

The usual model for multiple linear regression is written as

$$E(y|X) = 1\mu + X\beta \text{ and } Var(y|X) = \sigma^2 I \ , \tag{1.1}$$

where y is a $n \times 1$ vector of observed response values; 1 is a $n \times 1$ vector of ones; X is a given $n \times p$ matrix of non-constant regressor coordinates; μ is the unknown intercept; β is a $p \times 1$ vector of unknown regression coefficients; and the responses are stochastic and uncorrelated with constant, unknown variance, σ^2.

For simplicity, we assume that the data will be "centered" by subtracting off column means, so that $1'y = 0$ and $1'X = 0'$. In other words, centering allows the above model to be written succinctly as $E(y|X) = X\beta$ and $Var(y|X) = \sigma^2(I - 11'/n)$, where the μ intercept term from equation (1.1) is implicitly estimated by $\bar{y} - \bar{x}'\beta$.

Finally, we assume here that $r = rank(X)$ is at least 2, where $r \leq \min(p, n-1)$.

1.2 Principal Axis Rotation to Uncorrelated Components

The singular value decomposition of regressors is $X = H\Lambda^{1/2}G'$, where H is the $n \times r$ semiorthogonal matrix of *principal coordinates* of X, Λ is the $r \times r$ diagonal matrix of *eigenvalues* of $X'X = G\Lambda G'$, and G is the $p \times r$ semiorthogonal matrix of *principal axis direction cosines.* Note that $1'H = 0'$ because $1'X = 0'$. Furthermore, $G'G$ and $H'H$ are $r \times r$ identity matrices, while GG' will be a $p \times p$ identity matrix in the "full column rank" case, $r = p$. The principal axes are assumed here to be ordered such that the eigenvalues, $\lambda_1 \geq \cdots \geq \lambda_r > 0$, of the regressor inner products matrix are non-increasing.

The least-squares estimator, b^o, of β in equation (1.1) is not uniquely determined when $r < p$, so we adopt the convention here that

$$b^o \equiv X^+ y = Gc, \tag{1.2}$$

where X^+ is the (unique) Moore-Penrose inverse of X,

$$c \equiv \Lambda^{-1/2}H'y = (y'y)^{1/2}\Lambda^{-1/2}r^o \tag{1.3}$$

is the $r \times 1$ vector containing the (sample) *uncorrelated components* of b^o, and $r^o = H'y/\sqrt{y'y}$ is the $r \times 1$ vector of *principal correlations* between y and the columns of H. Note that $E[c|X] = G'\beta \equiv \gamma$, which is the $r \times 1$ vector of unknown *true components* of β, and that $Var[c|X] = \sigma^2\Lambda^{-1}$, which is a diagonal $r \times r$ matrix.

The F-ratio for testing $\gamma_i = 0$ is

$$F_i = \frac{c_i^2}{s^2/\lambda_i} = \frac{(n-r-1) \cdot r_i^{o2}}{(1-R^2)} \ , \tag{1.4}$$

where the familiar R-squared statistic is simply the sum-of-squares of the principal correlations, $R^2 = r_1^{o2} + \cdots + r_r^{o2}$; $(n-r-1)$ is the number of "degrees-of-freedom for error"; and $s^2 = y'(I-HH')y/(n-r-1) = (y'y) \cdot (1-R^2)/(n-r-1)$ is the unbiased estimator of σ^2 under normal distribution theory.

Note that equation (1.4) illustrates that the statistical significance of the uncorrelated components depends only upon their corresponding principal correlations. On the other hand, equation (1.3) shows that

$$c_i = r_i^o \cdot \sqrt{\frac{y'y}{\lambda_i}}. \tag{1.5}$$

Thus, an uncorrelated component can be relatively large, numerically, simply because its corresponding eigenvalue is relatively small, even when its principal correlation is not relatively large. In fact, components with relatively small eigenvalues also have relatively large variance, $V[c_i|X] = \sigma^2/\lambda_i$, and thus are primary candidates for shrinkage.

1.3 Linear Shrinkage Estimators

Our interest will focus on estimators that "shrink" least-squares coefficients along the r principal axes of $X'X = G\Lambda G'$ using multiplicative shrinkage factors, $\delta_1, \cdots, \delta_r$. Estimators of this form are linear in y if the $\delta-$factors are *nonstochastic* given X. The resulting shrinkage estimators of β are of the general form

$$b^* = G\Delta c, \tag{1.6}$$

where the direction cosines matrix, G, and uncorrelated components vector, c, are as in equation (1.2) and Δ is a $r \times r$ diagonal matrix containing the shrinkage factors, each of which is usually restricted to lie in the closed interval from zero to one, $0 \leq \delta_i \leq 1$.

The decomposition $b^o = Gc$ of equation (1.2) is also the basis for *principal components regression.* This is the special case of shrinkage regression, $b^* = G\Delta c$, in which each shrinkage factor is either $\delta_i = 0$ or $\delta_i = 1$. Massy(1965), page 241, lists two criteria for deciding which δ_i would be set to zero: (a) components with relatively small eigenvalues, λ_i, or (b) components with relatively small absolute principal correlations, $|r_i^o|$. We will see that the maximum likelihood approach to 2-parameter shrinkage regression uses both the λ_i and the $|r_i^o|$ to determine $\delta-$factors that frequently are strictly between zero and one.

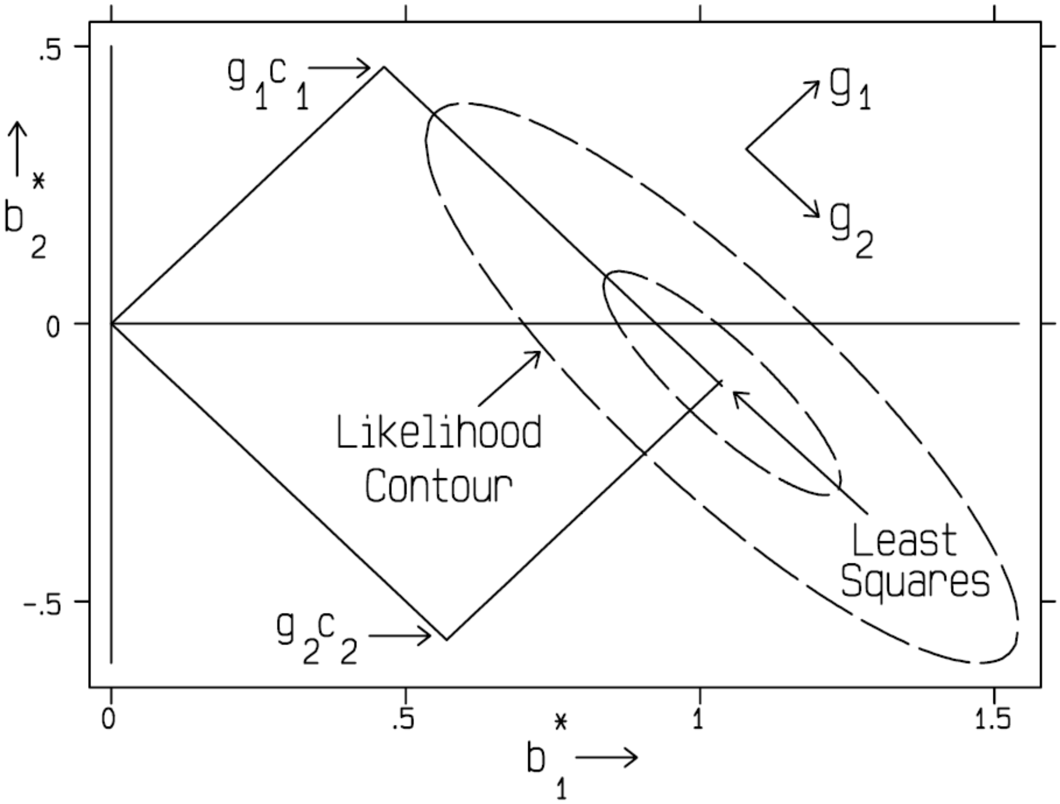

Figure 1: When the regressor X matrix is of rank $r = 2$, the $b^o = Gc$ decomposition is $b^o = g_1c_1 + g_2c_2$ where g_1 and g_2 are the columns of G. This principal axis rotation orients the edges of the rectangle containing all shrinkage estimators of the form $b^* = G\Delta c = g_1\delta_1c_1 + g_2\delta_2c_2$ with $0 \leq \delta_1, \delta_2 \leq 1$.

1.4 Optimal Shrinkage Factors

To apply normal-theory maximum likelihood to shrinkage estimation, we will need a *definition* for the Δ factors that make $b^* = G\Delta c$ the minimum MSE, linear estimator of β. We start by computing the risk of δ_ic_i as an estimator of the $i-$th true component γ_i:

$$MSE(\delta_ic_i) = E[(\delta_ic_i - \gamma_i)^2] = E\{[\delta_i(c_i - \gamma_i) - (1 - \delta_i)\gamma_i]^2\}. \tag{1.7}$$

Under the assumption that δ_i is nonstochastic given X, this risk expression can first be rewritten as

$$MSE(\delta_ic_i) = \delta_i^2E[(c_i - \gamma_i)^2] - 2\delta_i(1 - \delta_i)E(c_i - \gamma_i) + (1 - \delta_i)^2{\gamma_i}^2, \tag{1.8}$$

and then simplified, using $E(c_i) = \gamma_i$ and $V(c_i) = \sigma^2/\lambda_i$, to yield:

$$MSE(\delta_i c_i) = \delta_i^2 \sigma^2/\lambda_i + (1-\delta_i)^2 {\gamma_i}^2. \tag{1.9}$$

Next, we compute the partial derivatives of MSE risk with respect to changes in the $i-$th shrinkage factor:

$$\partial MSE(\delta_i c_i)/\partial \delta_i = 2\delta_i \sigma^2/\lambda_i - 2(1-\delta_i){\gamma_i}^2, \text{ and} \tag{1.10}$$

$$\partial^2 MSE(\delta_i c_i)/\partial \delta_i^2 = 2\sigma^2/\lambda_i + 2{\gamma_i}^2. \tag{1.11}$$

Since the second derivative will be strictly positive whenever $\sigma > 0$, the solution of $\partial MSE/\partial \delta_i = 0$ corresponds to minimum risk. This solution is $\delta_i = \delta_i^{MSE}$ where

$$\delta_i^{MSE} = \frac{\gamma_i^2}{\gamma_i^2 + (\sigma^2/\lambda_i)} = \frac{\lambda_i}{\lambda_i + (\sigma^2/\gamma_i^2)} \ . \tag{1.12}$$

Note that $0 \leq \delta_i^{MSE} \leq 1$. Furthermore, $\delta_i^{MSE} = 0$ only when $\gamma_i = 0$ or in the limit as σ^2 increases to $+\infty$. Similarly, $\delta_i^{MSE} = 1$ only when $\sigma^2 = 0$ or in the limit as $|\gamma_i|$ increases to $+\infty$.

Substituting δ_i^{MSE} into equation(1.7) and simplifying, we have shown that

$$MSE(\delta_i c_i) \geq \delta_i^{MSE} \sigma^2/\lambda_i \tag{1.13}$$

for all non-stochastic shrinkage factors $0 \leq \delta_i \leq 1$, with equality only when $\delta_i = \delta_i^{MSE}$. Of course, one never really knows when a non-stochastic choice for δ_i is equal to δ_i^{MSE}, and $\delta_i^{MSE} c_i = \lambda_i c_i/[\lambda_i + (\sigma^2/\gamma_i^2)]$ is a non-estimator of γ_i because its (σ^2/γ_i^2) factor is unknown. In fact, the OLS estimator, c_i, is the unique minimax estimator of γ_i under normal distribution theory; its risk, $MSE(c_i) = \sigma^2/\lambda_i$, is constant for all values of γ_i. Clearly, $\delta_i^{MSE} c_i$ would have the same risk as c_i when $\delta_i^{MSE} = 1$ and strictly smaller risk than c_i whenever $\delta_i^{MSE} < 1$; thus $\delta_i^{MSE} c_i$ would be a "better" minimax estimator of γ_i than is c_i alone if $\delta_i^{MSE} c_i$ were an operational estimate, which it isn't.

Another expression for the MSE optimal extent of shrinkage along the $i-$th principal axis of predictors is

$$\delta_i^{MSE} = \frac{\varphi_i^2}{\varphi_i^2 + 1} \ , \tag{1.14}$$

where $\varphi_i^2 = \gamma_i^2 \lambda_i/\sigma^2$ is the (unknown) noncentrality of the F-ratio for testing $\gamma_i = 0$, equation (1.4). In fact, $n \cdot F_i/(n-r-1)$ is the maximum likelihood estimator of the φ_i^2 noncentrality parameter under normal distribution theory.

Obenchain(1978) explored a number of alternative definitions for MSE optimal shrinkage but ultimately concluded that (1.12) is the most reasonable definition overall. An example of an unusable definition for MSE optimal estimation is the one due to Theil(1971), page 125. That non-estimator is $\tau \cdot \beta$ where the scalar $\tau = \beta' X'(X\beta\beta' X' + \sigma^2 I)^{-1} y$ is easily shown to be the linear combination with minimum MSE risk as an estimator of one, which is a known constant!

1.5 Unrestricted Maximum Likelihood

When no restrictions whatsoever are placed on the specific form of shrinkage, one is free to simply substitute maximum likelihood estimates for the unknowns in equation (1.12) or (1.14) to find the shrinkage estimator most likely to be $\delta_i^{MSE} c_i$. The resulting nonlinear, maximum likelihood estimator under normal distribution theory is of the "cubic" form

$$\hat{\delta}_i^{MSE} c_i = \frac{c_i^3}{c_i^2 + (\hat{\sigma}^2/\lambda_i)} = \frac{n \cdot r_i^{o3}}{n \cdot r_i^{o2} + (1 - R^2)} \cdot \sqrt{\frac{y'y}{\lambda_i}} \tag{1.15}$$

studied by Thompson(1968), where $\hat{\sigma}^2 = (y'y) \cdot (1 - R^2)/n$ is the maximum likelihood (rather than the unbiased) estimate of σ^2. The simulation results presented in Section 3 include some MSE risk profiles for this unrestricted shrinkage estimator.

1.6 A 2-parameter Shrinkage Family

Without loss of generality, non-stochastic shrinkage factors can be written as

$$\delta_i = \lambda_i/(\lambda_i + k_i), \tag{1.16}$$

where $k_1, \cdots, k_r$ are non-negative constants called *additive eigenvalue inflation factors* by Hoerl and Kennard(1970a). Our primary focus here will be on a 2-parameter family, Goldstein and Smith(1974), in which all r of these k_i-factors are restricted to be of the special form

$$k_i = k\lambda_i^q, \tag{1.17}$$

where k is non-negative and q is a finite power that determines the *shape* (or *curvature*) of the shrinkage path through $p-$dimensional b^*-estimate space.

The "ordinary" ridge estimator of Hoerl and Kennard(1970a) and Marquardt and Snee(1975) is the special case $q = 0$ of equation (1.17). The commonly seen formula for this estimator is $b^* = (X'X + kI)^{-1}X'y$, where the inverse matrix will always exist (even when the rank of X is $r < p$) as long as k is strictly positive.

It is not really appropriate to compare the $k-$factors for two shrinkage estimators corresponding to different $q-$shapes because the overall shrinkage pattern depends as much on q as it does on k. The most appropriate measure of overall *extent of shrinkage* is given by

$$m = r - \delta_1 - \cdots - \delta_r = rank(X) - trace(\Delta). \tag{1.18}$$

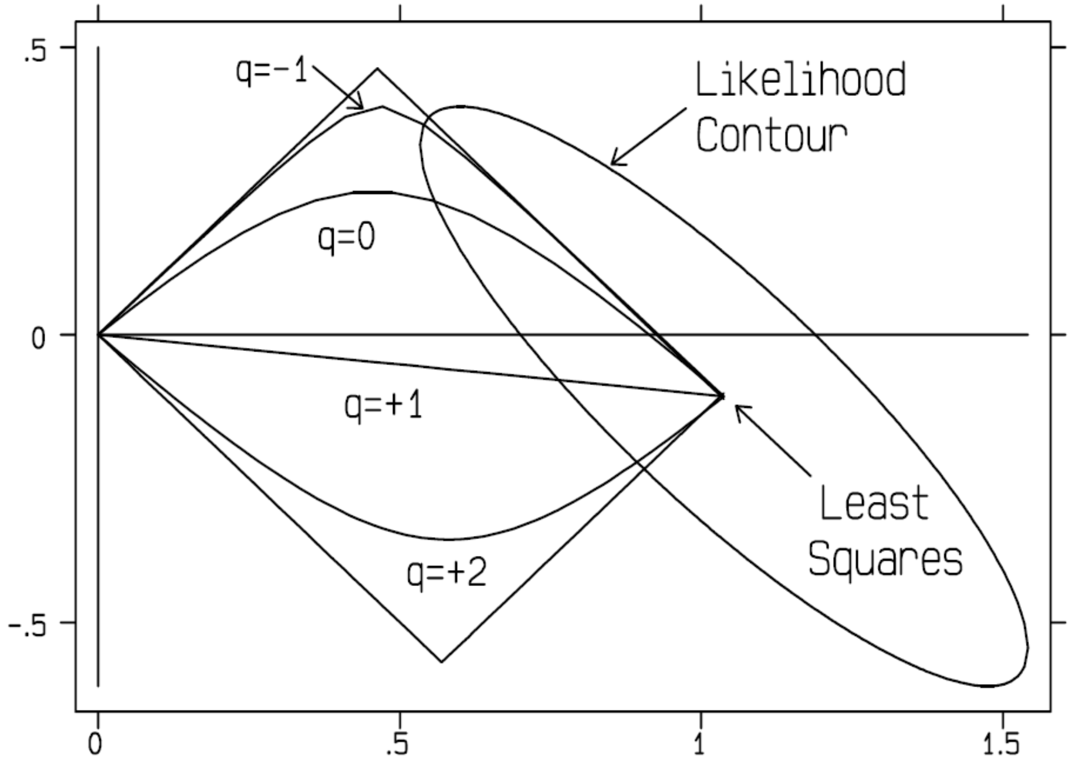

Figure 2: When the regressor X matrix is of rank $r = 2$, a $q-$shape shrinkage path can pass through any point within the shrinkage rectangle. But, when the rank of X is $r \geq 3$, the (q, k) shrinkage estimates form a surface of measure (volume) zero within the $r-$dimensional shrinkage hyper-rectangle.

This m is called the *multicollinearity allowance* parameter, introduced and discussed in Obenchain and Vinod (1974), Obenchain(1976), and Vinod (1976).

Note that $q = 1$ yields uniform shrinkage, $\delta_1 = \cdots = \delta_r = 1/(1 + k)$. In actual practice, ties among eigenvalues are rare (except in designed experiments.) The common situation is $\lambda_1 > \cdots > \lambda_r > 0$, and the first r shrinkage factors are then all unequal as long as $0 < m < r$ and $q \neq 1$ in equation (1.17). Note that $q > 1$ focuses initial shrinkage upon major principal axes, $\delta_1 < \cdots < \delta_r$, while $q < 1$ focuses initial shrinkage along minor axes, $\delta_1 > \cdots > \delta_r$. These $q < 1$ (*declining deltas*) shrinkage patterns, when favored by the response $y-$data, have much greater potential for reduction in MSE risk via variance-bias trade-offs than do the $q > 1$ patterns.

The limit as q approaches $+\infty$ is optimal for the Gibbons(1981) “unfavorable” case where the true β vector lies along the eigenvector corresponding to the smallest positive regressor eigenvalue, λ_r. Shrinkage to $m = r - 1$ ($\delta_1 = \cdots = \delta_{r-1} = 0$) then reduces all components of b^o orthogonal to the true β to zero! Similarly, the limit as q

approaches $-\infty$ contains all Massy(1965) "type (a)" principal component regression solutions. In both of these limiting cases, the shrinkage path travels along a series of edges of the generalized shrinkage hyper-rectangle. My experience is that the $q = \pm 5$ paths are frequently adequate to approximate these $q = \pm\infty$ limiting cases.

In numerical computations, values for the $k-$factor in (1.17) can be determined implicitly given values for m and q; my software uses Newtonian descent for this. For all (finite) choices of the $q-$shape parameter, the shrinkage estimate will coincide with least squares ($b^* = b^o$) at $m = 0$ [$k = 0$, $\delta_1 = \cdots = \delta_r = 1$], and all coefficient estimates will approach zero ($b^* = 0$) as m approaches its upper limit of $m = r$ [$k = +\infty$, $\delta_1 = \cdots = \delta_r = 0$]. When k and q are given, m is easily calculated using (1.18).

2 Restricted Maximum Likelihood Shrinkage

Note that equation (1.12) is easily solved to express the unknowns (γ and σ^2) as functions of the known eigenvalues ($\lambda_1, \cdots, \lambda_r$) and of the MSE optimal shrinkage factors ($\delta_1^{MSE}, \cdots, \delta_r^{MSE}$); this expression is

$$\gamma_i = \pm\sigma\sqrt{\delta_i^{MSE}/[\lambda_i(1-\delta_i^{MSE})]}\ . \tag{2.1}$$

If an estimator within the restricted, 2-parameter shrinkage factor family $\delta_i = 1/(1 + k\lambda_i^{q-1})$ of equation (1.17) were MSE optimal, equation (2.1) would then become

$$\gamma_i = \pm\sigma/\sqrt{k\lambda_i^q}. \tag{2.2}$$

Letting ρ^2 denote the common, unknown value of $\gamma_1^2\lambda_1^q = \cdots = \gamma_r^2\lambda_r^q = \sigma^2/k$ within our restricted 2-parameter search for MSE optimal shrinkage factors, the normal-theory likelihood that σ^2 and γ are of this highly restricted form for any given values of k (or m) and q is $L(\gamma, \sigma) = (2\pi\sigma^2)^{-n/2}e^{-u^2/2\sigma^2}$, where the intercept of (1) is implicitly $\hat{\mu} = \bar{y} - \bar{x}'\hat{\beta}$ and the quadratic form in the exponential is

$$u^2 = (y - X\beta)'(y - X\beta)$$

$$= y'y - 2\sqrt{y'y}\cdot\rho\cdot\sum|r_i^o|\lambda^{(1-q)/2} + \rho^2\cdot\sum\lambda_i^{(1-q)}, \tag{2.3}$$

because $y'X\beta = y'H\Lambda^{1/2}G'\beta = \sqrt{y'y}\cdot r^{o\prime}\Lambda^{1/2}\gamma$ and $\gamma_i = \pm\rho\lambda_i^{-q/2}$ under the restriction. Note, in particular, that the numerical sign of each $\hat{\gamma}_i$ has been taken to agree with its principal correlation, r_i^o, in equation (2.3); these sign choices make the middle term of (2.3) as negative as possible, thus reducing the quadratic form as long as $\rho \geq 0$. Once maximized by choice of this estimate for ρ, the $L(\gamma, \sigma)$ likelihood resulting from equation (2.3) is defined to be the likelihood that the given k (or m) and q yield MSE optimal $\delta-$factors.

Since the second derivative, $\partial^2[u^2]/\partial\rho^2 = +2 \cdot \sum \lambda_i^{(1-q)}$, is strictly positive, the minimum of the u^2 quadratic form is achieved at $\partial[u^2]/\partial\rho = 0$, which is $\rho = \sqrt{y'y} \cdot \sum |r_j^o| \lambda_j^{(1-q)/2} / \sum \lambda_j^{(1-q)}$. The corresponding minimum u^2 is $u^2 = y'y \cdot [1 - R^2 CRL^2(q)]$, where $R^2 = \sum r_j^{o2}$ and the *curlicue function* is

$$CRL(q) \equiv \frac{\sum \left| r_j^o \right| \lambda_j^{(1-q)/2}}{\sqrt{\sum r_j^{o2} \sum \lambda_j^{(1-q)}}}. \tag{2.4}$$

Note that $CRL(q)$ is the Cosine of the angle between the R-vector of absolute values of the *principal correlations* [r^o of equation(1.3)] and the L-vector of predictor eigenvalues raised to the power $(1-q)/2$.

Our next step is to minimize the minus-two-log-likelihood of $n \cdot \ln(2\pi\sigma^2) + \hat{u}^2/\sigma^2$ by choice of our estimate for σ^2, where n is again the number of observations. Differentiating as usual, we find that the best choice for $\hat{\sigma}^2$ is the minimum $\hat{u}^2$ divided by n. Furthermore, the corresponding MSE optimal $k-$factor, Obenchain(1981), is

$$\hat{k} = \hat{k}(q) = \hat{\sigma}^2/\hat{\rho}^2 = [\sum \lambda_j^{(1-q)}] \cdot \frac{[1 - R^2 \cdot CRL^2(q)]}{n \cdot R^2 \cdot CRL^2(q)}. \tag{2.5}$$

The final step is to further minimize the minimum $\hat{u}^2$ quadratic form (maximize the likelihood) by maximizing $CRL(q)$ over choice of $q-$shape for the shrinkage path. To approximate this MSE optimal $q-$shape, one may simply locate the maximum $CRL(q)$ over a lattice of values for q, typically covering the range $-5 \leq q \leq +5$ (and always including at least the range $-2 \leq q \leq +2$.)

Note that the arguments used here do not, technically, assume that each σ^2/γ_i^2 actually is equal to $k\lambda_i^q$. Rather, we are simply asking: "Which choice of k and q make it most likely that the unknown σ^2/γ_i^2 are of this $k\lambda_i^q$ form?" The corresponding minus-twice-log-likelihood-ratio for the maximum likelihood 2-parameter solution relative to the unrestricted solution of (1.15) is

$$T^2(q) = n \cdot \ln\{1 + \frac{R^2[1 - CRL^2(q)]}{(1 - R^2)}\}. \tag{2.6}$$

A large sample chi-squared test of the 2-parameter restriction has $(r-2)$ degrees-of-freedom when $T^2(q)$ has been minimized by choice of $q-$shape. A significantly large $T^2(q)$ then suggests that the 2-parameter family of (1.17) is too restrictive to contain the overall MSE optimal shrinkage $\delta-$factors.

3 Simulated Risk Profiles

The only published simulation results comparing the maximum likelihood approach described here with other methods for choosing an extent of shrinkage are those of Gibbons(1981). Her "O-method" results are remarkable in the sense that she generated them *before* the closed-form expression, (2.5), for the maximum likelihood $k-$factor was developed! Instead, Gibbons maximized $CRL(q)$ over the range $-5 \leq q \leq +1$ and then performed a second grid search for the optimal shrinkage $k-$extent using the general likelihood monitoring equations of Obenchain(1975). Gibbons found that the O-method can be much superior to Golub, Heath and Wahba(1979) generalized cross-validation when R^2 exceeds 0.5 and the MSE optimal shrinkage pattern corresponds to $q = -\infty$ (i.e. the true β vector lies along the first, major principal axis of regressors.) Gibbons also simulated some "unfavorable" cases where the MSE optimal shrinkage pattern corresponds to $q = +\infty$ (i.e. the true β vector lies along the last, minor principal axis of regressors). The O-method didn't perform well in these cases because Gibbons did not allow the fitted $q-$shape to exceed $+1$ (uniform shrinkage), but Gibbons did note the extreme allowed $q-$shape of $q = +1$ was always selected by the O-method in all of her "unfavorable" simulations.

The closed form expression, (2.5), for the maximum likelihood choice of $k-$extent for shrinkage along a path of given $q-$shape makes it possible to study the MSE risk of this estimator using a variety of techniques ranging from exact calculations to large sample approximations to numerical integration to Monte-Carlo simulation. The primary challenge in developing risk profiles for the maximum likelihood shrinkage estimator is in overcoming the potentially high dimensionality of the problem. Specifically, the extensive array of parameters that could be varied include the relative sizes of true γ components, spread in regressor λ eigenvalues, size of error σ^2 variance, number p of coefficients, rank r of components, number of degrees-of-freedom for error, etc.

A simple but interesting special case of (2.5) occurs for uniform shrinkage, $q = 1$; the maximum likelihood estimate for this common shrinkage factor is

$$\hat{\delta}_1 = \cdots = \hat{\delta}_r = 1/(1+\hat{k}) = \frac{n}{(n-r) + \overline{|r^o_{\cdot}|}^{-2}}, \tag{3.1}$$

where $\overline{|r^o_{\cdot}|} = \sum |r^o_j|/r$ is the average, absolute principal correlation. If we consider only the "null" case where all $p = r$ components are of equal noncentrality ($\varphi_i^2 = \gamma_i^2 \lambda_i / \sigma^2$), then $q = 1$ is also the MSE optimal path shape. The three parameters characterizing these special cases are thus: r = number of components to be uniformly shrunken, $(n - r - 1)$ = degrees-of-freedom for error, and $\delta^{MSE} = \varphi^2/(\varphi^2 + 1)$ = the MSE optimal extent of shrinkage for each component.

Figures 3 and 4 display the MSE risk profile associated with shrinking a single component and were generated by simulation. My simulation software uses a Monte-Carlo "swindle" that allows all estimators and all optimal extents of shrinkage to be simultaneously evaluated using the same sets of pseudo-random normal deviates, assuring that all simulated MSE risk values are as highly positively intercorrelated as is possible. I found that one million replications were sufficient to achieve 3+ decimal place accuracy in the simulated values for the **known** MSE risk of both least squares and the optimal shrinkage non-estimator, so similar accuracy is to be expected in estimates of the **unknown** risks of the non-linear estimators of equations (1.15) and (3.1).

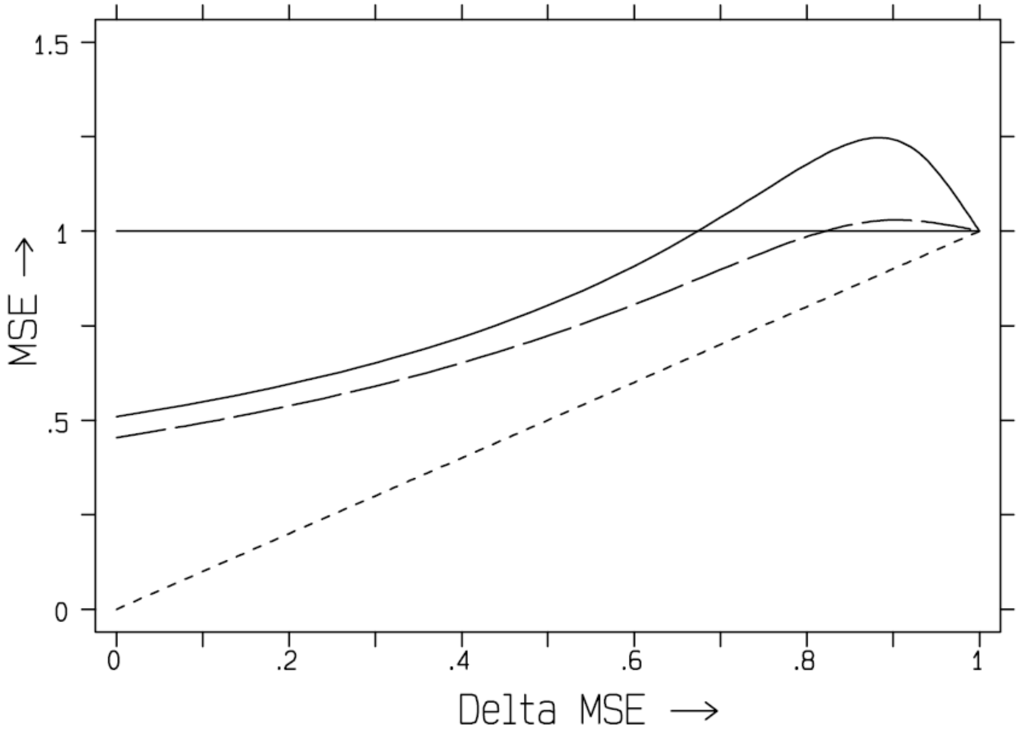

Figure 3: Simulated MSE risk with 5 degrees-of-freedom for error; the horizontal line denotes the constant MSE risk of least squares; the solid curve detotes the MSE risk of unrestricted (cubic) maximum likelihood; long dashes denote the MSE risk of maximum likelihood uniform shinkage of either one of p = r = 2 equal effects; short dashes denote the unobtainable MSE risk of the optimal non-estimator.

The top curves in Figures 3 and 4 display typical simulation results for the MSE risk profile of the unrestricted (cubic) maximum likelihood estimator. This estimator can yield an approximate 50% decrease in MSE risk when δ_i^{MSE} is close to zero, and its risk will be no larger than that of c_i as long as δ_i^{MSE} is no larger than about 0.66. But the MSE risk of this nonlinear estimator can also be as much as 25% larger than that of the least squares c_i when δ_i^{MSE} is in the 0.8 to 0.9 range.

The middle, long-dashed curves in Figures 3 and 4 display simulated MSE risk

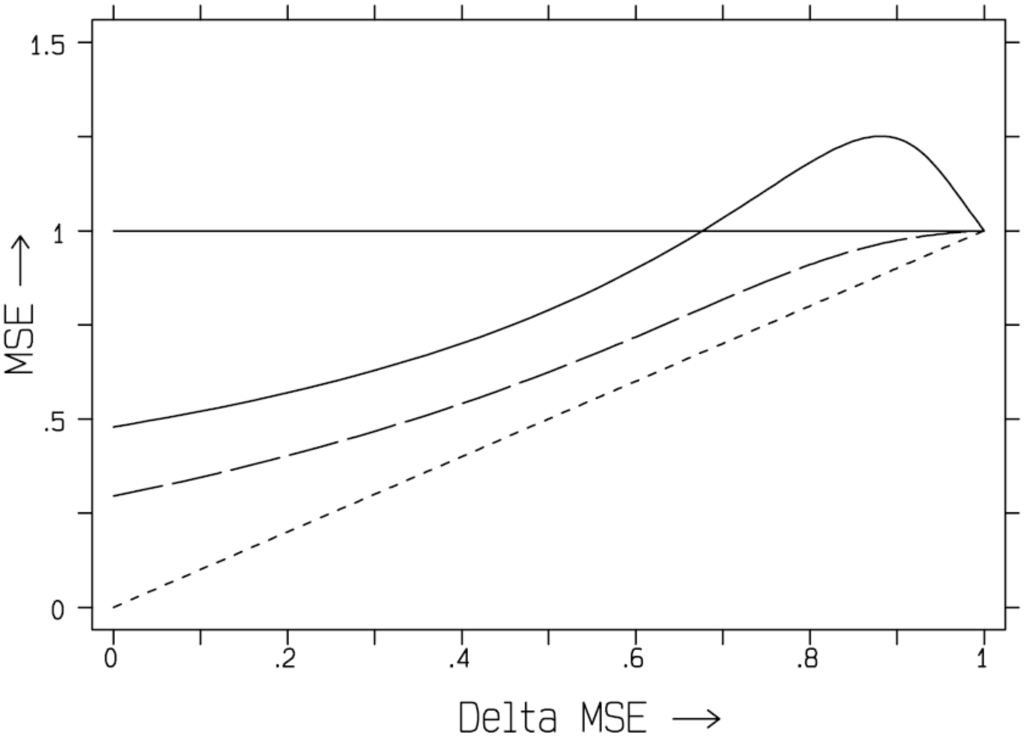

Figure 4: Simulated MSE risk with 20 degrees-of-freedom for error; the horizontal line denotes the constant MSE risk of least squares; solid curve denotes the MSE risk of unrestricted (cubic) maximum likelihood; long dashes denote the MSE risk of maximum likelihood uniform shinkage of any one of p = r = 4 equal effects; short dashes denote the unobtainable MSE risk of the optimal shrinkage non-estimator.

profiles for maximum likelihood uniform shrinkage ($q = 1$) of any one of $p = r = 2$ or 4 components, respectively, in the "null" case where the true effects have equal noncentrality. Note that this approach yields more than a 50% decrease in MSE risk when δ_i^{MSE} is close to zero, and the corresponding MSE risk when δ_i^{MSE} is large can be limited to being no more than a modest 3% increase.

4 Summary Remarks

Other likelihood-based methods besides the classical, fixed coefficient approach of Obenchain (1975, 1981) are available. For example, an empirical Bayes criterion [EBAY] was proposed by Efron and Morris(1977), and two equivalent "random coefficient" criteria [RCOF] were proposed by Golub, Heath and Wahba(1979) and by Shumway(1982). While these criteria can be monitored along any shrinkage path, they yield neither a closed form expression for an optimal extent of shrinkage nor insights into choice of path shape. In fact, the truly unique contribution of the classical, maximum likelihood approach [CLIK] described above is that it provides important insights on selection of both the $q-$shape and the $k-$extent of shrinkage.

My experience applying Normal theory likelihood-based shrinkage to example data sets is that the EBAY and RCOF criteria tend to agree rather closely, while the CLIK criterion typically suggests slightly more shrinkage (along paths of appropriate $q-$shape) as being "optimal" for scalar-valued measures of MSE risk, like equation (1.9). On the other hand, the full risk in estimating a $p-$vector of β coefficients for a centered $X-$matrix of rank $r > 1$ is matrix-valued. Thus, to have any chance of being a "good" estimate of β that dominates OLS in **every MSE** sense, the $2/r$ths "Rule-of-Thumb" of Obenchain(1978) calls for restricting the $m-$extent of shrinkage, equation (1.18), to be only $2/r$ths of the "optimal" $m-$extent.

Classical, Normal theory, maximum likelihood approachs to choosing an extent of shrinkage along a $q-$shape path can dramatically decrease MSE risk in "favorable" cases. More importantly, perhaps, this approach can also severely limit MSE risk increases in unfavorable cases. Even in its limiting "cubic" form, the potential decrease in MSE risk appears to always be at least twice the possible increase.

My new web site (2015) provides free downloads of matrix-language source code for maximum-likelihood shrinkage regression computations [CLIK, EBAY and RCOF] in R, SAS/IML, Stata and GAUSS. I also distribute a slightly modified version of the splendidly interactive routines for ridge shrinkage in XLisp-Stat by Walter(1994).

5 Dedication

Hoerl(1962) wrote "A maximum likelihood solution for the ridge analysis has not yet been theoretically derived." Art Hoerl died on December 13, 1994, and Robert Kennard died on September 11, 2011. This manuscript is dedicated to the memory of their pioneering spirits.

6 References

Efron, B. and Morris, C. (1977). Comment on "A simulation study of alternatives to ordinary least squares," by A. P. Dempster, Martin Schatzoff, and Nanny Wermuth. **Journal of the American Statistical Association** 72, 91-93.

Gibbons, D. G. (1981). "A simulation study of some ridge estimators." **Journal of the American Statistical Association** 76, 131-139.

Goldstein M. and Smith, A. F. M (1974). "Ridge-type estimators for regression analysis." **Journal of the Royal Statistical Society** B 36, 284-291.

Golub, G. H., Heath, M., and Wahba, G. (1979). “Generalized cross-validation as a method for choosing a good ridge parameter.” **Technometrics** 21, 215-223.

Hoerl, A. E. (1962). “Applications of ridge analysis to regression problems.” **Chemical Engineering Progress** 58, 54-59.

Hoerl, A. E. and Kennard, R. W. (1970a). “Ridge regression: biased estimation for non-orthogonal problems.” **Technometrics** 12, 55-67.

Hoerl, A. E. and Kennard, R. W. (1970b). “Ridge regression: applications to non-orthogonal problems.” **Technometrics** 12, 69-82. (errata: 723.)

Massy, W. F. (1965). “Principal components regression in exploratory statistical research.” **Journal of the American Statistical Association** 60, 234-246.

Marquardt, D. W. and Snee, R. D. (1975). “Ridge regression in practice.” **The American Statistician** 29, 3-20.

Obenchain, R. L. and Vinod, H. D. (1974). “Estimates of partial derivatives from ridge regression on ill-conditioned data.” **NBER-NSF Seminar on Bayesian Inference in Econometrics**, Ann Arbor, Michigan.

Obenchain, R. L. (1975). “Ridge analysis following a preliminary test of the shrunken hypothesis.” **Technometrics** 17, 431-441. (with discussion 443-445.)

Obenchain, R. L. (1976). “Methods of ridge regression.” **Proceedings of the Ninth International Biometrics Conference,** Invited Papers Volume 1, 37-57. Boston.

Obenchain, R. L. (1977). “Classical F-tests and confidence regions for ridge regression.” **Technometrics** 19: 429-439.

Obenchain, R. L. (1978). “Good and optimal ridge estimators.” **Annals of Statistics** 6, 1111-1121.

Obenchain, R. L. (1981). “Maximum likelihood ridge regression and the shrinkage pattern alternatives.” **Institute of Mathematical Statistics Bulletin** 10, 37; Abstract 81t-23.

Obenchain, R. L. (1984). “Maximum likelihood ridge displays.” **Communications in Statistics, A** 13, 227-240. (Proceedings of the Fordham Ridge Symposium, ed. H. D. Vinod.)

Obenchain, R. L. (2005 - 2011). “Maximum Likelihood Shrinkage via Generalized Ridge or Least Angle Regression.” **RXshrink** R package. Available on CRAN at http://cran.r-project.org/package=RXshrink

Obenchain, R. L. (2015). “Shrinkage” and “eBook” items on Main Menu. **RBstats & softRX freeware homepage.** http://localcontrolstatistics.org

Shumway, R. H. (1982). “Maximum likelihood estimation of the ridge parameter in linear regression.” Technical Report, Division of Statistics, University of California at Davis.

Theil, H. (1971). **Principles of Econometrics**. Amsterdam: North Holland Publishing Company.

Thompson, J. R. (1968). “Some shrinkage techniques for estimating the mean.” **Journal of the American Statistical Association** 63, 113-122.

Vinod, H. D. (1976). “Application of new ridge methods to a study of Bell System scale economies.” **Journal of the American Statistical Association** 71, 835-841.

Vinod, H. D. and Ullah, A. (1981). **Recent Advances in Regression Methods.** New York: Marcel Dekker.

Walter, B. (1994). Unpublished XLisp-Stat code for 2-parameter shrinkage regression. Technische Universitaet Muenchen, Weihenstephan, Germany.