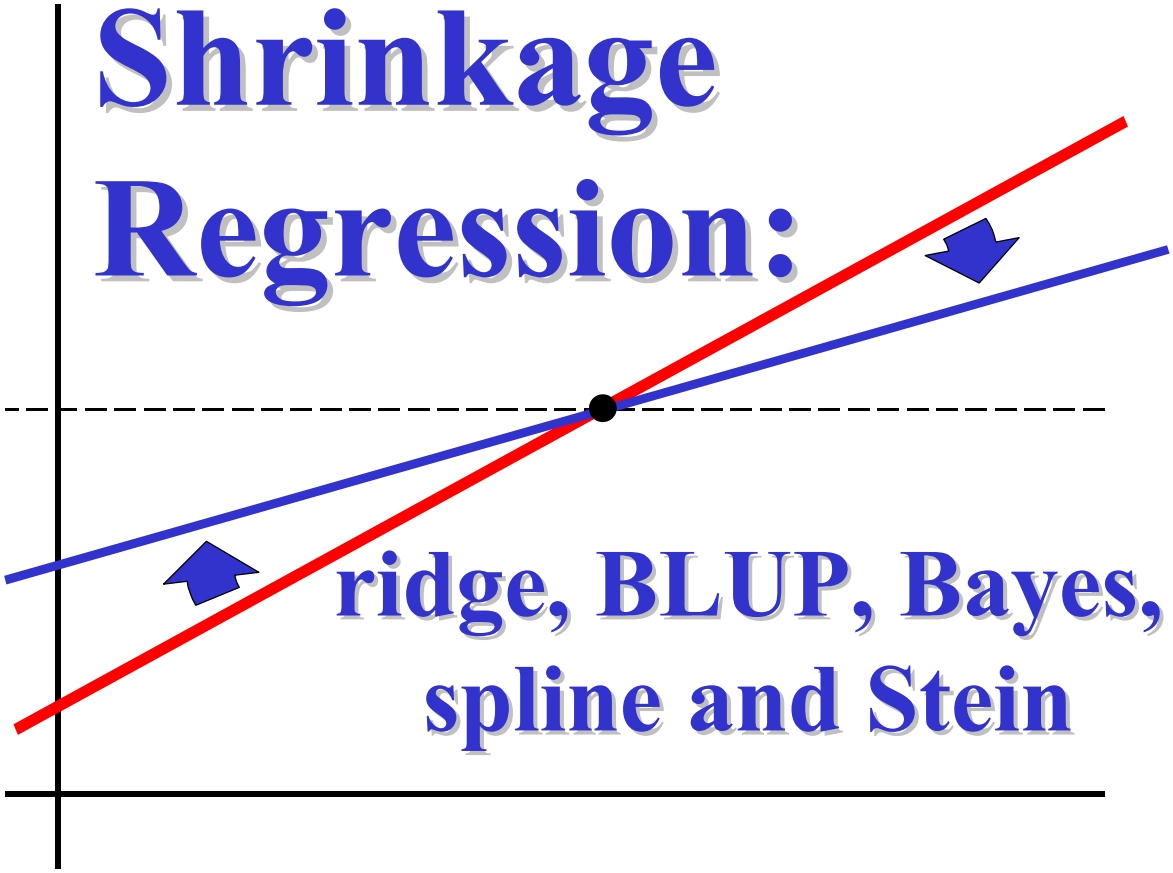


Chapter 04: The Risk of Shrinkage

Bob Obenchain, Ph.D.
softR$_X$ freeware
13212 Griffin Run
Carmel, Indiana 46033-8835

Chapter 4: THE RISK OF SHRINKAGE

What criterion does one use to determine an ideal amount of shrinkage in a particular situation?

In attempting to address this question, we consider two "standard" setups for general linear models. In sections §4.1 to §4.3, the symbols β and σ^2 will denote the unknown, true values for the classical (fixed effect) regression coefficient vector and the residual variance, respectively. On the other hand, when mixed (fixed and random coefficient) models are considered in section §4.4, the expected value of β will be denoted by β_o while its variance will be denoted by Σ_β .

The common thread that binds the arguments presented in this chapter is that desirable forms/extents of shrinkage are characterized as using variance-bias tradeoffs to reduce measures of the overall mean-squared-error risk in estimating β. The major advantage of adopting the β, σ^2, β_o , Σ_β notation described above is that risk formulas can utilize these unknown quantities essentially as if they had known values.

On the other hand, the truly pivotal, simplifying assumption that we make here in all sections of Chapter 4, except section §4.3, is that the regression shrinkage factors (the multiplicative δ_i terms) are known constants. This allows us to view shrinkage estimators as linear estimators and, thus, to develop simple, closed-form expressions for statistical characteristics (bias, variance, mean-squared-error risk, etc.) of shrinkage estimators. This "non-stochastic" shrinkage formulation is of questionable cogency in addressing the "full range" of practical questions that can arise in actual applications of shrinkage regression to ill-conditioned data. But it does provide us with a good starting-off point. In fact, we will see that the problem of selecting a/the "best" type of shrinkage is really a rather difficult question to address even in our possibly oversimplified, non-stochastic shrinkage formulation.

The main distinction between the "optimal" shrinkage formulations of section §4.1 and the "good" shrinkage formulations of section §4.2 is that the corresponding measures of overall mean-squared-error risk are scalar-valued and matrix-valued, respectively. Scalar-valued risk measures can usually be unambiguously minimized, so the corresponding shrinkage estimators are called "optimal" in section §4.1. Matrix-valued measures of risk are realistically "multivariate," and "orderings" of risk matrices can be ambiguous. Thus the arguments of section §4.2 simply compare shrinkage estimators to the least-squares estimator, labeling a shrinkage estimator "good" when it dominates least-squares in <u>every</u> mean-squared-error sense for a specified β, σ^2 pairing.

Features that the optimal and good estimators described in sections §4.1 and §4.2 share include the following two fundamental disclaimers:

(i) Whether or not a given set of shrinkage factors, Δ , yield an optimal or good $b^{\bigstar}$ depends upon β and σ^2. No fixed, given Δ can yield an optimal or good $b^{\bigstar}$ for every β and σ^2.

(ii) One never knows in practical applications which $b^{\bigstar}$'s actually are optimal or good. Again, these properties depend upon the unknowns, β and σ^2, that are to be estimated from the data at hand.

Because the formulas that define optimal and good estimators fail to yield "operational" versions of those estimators, it is probably best to think of these concepts as simply defining target values for appropriate forms and extents of shrinkage. Identification of shrinkage estimators "most likely" to be optimal or good in a practical application is the primary topic of our next chapter, Chapter 5. And Chapters 6 through 9 outline a spectrum of alternative methods for identifying desirable patterns of shrinkage.

4.1 Classical "Optimal" Shrinkage

The classical mean-squared-error risk matrix of $b^{\bigstar} = G \Delta c$ as an estimator of the unknown, fixed β vector, where Δ is a given diagonal matrix of non-stochastic generalized shrinkage factors, is:

$$\text{MSE}(b^{\bigstar}) = \text{E}[(b^{\bigstar} - \beta)(b^{\bigstar} - \beta)^{\text{T}}] = G \, \text{MSE}(\Delta c) \, G^{\text{T}} , \qquad \{ 4.1 \}$$

where G is the $P \times R$ semi-orthogonal matrix of principal axis direction cosines for the centered regressors matrix, X , of equation { 2.8 } and

$$\text{MSE}(\Delta c) = \sigma^2 \, \Delta^2 \, \Lambda^{-1} + (I - \Delta) \, \gamma \, \gamma^{\text{T}} \, (I - \Delta) \qquad \{ 4.2 \}$$

is the mean-squared-error matrix of Δc as an estimator of the true vector of uncorrelated components, γ of equation { 2.19 }. Note that MSE (Δc) is the sum of two matrices, namely:

(i) the diagonal variance matrix, $\sigma^2 \, \Delta^2 \, \Lambda^{-1}$, which is void of covariance terms because the components of Δc are uncorrelated when Δ is non-stochastic,

plus

(ii) the rank-one matrix, $(I - \Delta) \, \gamma \, \gamma^{\text{T}} \, (I - \Delta)$, with squared bias terms along its diagonal and bias cross-product terms off that diagonal.

4.1.1 Diagonal Elements of Mean Squared Error Matrices

For our first specific example of a scalar-valued measure of risk, let us focus upon any single diagonal element, say the i-th, of the mean-squared-error matrix of $\Delta\, c$:

$$\mathrm{MSE}(\delta_i c_i) \;=\; \sigma^2 \delta_i^2/\lambda_i \;+\; (1-\delta_i)^2 \gamma_i^2 \;. \qquad \{\, 4.3 \,\}$$

Now $\mathrm{MSE}(\delta_i c_i)$ of { 4.3 } will clearly change as the i-th δ-factor changes. In fact, the partial derivative of $\mathrm{MSE}(\delta_i c_i)$ with respect to δ_i is

$$\partial \mathrm{MSE}(\delta_i c_i)/\partial \delta_i \;=\; 2\sigma^2 \delta_i/\lambda_i \;-\; 2(1-\delta_i)\gamma_i^2 \;, \qquad \{\, 4.4 \,\}$$

while the second partial derivative is a non-negative constant...

$$\partial^2 \mathrm{MSE}(\delta_i c_i)/\partial \delta_i^2 \;=\; 2\sigma^2/\lambda_i \;+\; 2\gamma_i^2 \;. \qquad \{\, 4.5 \,\}$$

It follows from { 4.5 } that equating $\partial \mathrm{MSE}(\delta_i c_i)/\partial \delta_i$ of { 4.4 } to zero will yield a MINIMUM value for $\mathrm{MSE}(\delta_i c_i)$ as long as either $\sigma^2 > 0$ or $\gamma_i^2 > 0$. This optimal amount of shrinkage for the i-th uncorrelated component, c_i, is

$$\begin{aligned} \delta_i^{\mathrm{MSE}} \;&=\; \gamma_i^2 \,/\, [\, \gamma_i^2 + (\, \sigma^2 \,/\, \lambda_i \,) \,] \;=\; \lambda_i \,/\, [\, \lambda_i + (\, \sigma^2 \,/\, \gamma_i^2 \,) \,] \,, \qquad \{\, 4.6 \,\} \\ &=\; \phi_i^2 \,/\, (\, \phi_j^2 \;+\; 1 \,) \;=\; (\, 1 + \phi_j^{-2} \,)^{-1}, \end{aligned}$$

where $\phi_j^2 = \gamma_i^2 \lambda_i/\sigma^2$ of { 2.24 } is the noncentrality parameter of the F-ratio, $F_i = c_i^2 \lambda_i/s^2$ of { 2.22 }, for testing the hypothesis that the i-th true uncorrelated component, γ_i, of β is zero. It follows that

$$\mathrm{MSE}(\delta_i c_i) \;\geq\; \sigma^2 \cdot \lambda_i^{-1} \cdot \delta_i^{\mathrm{MSE}}, \qquad \{\, 4.7 \,\}$$

with equality only at $\delta_i \;=\; \delta_i^{\mathrm{MSE}}$. Note, in particular, that the lower bound on $\mathrm{MSE}(\delta_i c_i)$ of { 4.7 } involves the first power of δ_i^{MSE}...not its square.

See Figures 4.2, 4.3, and 4.4 in section §4.2 of this chapter for graphs that show how $\mathrm{MSE}(\delta_i c_i)$ changes as δ_i decreases from 1 to 0 in the cases where $\phi_i^2>1$, $\phi_i^2=1$, or $\phi_i^2<1$, respectively. Technically, these figures actually display "relative" mean-squared-errors, defined as $\mathrm{MSE}(\delta_i c_i)/\mathrm{MSE}(c_i) = \lambda_i \cdot \mathrm{MSE}(\delta_i c_i) \,/\, \sigma^2$.

Next, note that δ_i^{MSE} of { 4.6 } can never be negative nor larger than one,

$$0 \;\leq\; \delta_i^{\mathrm{MSE}} \;\leq\; 1 \,. \qquad \{\, 4.8 \,\}$$

On the other hand, δ_i^{MSE} will be zero only when γ_i is zero and/or σ^2 is plus infinity. Similarly, δ_i^{MSE} will be one only when γ_i^2 is plus infinity and/or σ^2 is zero. But cases where either γ_i^2 or σ^2 is infinite or else σ^2 is zero are of little or no practical interest in applications. Therefore, the optimal shrinkage range of PRACTICAL INTEREST is

$$0 \leq \delta_i^{MSE} < 1 ,$$

with equality only when γ_i is zero.

4.1.2 MSE Measures Depending Only Upon Diagonal Elements

It so happens that several well known, scalar-valued measures of the "overall" mse risk in $b^{\bigstar} = G \Delta c$ depend only upon the diagonal elements of the MSE (Δc) matrix. And all such measures are obviously minimized when $\delta_i = \delta_i^{MSE}$ of { 4.6 } for i = 1, ..., R [where R = rank(X).] Two such examples of overall mse risk that depend only upon the diagonal elements of MSE(Δc) are...

(i) Summed Mean Squared Error [Hoerl and Kennard(1970a)]:

$$\begin{aligned} SMSE(b^{\bigstar}) &= \text{trace}\{ E [(b^{\bigstar} - \beta)(b^{\bigstar} - \beta)^T] \} , \\ &= \text{trace}\{ G\, MSE(\Delta c)\, G^T \} , \\ &= \text{trace}\{ MSE(\Delta c)\, G^T G \} = \text{trace}\{ MSE(\Delta c) \}, \end{aligned} \qquad \{ 4.9 \}$$

and

(ii) Summed, Scaled Predictive Mean Squared Error [Mallows(1973)]:

$$\begin{aligned} PMSE(b^{\bigstar}) &= 1 + (E \| Xb^{\bigstar} - X\beta \|^2) / \sigma^2 , \\ &= 1 + [\sum_{i=1}^{R} \lambda_i\, MSE(\delta_i c_i)] / \sigma^2 . \end{aligned} \qquad \{ 4.10 \}$$

4.1.3 Weighted Mean Squared Error Measures

So far, we have only considered scalar-valued risk-of-shrinkage criteria that ignore the off-diagonal bias cross-product terms in the MSE (Δc) matrix. Scalar-valued risk measures that may (or may not) depend upon off-diagonal bias terms are forms of "weighted" mean-squared-error:

$$\begin{aligned} wmse(b^{\bigstar}, W) &= E [(b^{\bigstar} - \beta)^T W (b^{\bigstar} - \beta)] , \\ &= \sigma^2 \cdot \text{trace}(M \Delta^2 \Lambda^{-1}) + \gamma^T (I - \Delta) M (I - \Delta) \gamma , \end{aligned} \qquad \{ 4.11 \}$$

where W and M are matrices of weights. W is a $P \times P$ non-stochastic weight matrix that is always taken to be symmetric and either non-negative definite (i.e. $\alpha^T W \alpha \geq 0$ for every α) or positive definite (i.e. $\alpha^T W \alpha > 0$ for every $\alpha \neq 0$.) Similarly, $M = G^T W G$ is the corresponding weight matrix for the uncorrelated components of $b^\star$.

Note that { 4.9 } and { 4.10 } are both special cases of weighted mse:

$$\text{wmse}(b^\star, I) = E [(b^\star - \beta)^T (b^\star - \beta)] = \text{SMSE}(b^\star) ,$$

for the positive definite choice $W = I$, and

$$\text{wmse}(b^\star, X^T X) = \sigma^2 [\text{PMSE}(b^\star) - 1] ,$$

for the rank R choice $W = X^T X$.

Consider the following result, from Obenchain(1978), that applies whenever the weight matrix, W, is positive definite...

> WEIGHTED MEAN-SQUARED-ERROR OPTIMALITY: If σ^2 is strictly positive, β is finite, and W is positive definite, then wmse($b^\star$, W) of { 4.11 } is minimized by choice of non-stochastic shrinkage factors Δ at:
>
> $$\delta_i = \gamma_i \cdot \lambda_i \cdot \eta_i / (\sigma^2 \cdot m_{ii}) , \qquad \{ 4.12 \}$$
>
> where η_i is the i-th element of $\eta = (D + M^{-1})^{-1} \gamma$, $M = ((m_{ij})) = G^T W G$, and D is the diagonal matrix with i-th diagonal element ϕ_i^2 / m_{ii}. Equivalently, $\Delta \gamma = D \eta = D (D + M^{-1})^{-1} \gamma$.

Note from the second expression for wmse($b^\star$, W) in { 4.11 } that

$$\partial \text{wmse} / \partial \delta_i = 2 (\sigma^2 m_{ii} / \lambda_i) \cdot \delta_i - 2 \sum_{j=1}^{R} m_{ji} \gamma_i \gamma_j \cdot (1 - \delta_j) , \qquad \{ 4.13 \}$$

$$\partial^2 \text{wmse} / \partial \delta_i \partial \delta_j = 2 \, m_{ji} \gamma_i \gamma_j \quad \text{for } j \neq i,$$

and

$$\partial^2 \text{wmse} / \partial \delta_i^2 = 2 \, m_{ii} (\sigma^2 / \lambda_i + \gamma_i^2) .$$

The conditions that $\sigma^2 > 0$ and that W be positive definite (so that $m_{ii} > 0$) assure that the second derivatives matrix $((\partial^2 \text{wmse} / \partial \delta_i \partial \delta_j))$ will be positive definite and, therefore, that the MINIMUM wmse will occur at $\partial \text{wmse} / \partial \delta_i = 0$ for i = 1,...,R. If we define $\eta \equiv M (I - \Delta) \gamma$, then $\partial \text{wmse} / \partial \Delta = 0$ yields the “fixed point" version of equations { 4.12 }; in other words, the optimal Δ is defined in terms of an η vector that is, in turn, a function of this Δ .

When $\gamma_i = 0$, the corresponding optimal δ_i is clearly also zero, even from the fixed-point form of { 4.12 }. Any nonzero component must be finite because β is finite, so we can then

multiply { 4.12 } by γ_i , yielding $\delta_i\gamma_i = D_{ii}\eta_i$ where $D_{ii} = \phi_i^2/m_{ii}$ is the i-th diagonal element of the diagonal D matrix introduced above. As a result $\eta \equiv M (I - \Delta) \gamma$ can be rewritten as $\eta = M \gamma - M D \eta$ or $(I + M D) \eta = M \gamma$. Of course, $(I + M D)^{-1} = (D + M^{-1})^{-1} M^{-1}$ whenever M is positive definite, so we arrive at the closed-form solution $\eta = (D + M^{-1})^{-1}\gamma$ that we sought.

OBSERVATIONS ON OPTIMALLY WEIGHTED MEAN-SQUARED-ERROR:

Equation { 4.12 } shows that the optimal amount of wmse shrinkage will always be $\delta_i = \delta_i^{MSE}$ of { 4.6 } whenever the weight matrix, W, is such that $M = G^T W G$ is diagonal. After all, the terms in the summation of { 4.13 } with $j \neq i$ vanish when M is diagonal, and the two remaining terms both contain a m_{ii} factor that then cancels out!

Another special case where the optimal amount of wmse shrinkage will always be $\delta_i = \delta_i^{MSE}$ of { 4.6 } occurs when only one component, say γ_k , of γ is nonzero. And this statement holds for every positive definite choice of weight matrix, W . All terms in the summation of { 4.13 } with either $i \neq k$ or $j \neq k$ again vanish, leaving only the $-(m_{kk}\gamma_k^2) \cdot 2(1 - \delta_k)$ term in the k-th equation. Of course, $\delta_i^{MSE} = 0$ for $i \neq k$ in this case !

Once we have established some interesting results about mean-squared-error in specific directions of space in the next subsection, §4.1.4, we will show how weighted mse using the one-parameter family of weight matrices $W = I + (\zeta\text{-}1)\beta\beta^T$ provides some profound new insights about definitions of mean-squared-error optimal shrinkage.

4.1.4 The Mean Squared Error in Specific Directions

When α is a fixed vector of unit length (i.e. $\alpha^T\alpha = 1$), the rank-one weight matrix $W = \alpha\alpha^T$ yields...

$$\text{wmse}(b^\star, \alpha\alpha^T) = \alpha^T \text{MSE}(b^\star) \alpha = \text{MSE}(\pm \alpha^T b^\star) , \qquad \{ 4.14 \}$$

$$= \sigma^2 \xi^T\Delta^2\Lambda^{-1}\xi + [\xi^T(I - \Delta)\gamma]^2 .$$

MSE($\alpha^T b^\star$) measures the size of the component in the mean-squared-error of $b^\star$ in the direction parallel to $\pm\alpha$ in P-dimensional Euclidean space or, equivalently, the mean-squared-error of the linear combination $\alpha^T b^\star$ as an estimator of $\alpha^T \beta$. The last expression in { 4.14 } follows by writing $\xi^T = \alpha^T G$ to denote the unit vector that expresses the orientation of α relative to the principal axes of the centered regressors matrix, X .

Some additional results from Obenchain(1978) apply here...

> DIRECTIONAL MEAN-SQUARED-ERROR OPTIMALITY: If σ^2 is strictly positive, β is finite, and α is a fixed vector of unit length, then:

(i) MSE($\alpha^T b^\star$) does not depend upon δ_i whenever $\xi_i \equiv \alpha^T g_i = 0$, where g_i is the direction-cosine vector of the i-th principal axis of the centered regressors matrix.

(ii) MSE($\alpha^T b^\star$) depends upon δ_i whenever $\xi_i \equiv \alpha^T g_i \neq 0$ and, in fact, is minimized by choice of non-stochastic shrinkage factor, δ_i , at:

$$\delta_i = \delta_i(\alpha) = \alpha^T \beta \gamma_i \lambda_i / [\xi_i (\sigma^2 + \sum\nolimits^* \gamma_j^2 \lambda_j)] , \qquad \{ 4.15 \}$$

where $\sum^*$ denotes summation only over subscripts j such that $\xi_j \neq 0$.

The above results are demonstrated, first, by noting that $\xi^T = \alpha^T G$ implies that the i-th element of ξ will be $\xi_i \equiv \alpha^T g_i$. Thus the component of a generalized shrinkage estimator, $b^\star$, in the $+\alpha$ direction is $\alpha^T b^\star = \xi^T \Delta c = \sum \xi_j \delta_j c_j$. Thus, when an element of the ξ vector is null, there can be NO dependency of the component of $b^\star$ in the $\pm \alpha$ direction upon that δ_i factor. As a result, the corresponding "directional" mean-squared-error will also NOT depend in any way upon that δ_i factor.

It follows from { 4.14 } that

$$\partial MSE/\partial \delta_i = 2 (\sigma^2 \xi_i^2/\lambda_i) \cdot \delta_i - 2 [\xi^T(I - \Delta)\gamma] \cdot \xi_i \gamma_i , \qquad \{ 4.16 \}$$

$$\partial^2 MSE/\partial \delta_i \partial \delta_j = 2 \xi_i \gamma_i \xi_j \gamma_j \quad \text{for } j \neq i,$$

and

$$\partial^2 MSE/\partial \delta_i^2 = 2 \xi_i^2 (\sigma^2/\lambda_i + \gamma_i^2) .$$

The conditions that $\sigma^2 > 0$ and $\xi_i \neq 0$ assure that the relevant $((\partial^2 MSE/\partial \delta_i \partial \delta_j))$ sub-matrix will be positive definite and, therefore, that the MINIMUM directional MSE will occur at $\partial MSE/\partial \delta_i = 0$ for i = 1,...,R. As anticipated above, $\partial MSE/\partial \delta_i = 0$ reduces simply to $0 = 0$ and, thus, provides NO INFORMATION whenever $\xi_i = 0$. Thus the $\partial MSE/\partial \delta_i = 0$ equations for 1 = 1,...,R can be summarized as requiring that:

each optimal δ_i be proportional to $\gamma_i \lambda_i / \xi_i$ whenever its $\xi_i \neq 0$.

Furthermore, the common constant-of-proportionality is $k = [\xi^T(I - \Delta)\gamma]/\sigma^2$, which is an expression that depends upon these optimal δ_i shrinkage factors. Note, however, that we can rewrite k as: $k\sigma^2 = \xi^T\gamma - \sum \delta_j \xi_j \gamma_j = \xi^T\gamma - k\sum^* \gamma_j^2 \lambda_j$. Since $\alpha^T\beta = \xi^T\gamma$, the optimal k is thus $k(\alpha) = \alpha^T\beta / (\sigma^2 + \sum^* \gamma_i^2 \lambda_i)$ as in { 4.15 }.

A shrinkage formula equivalent to { 4.15 } requires, for i=1, ..., R, that

$$\delta_i \xi_i \gamma_i = \phi_i^2 \frac{\sum \xi_j \gamma_j}{(1 + \sum^* \phi_j^2)} \quad \text{whenever } \xi_i \neq 0.$$

OBSERVATIONS ABOUT OPTIMAL DIRECTIONAL MEAN-SQUARED-ERROR:

It is immediately clear from the form of equation { 4.15 } that the optimal shrinkage factors for the $-\alpha$ direction are identical to those for the $+\alpha$ direction.
Equation { 4.15 } also shows that the optimal amount of MSE shrinkage for direction $\alpha = g_i$ will always be: $\delta_i = \delta_i^{MSE}$ of { 4.6 } along the i-th axis. And, furthermore, δ_j will be completely undetermined for all axes $j \neq i$ when $\alpha = g_i$ because MSE($\delta_i c_i$) does not depend in any way upon shrinkage factors along orthogonal directions.

When $\beta = 0$, every α is orthogonal to β in the sense that $\alpha^T\beta = 0$, and { 4.15 } shows that drastic shrinkage ($\Delta = 0$) is optimal for all choices of α in this case. Even when $\beta \neq 0$, there still is a (R-1) dimensional space of α's that are orthogonal to β , and drastic shrinkage ($\Delta = 0$) is again optimal by { 4.15 } for all of these directions, α , such that $\alpha^T\beta = 0$. In other words,...

> Shrinking the least squares solution all of the way to ZERO by taking $\Delta = 0$, assures that NO ERRORS WHATSOEVER will be made in any direction orthogonal to the unknown, true β .

While it might be reassuring to have this knowledge of the desirability of drastic shrinkage along directions orthogonal to β, we don't want to make egregious errors in that one (unknown) direction that happens to be parallel to β.

When $\beta \neq 0$, taking α parallel to β yields $\xi_i = \gamma_i/\sqrt{\gamma^T\gamma}$. Note that no optimal shrinkage factor along the i-th principal axis, δ_i , is defined for minimizing MSE parallel to β whenever $\gamma_i = 0$. But, when $\gamma_i \neq 0$, optimal shrinkage for minimizing MSE parallel to β is necessarily of the form

$$\delta_i = k^{(=)}\lambda_i \quad \text{for} \quad k^{(=)} = (\gamma^T\gamma) / (\sigma^2 + \gamma^T\Lambda\gamma) . \qquad \{ 4.17 \}$$

In fact, it follows that MSE[$\beta^T b^\bigstar / \sqrt{\beta^T\beta}$] $\geq \sigma^2 k^{(=)}$ with equality only at $\Delta = k^{(=)} \Lambda$ of { 4.17 }.

Equation { 4.17 } provides the following, almost astounding insight ! The generalized shrinkage estimator with $\Delta = k^{(=)}\Lambda$ is:

$$b^\bigstar = k^{(=)} G \Lambda c = k^{(=)} X^T y ,$$

where $X^T y$ is the $(P \times 1)$ vector of inner products between the centered regressor matrix and the centered response vector. In other words,...

> Although the true β is unknown, we know that the generalized shrinkage estimator achieving minimal mean-squared-error parallel to β has the SAME RELATIVE MAGNITUDES as does the vector of MARGINAL INNER PRODUCTS of the regressors with the response.

Furthermore, these marginal relative magnitudes are generally different from those of the least-squares coefficients ...unless regressors are uncorrelated as in { 2.7 }.

One curious property of the optimal factors $\Delta = k^{(=)}\Lambda$ of { 4.17 } is that some of them may exceed 1. As we shall see in Chapter 6, the normal-theory maximum likelihood estimator of $k^{(=)}$ is $\sum r_{yi}^2 \lambda_i^{-1}/[R^2+(1-R^2)/n]$, where the r_{yi} are the principal correlations of { 2.23 }. Thus the true values of the optimal $\delta_i = k^{(=)}\lambda_i$ and also their normal-theory estimators are both potentially quite sensitive to the eigenvalue spectrum, Λ, of regressors; the largest and smallest λ's cause two different sorts of sensitivity to ill-conditioned regressors!

In subsection §4.1.5 we will seek a balance between the conflicting objectives of minimizing mean squared error parallel to and orthogonal to the unknown, true β. But let us first comment on some of the LESS intuitively satisfying implications of equation { 4.15 }. Specifically, for directions, α, that are "oblique" not only to the principal axes of the given, centered regressors matrix, X, but also "oblique" to the true coefficient vector, β, it turns out that the optimal shrinkage factors of { 4.15 } may NOT fall within the range $0 \leq \delta_i(\alpha) \leq 1$. By "oblique" here I simply mean that angles between directions are neither zero nor an exact multiple of 90°. Anyway, we are NOT talking about "trivial" violations of the $0 \leq \delta_i(\alpha) \leq 1$ range restriction that arise simply because principal axis i is strictly orthogonal to the chosen α; shrinkage factor δ_i remains undetermined by { 4.15 } in these cases, and δ_i could take on any numerical value without making any real difference in MSE($\alpha^T b^\star$).

To be specific, consider the special case of MSE(β_1+β_2) where $\beta = \gamma$ (i.e. G = I), $\alpha^T = (+1, +1, 0^T)/\sqrt{2}$, $\lambda_1 = \lambda_2$, and the first two components of β are of the form $\beta_1 \equiv f \cdot \beta_2$ ...where $\beta_2 \neq 0$ and the constant factor, f, is not 0, +1, or -1. In this case, relationship { 4.15 } determines only the first two shrinkage factors, δ_1 and δ_2. And the values for these factors that minimize MSE($b_1^\star$+$b_2^\star$) are...

$$\delta_1 = \frac{(f+1)\cdot f}{(a^2+f^2+1)} \text{ and } \delta_2 = \frac{(f+1)}{(a^2+f^2+1)} ,$$

where $a^2 = \sigma^2/\beta_2^2$. Note that the cases we have explicitly excluded give "reasonable" answers, namely...

f= -1 implies that α is orthogonal to β, and the optimal $\delta_1 = \delta_2$=0,
f=0 implies $\beta_1 = 0$, $\delta_1 = \delta_1^{MSE} = 0$, and δ_2=δ_2^{MSE},
and f=+1 implies the nonzero components of α lie parallel to their β components and $\delta_1 = \delta_2$=2/(a^2+2).

But the "bad news" is that...

f more negative than -1 gives a negative optimal value for δ_2,
f within (-1, 0) gives a negative optimal value for δ_1,
f within $0.5 \pm \sqrt{0.25 - a^2}$ yields an optimal δ_2 larger than +1 when a^2<0.25,

and f larger than $1+a^2$ gives an optimal δ_1 larger than +1.

Results of the above sort are, intuitively speaking, really not very pleasing or insightful. Apparently, equation { 4.15 } can exploit highly specialized [and potentially "weird"] forms of shrinkage patterns (like the $\delta_1 = f \cdot \delta_2$ pattern in the above example were $\beta_1 \equiv f \cdot \beta_2$) to gain reductions in mean-squared-error. Unfortunately, the kind of "information" exploited in these special cases is unrealistic in the sense that it is either "not available" or is potentially "incorrect/misleading" in most actual applications of shrinkage regression to real data.

It was probably quite "intrepid" of us to even consider the problem of optimizing a scalar valued function, $MSE(\alpha^T b^\bigstar)$, by choice of a relatively "large" number of shrinkage parameters, δ_1, δ_2, ..., δ_R, when the fundamental parameters involved (β and σ^2) are actually unknowns. So let us be content here with the relatively simple and intuitive results we obtained for the special cases where α corresponds to a principal regressor axis or is either strictly orthogonal to or strictly parallel to β.

4.1.5 Balancing Components of MSE Risk Parallel To and Orthogonal To the Unknown True Coefficient Vector

Any generalized shrinkage estimator can be written as

$$b^\bigstar = b^{(=)} \cdot \beta + b^{(\perp)}, \qquad \{ 4.18 \}$$

where the scalar $b^{(=)}$ determines the size of the component of $b^\bigstar$ parallel to β and $b^{(\perp)}$ is the component of $b^\bigstar$ orthogonal to β. To display explicit formulas, let β^+ denote the row-vector defining the Moore-Penrose inverse of the column-vector of regression coefficients, β. Thus $\beta^+ = 0$ when $\beta = 0$, and otherwise $\beta^+ = \beta^T / \beta^T\beta$.

Orthogonal projection in Euclidean space is accomplished by a linear operator that can be represented as a symmetric, idempotent, and uniquely determined matrix, Rao(1973) pp 46-47. Thus...

$\beta\beta^+$ is the projection for the space, of dimension 1 or 0, parallel to β,

and

$I - \beta\beta^+$ is the projection for the space, of dimension P-1 or P, orthogonal to β.

Thus equation { 4.18 } implies that

$$b^{(=)} = \beta^+ b^\bigstar = \gamma^+ \Delta c, \qquad \{ 4.19 \}$$

and

$$b^{(\perp)} = (I - \beta\beta^+) b^\bigstar = G (I - \gamma\gamma^+) \Delta c . \qquad \{ 4.20 \}$$

It would clearly be desirable for generalized shrinkage estimators based upon non-stochastic Δ factors to have the features that $b^{(=)}$ tends to be close to 1, at least when $\beta \neq 0$, and that $b^{(\perp)}$ tends to be small.

What non-stochastic choice of Δ in { 4.19 } makes $b^{(=)}$ a minimum mean-squared-error estimator of ONE ? Clearly, no choice of Δ can be of any real help if $\beta = \gamma = 0$; $b^{(=)} \equiv 0$ in this case. So suppose that $\gamma \neq 0$, and note that we then have

$$MSE(\gamma^{+}\Delta\, c) = E[(\gamma^{+}\Delta\, c - 1)^2] \; = \{\sigma^2\gamma^T\Delta^2\Lambda^{-1}\gamma + [\gamma^T(I-\Delta)\gamma]^2\} / (\gamma^T\gamma)^2 \,, \qquad \{ 4.21 \}$$

where we used the fact that $1 = \gamma^{+}\gamma \; = \gamma^T\gamma/\gamma^T\gamma$. Now { 4.21 } is identical to { 4.14 } except, of course, that γ may not be of unit length like ξ . However, the mse optimal shrinkage factors are again $\Delta = k^{(=)}\Lambda$ for $k^{(=)} = (\gamma^T\gamma) / (\sigma^2 + \gamma^T\Lambda\gamma)$ as in { 4.17 }, where the minimum value of { 4.21 } thereby attained is $\sigma^2/(\sigma^2 + \gamma^T\Lambda\gamma)$. Therefore, our only new insight so far is that $k^{(=)}\gamma^{+}\Lambda\, c$ is, when viewed as a known linear function of the uncorrelated components, c, a minimum mean-squared-estimator of $\gamma^{+}\gamma$, which is either 1 or 0. [It's perhaps unfortunate we restricted attention to estimators that have to depend upon the sample estimate, c , of γ ; if we hadn't, we might have found even "better" estimates of 0 and 1 ..., namely, 0 and 1 !]

What non-stochastic choice of Δ in { 4.20 } makes $b^{(\perp)}$ as small as possible ? Well $\Delta = 0$ clearly provides the global minimum! Unfortunately, this choice makes $MSE(b^{(=)}) = 1$ when $\beta \neq 0$, which can be considerably larger than the minimum, $\sigma^2/(\sigma^2 + \gamma^T\Lambda\gamma)$, attained at $\Delta = k^{(=)}\Lambda$, or the value $\sigma^2\gamma^T\Lambda^{-1}\gamma/(\gamma^T\gamma)^2$, attained at the unbiased, least-squares solution, $\Delta = I$.

Therefore let us now consider the problem of minimizing the weighted mean-squared-error orthogonal to β, $wmse(b^{\bigstar}, I - \beta\beta^{+})$, under a restriction on the amount of mean-squared-error allowed parallel to β, $wmse(b^{\bigstar}, \beta\beta^{+})$. The optimal solution will, again, obviously be $\Delta = 0$ when $\beta = 0$. So suppose that $\beta \neq 0$, and note that we can then write the Lagrange multiplier equation for the constrained optimization problem as...

$$\Psi(\Delta) \; = wmse(b^{\bigstar}, I - \beta\beta^{+}) \; + \; \zeta \cdot [\, wmse(b^{\bigstar}, \beta\beta^{+}) + \eta^2 \; - \; U \,] \,, \qquad \{ 4.22 \}$$

where ζ is the Lagrange multiplier, η^2 is a slack variable, and U is the desired upper limit on $wmse(b^{\bigstar}, \beta\beta^{+})$. The range of interest for U will be $\sigma^2(\gamma^T\gamma) / (\sigma^2 + \gamma^T\Lambda\gamma) \leq U \leq \sigma^2(\gamma^T\gamma)$, where the lower limit is achieved at $\Delta = k^{(=)}\Lambda$ and the upper limit is achieved at $\Delta = 0$. As usual, $\partial\Psi/\partial\eta = 0$ gives us the familiar condition that the optimum must occur where either the slack is zero [$\eta = 0$] or the multiplier is zero [$\zeta = 0$]; and $\partial\Psi/\partial\zeta = 0$ gives us the familiar condition that the constraint must be satisfied [$wmse(b^{\bigstar}, \beta\beta^{+}) \leq U$] .

Let us denote the minimal value of $wmse(b^{\bigstar}, I - \beta\beta^{+})$ achievable in { 4.22 } by choice of Δ subject to the restriction $wmse(b^{\bigstar}, \beta\beta^{+}) \leq U$ by the function h(U). Then $\partial\Psi/\partial U = 0$ implies

$$\partial h(U)/\partial U \; = \; -\zeta \;. \qquad \{ 4.23 \}$$

The larger is U, the more Δ can deviate from $k^{(=)}\Lambda$ and effect a reduction in h(U), yet still satisfy wmse($b^{\bigstar}$, $\beta\beta^{+}$) $\leq$ U .

Consider, now, the special case where $\zeta = 1$ in { 4.23 }. Slack must clearly be zero for any optimum that occurs here because $\zeta \neq 0$. Thus the primary message of { 4.23 } is that an EXACT BALANCE is struck at $\zeta = 1$ between the rate of change (decrease or increase) in weighted mean-squared-error orthogonal to β , h(U), and the rate of change (increase or decrease) in mean-squared-error parallel to β , U .

Now that we have gained the fundamental insights provided by equations { 4.18 } through { 4.23 }, we can skip over a great deal of other technical details [such as establishing the convexity of the minimization problem in { 4.21 }] by noting that our equation { 4.12 } actually provides an almost complete solution to the optimal tradeoffs problem. After all, the combined (orthogonal plus parallel) weight matrix in the Lagrange equation, namely...

$$W = I + (\zeta - 1) \cdot \beta\beta^{+} ,$$

will be positive definite as long as $\zeta > 0$. And we also already know that $\Delta = 0$ is optimal for $\zeta = 0$. Therefore, we can summarize all of our findings as follows:

Choice of the ζ hyper-parameter is critical in establishing tradeoffs between the mean-squared-error component parallel to β and those orthogonal to β . Small numerical values of ζ emphasize reductions in mean-squared-error orthogonal to β ; large numerical values of ζ emphasize reduction in mean-squared-error parallel to β .
In the limit as ζ approaches zero, there is effectively no constraint on the amount of mean-squared-error allowed parallel to β, and the optimal SHRINKAGE TARGET is declared to be $\Delta = 0$. This strategy would be REALLY EASY to implement in actual practice, but it is clearly also a rather EXTREME strategy.

In the limit as ζ approaches plus infinity ($+\infty$), there is effectively no constraint on the amount of mean-squared-error allowed orthogonal to β, and the optimal SHRINKAGE TARGET is declared to be the $\Delta = k^{(=)}\Lambda$ of { 4.17 }. The resulting shrinkage $b^{\bigstar}$ will be parallel to the marginal inner products vector, $X^{T}y$. This strategy would require an estimate of $k^{(=)}$ to be derived from the data. (Alternatively, one might simply adjust the length of this $b^{\bigstar}$ to maximize its likelihood of being β , i.e. minimize the resulting residual sum-of-squares.) But this, too, would be a rather EXTREME strategy.

The choice $\zeta = 1$ establishes EQUILIBRIUM in the tradeoff between mean-squared-error components orthogonal to β and the mean-squared-error parallel to β . And ζ=1 corresponds to the weight matrix $W = I$ in { 4.23 }. As a result, the optimal SHRINKAGE TARGET values are the $\delta_{i}^{MSE} = \phi_{i}^{2} / (\phi_{j}^{2} + 1)$ of { 4.6 }. This strategy is relatively difficult to implement in actual practice because it could require estimates for all R of the canonical signal-to-noise ratios, i.e. the $\phi_{j}^{2} = \gamma_{i}^{2}\lambda_{i}/\sigma^{2}$ noncentrality parameters of { 2.24 }. But, of the three strategies discussed here, this also appears to be the MOST REASONABLE overall strategy.

4.1.6 Canonical Form for Optimal Shrinkage of a Single Fixed-Effect Estimator

Because the different components of a vector of regression coefficients can be of different numerical sizes and can have different variances, consider the possibility of rescaling each individual component [using the appropriate unknown, multiplicative constant] so that the variance of its least-squares estimate will equal one. Once placed in this canonical form, we will see that the rescaled size of each fixed-effect component plays a pivotal role.

The j-th uncorrelated component, γ_j , of β is placed in its canonical form by dividing it by its standard deviation, $\sigma \cdot \lambda_j^{-1/2}$. An additive – error model for this rescaled component is thus of the form:

$$\begin{matrix}\text{FIXED-EFFECT}\\ \text{ESTIMATE}\end{matrix} = \begin{matrix}\text{FIXED-EFFECT}\\ \text{SIGNAL}\end{matrix} + \begin{matrix}\text{STANDARDIZED}\\ \text{NOISE}\end{matrix} ,$$

where the standardized noise has mean zero and variance one. Note that $\phi_j = \gamma_j \cdot \lambda_j^{1/2} / \sigma$ is then both the size of the fixed-effect signal and the expected value of the fixed-effect estimate. In other words, the rescaled component has variance one and expected value $\phi_j = \gamma_j \cdot \lambda_j^{1/2} / \sigma$. Note also that the optimal extent of shrinkage for this canonical fixed-effect component is $\delta_i^{MSE} = \phi_i^2 / (\phi_j^2 + 1)$, as in { 4.6 }.

Canonical forms will be used at the end of Section §4.4 to display an exact analogy between the fixed-effect and random-effect formulations for optimal shrinkage. These canonical forms will also be used extensively in the risk simulations of Chapter §6.

4.2 Classical "Good" Shrinkage

Matrix-valued mean-squared-error risk criteria are much more in keeping with the primary theme of our book than are scalar-valued criteria. After all, we always prefer to stress our theme: Simultaneous estimation of 2 or more regression coefficients is nothing less than a full-blown problem in multivariate analysis.

Swindel and Chapman(1973) originally defined “good" ridge estimators only within the one-parameter ridge family of Hoerl and Kennard(1970a). But we will apply their definition to all generalized shrinkage estimators as follows:

> **GOOD SHRINKAGE ESTIMATORS:** A generalized shrinkage estimator, $b^{\bigstar} = G \Delta c$ of { 3.1 } with non-stochastic shrinkage factors Δ , will be said to be GOOD for a specified β , σ^2 pairing if and only if it dominates the least-squares estimator $b^o \equiv X^+ y = G c$ in EVERY mean-squared-error sense.

To explore mathematically equivalent formulations, consider the following...

The EXCESS Mean-Squared-Error of b^o relative to $b^\bigstar$ is defined to be simply the corresponding algebraic difference in $P \times P$ Mean-Squared-Error matrices (the least-squares variance matrix minus a shrinkage estimator MSE matrix.) This difference can be thought of as depending only upon the form and extent of shrinkage applied to $b^\bigstar$; in fact, we will commonly think of this difference as being primarily a function of the diagonal matrix of shrinkage factors, Δ . In any case, we will denote this difference matrix by

$$\text{EMSE}(b^\bigstar) \;=\; \text{MSE}(b^o) \;-\; \text{MSE}(b^\bigstar) \;=\; G\, \text{EMSE}(\Delta c)\, G^T , \qquad \{ 4.24 \}$$

where G is the $P \times R$ semi-orthogonal matrix of principal axis direction cosines for the centered regressors matrix, X , of equation { 2.8 } and

$$\text{EMSE}(\Delta c) \;=\; \sigma^2 (I \;-\; \Delta^2) \Lambda^{-1} \;-\; (I - \Delta) \gamma \gamma^T (I - \Delta) . \qquad \{ 4.25 \}$$

Three equivalent conditions that assure that the generalized shrinkage estimator $b^\bigstar = G \Delta c$ with non-stochastic shrinkage factors Δ is GOOD for a specified β , σ^2 pairing can now be stated as follows:

(i) $\text{EMSE}(b^\bigstar)$ is a positive definite matrix, or { 4.26 }

(ii) $\text{MSE}(\alpha^T b^o) > \text{MSE}(\alpha^T b^\bigstar)$ for every unit vector α , or { 4.27 }

(iii) $\text{wmse}(b^o, W) > \text{wmse}(b^\bigstar, W)$ for every { 4.28 }

nonzero, non-negative definite weight matrix, W. The equivalence of { 4.26 } and { 4.27 } follows immediately from the very definition of a positive-definite matrix. Theobald(1974), Theorem 1, claimed it was sufficient to consider only positive definite matrices, W, in { 4.28 }. But, in a 1979 letter to me, Masashi Okamoto showed that the stronger condition that all nonzero, non-negative definite weight matrices be considered is necessary to imply equivalence with the other two definitions.

The highly specialized form of the EMSE matrix of { 4.25 } [namely a diagonal matrix minus a symmetric, rank-one matrix], turns out to imply some key results about GOOD shrinkage estimators. We will need the following lemma from Obenchain(1978)...

> **OBENCHAIN'S LEMMA:** If $D = \text{Diag}(d_1, d_2, ..., d_p)$ is a $p \times p$ positive-definite diagonal matrix and $z^T = (z_1, z_2, ..., z_p)$ is a row vector with $p > 2$ elements, then the length of the z-vector is critical in determining whether or not matrices of the general form $A \;=\; D (I \;-\; z z^T) D$ are positive definite. Specifically,
>
> (i) A will be positive definite iff $z^T z < 1$,

(ii) A will be non-negative definite of rank $(p-1)$ iff $z^T z = 1$, and

(iii) A will have $(p-1)$ positive eigenvalues and 1 negative eigenvalue iff $z^T z > 1$.

Furthermore, the eigenvector, τ, of A corresponding to a eigenvalue, λ, is of the general form:

(a) $\tau =$ (i-th column of the identity matrix) and $\lambda = d_i^2$ if $z_i = 0$ for some i, and

(b) $\tau \propto (D^2 - \lambda \cdot I)^{-1} D z$ if $\lambda \neq d_i^2$ for any $i = 1, 2, ..., p$.

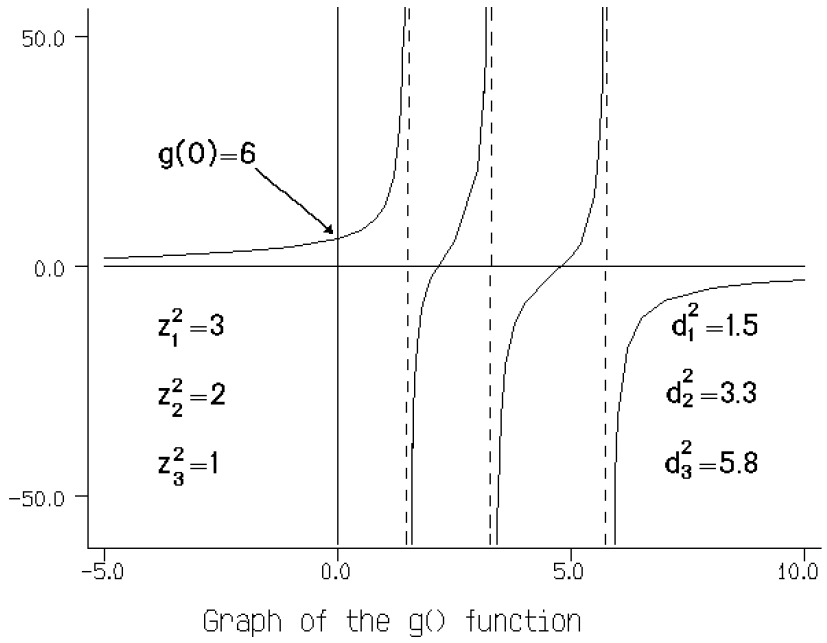
Since $\tau \neq 0$ is to be an eigenvector of A with eigenvalue λ, we know that $A\tau = D^2\tau - D z$ $z^T D \tau$ must be of the special form $\lambda \cdot \tau$. We can rewrite this condition as $(D^2 - \lambda \cdot I) \tau = k \cdot D z$ where k is the scalar $k = z^T D \tau$. The problem of finding a complete set of eigenvectors and eigenvalues for A becomes simple [as in case (a), above] when $z = 0$, so suppose that $z \neq 0$. Furthermore, if $\lambda \neq d_i^2 > 0$ for i=1, 2, ..., p, then $(D^2 - \lambda \cdot I)$ will be invertible, and τ will necessarily be of the special form $\tau = k \cdot (D^2 - \lambda \cdot I)^{-1} D z$ of case (b) above, at least when k is non-zero. But this means that $k = z^T D \tau = k \cdot z^T D (D^2 - \lambda \cdot I)^{-1} D z = k \cdot g(\lambda)$, where g denotes the scalar valued function

$$g(\lambda) = \sum_{i=1}^{p} z_i^2 d_i^2 (d_i^2 - \lambda)^{-1}. \qquad \{ 4.29 \}$$

It follows that each eigenvalue of A must either coincide with one of the positive d_i^2 values or else be a solution of $g(\lambda) = 1$, which excludes the possibility that $k = z^T D \tau = 0$ in case (b).

Letting d_{MIN}^2 denote the smallest numerical value of d_i^2 for which a corresponding $z_i^2 > 0$, it is clear that $g(\lambda) \geq 0$ and is strictly increasing on $-\infty < \lambda < d_{MIN}^2$ and, furthermore, that $g(\lambda) = 1$ has exactly one solution in this interval. After all, if this minimal solution to $g(\lambda) = 1$ is denoted by λ_{MIN} , then it cannot be a multiple eigenvalue of A because the eigenspace of $(D^2 - \lambda_{MIN} \cdot I)^{-1} D z$ is clearly of rank one. It follows that the numerical value of g() at $\lambda = 0$, namely $g(0) = z^T z$, is critical in determining whether the smallest eigenvalue of A is negative, zero, or positive. Specifically, $z^T z < 1$ implies that λ_{MIN} is strictly positive because $g(\lambda)$ has not yet reached the critical numerical value of 1 at $\lambda = 0$. Similarly, $z^T z = 1$ implies $\lambda_{MIN} = 0$, and $z^T z > 1$ implies $\lambda_{MIN} < 0$, as was to be shown.

Figure 4.1: g(λ) Function Numerical Example



In applying **OBENCHAIN'S LEMMA** to identify "good" shrinkage estimators, the following function will play the role of g(λ) of { 4.29 }.

RIDGE FUNCTION: The following scalar valued function of the generalized shrinkage factors, Δ, is called the Ridge Function:

$$\text{RF}(\Delta) = \sum_{j=1}^{R} \phi_j^2 \cdot (1 - \delta_j)/(1 + \delta_j), \qquad \{ 4.30 \}$$

$$= \sum_{j=1}^{R} [\delta_j^{MSE} \cdot (1 - \delta_j)] / [(1 - \delta_j^{MSE}) \cdot (1+\delta_j)],$$

where $\phi_j^2 = \gamma_j^2 \lambda_j / \sigma^2$ is, again, the unknown noncentrality of the F-ratio for the hypothesis that $\gamma_j = 0$ [i.e. the hypothesis that the j-th true component of β is zero.]

We can now state a theorem which is a mild generalization of the main result of Obenchain(1978)...

RIDGE FUNCTION THEOREM: If the parameters β and σ^2 of a classical, fixed effects linear model are such that $\beta^T\beta < \infty$ and $0 < \sigma^2 < \infty$, the given matrix of centered regressor coordinates X is of rank $R \geq 1$, and the generalized shrinkage factors Δ are non-stochastic on the range $0 \leq \delta_i < 1$ for i = 1, 2, ..., R, then

(i) the (R-1) largest eigenvalues of EMSE($b^{\bigstar}$) will always be positive,

(ii) the smallest eigenvalue of EMSE($b^{\bigstar}$) will also be positive iff $RF(\Delta) < 1$,

(iii) the smallest eigenvalue of EMSE($b^{\bigstar}$) will be zero iff $RF(\Delta) = 1$, and

(iv) the smallest eigenvalue of EMSE($b^{\bigstar}$) will be negative iff $RF(\Delta) > 1$. In this case, the eigenvector corresponding to the negative eigenvalue, ξ_R, has elements of the general form:

$$\alpha_i \propto \sum_{j=1}^{R} [g_{ij} \cdot (1 - \delta_j) \cdot \gamma_j] / [\, \sigma^2 \lambda_j (1 - \delta_j^2) + |\xi_R| \,] , \qquad \{ 4.31 \}$$

for i = 1, 2, ..., P, which defines the **INFERIOR DIRECTION** of P-dimensional space along which MSE($b^{\bigstar}$) exceeds MSE(b^{o}).

It will only be necessary to show that EMSE($b^{\bigstar}$) can be rewritten in a form to which OBENCHAIN'S LEMMA applies. But EMSE($b^{\bigstar}$) of { 4.24 } is $G\,A\,G^T$ where A of { 4.25 } is of the desired form with $D = (I - \Delta^2)^{1/2} \Lambda^{-1/2} \sigma$ and $z = \Lambda^{1/2} (I - \Delta^2)^{-1/2} (I - \Delta) \gamma / \sigma$. Note, in particular, that RF(Δ) of { 4.30 } is then $z^T z$, and that the α vector of { 4.31 } is necessarily of the form $G \tau$ for an eigenvector of A specified by case (b) of OBENCHAIN'S LEMMA because its eigenvalue is negative.

COMMENTS ON GOOD SHRINKAGE ESTIMATORS:

The numerical value attained by $RF(\Delta)$ is critical in determining the matrix mean-squared-error characteristics of generalized shrinkage estimators, $b^{\bigstar}$. Specifically, $RF(\Delta) < 1$ implies that, if the corresponding $b^{\bigstar}$ differs from b^{o}, this $b^{\bigstar}$ is “good" in the sense that it dominates b^{o} in ALL mean-squared-error senses.

On the other hand, if the ridge function EXCEEDS one, then there is at most one direction in P-dimensional space along which $b^{\bigstar}$ has larger mean squared error than does b^{o}.

In the one-parameter “ordinary ridge" family of Hoerl and Kennard(1970a), the shrinkage factors are restricted to be of the form $\delta_i = \lambda_i / (\lambda_i + k)$ for i = 1, 2, ..., p where k is a non-negative scalar. In this case, the ridge function is of the special form $RF(k) = \sum \phi_i^2 / (1 + 2\lambda_i k^{-1})$ and the main results of Swindel and Chapman(1973) follow as a special case of part (ii) of the ridge function theorem. Namely, every positive k value yields a good ordinary ridge estimator if $\sum \phi_i^2 < 1$; otherwise, the good range is $0 < k < 2 / |\eta_p|$ where η_p is the negative eigenvalue of $(X^T X)^{-1} - \beta\beta^T / \sigma^2$. As a result, the sufficient condition of Theobald(1974), Theorem 2, that $0 < k < 2\sigma^2 / \beta^T \beta$ tends to be much more stringent than is necessary.

THE (2/R)THS RULE-OF-THUMB: P, the number of (non-constant) predictor variables in our regression equation, is an upper bound for R = rank(X). Obenchain(1978) described a "(2/P)ths" guideline for GOOD shrinkage under the assumption that R = P. When R is less that P, my original quideline is more accurately described as a "(2/R)ths rule." Consider the problem of limiting shrinkage along each of the R principal regressor axes so that each of the R terms in { 4.30 } will not exceed 1/R. This is certainly one way of guaranteeing that the ridge function will not exceed one. This sufficient condition for "good"ness implies, for axis j, that

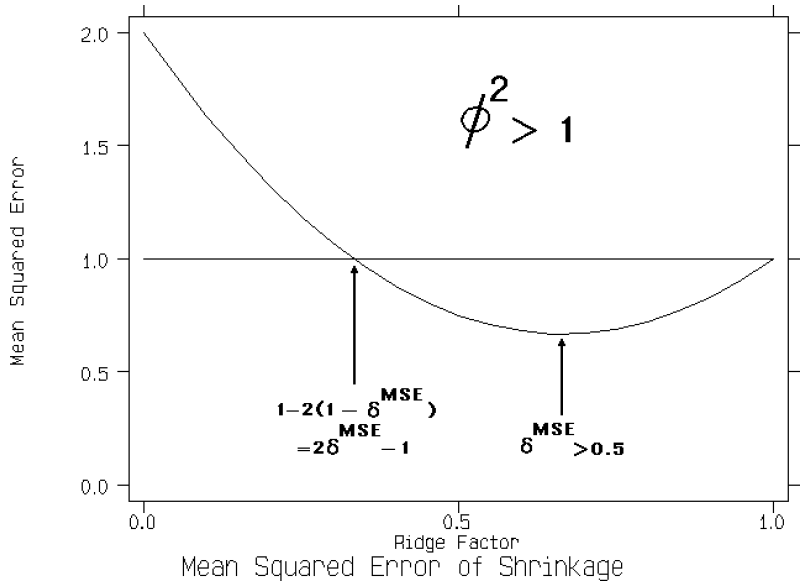
$$\delta_j^{MSE} \cdot (1 - \delta_j) \;\leq\; (1 - \delta_j^{MSE}) \cdot (1+\delta_j) / R$$

or, equivalently,

$$\delta_i \;\geq\; 1 \;-\; 2\,(1 - \delta_i^{MSE}) \cdot [\; 1 + (R - 1)\delta_i^{MSE} \;]^{-1}.$$

A set of sufficient conditions that are somewhat weaker, at least when R > 1, can thus be written as $\delta_i \geq 1 - 2\,(1 - \delta_i^{MSE}) / R$ for i=1,...,R, which means that the "good" shrinkage range will always be AT LEAST (2/R)ths of the "optimal" shrinkage range.

Figure 4.2: "Good" Range for Phi-Squared Greater Than 1.



When the rank of X is R=1, $\delta_1 c_1$ will be a "good" estimator of $\beta_1=\gamma_1$ even for shrinkage extents as much as TWICE the "optimal" extent; the good range will be $\delta_1^{MIN} \leq \delta_1 < 1$, where $\delta_1^{MIN} = \max(\, 0,\, 2 \cdot \delta_1^{MSE} - 1\,)$, as shown in Figures 4.2-4.4.

Figure 4.3: "Good" Range when Phi-Squared Equals 1.

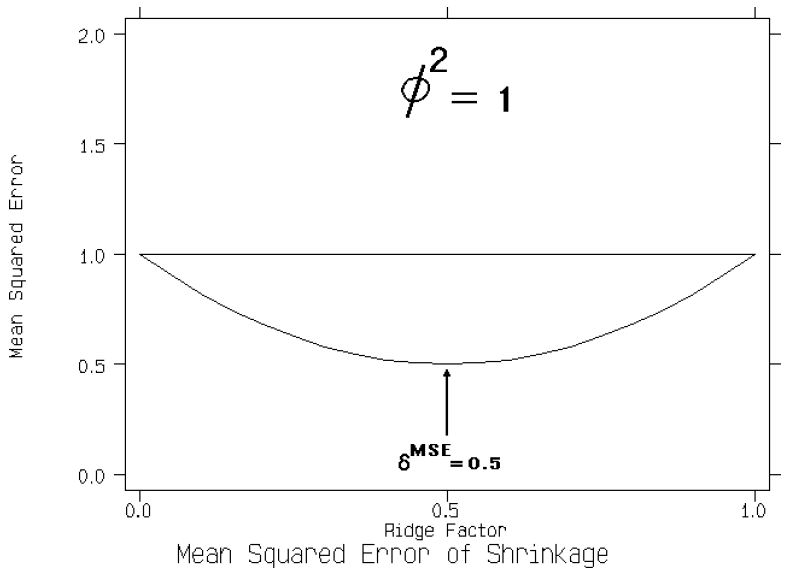
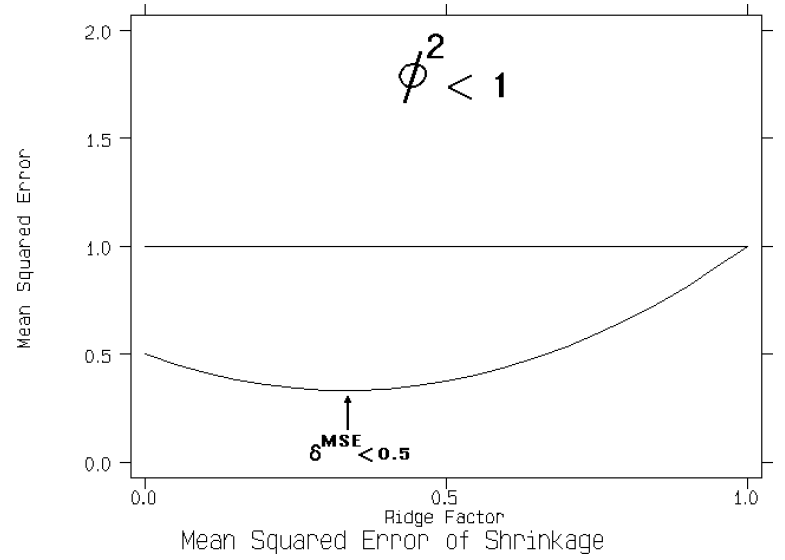


Figure 4.4: "Good" Range for Phi-Squared Less Than 1.



When the rank of X is two, RF(Δ^{MSE}) cannot exceed 1, and the good shrinkage range it thus at least $\delta_1^{MSE} \leq \delta_1 < 1$ and $\delta_2^{MSE} \leq \delta_2 < 1$.

When the rank of X exceeds two, RF(Δ^{MSE}) can exceed 1.

When the overall EXTENT of shrinkage is measured on the "multicollinearity allowance" scale, MCAL = $P - \delta_1 - \delta_2 - ... - \delta_R$, then the (2/R)ths Rule-of-Thumb for shrinkage can be quite simply stated as:

The "good" shrinkage range always includes AT LEAST (2/R)ths of the "optimal" range, namely...

$$0 \leq \text{MCAL} \leq 2 \cdot \text{MCAL}^{\text{MSE}} / R .$$

4.3 Classical "Ultimate" Shrinkage

Unlike the arguments we have explored so far here in the first two sections of Chapter 4, let us now specifically address the question: "What TARGET values should we use when shrinkage-factors are specifically recognized as being stochastic?"

After all, when we begin a new application of regression, we usually do not know in advance exactly how much of exactly which kind of shrinkage we might end up using. When our approach is "classical," the observed regressor values (the centered X coordinates) are assumed given, and we wish to make appropriate inferences about the conditional mean and variance of the distribution of responses, y , given those X realizations. When we "examine" the observed response values (using ridge trace displays and/or maximum likelihood calculations) to decide upon a form and extent for shrinkage, we end up using shrinkage factors that depend upon the observed y values. Therefore, the shrinkage factors we usually end up using are stochastic (given X .)

Our target values for stochastic shrinkage could still be the δ_i^{MSE} factors of { 4.6 }, of course, and I personally feel that the shrinkage target should be unchanged. However, let us first examine an alternative target.

What shrinkage factors would result from equating a generalized shrinkage estimator, $b^{\bigstar} = G \Delta c$, with the vector of true (fixed) coefficients, $\beta = G \gamma$? In other words, instead of merely reducing MSE risk (expected quadratic loss) of estimation, this choice would actually achieve zero loss! The equations that result state: $\Delta c = \gamma$. Therefore, whenever the i-th uncorrelated component of b^o is nonzero ($c_i \neq 0$), the corresponding ultimate choice for the i-th shrinkage factor would be:

$$\delta_i^{\text{ULT}} = \gamma_i / c_i . \qquad \{ 4.32 \}$$

Note that a strictly positive contribution to quadratic loss, $(\delta_i^{\text{ULT}} c_i - \gamma_i)^2$, can then occur only when an uncorrelated component of b^o happens to be zero, $c_i = 0$, while the corresponding true component is nonzero, $\gamma_i \neq 0$. Of course, the probability of observing an exactly null component, $c_i = 0$, is zero for any continuous (non-atomic) probability distribution of regression disturbance terms.

The formula for ultimate shrinkage, { 4.32 }, results in a stochastic choice for each ridge shrinkage factor. After all, the denominator c_i terms are assumed to be random given the observed regressor coordinates, X , so the δ_i^{ULT} factors are then also random.

The δ_i^{ULT} factors of { 4.32 } are clearly highly desirable target values. One might argue, perhaps, that the "natural" estimate of δ_i^{ULT} would be 0 when one believes that $\gamma_i = 0$ and 1 otherwise. On the other hand, an actual δ_i^{ULT} value could, of course, be negative! And no finite lower or upper bounds can be placed upon potential realizations of δ_i^{ULT} factors! In other words, the δ_i^{ULT} factors can represent sign-changes and/or expansions of the uncorrelated components of b^o rather than shrinkages. For example, it can be shown that the distribution of δ_i^{ULT} under normal-theory is bimodal with a less-likely, negative mode and a more-likely, positive mode at $2 / \sqrt{1 + (8/\phi_i^2)}$.

Vinod(1976) argued that the δ_i^{ULT} factors of equation { 4.32 } provide a "global minimum" in mean-squared-error of estimation. Actually, Vinod(1976) expressed results in terms of "additive eigenvalue inflation konstants," k_i . In this notation, shrinkage factors are of the form $\delta_i = \lambda_i / (\lambda_i + k_i)$, which is the generalization of equation { 3.6 } in which a potentially different amount is added to each eigenvalue of X^TX ; see Hoerl and Kennard(1970), sections §5 and §7. Using this notation, Vinod(1976) called

$$k_i^{MSE} = \sigma^2/\gamma_i^2 \qquad \{ 4.33 \}$$

"suboptimal" compared to

$$k_i^{ULT} = (c_i - \gamma_i) \cdot \lambda_i / \gamma_i \, . \qquad \{ 4.34 \}$$

In his response, Kennard(1976) said that { 4.34 } (or { 4.32 }) expresses a simple tautology: zero risk (or loss) can only be achieved when the γ_i components essentially have known values. In fact, Kennard points out that Vinod's "recommendation is just that of using the parameter value as its estimate."

In summary, equations { 4.32 } and { 4.34 } both essentially say... "Perform the exactly correct adjustment to c_i so that $\delta_i \cdot c_i$ will coincide exactly with γ_i ." Unfortunately, no advice on how to actually accomplish anything like this has been (or can be) given. After all, a stochastic target is (by its very definition) a constantly moving target!

> Actual reductions in RISK (over at least some parts of parameter space) can result from "aiming" at the fixed (but unknown) minimal MSE shrinkage target, { 4.6 } . Specific examples of this type are given in Chapter 6.
>
> No systematic reduction in LOSS has ever been demonstrated to be possible in any even remotely realistic situation.

4.4 Random Coefficient Shrinkage

While a general discussion of "mixed" linear models (models that contain both fixed and random regression coefficients) is best postponed until Chapter 7, we now discuss the generalization of some of the fixed-effect results of Section §4.1 to the case where β is random with expected value β_o and variance-covariance matrix Σ_β .

In the following discussion, it is quite important to remember that generalized shrinkage vector, $b^{\bigstar} = G\,\Delta\,c$, is still to be viewed as an estimator of the unknown, random β vector rather than as an estimator of the expectation vector, β_o. The uncorrelated components of the least squares estimator are still defined here to be $c = G^T b^o$, but the corresponding true components, γ, will now be random with expected value vector $E(\gamma) = \gamma_o \equiv G^T \beta_o$ and variance-covariance matrix $V(\gamma) = \Sigma_\gamma \equiv G^T \Sigma_\beta G$. The arguments leading to equation { 4.2 } of Section §4.1 still hold as long as the resulting mean-squared-error matrix is viewed as being conditional given a specific realization for γ :

$$MSE(\Delta c \mid \gamma) = \sigma^2 \Delta^2 \Lambda^{-1} + (I - \Delta)\,\gamma\,\gamma^T (I - \Delta). \qquad \{ 4.35 \}$$

The corresponding unconditional mean-square-error matrix is then of the general form

$$MSE(\Delta c) = \sigma^2 \Delta^2 \Lambda^{-1} + (I - \Delta)[\gamma_o \gamma_o^T + \Sigma_\gamma](I - \Delta), \qquad \{ 4.36 \}$$

in which the second term is no longer necessarily of rank one.

Two extreme, special cases of { 4.36 } will be of primary interest to us...

First of all, when $\Sigma_\beta = 0$ (so that $\beta \equiv \beta_o$ and $\gamma \equiv \gamma_o$) all random-effect risk measures revert to their somewhat more simple fixed-effect forms.

Secondly, the completely random coefficients case results when $\beta_o = 0$ and $\gamma_o = 0$.

The weighted mean-squared-error measure of equation { 4.11 } and the directional mean-squared-error measure of equation { 4.14 } take on the following forms when coefficients are random:

$$wmse(b^{\bigstar}, W) = \sigma^2 \cdot \text{trace}(M\Delta^2\Lambda^{-1}) + \gamma_o^T(I - \Delta)M(I - \Delta)\gamma_o + \text{trace}[\Sigma_\gamma(I - \Delta)M(I - \Delta)] \qquad \{ 4.37 \}$$

and

$$wmse(b^{\bigstar}, \alpha\alpha^T) = \sigma^2 \xi^T [\Delta^2\Lambda^{-1} + (I - \Delta)(\gamma_o \gamma_o^T + \Sigma_\gamma)(I - \Delta)]\xi, \qquad \{ 4.38 \}$$

where, again, $M = G^T W G$ and $\xi^T = \alpha^T G$. The corresponding generalizations of equations { 4.13 } and { 4.16 } for risk partial derivatives are straightforward, but closed form solutions like those of { 4.12 } and { 4.15 } for the shrinkage factors that minimize weighted or directional risks of random-coefficient shrinkage estimates are not obvious, except in certain special cases.

4.4.1 Shrinkage Risk of a Single Random Coefficient

Suppose that we start with an unbiased estimate, c , of an unknown, scalar-valued effect, γ, that has variance σ^2. In other words, given γ and σ^2, the conditional moments of c are E(c | γ, σ) = γ and V(c | γ, σ) = σ^2. [Note that, in the notation of section §4.1.1, the variance of the i-th uncorrelated component of the least-squares estimator would be written as σ^2/λ_i; here, that variance is simply being called σ^2.]

Next suppose that σ^2 has a fixed, unknown value while γ is random. Specifically, suppose that the expected value of γ is γ_o and its variance is σ_γ^2 :

$$E(\gamma) = \gamma_o \quad \text{and} \quad V(\gamma) = \sigma_\gamma^2 . \qquad \{ 4.39 \}$$

With δ denoting a known, non-stochastic shrinkage factor value, what are the mean-squared-error properties of $\delta \cdot c$ as an estimator of γ ? We again stress that we are viewing $\delta \cdot c$ as an estimator of γ itself, which is random when $\sigma_\gamma^2 > 0$, rather than as an estimator of γ_o, the fixed-effect, expected value of γ. The mean-squared-error of interest is thus

$$\begin{aligned} MSE(\delta \cdot c) &= E[(\delta \cdot c - \gamma)^2] \\ &= \delta^2 \cdot E(c^2) + E(\gamma^2) - 2 \cdot \delta \cdot E(c \cdot \gamma) \\ &= \delta^2 \cdot E(c^2) + (1 - 2 \cdot \delta) \cdot E(\gamma^2) \\ &= \delta^2 \cdot [\gamma_o^2 + \sigma_\gamma^2 + \sigma^2] + (1 - 2 \cdot \delta) \cdot [\gamma_o^2 + \sigma_\gamma^2] \\ &= \delta^2 \cdot \sigma^2 + (1 - \delta)^2 \cdot [\gamma_o^2 + \sigma_\gamma^2] . \qquad \{ 4.40 \} \end{aligned}$$

Now MSE($\delta \cdot c$) of { 4.40 } clearly changes as the δ-factor changes. In fact, the partial derivative of MSE($\delta \cdot c$) with respect to δ is

$$\partial MSE(\delta \cdot c) / \partial \delta = 2 \cdot \sigma^2 \cdot \delta - 2 \cdot (1 - \delta) \cdot [\gamma_o^2 + \sigma_\gamma^2] , \qquad \{ 4.41 \}$$

while the second partial derivative is a non-negative constant...

$$\partial^2 MSE(\delta \cdot c) / \partial \delta^2 = 2 \cdot [\sigma^2 + \gamma_o^2 + \sigma_\gamma^2] . \qquad \{ 4.42 \}$$

Equation { 4.42 } implies that equating $\partial MSE(\delta \cdot c) / \partial \delta_i$ of { 4.41 } to zero will yield a MINIMUM value for MSE($\delta \cdot c$) as long as $\sigma^2 > 0$ or $\gamma_o^2 > 0$ or $\sigma_\gamma^2 > 0$. This optimal amount of shrinkage is

$$\begin{aligned} \delta^{MSE} &= (\gamma_o^2 + \sigma_\gamma^2) / (\gamma_o^2 + \sigma_\gamma^2 + \sigma^2) , \qquad \{ 4.43 \} \\ &= \phi^2 / (\phi^2 + 1) = (1 + \phi^{-2})^{-1}, \end{aligned}$$

where $\phi^2 = (\gamma_o^2 + \sigma_\gamma^2) / \sigma^2$.

Again, the extreme cases of { 4.43 } are of special interest to us...

The fixed-effect results of equations { 4.3 } through { 4.6 } correspond to the special case of { 4.40 } through { 4.43 } where $\sigma_\gamma^2 = 0$ (so that $\gamma \equiv \gamma_o$.) Here $\phi^2 = \gamma_o^2 / \sigma^2$ is the unknown noncentrality parameter of the F-statistic for testing $\gamma_o = 0$ of { 2.22 } and { 2.23 }.

The completely random coefficient case results when $\gamma_o \equiv 0$. In this special case, $\phi^2 = \sigma_\gamma^2 / \sigma^2$ is an unknown, true ratio of variances, while an F-statistic is the corresponding ratio of sample variances.

Of course, we may also find ourselves in an "intermediate" situation where both $\sigma_\gamma^2 > 0$ and $\gamma_o \neq 0$. But the risk simulations of Chapter §5 will at least pin-down the extremes.

4.4.2 Canonical Form for Optimal Shrinkage of a Single, Completely-Random Effect

By dividing each component of a random coefficient vector by its noise standard deviation, we can place random-coefficient estimation problems in a canonical form analogous to that of Section §4.1.6 for fixed-effect models. The resulting "relative" standard deviation, $\phi = \sigma_\gamma / \sigma$ then plays a pivotal role.

An additive – error model for a rescaled random-effect estimate would be:

$$\begin{matrix}\text{RANDOM-EFFECT} \\ \text{ESTIMATE}\end{matrix} = \begin{matrix}\text{RANDOM-EFFECT} \\ \text{SIGNAL}\end{matrix} + \begin{matrix}\text{STANDARDIZED} \\ \text{NOISE}\end{matrix} ,$$

where the standardized noise has mean zero and variance one. Note that the random-effect signal has mean zero and variance ϕ^2 , while the random-effect estimate has mean zero and variance $\phi^2 + 1$. Note also that the optimal extent of shrinkage for this canonical random-effect would be $\delta_i^{MSE} = \phi^2 / (\phi^2 + 1)$, as in { 4.43 }.

4.5 Summary

In this chapter, we have used a wide variety of rather technical arguments to address an extremely important practical issue, that of selecting TARGET VALUES for shrinkage in regression models. The first-time reader may well ask "What do all of those theorems and special cases have to say about what to do in general, shrinkage regression practice?" Here are my personal opinions...

The δ^{MSE} generalized shrinkage factors, defined as in either { 4.6 } or { 4.43 }, seem to be the target values that make the most sense from the widest selection of alternative points-of-view. These factors establish optimal variance-bias tradeoffs that minimize a wide variety of univariate measures of mean-squared-error.

Once one adopts a truly multivariate (maxtix-valued risk) point-of-view, one still probably wishes to investigate shrinkage path shapes that lead generally "toward" (if not exactly "through") Δ^{MSE} on their way from $\Delta = I$ to $\Delta = 0$. However, a cautious practitioner might well wish to stop well short of Δ^{MSE} as his/her conservative choice for an extent of shrinkage. Objective shrinkage practitioners may ultimately find that the most important concepts introduced in this chapter are: (i) the "inferior direction" associated with excessive shrinkage, equation { 4.31 }, [as well as its associated excess-MSE eigenvalue spectrum] and (ii) the "(2/R)ths Rule-of-Thumb" given at the end of section §4.2.

In both our fixed-effect and random-coefficient formulations, shrinkage results from multiplying an unbiased estimator by a non stochastic factor, δ, on the range $0 \leq \delta \leq 1$. Bias is introduced when $\delta < 1$, but the corresponding variance is thereby reduced by a multiplicative factor of δ^2. A lower bond on the MSE risk (variance plus squared-bias) associated with this shrinkage results from multiplying the variance of the unbiased estimator by δ ; potential minimum risk decreases linearly with δ, { 4.7 }. On the other hand, this lower limit on risk is actually achieved only when applying the "right" extent of shrinkage, $\delta = \delta^{MSE}$ of { 4.6 } or { 4.43 }.

There is an exact analogy between the fixed-effect and completely-random-effect formulations for optimal shrinkage. The ϕ parameter takes the form of either a standardized fixed-effect when $\sigma_\gamma = 0$,

$$\phi = \gamma / \sigma = \text{(expected signal) / (standard deviation of additive noise)},$$

or a ratio of standard deviations when $E(\gamma) = 0$,

$$\phi = \sigma_\gamma / \sigma = \text{(standard deviation of signal) / (standard deviation of additive noise)}.$$

In fact, the optimal shrinkage target (in both extreme and all intermediate cases) is always of the general form $\delta^{MSE} = \phi^2 / (\phi^2 + 1)$ for $\phi^2 = (\gamma^2 + \sigma_\gamma^2) / \sigma^2$.

References for Chapter Four

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

Kennard, R. W. (1976). Letter to the Editor. **Technometrics**, 18, 504-505.

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

Okamoto, M. (1979). Personal communication.

Rao, C. R. (1973). **Linear Statistical Inference and Its Applications, Second Edition**. New York: John Wiley and Sons.

Swindel, B. F. and Chapman, D. D. (1973). "Good ridge estimators." Abstracts Booklet, New York Joint Statistical Meetings, page 126.

Theobald, C. M. (1974). "Generalizations of mean square error applied to ridge regression." **Journal Royal Statistical Society B**, 36, 103-105.

Vinod, H. D. (1976). Letter to the Editor. **Technometrics**, 18, 504.

Further Reading for Chapter Four

Farebrother, R. W. (1975). "The minimum mean square error linear estimator and ridge regression." **Technometrics**, 17, 127-128.

Farebrother, R. W. (1976). "Further results on the mean squared error of ridge regression." **Journal Royal Statistical Society**, B, 38, 248-250.

Farebrother, R. W. (1978). "A class of shrinkage estimators." **Journal Royal Statistical Society**, B, 40, 47-49.

Kawai, N. and Okamoto, M. (1979). "A generalization of the ridge function theorem." **Math. Japonica** 24, No.2, 175-178. [Abstract 79t-123. **Institute of Mathematical Statistics Bulletin** 8, No. 4.]

Trenkler, G. and Trenkler, D. (1981). "Estimable functions and reduction of mean squared error." **Methods of Operations Research** 44, 225-234. Oelgeschlager, Gunn & Hain, Cambridge, Mass.