More on Regularization and (Generalized) Ridge Operators

Yoshio Takane,[1]
(1) Department of Psychology, McGill University, 1205 Dr. Penfield Ave., Montreal, QC, H3A 1B1 Canada

Abstract

Regularization is a useful technique for supplementing insufficient data by prior knowledge. In the ridge type of regularization, prior knowledge comes as a belief that parameters in statistical models can never be too far away from zero, and the usual estimates of the parameters are shrunk toward zero. This tends to produce estimates which are on average closer to their population values. In this paper, the effect of shrinkage estimation was first demonstrated, starting from the simplest case of estimating a mean and working up to more complicated cases. The notion of (generalized) ridge operators (GRO) was then introduced, and their properties systematically investigated.

1. Introduction

Regularization is a useful technique as a way of supplementing insufficient data by prior knowledge and/or incorporating certain desirable properties (e.g., smoothness) in the estimates of model parameters (Takane and Hwang, 2006). In the ridge type of regularization, prior knowledge takes the form of a conviction that parameters in statistical models can never be too far away from zero, and consequently their estimates should be shrunk toward zero. This tends to produce estimates of parameters which are on average closer to the true population values (Hoerl and Kennard, 1970).

For quite some time now, Takane and his collaborators (Takane and Hwang, 2006, 2007; Takane, Hwang, and Abdi, 2006; Takane and Jung 2006, 2007; Takane and Yanai, 2006) have been working on a project incorporating the ridge regularization into a variety of multivariate analysis (MVA) techniques. These techniques include multiple-set canonical correlation analysis (GCANO), redundancy analysis (RA), etc., each of which in turn subsumes a number of representative techniques of MVA as its special cases. In this paper, we first demonstrate the effect of shrinkage estimation, starting from the simplest case of estimating a mean and working up to more complicated cases of MVA. We then introduce the notion of (generalized) ridge operators (GRO) and systematically investigate their mathematical properties. For a fixed value of ridge parameter, this class of operators are linear and characterized as "contractions", represented by matrices whose eigenvalues are all between 0 and 1 inclusive, and generalize the notion of projectors. We show how the ridge regularization methods developed earlier for GCANO and RA can be regarded as special cases of this class of operations.

2. Simple Demonstrations of the Effect of Regularization

We begin with perhaps the simplest possible demonstration of the effect of ridge regularization. Consider estimating a mean based on n observations from a population with mean μ and variance σ^2. Note that no assumptions are made about the shape of the distribution. An estimate of μ that immediately comes to our mind is the sample mean. Let us call this estimation method "Method 1". In Method 2, we shrink the sample mean by a factor of a. We evaluate the goodness of estimators by mean square error (MSE), the expected value of the squared

discrepancy between an estimator and the population value. Let $\hat{\mu}$ represents an estimator of μ. Then,

$$\mathrm{MSE} = \mathrm{E}[(\mu - \hat{\mu})^2], \tag{1}$$

where E indicates an expectation operation. MSE can be split into two parts, squared bias and variance, i.e.,

$$\begin{array}{rccc} \mathrm{MSE} = & [\mu - \mathrm{E}(\hat{\mu})]^2 & + & \mathrm{E}[(\hat{\mu} - \mathrm{E}(\hat{\mu}))^2]. \\ & \text{(squared bias)} & & \text{(variance)} \end{array} \tag{2}$$

The sample mean (Method 1) has no bias, but has the variance of σ^2/n, so that its MSE is equal to:

$$\mathrm{MSE}_1 = \sigma^2/n. \tag{3}$$

The shrinkage estimator (Method 2), on the other hand, has some bias since its expected value is $a\mu$, but the variance is smaller by a factor of a^2, so that its MSE is given by

$$\mathrm{MSE}_2 = \mu^2(1-a)^2 + a^2\sigma^2/n. \tag{4}$$

A question is for what values of a the shrinkage estimator gives a smaller MSE than the sample mean.

This of course depends on the values of n, μ and σ^2. We tentatively assumed $n = 10$, $\mu^2 = 4$, and $\sigma^2 = 9$. Figure 1 shows MSE functions for the two estimation methods. MSE_1 stays the same at .9 no matter what the value of a is, while MSE_2 is a quadratic function of a that passes through and crosses with MSE_1 at point (1, .9). A little calculation shows that this curve also crosses with MSE_1 at (.63, .9) and takes a minimum value of .73 at $a = .82$. This indicates that for values of a between .63 and 1, the shrinkage estimator yields a smaller value of MSE than the sample mean. (This range becomes wider or narrower depending on n, μ and σ^2. In general, it gets narrower as n and μ gets larger, and it gets wider as σ^2 gets larger.) This means that the shrinkage estimator is on average closer to μ, if an appropriate value of a is known.

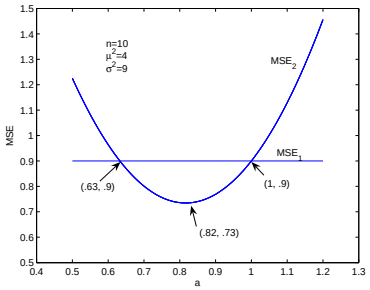

Fig. 1 MSE as a function of a for the two estimators.

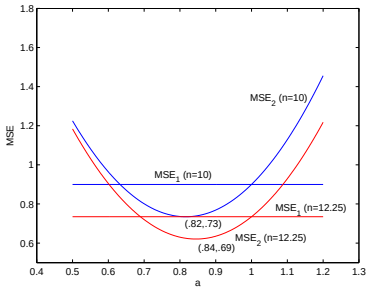

Fig. 2 MSE reduction and sample size increase.

We also calculated how many more observations are needed to achieve the level of MSE (.73) achieved by the optimal shrinkage factor of .82 without shrinking. This turned out to be 12.25. This means that more than 20% more observations

are necessary, indicating that this many observations can be saved by shrinking. Of course, if we have this many observations and optimally shrink, we can achieve even a smaller MSE (.69). This is shown in Figure 2.

Figure 3 breaks down MSE_2 into squared bias and variance as functions of a. In this figure, however, the shrinkage effect gets larger as a gets smaller. This is opposite to the conventional ridge regression situation, where a larger value of the shrinkage factor (ridge parameter λ) yields a larger shrinkage effect. We thus transformed a into λ by $\lambda = n(1/a - 1)$, and redrew the graph in Figure 4. The squared bias gets consistently larger as λ gets larger, while the variance gets consistently smaller. The sum of the two, MSE, takes a minimum value at about $\lambda = 2.2$.

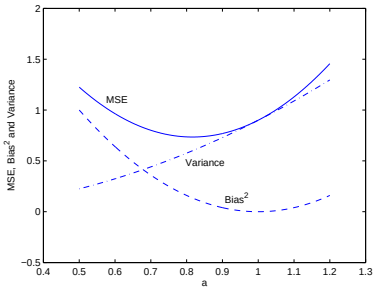

Fig. 3 A breakdown of MSE_2 into bias^2 and variance as a function of a.

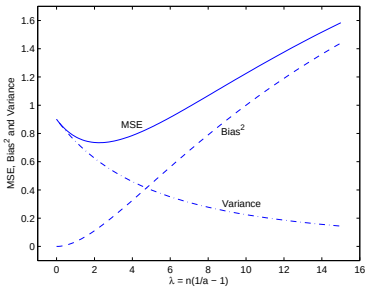

Fig. 4 The same as Figure 3, but as a function of $\lambda = n(1/a - 1)$.

Figure 4 is strikingly similar to the result obtained by Hoerl and Kennard (1970, Figure 1), who for the first time in history demonstrated the effect of the ridge type of shrinkage estimation in regression analysis. Our case is in fact a special case of theirs. As is well known, calculating a sample mean is equivalent to applying a regression analysis with a single constant predictor variable. So there is nothing surprising in our result. An important thing is that this effect is apparently very robust and can be observed in many other situations. We have conducted many simulation studies in the contexts of other MVA techniques (Takane and Hwang, 2006, 2007; Takane, Hwang, and Abdi, 2006; Takane and Jung, 2006, 2007), and have repeatedly found essentially the same results. In these simulations, we generate a number of data sets according to a population model, estimate parameters by the regularized estimation method, and calculate MSE's, squared bias, and variances as functions of the ridge parameter λ and the sample size. Figures 5 and 6 depict these functions for multiple-set CANO (GCANO) and redundancy analysis (RA), respectively. These are just two examples (although each of these techniques subsumes a number of representative techniques of MVA as its special cases). In all cases we have tried so far, we have observed essentially the same results.

3. (Generalized) Ridge Operators

Given that the ridge regularization is useful, there are two important topics that remain to be addressed. One is how to implement the regularized estimation into various MVA techniques, and the other is to investigate some mathematical properties of this type of estimation. In this paper, we focus on the second topic, assuming that an "optimal" value of the ridge parameter λ is already known. A number of strategies exist for an optimal choice of λ including cross validation. See,

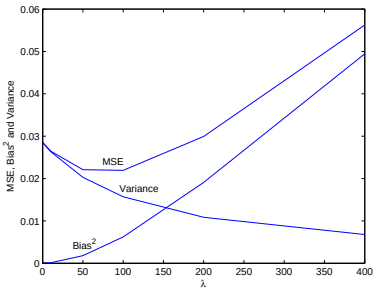

Fig. 5 MSE, bias^2, and variance as a function of λ and n for GCANO.

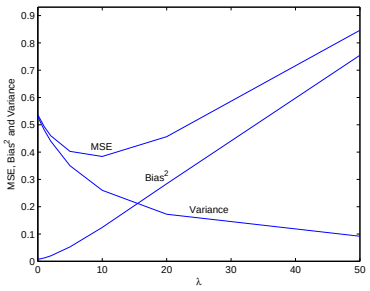

Fig. 6 The same as Figure 5, but for RA.

for example, Gruber (1998) for other possibilities. Solutions to the implementation problem follow from the general theory we present.

3.1. Projectors

We begin with a brief review of projection operators (or predictors for short), which are closely related to ridge operators, and which play an important role in least squares (LS) estimation. Let

$$f(B) = \mathrm{SS}(Y - XB) = \mathrm{tr}[(Y - XB)'(Y - XB)], \tag{5}$$

be the LS criterion, where Y, X $(n \times p)$, and B are matrices of criterion variables, predictor variables, and regression coefficients, respectively, in multivariate multiple regression analysis. We estimate B in such a way as to minimize (5), which leads to the following LS estimate:

$$\hat{B} = (X'X)^{-}X'Y, \tag{6}$$

where $^{-}$ indicates a generalized inverse (g-inverse). The matrix of predictions is obtained by

$$\hat{Y} = X\hat{B} = X(X'X)^{-}X'Y = P_X Y, \tag{7}$$

where $P_X = X(X'X)^{-}X'$ is called a projector, the orthogonal projector onto the column space of X (denoted as $\mathrm{Sp}(X)$). This matrix is invariant over the choice of a g-inverse $(X'X)^{-}$, symmetric, and idempotent (i.e., $P_X^2 = P_X$).

The unweighted LS criterion (5) can readily be extended to the weighted LS criterion

$$f^{(W)}(B) = \mathrm{SS}(Y - XB)_W = \mathrm{tr}[(Y - XB)'W(Y - XB)], \tag{8}$$

where W is *nnd* (nonnegative-definite) and such that $\mathrm{rank}(WX) = \mathrm{rank}(X)$. Minimizing this criterion leads to the weighted LS (WLS) estimate of B given by

$$\hat{B} = (X'WX)^{-}X'WY. \tag{9}$$

The matrix of predictions is then obtained by

$$\hat{Y} = X(X'WX)^{-}X'WY = P_{X/W}Y, \tag{10}$$

where

$$P_{X/W} = X(X'WX)^{-}X'W \tag{11}$$

is called a W-orthogonal projector, the projector onto $\mathrm{Sp}(X)$ along $\mathrm{Ker}(X'W)$, where Ker indicates a null space. This projector has similar properties to P_X (invariant over the choice of a g-inverse $(X'WX)^{-}$, symmetric with respect to W, and idempotent).

3.2. Ridge operators (RO)

We define a ridge LS (RLS) criterion by

$$f_\lambda(B) = \mathrm{SS}(Y - XB) + \lambda \mathrm{SS}(B)_{P_{X'}}, \tag{12}$$

where $\lambda \geq 0$ is the ridge parameter, $\mathrm{SS}(B)_{P_{X'}} = \mathrm{tr}(B'P_{X'}B)$, and $P_{X'} = X'(XX')^{-}X$ is the orthogonal projector onto $\mathrm{Sp}(X')$, the row space of X. We minimize (12) to obtain an RLS estimate of B

$$\tilde{B} = (X'X + \lambda P_{X'})^{-}X'Y, \tag{13}$$

which leads to the matrix of predictions given by

$$\tilde{Y} = X\tilde{B} = X(X'X + \lambda P_{X'})^{-}X'Y = R_X(\lambda)Y, \tag{14}$$

where

$$R_X(\lambda) = X(X'X + \lambda P_{X'})^{-}X' \tag{15}$$

is called a ridge operator. This matrix is again invariant over the choice of a g-inverse $(X'X + \lambda P_{X'})^{-}$, symmetric, but generally not idempotent. Note that $P_{X'}$ reduces to an identity matrix, when X is columnwise nonsingular. However, $(X'X + \lambda I)^{-1}$ is often used for $(X'X + \lambda P_{X'})^{-}$ even when X is not columnwise nonsingular. This can be justified by the fact that $(X'X + \lambda I)^{-1}$ is a kind of g-inverse of $X'X + \lambda P_{X'}$. (We write $(X'X + \lambda I)^{-1} \subset \{(X'X + \lambda P_{X'})^{-}\}$.) This choice of a g-inverse affects an estimate of B, but not the matrix of predictions.

Let $X = UDV'$ represent the SVD (singular value decomposition) of X, where U is the matrix of left singular vectors such that $U'U = I_r$, V is the matrix of right singular vectors such that $V'V = I_r$, D is the positive-definite (*pd*) diagonal matrix of order $r = \mathrm{rank}(X)$. Then, $R_X(\lambda)$ can be expressed as

$$R_X(\lambda) = UD^2(D^2 + \lambda I_r)^{-1}U'. \tag{16}$$

This is called a canonical form representation and is extremely useful in characterizing a number of interesting properties of $R_X(\lambda)$. For example, it is immediately clear from the above representation that $R_X(\lambda)$ is a contraction matrix, whose eigenvalues are all between 0 and 1 inclusive, is semi-simple (i.e., $\mathrm{rank}(R_X(\lambda)^2) = \mathrm{rank}(R_X(\lambda))$, and is consequently diagonalizable by a similarity transformation (U). (This situation may be contrasted with that of P_X, where $P_X = UU'$; it is also a contraction matrix with all the eigenvalues either 0 or 1, again semi-simple, and diagonalizable by U.)

An alternative way of characterizing a contraction matrix is:

$$R_X(\lambda) - R_X(\lambda)^2 = \lambda X(X'X + \lambda P_{X'})^{+2}X' \geq 0, \tag{17}$$

where $(X'X + \lambda P_{X'})^+$ indicates a Moore-Penrose g-inverse, $(X'X + \lambda P_{X'})^{+2}$ its square, and ≥ 0 indicates that the matrix on the left hand side is *nnd*. The contraction matrix generalizes the notion of projectors, which satisfy the lower limit of the above inequality. Let $S_X(\lambda) = I - R_X(\lambda)$, a "complement" of $R_X(\lambda)$. Then,

$$R_X(\lambda) - R_X(\lambda)^2 = R_X(\lambda)S_X(\lambda) = S_X(\lambda)R_X(\lambda) = S_X(\lambda) - S_X(\lambda)^2 \geq 0. \quad (18)$$

3.3. The ridge metric matrix

We define a ridge metric matrix, which plays a crucial role in the development to follow. Let

$$M_X(\lambda) = J_n + \lambda(XX')^+, \quad (19)$$

where J_n is any matrix such that $X'J_nX = X'X$ (e.g., $J_n = I_n$, $J_n = P_X$, etc.). Note that $(XX')^+$ has the following expression:

$$(XX')^+ = X(X'X)^{+2}X'. \quad (20)$$

Using the ridge metric matrix defined above, we can rewrite $X'X + \lambda P_{X'}$ as

$$X'X + \lambda P_{X'} = X'M_X(\lambda)X, \quad (21)$$

so that $R_X(\lambda)$ can be rewritten as

$$R_X(\lambda) = X(X'M_X(\lambda)X)^-X'. \quad (22)$$

The above expression of $R_X(\lambda)$ is interesting for two reasons. One is

$$R_X(\lambda)M_X(\lambda)R_X(\lambda) = R_X(\lambda). \quad (23)$$

That is, although $R_X(\lambda)$ itself is not idempotent, it is indeed idempotent with respect to the metric matrix $M_X(\lambda)$. Another way of looking at the above identity is that $M_X(\lambda) \subset \{(R_X(\lambda)^-\}$. In fact, it can easily be verified that $M_X(\lambda)$ is the Moore-Penrose g-inverse of $R_X(\lambda)$, which also implies that the latter is the Moore-Penrose g-inverse of $M_X(\lambda)$.

The other is that $R_X(\lambda)M_X(\lambda)$ is an $M_X(\lambda)$-orthogonal projector under the condition that $\text{rank}(M_X(\lambda)X) = \text{rank}(X)$. More specifically, it is the projector onto $\text{Sp}(X)$ along $\text{Ker}(X'M_X(\lambda))$. However, this is not all. It is also the usual orthogonal projector onto $\text{Sp}(X)$ (i.e., $R_X(\lambda)M_X(\lambda) = P_X$). This is because $\text{Sp}(M_X(\lambda)X) = \text{Sp}(X)$ (i.e., premultiplying X by $M_X(\lambda)$ does not change the column space of X). Define

$$N_X(\lambda) = J_p + \lambda(X'X)^+, \quad (24)$$

where J_p is any matrix such that $XJ_p = X$ (e.g., $J_p = I_p$, $J_p = P_{X'}$, etc.). Then,

$$M_X(\lambda)X = XN_X(\lambda). \quad (25)$$

This implies $\text{Sp}(M_X(\lambda)X) = \text{Sp}(XN_X(\lambda)) \subset \text{Sp}(X)$, but because $\text{rank}(XN_X(\lambda))$ $= \text{rank}(X)$, we have $\text{Sp}(XN_X(\lambda)) = \text{Sp}(X)$, which implies $\text{Sp}(M_X(\lambda)X) = \text{Sp}(X)$. This in turn implies $\text{Ker}(X'M_X(\lambda)) = \text{Ker}(X')$. It also shows that $\text{rank}(M_X(\lambda)X) = \text{rank}(X)$, the condition required for $R_X(\lambda)M_X(\lambda)$ to be a projector.

3.4. When X is partitioned into K disjoint subsets

So far we've been treating X as a single matrix. What happens if it is partitioned into K disjoint subsets? Let $X = [X_1, \ldots, X_K]$ be an n by p row block matrix, where the X_k (n by p_k and $k = 1, \ldots, K$), are assumed to satisfy the disjointness condition

$$\mathrm{rank}(X) = \sum_{k=1}^{K} \mathrm{rank}(X_k). \tag{26}$$

Then,

$$X_k' M(\lambda) X_j = \begin{cases} X_k' X_k + \lambda P_{X_k'} & (k = j), \\ X_k' X_j & (k \neq j), \end{cases} \tag{27}$$

where $P_{X_k'} = X_k'(X_k X_k')^- X_k$ is the orthogonal projector onto $\mathrm{Sp}(X_k')$. (Note that $P_{X_k'}$ reduces to I_{p_k} if X_k is of full column rank.)

This follows from Theorem 1.2 of Anderson and Styan (1982), which states that $A_k A^- A_k = A_k$ and $A_k A^- A_j = 0$ for $k \neq j$ if and only if $\mathrm{rank}(A) = \sum_{k=1}^{K} \mathrm{rank}(A_k)$, where $A = \sum_{k=1}^{K} A_k$. By setting $A_k = X_k X_k'$ in this theorem, we obtain $A = XX' = \sum_{k=1}^{K} X_k X_k' = \sum_{k=1}^{K} A_k$, so that $\mathrm{rank}(A) = \sum_{k=1}^{K} \mathrm{rank}(A_k)$ is equivalent to $\mathrm{rank}(X) = \sum_{k=1}^{K} \mathrm{rank}(X_k)$. It also holds that

$$X_k X_k'(XX')^- X_j X_j' = \begin{cases} X_k X_k' & (k = j), \\ 0 & (k \neq j). \end{cases} \tag{28}$$

By pre- and postmultiplying (28) by $(X_k' X_k)^+ X_k'$ and $X_j(X_j' X_j)^+$, respectively, we obtain

$$X_k'(XX')^- X_j = \begin{cases} P_{X_k'} & (k = j), \\ 0 & (k \neq j), \end{cases} \tag{29}$$

It is interesting to observe that $M_X(\lambda)$ has no effect if k and j are distinct, but it adds an extra term $\lambda P_{X_k'}$ if k and j coincide. When the X_k are not disjoint (do not satisfy (26)), we have $P_{X_k'} \geq X_k'(XX')^- X_k$ in general (Yanai and Mayekawa, 1988), where $A \geq B$ means $A - B \geq 0$.

3.5. When $K = 2$

We focus on the special case of the above in which $K = 2$, and derive decompositions of $R_X(\lambda)$ analogous to the well-known decompositions of projectors (Rao and Yanai, 1979; Takane and Yanai, 1999; Yanai, 1990). We still assume that X_1 and X_2 are disjoint. Let $X = [X_1, X_2]$, and let $R_{X_1}(\lambda)$ and $R_{X_2}(\lambda)$ be as defined analogously to (15). Then,

$$R_X(\lambda) = R_{X_1}(\lambda) + R_{X_2}(\lambda), \tag{30}$$

if and only if $X_1' M(\lambda) X_2 = X_1' X_2 = 0$. By a standard decomposition of a projector, we have $R_X(\lambda) M_X(\lambda) = R_{X_1}(\lambda) M_X(\lambda) + R_{X_2}(\lambda) M_X(\lambda)$. We can

then get rid of $M_X(\lambda)$ at the end of each term by postmultiplying both sides by $M_X(\lambda)^+ = R_X(\lambda)$.

When X_1 and X_2 are not mutually orthogonal, we first orthogonalize them and then apply the above decomposition. We have

$$R_X(\lambda) = R_{X_1}(\lambda) + R_{S_{X_1}(\lambda)X_2}(\lambda) = R_{X_2}(\lambda) + R_{S_{X_2}(\lambda)X_1}(\lambda), \tag{31}$$

where $R_{S_{X_1}(\lambda)X_2}(\lambda) = S_{X_1}(\lambda)X_2(X_2'S_{X_1}(\lambda)X_2 + \lambda R_{X_2'})^- X_2'S_{X_1}(\lambda)$ and $R_{S_{X_2}(\lambda)X_1}(\lambda)$ is analogously defined. Note that X_1 and $S_{X_1}(\lambda)X_2$ are orthogonal with respect to $M_X(\lambda)$ (We say that X_1 and $S_{X_1}(\lambda)X_2$ are $M_X(\lambda)$-orthogonal.), and so are X_2 and $S_{X_2}(\lambda)X_1$. This decomposition is useful when we fit X_1 first and then X_2 to residuals from X_1, or vice versa.

We may also split $\mathrm{Sp}(X)$ into two $M_X(\lambda)$-orthogonal subspaces and obtain

$$R_X(\lambda) = R_{XT}(\lambda) + R_{XH}(\lambda), \tag{32}$$

where H is such that $\mathrm{Sp}(H) = \mathrm{Ker}(T'X'M(\lambda)X)$ for a given T, or T is such that $\mathrm{Sp}(T) = \mathrm{Ker}(H'X'M(\lambda)X)$ for a given H. It is clear that XT and XH are $M_X(\lambda)$-orthogonal. This decomposition is useful when we have an additional restriction on B in the form of $B = TB^*$ for some B^* and for a given constraint matrix T, thereby splitting the effect of X into the effects of XT and its $M_X(\lambda)$-orthogonal component XH.

We may combine (31) and (32) to obtain more complicated decompositions, e.g.,

$$\begin{aligned} R_X(\lambda) = {} & R_{R_{X_2}(\lambda)X_1T}(\lambda) + \\ & R_{R_{X_2}(\lambda)X_1A}(\lambda) + R_{S_{X_2}(\lambda)X_1T}(\lambda) + R_{S_{X_2}(\lambda)X_1G}(\lambda) + R_{X_2C}(\lambda), \end{aligned} \tag{33}$$

where A, G, and C are such that $\mathrm{Sp}(A) = \mathrm{Ker}(T'X_1'R_{X_2}(\lambda)M_X(\lambda)X_1)$, $\mathrm{Sp}(G) = \mathrm{Ker}(T'X_1'S_{X_2}(\lambda)M_X(\lambda)X_1)$, and $\mathrm{Sp}(C) = \mathrm{Ker}(X_1'X_2)$.

3.6. Generalized ridge operators (GRO)

The ridge operator and the ridge metric matrix can be generalized by replacing $P_{X'}$ in (12) by L, and by introducing a weight matrix W as in (8). Let L be any *nnd* matrix such that $\mathrm{Sp}(L) = \mathrm{Sp}(X')$, and let W be any *nnd* matrix such that $\mathrm{rank}(WX) = \mathrm{rank}(X)$. Let

$$f_\lambda^{(W,L)}(B) = \mathrm{SS}(Y - XB)_W + \lambda \mathrm{SS}(B)_L \tag{34}$$

be the weighted ridge LS (WRLS) criterion. Then, a generalized ridge operator (GRO) is obtained by

$$R_X^{(W,L)}(\lambda) = X(X'WX + \lambda L)^- X'W. \tag{35}$$

Mitra (1975) called the $(X'WX+\lambda L)^- X'W$ part of this matrix an optimal inverse. This matrix has similar properties as $R_X(\lambda)$. This generalization is useful when we need a regularization term more complicated than $P_{X'}$. Such cases arise, for example, when we wish to incorporate certain degrees of smoothness in the estimated curves by way of regularization (Adachi, 2002; Ramsay and Silverman, 2006). Define the generalized ridge metric matrix by

$$M_X^{(W,L)}(\lambda) = J_n + \lambda(XL^-X'W)^+_{W,W}, \tag{36}$$

where $(XL^-X'W)^+_{W,W} = X(X'WX)^-L(X'WX)^-X'W$ is the weighted Moore-Penrose g-inverse of $XL^-X'W$ with respect to the weight matrices W and W. (The symmetry conditions among the four Penrose conditions are satisfied only with respect to these weight matrices.) Then, $X'WX + \lambda L = X'WM_X^{(W,L)}(\lambda)X$, and the GRO can be rewritten as

$$R_X^{(W,L)}(\lambda) = X(X'WM_X^{(W,L)}(\lambda)X)^-X'W. \tag{37}$$

Note that $M_X^{(W,L)}(\lambda)$ is itself not symmetric, but it is always used in the form of $WM_X^{(W,L)}(\lambda)$, which is always symmetric.

A canonical form representation of the GRO is useful in characterizing various properties of the operators. Let $X = UDV'$ represent the generalized SVD (GSVD) of X with respect to metric matrices W and L^- (which we write as $\mathrm{GSVD}(X)_{W,L^-}$), where U and V are such that $U'WU = I_r$, $V'L^-V = I_r$, and D is a *pd* diagonal matrix of order r. Then,

$$R_X^{(W,L)}(\lambda) = UD^2(D^2 + \lambda I_r)^{-1}U'W. \tag{38}$$

3.7. WRLS in terms of WLS

The WRLS criterion can be reformulated in the form of a WLS by redefining Y, X, and W as follows (e.g., Ramsay and Silverman, 2006, section 5.2.7). Let

$$\tilde{Y} = \begin{bmatrix} Y \\ 0 \end{bmatrix}, \quad \tilde{X} = \begin{bmatrix} X \\ (\lambda L)^{1/2} \end{bmatrix}, \quad \text{and} \quad \tilde{W} = \begin{bmatrix} W & 0 \\ 0 & I \end{bmatrix}.$$

(This is analogous to partitioning $\tilde{X}$ into two blocks rowwise as opposed to columnwise as we have done previously.) Then, (34) can be rewritten as

$$f_\lambda^{(W,L)}(B) = \mathrm{SS}(\tilde{Y} - \tilde{X}B)_{\tilde{W}}. \tag{39}$$

This leads to a partitioned $\tilde{W}$-orthogonal projector

$$P_{\tilde{X}/\tilde{W}} = \begin{bmatrix} R_X^{(W,L)}(\lambda) & A \\ A'W & C \end{bmatrix}, \tag{40}$$

where $A = (\lambda)^{1/2}X(X'WX + \lambda L)^-L^{1/2}$, and $C = \lambda L^{1/2}(X'WX + \lambda L)^-L^{1/2}$. From the idempotency of $P_{\tilde{X}/\tilde{W}}$, it follows that

$$R_X^{(W,L)}(\lambda) - (R_X^{(W,L)}(\lambda))^2 = AA'W \geq 0, \tag{41}$$

$$C - C^2 = A'WA \geq 0, \tag{42}$$

and

$$R_X^{(W,L)}(\lambda)A + AC = A. \tag{43}$$

(41) generalizes (17).

4. Concluding remarks

Much has been done on the utility of the ridge type of regularization in regression analysis (Gruber, 1998), but not so much in other contexts until recently. This situation is rapidly changing with the effect of regularization amply demonstrated in other techniques of MVA, and there is now a solid mathematical foundation behind the operation.

5. Acknowledgement

This paper is dedicated to Professor Haruo Yanai for his friendship and guidance in the past forty years. The work reported in this paper is supported by Grant A6294 from the Natural Sciences and Engineering Research Council of Canada. I would like to thank Sunho Jung for his careful proofreading.

References

Adachi, K. (2002). Homogeneity and smoothness analysis for quantifying a longitudinal categorical variable. In S. Nishisato, Y. Baba, H. Bozdogan, & K. Kanefuji (Eds.), *Measurement and multivariate analysis*. Tokyo: Springer, pp. 47–56.

Anderson, T. W., & Styan, G. P. H. (1982). Cochran's theorem, rank additivity and tripotent matrices. In G. Kallianpur, P. R. Krishnaiah, & J. K. Ghosh (Eds.), *Statistics and probability: Essays in honor of C. R. Rao*. Amsterdam: North Holland, pp. 1–23.

Gruber, M. H. J. (1998). *Improving efficiency by shrinkage, the James-Stein and ridge regression estimators*. New York: Marcel Dekker.

Hoerl, A. F., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*, 55–67.

Mitra, S. K. (1975). Optimal inverse of a matrix. *Sankhyā, Series A, 37*, 550–563.

Ramsay, J. O. & Silverman, B. (2006). *Functional data analysis*, Second Edition. New York: Springer.

Rao, C. R., & Yanai, H. (1979). General definition and decomposition of projectors and some applications to statistical problems. *Journal of Statistical Planning and Inference, 3*, 1–17.

Takane, Y., & Hwang, H. (2006). Regularized multiple correspondence analysis. In M. J. Greenacre, & J. Blasius (Eds.), *Multiple correspondence analysis and related methods*. London: Chapman and Hall, pp. 259–279.

Takane, Y., & Hwang, H. (2007). Regularized linear and kernel redundancy analysis. *Computational Statistics and Data Analysis, 52*, 394–405.

Takane, Y., Hwang, H., & Abdi, H. (2006). Regularized multiple-set canonical correlation analysis. Submitted for publication.

Takane, Y., & Jung, S. (2006). Regularized partial and/or constrained redundancy analysis. Submitted for publication.

Takane, Y., & Jung, S. (2007). Regularized nonsymmetric correspondence analysis. Submitted for publication.

Takane, Y., & Yanai, H. (1999). On oblique projectors. *Linear Algebra and Its Applications, 289*, 297–310.

Takane, Y., & Yanai, H. (2006). On ridge operators. Submitted for publication.

Yanai, H. (1990). Some generalized forms of least squares g-inverse, minimum norm g-inverse, and Moore-Penrose inverse matrices. *Computational Statistics and Data Analysis, 10*, 251–260.

Yanai, H. & Mayekawa, S. (1988). Some extensions of inequalities concerning diagonal elements of orthogonal projectors and conditions for equalities. *Japanese Journal of Applied Statistics, 17*, 131–138 (in Japanese).