Regularized Linear and Kernel Redundancy Analysis

Yoshio Takane *, Heungsun Hwang

McGill University, Department of Psychology, 1205 Dr. Penfield Avenue, Montreal, QC, H3A 1B1, Canada

Abstract

Redundancy analysis (RA) is a versatile technique used to predict multivariate criterion variables from multivariate predictor variables. The reduced-rank feature of RA captures redundant information in the criterion variables in a most parsimonious way. A ridge type of regularization was introduced in RA to deal with the multicollinearity problem among the predictor variables. The regularized linear RA was extended to nonlinear RA using a kernel method to enhance the predictability. The usefulness of the proposed procedures was demonstrated by a Monte Carlo study and through the analysis of two real data sets.

Key words: Ridge regression, Reduced rank approximation, Generalized singular value decomposition (GSVD), Kernel methods, Gaussian kernel, Permutation tests, J-fold cross validation, Bootstrap method

1 Introduction

Redundancy analysis (RA; van den Wollenberg, 1967) is often used to investigate the directional relationship between two sets of variables. Given multivariate criterion and predictor variables, RA aims to find a subspace of a prescribed dimensionality in the space of predictor variables that is best predictive of the criterion variables as a whole. RA is a popular multivariate data analysis technique in many scientific disciplines, especially in biology and ecology (Legendre and Legendre, 1998; ter Braak, 1987), psychology (e.g., van der Leeden, 1990), and econometrics (Reinsel and Velu, 1998). A variant of

* Corresponding author (tel: 514-398-6125; fax: 514-398-4896).
Email addresses: takane@psych.mcgill.ca (Yoshio Takane), heungsun.hwang@mcgill.ca (Heungsun Hwang).

RA is also known as reduced rank regression (Anderson, 1951) and principal components of instrumental variables (Rao, 1964).

Canonical correlation analysis (CANO) also analyzes the relationships between two sets of multivariate data. In CANO, however, the two sets of variables are treated symmetrically with no distinction between criterion and predictor variables. Furthermore, a high canonical correlation between two sets of variables does not necessarily imply that the entire sets of variables are highly correlated. RA, on the other hand, attempts to explain as much variation in the criterion variables as possible by as few components of predictor variables as possible. RA thus reflects the relationships between two sets of variables more faithfully (Lambert, et al., 1988).

In this paper we incorporate a ridge type of regularization into RA. This is a natural extension of ridge regression originally developed for univariate regression problems. Let y, X, and b represent a vector of observations on the criterion variable, a matrix of predictor variables, and a vector of regression coefficients, respectively. In ridge regression (Hoerl and Kennard, 1970), the estimate of b (which we call the ridge least squares (RLS) estimate) is obtained by

$$\tilde{b} = (X'X + \lambda I)^{-1} X'y, \tag{1}$$

where $\lambda\,(\geq 0)$ is called the ridge parameter. The ridge parameter has the effect of shrinking the usual least squares (LS) estimate of b toward 0. It is known (Hoerl and Kennard, 1970) that for a certain range of values of λ, the RLS estimate is, on average, closer to the true population value than the LS estimate. This effect is more pronounced when the sample size is small and/or X is ill-conditioned due to high collinearity among the predictor variables. (See Groß (2003) for an up-to-date account of ridge regression.) Note that the reduced rank aspect of RA does not correct for multicollinearity in X, and the regularization is still needed to remedy the situation. In this paper we extend the basic methodology of ridge regression to RA and illustrate its use. We also consider a nonlinear extension of RA by a kernel method.

The rest of this paper is organized as follows. In section 2.1 we discuss the model and the parameter estimation procedure for linear RA. We first briefly review the LS estimation, and then present a parallel development for the RLS estimation. We then extend linear RA to a nonlinear RA called kernel RA (section 2.2). This is followed by a cross validation method for choosing an "optimal" value of the ridge parameter, and permutation tests for identifying the best dimensionality in the solution (section 2.3). In section 3 we give empirical demonstrations of the usefulness of the proposed methods.

2 The Method

2.1 Linear Redundancy Analysis

Let Y be an n (cases) by p (variables) matrix of criterion variables, and let X be an n by q matrix of predictor variables. We assume that both X and Y are at least columnwise centered, and that they are further standardized if there is no common measurement scale across the variables. We write the model for RA as

$$Y = XB + E, \tag{2}$$

where the q by p matrix of regression coefficients B is subject to a rank restriction,

$$\text{rank}(B) = r \leq m = \text{rank}(X'Y) \leq \min(\text{rank}(X), \text{rank}(Y)), \tag{3}$$

and E is an n by p matrix of disturbance terms. Let

$$\phi(B) = \text{SS}(Y - XB) \tag{4}$$

be the least squares (LS) criterion, where $\text{SS}(A) = \text{tr}(A'A)$. We estimate B so as to minimize $\phi(B)$ subject to the rank restriction (3). To achieve this goal, we first rewrite $\phi(B)$ as (ten Berge, 1993; see also Takane and Shibayama, 1991):

$$\phi(B) = \text{SS}(Y - X\hat{B}) + \text{SS}(\hat{B} - B)_{X'X}, \tag{5}$$

where $\text{SS}(A)_M = \text{tr}(A'MA)$, and

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

is a LS estimate of B without rank restriction, where $(X'X)^-$ is a generalized inverse (g-inverse) of $X'X$. (While $\hat{B}$ in (6) is not unique if X is singular, the decomposition (5) is unique. To obtain a unique estimate of $\hat{B}$, we can use the Moore-Penrose inverse for $(X'X)^-$.) Since the first term on the right hand side of (5) is unrelated to B, the reduced rank estimate of B can be obtained by minimizing the second term. This can be done through the generalized singular value decomposition (GSVD) of $\hat{B}$ with row metric $X'X$ (Takane and Hunter, 2001; Takane and Shibayama, 1991), which is written as $\text{GSVD}(\hat{B})_{X'X,\,I}$.

Here GSVD refers to an SVD under nonidentity metrics (Cailliez and Pages, 1976; Greenacre, 1984). This should not be confused with the use of the same GSVD linear algebra terminology (e.g. Golub and Van Loan, 1989). The latter refers to a pair of matrix decompositions that in effect obtain the generalized eigenvalue decomposition of $A'A$ with respect $B'B$ without calculating these products. Further details can be found in Takane and Hunter (2001, p. 415) and Takane (2002).

Let

$$\hat{B} = U_B D_B V_B' \tag{7}$$

represent $\text{GSVD}(\hat{B})_{X'X,I}$, where U_B is the q by m matrix of generalized singular vectors (such that $U_B' X'X U_B = I_m$, where $m = \text{rank}(\hat{B})$), V_B is the p by m matrix of right singular vectors (such that $V_B' V_B = I_m$), and D_B is an order m positive-definite (pd) diagonal matrix of (generalized) singular values as its diagonal elements in descending order of magnitude. Then, the reduced rank estimate of B is obtained by retaining only the portions of U_B, V_B, and D_B pertaining to the $r(\leq m)$ dominant singular values. Specifically, let $\tilde{U}_B$ and $\tilde{V}_B$ denote the q by r and p by r matrices formed from the first r columns of U_B and V_B, respectively, and let $\tilde{D}_B$ denote the diagonal matrix of order r formed from the first r rows and r columns of D_B. Then, the reduced rank estimate of B is obtained by

$$\tilde{B} = \tilde{U}_B \tilde{D}_B \tilde{V}_B'. \tag{8}$$

Quantities typically given in the output from RA are obtained by simple manipulations of $\tilde{U}_B$, $\tilde{V}_B$, and $\tilde{D}_B$. The matrix of component weights (weights applied to X to derive redundancy components) is obtained by $W = n^{1/2}\tilde{U}_B$. The matrix of redundancy components is obtained by $F = XW = n^{1/2}X\tilde{U}_B$. The structure matrix (correlations (or covariances) between X and F) is obtained by $n^{-1}X'F = n^{-1}X'XW = n^{-1/2}X'X\tilde{U}_B$. Finally, the cross loading matrix (correlations between Y and F) is obtained by $n^{-1}Y'F = n^{-1/2}\tilde{V}_B\tilde{D}_B$.

We now extend the above method to the ridge LS (RLS) estimation. As in the LS case, the solution can be derived in closed form. Let

$$\phi_\lambda(B) = \text{SS}(Y - XB) + \lambda \text{SS}(B)_{P_{X'}} \tag{9}$$

denote the RLS criterion, where λ is called the ridge parameter, $P_{X'} = X'(XX')^- X$ is the orthogonal projector onto the row space of X, and $\text{SS}(B)_{P_{X'}} = \text{tr}(B'P_{X'}B) = \text{tr}(B'B) = \text{SS}(B)$. (Without loss of generality, we may assume B is in the row space of X. See the next subsection for details.) The ridge parameter typically takes a small positive value (Hoerl and Kennard,

1970), which we tentatively assume known. (A way to determine an optimal value of λ will be discussed later.) We minimize the RLS criterion under the rank restriction (3). To achieve this goal, we first rewrite $\phi_\lambda(B)$ as

$$\phi_\lambda(B) = \mathrm{SS}(Y)_{Q_X(\lambda)} + \mathrm{SS}(\hat{B}(\lambda) - B)_{X'X+\lambda P_{X'}}, \tag{10}$$

where

$$\hat{B}(\lambda) = (X'X + \lambda P_{X'})^- X'Y \tag{11}$$

is an RLS estimate of B without rank restriction, and

$$Q_X(\lambda) = I - P_X(\lambda), \tag{12}$$

where

$$P_X(\lambda) = X(X'X + \lambda P_{X'})^- X' \tag{13}$$

is called a ridge operator (Takane and Yanai, 2006). Since the first term of (10) is unrelated to B, $\phi_\lambda(B)$ can be minimized with respect to B by minimizing the second term, which is obtained by $\mathrm{GSVD}(\hat{B}(\lambda))_{X'X+\lambda P_{X'},I}$. The rest of the procedure remains essentially the same as in the LS estimation.

A few remarks are in order regarding the above procedure. First of all, decomposition (10) is analogous to decomposition (5). To see (10), we note that $\phi_\lambda(B) = \mathrm{tr}(Y'Y - Y'X(X'X + \lambda P_{X'})^- X'Y + Y'X(X'X + \lambda P_{X'})^- X'Y - 2Y'XB + B'(X'X + \lambda P_{X'})B) = \mathrm{tr}(Y'Y - 2Y'XB + B'X'XB + \lambda B'B)$, which is equal to the expression obtained by expanding $\phi_\lambda(B)$ in (9). Note, however, that $Q_X(\lambda)$ is in general not idempotent, so that the first term in decomposition (10) cannot be written as $\mathrm{SS}(Q_X(\lambda)Y) = \mathrm{SS}(Y - X\hat{B}(\lambda))$, unlike the first term in decomposition (5). Secondly, the RLS estimate of B given in (11) is not unique, unless $X'X$ is nonsingular. However, the matrix of predictions $X\hat{B}(\lambda)$ is unique, and so are the ridge operators. Finally, $X'X + \lambda P_{X'}$ reduces to $X'X + \lambda I_q$ when $X'X$ is nonsingular. Often, however, $(X'X + \lambda I_q)^{-1}$ is used for $(X'X + \lambda P_{X'})^-$ even when $X'X$ is singular. This can be justified by noting that $(X'X + \lambda I_q)^{-1}$ is a g-inverse of $X'X + \lambda P_{X'}$ (Takane and Yanai, 2006).

2.2 Kernel Redundancy Analysis

In linear RA, the space of predictor variables does not go beyond the range space of X, whether we use the LS or the RLS estimation. In this section

we develop a method of RA that expands the prediction space by nonlinear transformations of X. Using a "kernel" trick (Schölkopf et al., 1997), this can be done without explicitly specifying the nonlinear transformations to be applied to X.

We again take model (2) as a point of departure. As alluded to earlier, we may assume without loss of generality that B is in the row space of X, that is, $\mathrm{Sp}(B) \subset \mathrm{Sp}(X')$, where $\mathrm{Sp}(X')$ indicates the range space of X'. (This can be readily seen as follows: Let $B = B_1 + B_2$, where $\mathrm{Sp}(B_1) \subset \mathrm{Sp}(X')$ and $\mathrm{Sp}(B_2) \subset \mathrm{Ker}(X)$, where $\mathrm{Ker}(X)$ indicates the null space of X. Then, $XB = XB_1 + XB_2 = XB_1$, so that we can reset $B = B_1$ without affecting the matrix of predictions.) This implies that B can be rewritten as $B = X'G$ for some n by p matrix G. The model (2) can then be rewritten as

$$Y = XX'G + E. \tag{14}$$

Matrix XX' is called a kernel matrix. It is a special kind of kernel matrix called the covariance kernel. More generally, a kernel matrix is a kind of similarity matrix among cases in the data set on the predictor variables. Note that (14) is merely a restatement of (2), and it is no more interesting than the original model (2). This form of the model will become interesting as we consider nonlinear transformations of the predictor variables, say, by $H(X)$. The kernel matrix then becomes $H(X)H(X)'$. The gist of kernel methods is that we do not explicitly define $H(X)$ (which presupposes exact knowledge of the nonlinear transformations to be applied to X), but instead we directly define a kernel matrix K (n by n, and non-negative definite (nnd)), which could have followed from a certain desirable $H(X)$.

We tentatively assume that K is known. (We will discuss a way to define K shortly.) Then, model (14) can be recast in the form of:

$$Y = KG + E. \tag{15}$$

This model, however, can easily be abused, since for any nonsingular matrix K, we obtain $\hat{G} = K^{-1}Y$, so that $K\hat{G} = KK^{-1}Y = Y$. This means that observed Y can always be perfectly predicted. In data analysis, however, the predictability for future observations is more important. To achieve this goal, we introduce the ridge type of regularization. Let

$$\psi_\lambda(G) = \mathrm{SS}(Y - KG) + \lambda \mathrm{SS}(G)_K \tag{16}$$

be the ridge LS criterion for kernel RA. Then, the RLS estimate of G that minimizes this criterion is given by

$$\hat{G} = (K + \lambda I)^{-1}Y. \tag{17}$$

The matrix of predictions is obtained by $K\hat{G} = K(K + \lambda I)^{-1}Y$. An optimal value of λ may be determined by cross validation in a manner similar to the linear case. The reduced rank estimate of G is obtained by $\text{GSVD}(\hat{G})_{K^2+\lambda K, I}$.

There are a number of possible ways of defining the kernel matrix K (Schölkopf et al., 1997). We use the Gaussian kernel because it is easy to calculate, it is guaranteed to be positive-definite (*pd*), and it is known to work for a wide range of problems. It is defined as

$$k_{ij} = \exp(-d_{ij}^2/\sigma), \tag{18}$$

where k_{ij} is the ij^{th} element of K, $d_{ij}^2 = (x_i - x_j)'(x_i - x_j)$ is the squared Euclidean distance between the i^{th} row (x_i') and the j^{th} row (x_j') of X, and σ is a scaling factor. The negative exponential function turns the squared distance into similarity. We then double center K to take into account the fact that Y is centered. This will make K singular. However, $K + \lambda I$ is usually nonsingular for $\lambda > 0$, and we may use the Moore-Penrose inverse of K when $\lambda = 0$. The prediction of a new case with predictor vector $x^{(n)\prime}$ is obtained by

$$y^{(n)\prime} = k^{(n)\prime}\tilde{G}, \tag{19}$$

where $k^{(n)\prime}$ is the kernel vector indicating the similarity between $x^{(n)\prime}$ and the rows of X calculated in the same way as for the kernel matrix K defined in (18), and $\tilde{G}$ is the reduced rank estimate of G.

The scale factor σ modulates the speed with which similarity should decay as a function of d_{ij}^2. An optimal value of σ can be chosen in much the same way as the value of λ is chosen. We systematically vary this value and choose the "best" one by monitoring the prediction error. (See the next subsection).

For pure prediction purposes kernel RA is often superior to linear RA. However, as a method for "understanding" the data, kernel RA (or kernel methods in general) has one potential drawback. It is often extremely difficult to identify the nature of the redundancy components obtained by kernel RA. Matrix G, which is an n by m matrix of regression coefficients, is usually not directly interpretable, although some attempts have been made to develop some techniques to facilitate the interpretation (Schölkopf et al., 1997). It is often helpful to correlate redundancy components obtained by kernel RA with the original predictor variables (X) and the criterion variables (Y) to understand the nature of the components. (See the food and cancer example in the empirical demonstrations section.)

There are areas in which interpretation is not as important as prediction. In engineering, for example, solving practical problems is often the most important

concern. Developing a machine for automatic discrimination of handwritten characters is a case in point. The machine has to be capable of discrimination whether or not we fully understand the nature of the discrimination. Here, prediction plays a primary role. Kernel methods have been popularized by engineers who constantly deal with practical problems (e.g., Herbrich, 2002; Suykens et al., 2002).

2.3 The Choice of Dimensionality, λ, and σ

There are three important choices to be made in application of regularized RA: the choice of dimensionality, the choice of an "optimal" value of the regularization parameter λ, and the choice of an "optimal" value of σ in kernel RA. We discuss these topics in turn.

We use permutation tests to choose the best dimensionality (the number of components, r) in the solution. In the permutation tests, rows of X are randomly permuted many times. RA is applied to each permuted X and the original Y repeatedly to obtain the null distribution of the largest singular value. If these singular values are smaller than the largest singular value obtained from the original X and Y $100(1-\alpha)\%$ of times, the first redundancy component is considered statistically significant at the α level. Alternatively, we may count the number of times this occurs and calculate the p-value. If the first component is significant, we eliminate the effect of the first component from X and apply the same procedure as above to test the significance of the second component, and so on. We continue this procedure until we find a nonsignificant component or reach the maximum possible number of components. See Legendre and Legendre (1998), ter Braak and Šmilauer (1998), and Takane and Hwang (2002) for more general discussions on the permutation tests in similar contexts. We may apply the above procedure with different values of λ in cases where the best dimensionality depends on the value of λ.

We use the J-fold cross validation method (Hastie, et al., 2001) to choose an optimal value of λ. A similar strategy can also be used to choose the value of σ in kernel RA. In this method, the data at hand are randomly divided into J subsets. One of the J sets is set aside as test samples, and model parameters are estimated from the remaining $J-1$ subsets (called calibration samples). These estimates are then applied to the test samples to estimate prediction errors. This is repeated J times with the test samples changed systematically, and the prediction errors accumulated over the J sets of test samples.

Let $Y^{(-j)}$ and $X^{(-j)}$ represent matrices of criterion variables and predictor variables, respectively, with the j^{th} subset of cases, $Y^{(j)}$ and $X^{(j)}$, removed from the original data sets. We obtain the RLS estimates of reduced rank

regression coefficients based on $Y^{(-j)}$ and $X^{(-j)}$, following the procedure described in section 2.1. Let $\tilde{B}^{(-j)}(\lambda)$ denote the matrix of estimated regression coefficients. We then evaluate

$$\epsilon(\lambda) = \sum_{j=1}^{J} \mathrm{SS}(Y^{(j)} - X^{(j)}\tilde{B}^{(-j)}(\lambda))/\mathrm{SS}(Y) \tag{20}$$

with the value of λ systematically varied. We choose the value of λ associated with the smallest value of $\epsilon(\lambda)$. Essentially the same procedure can be used for choosing an "optimal" value of σ in kernel RA (with $X^{(-j)}$ replaced by $K^{(-j)}$, $X^{(j)}$ replaced by $K^{(j)}$, and $\tilde{B}^{(-j)}(\lambda)$ replaced by $\tilde{G}^{(-j)}(\lambda)$).

One cautionary remark is in order, however, when the J-fold cross validation is used in situations where the space of criterion variables is totally contained in the predictor space. Such is always the case in kernel RA, and when $n \leq p$ in linear RA. If we take $J = n$, the J-fold cross validation reduces to the well-known leaving-one-out (LOO) method (or the Jackknife method). This particular choice of J should be avoided in such situations. Due to the double centering of the predictor set (X or K), the test sample in the LOO method is always perfectly predicted based on the remaining $n - 1$ cases when $\lambda = 0$. This inevitably leads to the conclusion that the non-regularized case always cross-validates best (which is usually not true). This may be seen from the fact that $y^{(j)\prime} = -1'_{n-1}Y^{(-j)}$, $x^{(j)\prime} = -1'_{n-1}X^{(-j)}$, and $\hat{B}^{(-j)} = (X^{(-j)\prime}X^{(-j)})^{+}X^{(-j)\prime}Y^{(-j)}$, where 1_{n-1} is the $(n-1)$-component vector of ones, so that

$$x^{(j)\prime}\hat{B}^{(-j)} = -1'_{n-1}X^{(-j)}(X^{(-j)\prime}X^{(-j)})^{+}X^{(-j)\prime}Y^{(-j)} = y^{(j)\prime},$$

since $X^{(-j)}(X^{(-j)\prime}X^{(-j)})^{+}X^{(-j)\prime} = I_{n-1}$. The case of kernel RA is similar.

A bootstrap method (Efron, 1982) is used to assess the reliability of parameter estimates. In this method, random samples of size n (equal to the size of the original data set) are repeatedly sampled from the original data set with replacement. Estimates of parameters are obtained for each bootstrap sample. We then calculate the means and the variances of the estimates across the bootstrap samples to estimate biases and standard errors of the estimates derived from the original data set. Significance tests of estimated coefficients may also be performed as a by-product of the bootstrap procedure. We simply count the number of times bootstrap estimates "cross" over zero (if the original estimate is positive, we count the number of times the bootstrap estimates turn out to be negative, and vice versa). If the relative frequency of cross-overs is less than a prescribed value of α, we conclude that the coefficient is significantly positive (or negative).

3 Empirical Demonstrations

In this section we provide empirical demonstrations of the usefulness of the proposed methods. The first study investigates the positive effect of regularization using a Monte Carlo technique. The second and third studies pertain to the analysis of real data sets.

3.1 A Monte Carlo Study

We first demonstrate the better quality of the RLS estimator using a Monte Carlo technique. The quality of an estimator can be measured by how close it is on average to its population counterpart. We may use the expected mean square error (MSE) for this purpose. MSE indicates the average squared Euclidean distance between population parameters and their estimators. Let θ and $\hat{\theta}$ represent vectors of generic population parameters and their estimators, respectively. Then,

$$\text{MSE} = \text{E}[\text{SS}(\theta - \hat{\theta})], \tag{21}$$

where E takes the expectation. The estimator associated with a smaller value of MSE is considered as a better estimator. The RLS estimator with a small positive value of λ is often associated with a smaller MSE than its LS counterpart. MSE in (21) can be decomposed into two parts:

$$\text{MSE} = \text{E}[\text{SS}(\theta - \text{E}(\hat{\theta}))] + \text{E}[\text{SS}(\hat{\theta} - \text{E}(\hat{\theta}))]. \tag{22}$$

The first term on the right hand side is the squared bias, and the second term is the variance of the estimator $\hat{\theta}$. While the LS estimator is an unbiased estimator (and consequently, squared bias = 0), it tends to have a large variance. The RLS estimator, on the other hand, is biased (*albeit* usually slightly), while it has a much smaller variance, resulting in a smaller MSE.

To confirm that the above expectation indeed holds for RA, a small Monte Carlo study was conducted. First, a population RA model was postulated, from which many replicated data sets of varying sample sizes ($N = 20, 50, 100,$ 200) were generated. RA was then applied to these data sets to derive the RLS estimates of regression coefficients with the value of λ systematically varied ($\lambda = 0, 1, 5, 10, 20, 50$). Average MSE, squared bias, and variance were calculated in reference to the assumed population values of regression parameters. In the assumed population model, the number of criterion variables was set to 3, that of predictor variables to 4, and each row of Y was generated according to $y_j' = x_j' B + e_j'$, where $x_j' \sim \text{N}(0, \Sigma)$, and $e_j' \sim \text{N}(0, \sigma^2 I_p)$ for $j = 1, \cdots, N$.

The diagonal elements of Σ were set to unity, and off-diagonal elements varied at three levels (0, .5, and .9). The value of σ^2 was varied at three levels (.5, 1, and 2). For each data set, elements of B were generated by uniform random numbers initially. Matrix B was then subjected to GSVD to reduce its rank to 1 or 2.

Figures 1 and 2 present the main results of the study for a particular combination of settings, namely the medium level of error variance ($\sigma^2 = 1$), the medium level of correlations among predictor variables (off-diagonal elements of $\Sigma = .5$), and rank$(B) = 2$. (Other combinations yielded similar patterns of results.) Figure 1 shows MSE for regression coefficients in RA as a function of the sample size (N) and the ridge parameter (λ). In all cases, MSE is high at $\lambda = 0$, it decreases rapidly as soon as λ gets larger than 0, but then rises again gradually. This tendency is clearer for small sample sizes, although it can still be observed for larger sample sizes. (The effect of regularization looks negligible for $N = 200$. However, this is all relative to the case of $N = 20$. It looks negligible because the $N = 20$ case, where the effect is much larger, is drawn in the same figure.) This means that better estimates of regression parameters can be obtained by the ridge estimation. Figure 2 breaks down the MSE for $N = 50$ into squared bias and variance. The squared bias increases monotonically as the value of λ increases, while the variance decreases. The sum of these two (= MSE) takes a minimum value somewhere in the middle. These results are consistent with our expectations, as discussed above. Similar observations have been made in related contexts, univariate regression (Hoerl and Kennard, 1970), multiple correspondence analysis (Takane and Hwang, 2006), and partial and/or constrained RA (Takane and Jung, 2006).

3.2 Car Attributes and Preferences

The first real data set we analyze comes from a marketing research study (Lilien and Rangaswamy, 2003, p. 149). Ten different makes of cars were rated on fifteen attributes: 1. Attractive, 2. Quiet, 3. Unreliable, 4. Poorly Built, 5. Interesting, 6. Sporty, 7. Uncomfortable, 8. Roomy, 9. Easy to Service, 10. High Prestige, 11. Common, 12. Economical, 13. Successful, 14. Avant-garde, and 15. Poor Value. These ratings were used as predictor variables. Three separate ratings were also made on the preference for the same set of cars by three groups of subjects. These groups represent three distinct segments of consumers targeted for the promotion of cars and are described as: I. Western Yuppie, II. Upwardly Mobile Families, and III. American Dreamers. Average preference ratings of the three groups were used as the criterion variables. This data set clearly involves an extreme case of multicollinearity with the number of cases smaller than the number of predictor variables.

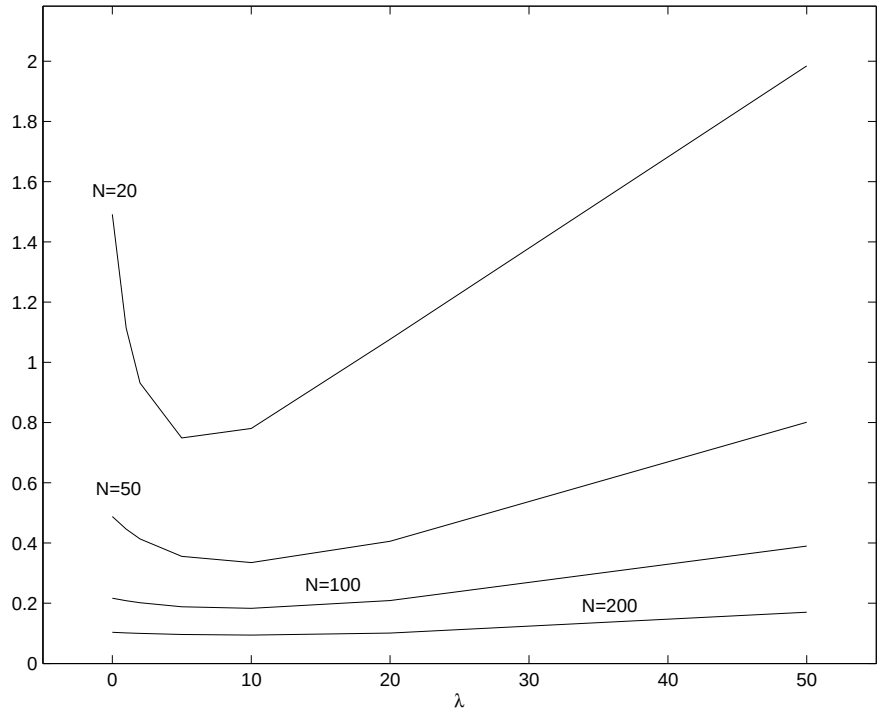

Fig. 1. Plot of MSE as functions of sample size N and the regiularization parameter λ.

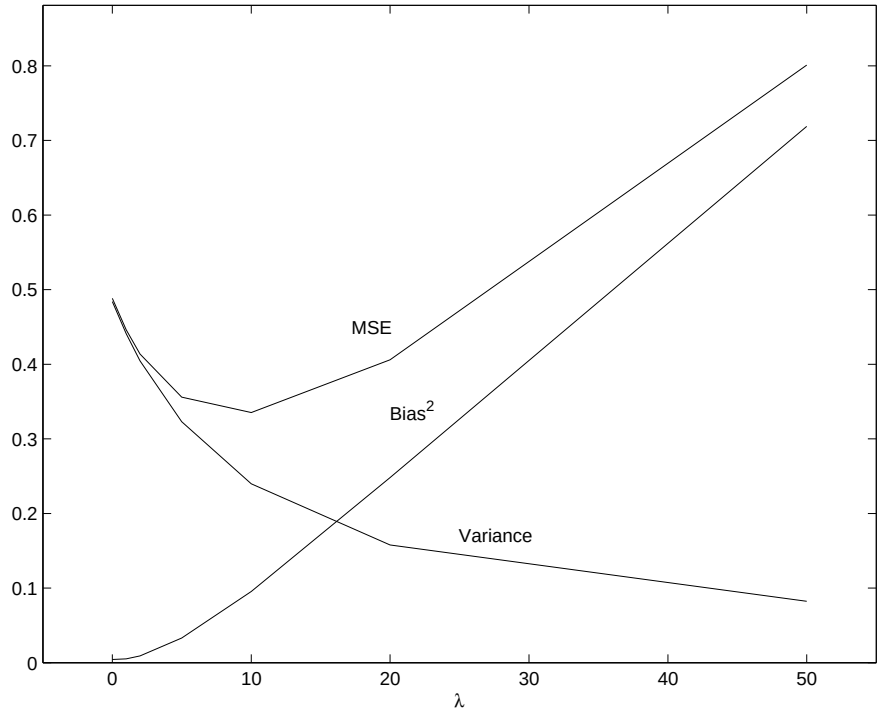

Fig. 2. Breakdown of MSE into squared bias and variance for $N = 50$.

Permutation tests were first applied, which consistently indicated one and only one significant dimension for $\lambda \geq 1$. No dimensions were found significant at $\lambda = 0$. That at least one dimension is found significant in the regularized case, while it is not in the non-regularized case, is one indication that more reliable

estimates tend to be obtained by RLS. The five-fold cross validation method was then applied, assuming the dimensionality was one. (Since $p > n$ in this data set, the leaving-one-out method could not be used. The data set was repartitioned several times to increase the number of validation samples.) The $\lambda = 20$ has been found best with the smallest prediction error. (The prediction error is 1.040 for $\lambda = 0$, .741 for $\lambda = 5$, .706 for $\lambda = 10$, .685 for $\lambda = 20$, and .709 for $\lambda = 50$.)

Table 1 compares the LS estimates and the best RLS estimates of redundancy weights (weights applied to the predictor variables to derive the first redundancy component), and cross loadings (correlations between the redundancy component and the criterion variables). While this comparison is somewhat moot, since the first dimension was already found insignificant in the LS case ($\lambda = 0$) by the permutation test, it presents an interesting case of what happens in the LS estimation when the data are very weak. The LS estimates are much larger in absolute size than the RLS estimates, but they are also much more variable than the latter. Only one coefficient (for y_1) has turned out to be significant by the LS estimation, whereas many more coefficients are significant by the RLS estimation. There are also a number of LS estimates of the weights whose signs are contrary to our intuition (e.g., x_2, x_5, x_8, x_{10}, x_{11}, and x_{13}). This is due to the extreme case of multicollinearity among the predictor variables in this data set. However, none of the RLS estimates are sign reversed.

Since in this data set the linear predictor space already contains the criterion space ($n < p$), there is little point of applying kernel RA.

3.3 Food and Cancer Data

The second data set we analyze concerns the prediction of mortality rates by cancer in lower digestive organs from food variables. Information on mortality rate by large intestine cancer (y_1) and that by rectum cancer (y_2) for 47 countries in the world was initially gathered by WHO, and information on food variables, total amount of calories per day per capita (x_1), the amount of meat supply (x_2), the amount of milk consumption (x_3) and the amount of alcohol consumption (x_4), was originally collected by FAO for the same 47 countries. The data used in the present analysis were taken from Yanai and Takagi (1986), who compiled the data in the present form.

We applied both linear and kernel RA to the data. Permutations tests were first applied with varying values of λ, which unanimously found that the full rank model ($r = 2$) was the best. We then applied the 15-fold cross validation to find an optimal value of λ. In kernel RA, this was done in combination with

Table 1
The LS and RLS estimates of component weights and cross loadings and their Bootstrap standard error estimates from the car attribute data obtained by linear RA. ("**" in the table indicates a significance at the 1% level, and "*" at the 5% level.)

		L S		R L S	
	Vari.	Estimate	Std. Err.	Estimate	Std. Err.
Weights	x_1	.836	.425	*.115	.035
	x_2	-.382	.233	*.080	.027
	x_3	-.052	.139	*-.082	.037
	x_4	-.264	.151	**-.093	.019
	x_5	.462	.357	*-.106	.031
	x_6	-.378	.290	-.028	.053
	x_7	-.524	.361	*-.072	.036
	x_8	-.592	.316	.035	.042
	x_9	-.372	.210	-.049	.054
	x_{10}	-.182	.241	*.103	.024
	x_{11}	.119	.325	*-.098	.034
	x_{12}	.091	.166	.055	.033
	x_{13}	-.028	.201	*.118	.032
	x_{14}	.096	.249	.022	.075
	x_{15}	-.649	.394	-.125	.059
Loadings	y_1	*.913	.324	*.807	.220
	y_2	.056	.671	.346	.423
	y_3	.887	.454	*.651	.295

the search for an optimal value of σ. Tables 2 and 3 summarize the results. In Table 2, prediction errors are reported as a function of σ varied from .1 to 1 in steps of .1. These values are reported only for the value of λ at which the prediction error took the smallest value (λ^* in the last column). It was found that the combination of $\sigma = .4$ and $\lambda = .01$ gave the best kernel RA solution.

Table 3 compares the results of linear RA and kernel RA as a function of λ. For kernel RA, the value of σ was fixed at its optimal value (.4) in all cases. Kernel RA yields better (cross validated) predictions than linear RA. The best overall solution is obtained by kernel RA with $\lambda = .01$.

Table 2
Cross validated prediction error as a function of σ in kernel RA of Yanai and Takagi's data. ("*" indicates an optimal value, and λ^* indicates the value of λ at which the prediction error is smallest for each σ.)

σ	Pred. Err.	λ^*
.1	.455	0
.2	.285	.0001
.3	.232	.005
.4	*.219	.01
.5	.222	.01
.6	.230	.01
.7	.239	.05
.8	.246	.05
.9	.254	.05
1	.263	.05

Table 3
Cross validated prediction error as a function of λ by linear and kernel RA of Yanai and Takagi's data. ("**" indicates the best overall solution, and "*" indicates the best solution within a method. $\sigma = .4$ in all cases for kernel RA.)

Analysis	λ	Pred. Err.
	0	.312
	1	*.307
Linear	5	.309
	10	.318
	20	.339
	50	.396
	0	.231
	.001	.228
Kernel	.005	.221
($\sigma = .4$)	.01	**.219
	.05	.243
	.1	.269
	1	.411

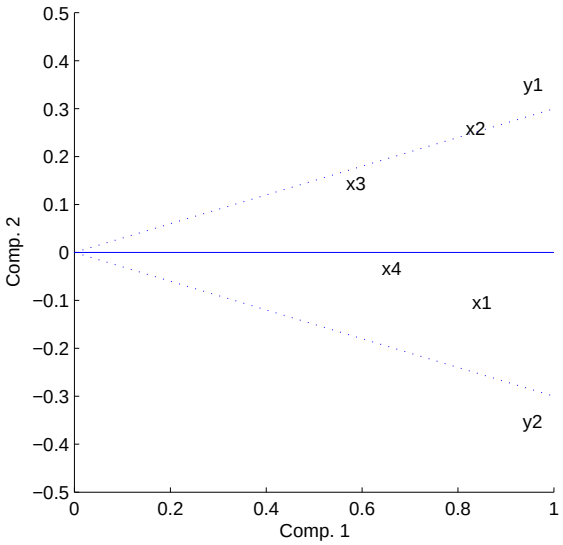

Fig. 3. Plot of cross loadings (correlations between redundancy components and Y) and predictor loadings (correlations between redundancy components and X) with the redundancy components obtained by kernel RA of Yanai and Takagi's data.

To understand the nature of the redundancy components in the best kernel solution, they were correlated with the original predictor and criterion variables. These correlations are analogous to predictor loadings (correlations between the redundancy components and X), and cross loadings (correlations between the redundancy components and Y) in the linear case. Figure 3 displays the plot of the loadings. The first component is clearly predominant, correlated positively with all the six observed variables. This component, characterized by high fat, high calorie foods, is considered to affect both large intestine and rectum cancers in a similar manner. The second component, on the other hand, differentiates the two types of cancer. Interestingly, meat (x_2) and milk products (x_3) are more closely related to large intestine cancer (y_1), while the total calorie (x_1) and alcohol (x_4) are more closely related to rectum cancer. We have also drawn an analogous picture for the linear RA solution. This has turned out to be strikingly similar to Figure 3. Apparently, nonlinear components are only slightly different from the linear ones, although an improvement in predictability (of .088 in terms of normalized prediction error) is substantial.

4 Concluding Remarks

In this paper we developed and evaluated a simple regularization technique for redundancy analysis (RA). This is a straightforward extension of the ridge

regression originated by Hoerl and Kennard (1970). As in the non-regularized LS case, the solution could be obtained in closed form for a fixed value of the ridge parameter λ. An optimal value of λ in turn can be selected by a cross validation procedure. The closed form solution is a big advantage in the cross validation process. The closed form solution is enabled by the decomposition (10) of the RLS criterion into the sum of two terms, one of which is unrelated to unknown parameters. Consequently the entire RLS criterion can be minimized by minimizing the other term, which is achieved by generalized singular value decomposition (GSVD). We also extended linear RA to kernel RA. The proposed methods were evaluated by a Monte Carlo study and through the analysis of two real data sets.

The ridge regression has been extended to multivariate multiple regression analysis without rank restriction (Haitovsky, 1987). However, these are rather routine extensions of the univariate case because the multivariate multiple regression is merely a collection of separate univariate regressions without the rank restriction. Aldrin (2000) proposed a somewhat different technique for regularized RA. His procedure first estimates the matrix of regression coefficients without rank restriction by simple (non-regularized) LS, and then applies SVD to $X\hat{B}$ to obtain $X\hat{B} = \sum_{j=1}^{m} d_j u_j v_j'$. Then, the m terms in the SVD of $X\hat{B}$ are re-weighted by regressing Y onto the set of $d_j u_j v_j'$ $(j = 1, \cdots, m)$'s. The ridge regression is used in this last phase. This procedure is much more complicated than ours, although a systematic comparison of the performance between the two approaches would undoubtedly be interesting.

For further extensions of the proposed methods, we might consider ridge estimation of partial and/or constrained RA. This is already in progress (Takane and Jung, 2006). We might also consider a more generalized form of ridge regression that incorporates λL (instead of $\lambda P_{X'}$) as the regularization term, where L is a q by q nonnegative-definite (nnd) matrix such that $\mathrm{Sp}(L) = \mathrm{Sp}(X')$. This generalized form of ridge regression is useful for incorporating more complicated forms of regularization such as smoothness (Ramsay and Silverman, 2006).

5 Acknowledgement

The work reported in this paper has been supported by by Grant A6394 to Yoshio Takane, and Grant 290439 to Heungsun Hwang from the Natural Sciences and Engineering Research Council of Canada.

References

Aldrin, M., 2000. Multivariate prediction using softly shrunk reduced-rank regression. The American Statistician, 54, 29-34.

Anderson, T.W., 1951. Estimating linear restrictions on regression coefficients for multivariate normal distributions. Annals of Mathematical Statistics, 22, 327-351.

Cailliez, F., Pages, J.P., 1976. Introduction à l'Analyse des Données, Societe de Mathematique Appliquees et de Sciences Humaines, Paris.

Efron, B., 1982. The Jackknife, the Bootstrap and Other Resampling Plans, SIAM, Philadelphia.

Golub, G.H., Van Loan, C.F., 1989. Matrix Computations, (Second Edition), Johns Hopkins University Press, Baltimore.

Greenacre, M.J., 1984. Theory and Applications of Correspondence Analysis, Academic Press, London.

Groß, J., 2003. Linear Regression, Springer, Berlin.

Haitovsky, Y., 1987. On multivariate ridge regression. Biometrika, 74, 563-570.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York.

Herblich, R., 2002. Learning Kernel Classifiers, MIT Press, Cabridge, MA.

Hoerl, K.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12, 55-67.

Lambert, Z.V., Wildt, A.R., Durand, R.M., 1988. Redundancy analysis: An alternative to canonical correlation and multivariate multiple regression in exploring interset associations. Psychological Bulletin, 104, 282-289.

Legendre, P., Legendre, L., 1998. Numerical Ecology, (Second English Edition), Elsevier, Oxford.

Lilien, G.L., Rangaswamy, A., 2003. Marketing Engineering, (Second Ediion), Prentice Hall, Upper Saddle River.

Ramsay, J.O., Silverman, B.W., 2006. Functional Data Analysis, (Second Edition), Springer, New York.

Rao, C.R., 1964. The use and interpretation of principal component analysis in applied research. Sankha, A 26, 329-358.

Reinsel, G.C., Velu, R.P., 1998. Multivariate Reduced-rank Regression, Springer, New York.

Schölkopf, B., Burges, C.J.C., Smola, A.J., 1997. Advances in Kernel Method, MIT Press, Cambridge, MA.

Suykens, J.A.K., Van Gestel, T., de Brabanter, J., De Moor, B., Vandewalle, J., 2002. Least Squares Support Vector Machines, World Scientific, Singapore.

Takane, Y. (2002). Relationships among various kinds of eigenvalue and singular value decompositions. In: H. Yanai, A. Okada, K. Shigemasu, Y. Kano, J. Meulman, (Eds.), New Developments in Psychometrics (pp. 45-56), Springer Verlag, Tokyo.

Takane, Y., Hunter, M.A., 2001. Constrained principal component analysis: A

comprehensive theory. Applicable Algebra in Engineering, Communication and Computing, 12, 391-419.
Takane, Y., Hwang, H., 2002. Generalized constrained canonical correlation analysis. Multivariate Behavioral Research, 37, 163-195.
Takane, Y., Hwang, H., 2006. Regularized multiple correspondence analysis. In: J. Blasius, M.J. Greenacre (Eds.), Multiple Correspondence Analysis and Related Methods (pp. 259-279), Chapman and Hall, London.
Takane, Y., Jung, S., 2006. Regularized partial and/or constrained redudndacy analysis. Submitted to Psychometrika.
Takane, Y., Shibayama, T., 1991. Principal component analysis with external information on both subjects and variables. Psychometrika, 56, 97-120.
Takane, Y., Yanai, H., 2006. On ridge operators. Submitted to Linear Algebra and Its Applications.
ten Berge, J.M.F., 1993. Least Squares Optimization in Multivariate Analysis, DSWO Press, Leiden, The Netherlands.
ter Braak, C.J.F., 1987. Unimodal Models to Relate Species to Environment, The Agricultural Mathematics Group, Wegeningen, The Netherlands.
ter Braak, C.J.F., Šmilauer, P., 1998. CANOCO Reference Manual and User's Guide to Canoco for Windows. Microcomputer Power, Ithaka, N.Y.
Van den Wollenberg, A.L., 1977. Redundancy analysis: An alternative for canonical analysis. Psychometrika, 42, 207-219.
van der Leeden, R., 1990. Reduced Rank Regression with Structured Residuals, DSWO Press, Leiden, The Netherlands.
Yanai, H., Takagi, H. (Eds.), 1986. Tahenryokaiseki Handobukku [Handbook of Multivariate Analysis], Gendaisuugakusha, Kyoto, Japan.