REGULARIZED PARTIAL AND/OR CONSTRAINED REDUNDANCY ANALYSIS

Yoshio Takane and Sunho Jung

mcgill university

We thank Jim Ramsay for his insightful comments on an earlier draft of this paper. Correspondence should be sent to Yoshio Takane, Department of Psychology, McGill University, 1205 Dr. Penfield Avenue, Montreal, QC, H3A 1B1, Canada. Matlab programs that carried out the computations reported in this paper are available from the first author upon request.
E-Mail: takane@psych.mcgill.ca
Phone: 514-398-6125
Fax: 514-398-4896
Website: http://takane.brinkster.net/Yoshio/

REGULARIZED PARTIAL AND/OR CONSTRAINED REDUNDANCY ANALYSIS

Abstract

Methods of incorporating a ridge type of regularization into partial redundancy analysis (PRA), constrained redundancy analysis (CRA), and partial and constrained redundancy analysis (PCRA) were discussed. The usefulness of ridge estimation in reducing MSE (mean square error) has been recognized in multiple regression analysis for some time, especially when predictor variables are nearly collinear, and the ordinary least squares estimator is poorly determined. The ridge estimation method was extended to PRA, CRA, and PCRA, where the reduced rank ridge estimates of regression coefficients were obtained by minimizing the ridge least squares criterion. It was shown that in all cases they could be obtained in closed form for a fixed value of ridge parameter. An optimal value of the ridge parameter is found by G-fold cross validation. Illustrative examples were given to demonstrate the usefulness of the method in practical data analysis situations.

Key words: Reduced rank approximations, Covariates, Linear constraints, Least squares estimation, Ridge least squares estimation, Generalized singular value decomposition (GSVD), G-fold cross validation, The bootstrap method

Introduction

Redundancy analysis (RA; Van den Wollenberg, 1977) is a useful technique for analyzing a directional relationship between two sets of multivariate data (Lambert, Wildt, and Durand, 1988). RA aims to extract components of predictor variables that are most predictive of criterion variables as a whole. A series of components called redundancy components are mutually orthogonal and successively account for the maximum variance in the criterion variables. Let $\mathbf{Y}$ and $\mathbf{X}$ be n by p and n by q matrices of criterion and predictor variables, respectively. The model for RA can be written as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \tag{1}$$

where B is the q by p matrix of regression coefficients, and E is the n by p matrix of disturbance terms. In RA, a rank restriction is imposed on $\mathbf{B}$, i.e., $\text{rank}(\mathbf{B}) = r \leq \text{rank}(\mathbf{X}'\mathbf{Y}) \leq \min(\text{rank}(\mathbf{X}), \text{rank}(\mathbf{Y})) \leq \min(p, q)$.

Takane and Hwang (2007) recently proposed a ridge type of regularized estimation for RA. A rank-free LS estimate of $\mathbf{B}$ (an estimate of $\mathbf{B}$ without rank restriction) is obtained by $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Y}$, where $(\mathbf{X}'\mathbf{X})^-$ is a generalized inverse (g-inverse) of $\mathbf{X}'\mathbf{X}$, while a rank-free ridge least squares (RLS) estimate of B by $\hat{\mathbf{B}}(\lambda) = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{P}_{X'})^-\mathbf{X}'\mathbf{Y}$, where λ is the ridge parameter, and $\mathbf{P}_{X'} = \mathbf{X}'(\mathbf{X}\mathbf{X}')^-\mathbf{X}$ is the orthogonal projector onto the row space of $\mathbf{X}$. Note that $\mathbf{P}_{X'}$ reduces to the identity matrix of order q when $\mathbf{X}$ is columnwise nonsingular. A small positive value of λ typically obtains estimates of regression coefficients that are on average closer to true parameter values than their LS counterparts (Hoerl and Kennard, 1970; see Groß(2003) for an up-to-date account of ridge regression). Takane and Hwang (2007) have shown that the reduced rank RLS estimate of $\mathbf{B}$ is obtained by the generalized singular value decomposition (GSVD) of $\hat{\mathbf{B}}(\lambda)$ with metric matrices $\mathbf{X}'\mathbf{X} + \lambda\mathbf{P}_{X'}$ and $\mathbf{I}_p$. This is analogous to the reduced rank LS estimate of $\mathbf{B}$, which is obtained by the GSVD of $\hat{\mathbf{B}}$ with metric matrices $\mathbf{X}'\mathbf{X}$ and $\mathbf{I}_p$. The reduced rank feature of RA captures redundant information in criterion variables in a most parsimonious way. The rank reduction, however, does not correct for redundancy (multicollinearity) among predictor variables. The regularization is introduced to deal with the multicollinearity problem among the predictor variables that often exists in varying degrees in multiple regression situations.

In applying RA to predict $\mathbf{Y}$ from $\mathbf{X}$, we may want to take into account the effects of extraneous variables. For example, we may wish to evaluate the effects of food variables on cancer mortality rates, while eliminating the effects of economic variables. A model suitable for this has already been proposed by Anderson (1951; see also Velu, 1991). In this model $\mathbf{X}$ is divided into two subsets, one of which is regarded as a set of predictor variables with reduced rank coefficients, while the other with full rank coefficients is regarded as a set

covariates whose effects we wish to eliminate in predicting $\mathbf{Y}$. This may be called partial RA (PRA).

We may also have some additional information on the regression coefficients $\mathbf{B}$. This often comes in the form of an hypothesis about $\mathbf{B}$. For example, there may be a good theoretical reason to believe that a particular element in $\mathbf{B}$ should be equal to another element. Such an hypothesis can generally be expressed as a linear constraint, $\mathbf{R}'\mathbf{B} = \mathbf{0}$, where $\mathbf{R}$ is a given matrix. To evaluate empirical validity of the hypothesis, an estimate of $\mathbf{B}$ has to be obtained under the constraint. In the context of RA, the linear and rank restrictions on $\mathbf{B}$ can be combined to yield another variant of RA, which might be called constrained RA (CRA). When the imposed constraints are consistent with the process that generated the data, CRA may provide more stable estimates of regression coefficients than their unconstrained counterparts.

When the linear constraint on $\mathbf{B}$ is combined with PRA, we can derive a mixed type of partial and constrained RA (PCRA). Here, $\mathbf{X}$ is decomposed into two parts, and linear and rank restrictions are imposed only on one set of predictor variables.

In this paper, we develop methods for incorporating a ridge type of regularized estimation in PRA, CRA, and PCRA. The ridge estimators are particularly attractive when the columns of $\mathbf{X}$ are nearly collinear and/or the sample size is small (Hoerl and Kennard, 1970; Groß, 2003). One way of assessing the quality of estimators is in terms of mean squared error (MSE). MSE is the expected value of $\mathrm{SS}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the vector of true parameters and $\hat{\boldsymbol{\theta}}$ is their estimators. MSE can be decomposed into two parts, the squared bias (the squared distance between the true parameters and the mean of the estimates) and the variance (the average squared distance between estimates and the mean of the estimates). The LS estimate of $\mathbf{B}$ has zero bias and the smallest variance among all linear unbiased estimators of $\mathbf{B}$. However, their variance may not be the smallest among the biased estimators. The ridge estimator, on the other hand, is usually biased (*albeit* often only slightly), but provides more accurate estimates of $\mathbf{B}$ than the ordinary LS estimator by having a much smaller variance.

Figures 1 and 2 illustrate a case in point. (See the next paragraph for a detailed description of how these figures were generated.) Figure 1 displays MSE for regression parameters in PRA, as a function of the sample size ($n = 20, 50, 100, 200$) and the ridge parameter ($\lambda = 0, 1, 5, 10, 20, 50$). In all cases, MSE goes down quickly as soon as the value of λ departs from zero ($\lambda = 0$ corresponds with the LS estimation), and then rises gradually. This tendency is clearer for small sample sizes, although it can still be observed for larger sample sizes. This means that better estimates of regression parameters may be obtained by the ridge estimation. It is interesting to point out that to achieve the level of MSE attained at a near optimal value of $\lambda(= 10)$ for $n = 50$ using the LS estimation,

roughly twice as many observations are necessary. Figure 2 breaks down the MSE function 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. Similar observations have been made in univariate regression (Hoerl and Kennard, 1970) and in multiple correspondence analysis (Takane and Hwang, 2006).

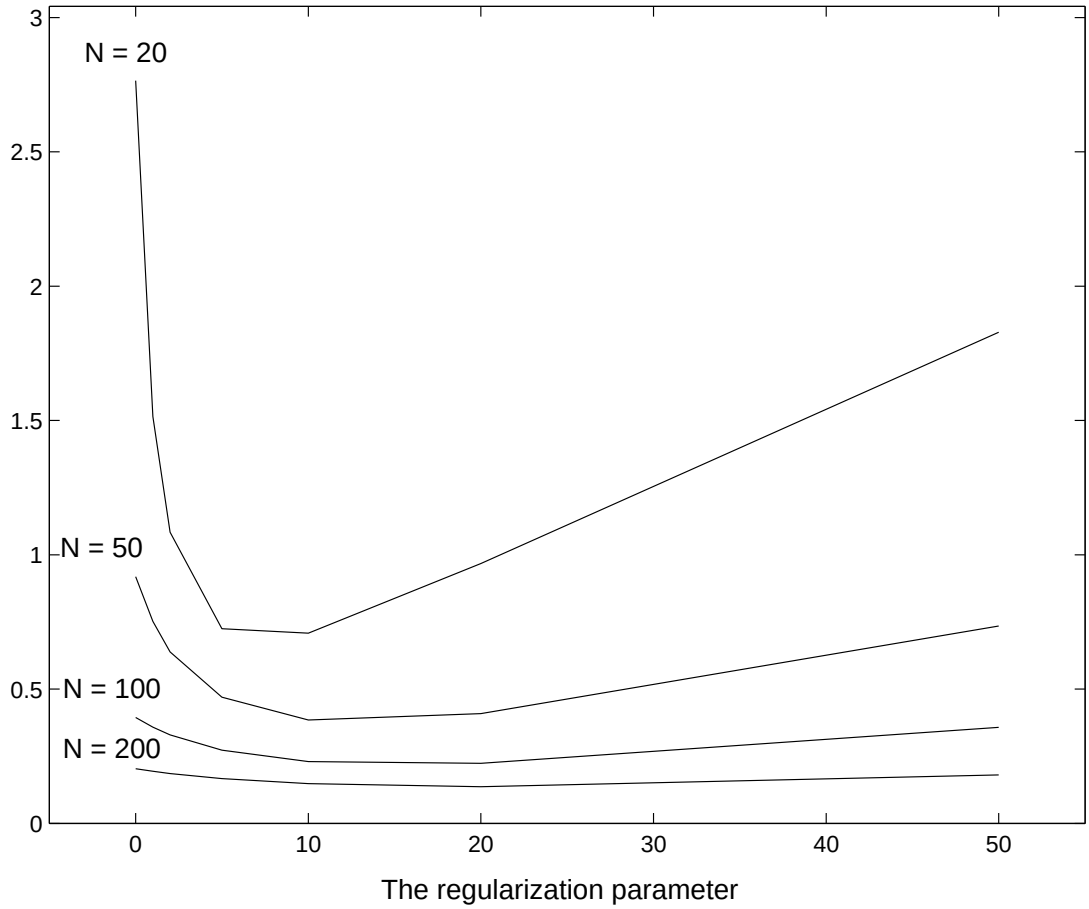

FIGURE 1.
MSE as a function of the ridge parameter (λ) and sample size (n).

These figures were obtained as follows. First, a population PRA model was postulated, from which many replicated data sets of varying sample sizes were generated. PRA was then applied to these data sets to derive reduced rank ridge estimates of regression coefficients with the value of λ systematically varied. Average MSE, squared bias, and variance were calculated in reference to the assumed population regression coefficients. In the assumed population model, the number of criterion variables was set to 3, that of

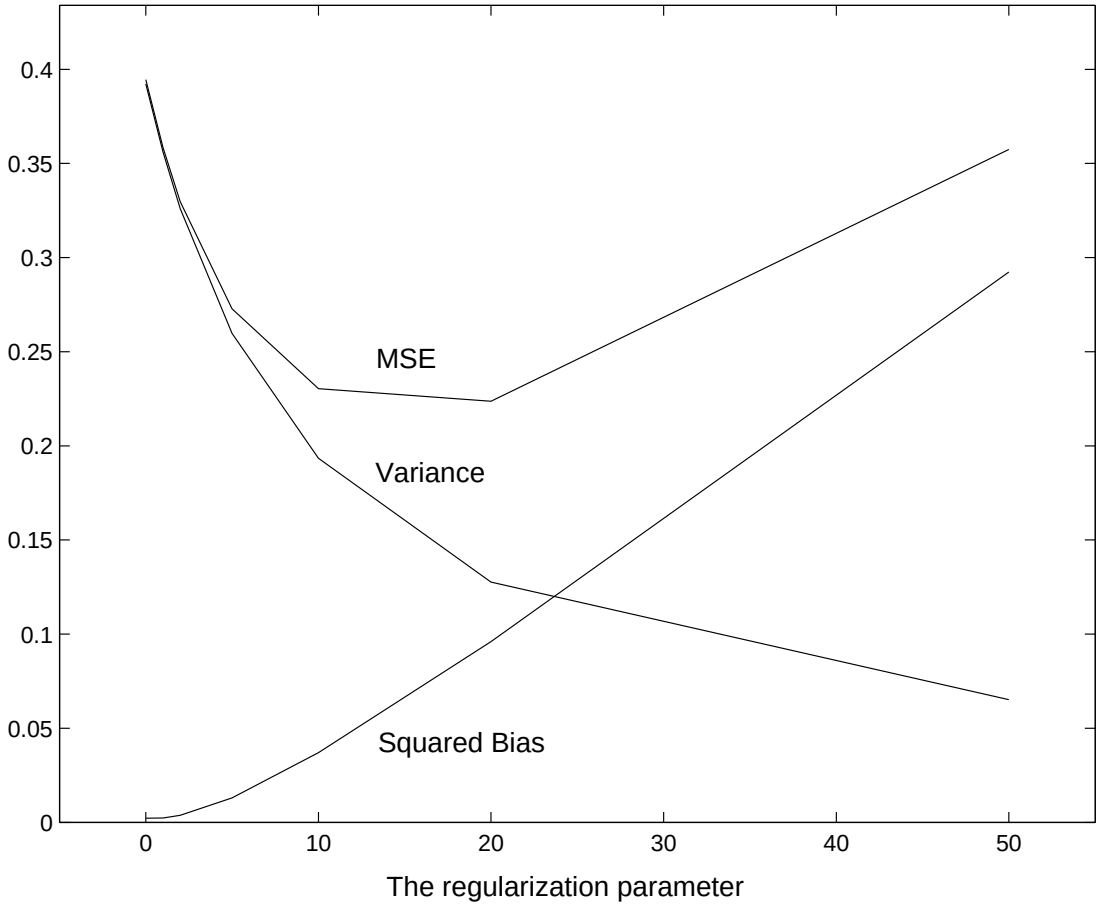

FIGURE 2.
MSE, Squared Bias and Variance as function of the ridge parameter (λ) for $n = 50$.

predictor variables to 4, and that of covariates to 1, and each row of $\mathbf{Y}$ was generated according to $\mathbf{y}_j' = \mathbf{x}_j'\mathbf{B} + \mathbf{e}_j'$, where $\mathbf{x}_j' \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma})$, and $\mathbf{e}_j' \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}_p)$ for $j = 1, \cdots, n$. The diagonal elements of $\mathbf{\Sigma}$ were set to unity and off-diagonal elements to .5. The value of σ^2 was set to 1. (We also tried many other combinations of values of off-diagonal elements of $\mathbf{\Sigma}$, e.g., 0 and .9, and of σ^2, e.g., .5 and 2, but observed similar tendencies.) The matrix of regression coefficients $\mathbf{B}$ consisted of two parts, $\mathbf{B}_1$ and $\mathbf{B}_2$, where $\mathbf{B}_1$ pertained to the predictor variables with reduced rank coefficients, and $\mathbf{B}_2$ the covariate with full rank coefficients. Elements of $\mathbf{B}$ were generated by uniform random numbers initially, the $\mathbf{B}_1$ part of which was then subjected to GSVD to reduce its rank to 2. Although the displayed results are only for PRA, similar results were obtained for CRA as well as PCRA.

Since the first publication of ridge regression (Hoerl and Kennard, 1970), the ridge estimator has received much attention in statistical literature as an alternative to LS

estimation. Of particular interest in this paper is to introduce the ridge estimation in partial and/or constrained RA. G-fold cross validation will be discussed, which is one of the most widely used methods for assessing cross-validated prediction performance. Two illustrative examples are given to demonstrate that the above proposed procedures perform well, and some possible extensions are suggested for future research.

The Methods

We describe how the ridge LS (RLS) estimator of $\mathbf{B}$ is derived in PRA, CRA, and PCRA subject to a rank restriction. In each case, we first discuss the LS estimation, and then its extension to the RLS estimation. In the following subsection, we briefly review the ordinary RA (Takane and Hwang, 2007), which serves as a benchmark for our extensions. It will be shown that in all cases reduced rank ridge estimates of regression coefficients are obtained in closed form for a fixed value of ridge parameter.

Ordinary Redundancy Analysis (ORA)

A standard LS solution to ordinary RA (ORA) is well known (e.g., Takane and Shibayama, 1991; ten Berge, 1993; Van den Wollenberg, 1977). This solution has many aspects in common with the variants of RA we address in this paper. We begin by reviewing the solution for ordinary RA, and will focus only on unique aspects of the solution for each variant.

Let $\mathbf{Y}$, $\mathbf{X}$, $\mathbf{B}$, and $\mathbf{E}$ be as introduced in (1). Throughout this paper we assume that both $\mathbf{Y}$ and $\mathbf{X}$ are columnwise standardized. In the LS estimation, we fit $\mathbf{XB}$ to $\mathbf{Y}$ in such a way that

$$\phi(\mathbf{B}) = \mathrm{SS}(\mathbf{E}) = \mathrm{SS}(\mathbf{Y} - \mathbf{XB}) \tag{2}$$

is minimized with respect to $\mathbf{B}$ subject to a rank restriction $\mathrm{rank}(\mathbf{B}) = r$. Let $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$ be a rank-free estimate of $\mathbf{B}$ (a LS estimate of $\mathbf{B}$ without rank restriction). To derive a reduced rank estimate of $\mathbf{B}$, we rewrite (2) as

$$\phi(\mathbf{B}) = \mathrm{SS}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}) + \mathrm{SS}(\hat{\mathbf{B}} - \mathbf{B})_{X'X}, \tag{3}$$

where $\mathrm{SS}(\mathbf{A})_K = \mathrm{tr}(\mathbf{A}'\mathbf{K}\mathbf{A})$. Since the first term on the right hand side of (3) is unrelated to $\mathbf{B}$, (3) can be minimized by minimizing the second term subject to the rank restriction. A reduced rank estimate of $\mathbf{B}$ is obtained by the generalized singular value decomposition (GSVD) of $\hat{\mathbf{B}}$ with respect to metric matrices $\mathbf{X}'\mathbf{X}$ and $\mathbf{I}$. This is written as $\mathrm{GSVD}(\hat{\mathbf{B}})_{X'X,\ I}$.

Let $\hat{\mathbf{B}} = \mathbf{U}\mathbf{D}\mathbf{V}'$ represent the $\mathrm{GSVD}(\hat{\mathbf{B}})_{X'X,\ I}$. Then a reduced rank estimate of $\mathbf{B}$ is obtained by retaining only the portions of $\mathbf{U}$, $\mathbf{D}$, and $\mathbf{V}$ pertaining to the r largest

(generalized) singular values (assuming that the rank of $\hat{\mathbf{B}}$ is at least as large as r). Let these reduced matrices be denoted as $\tilde{\mathbf{U}}$, $\tilde{\mathbf{D}}$, and $\tilde{\mathbf{V}}'$. Then, a reduced rank estimate of $\mathbf{B}$, denoted as $\tilde{\mathbf{B}}$, is obtained by $\tilde{\mathbf{B}} = \tilde{\mathbf{U}}\tilde{\mathbf{D}}\tilde{\mathbf{V}}'$. This process of obtaining a reduced rank estimate from a rank-free estimate remains essentially the same for all the methods to be discussed in this paper. Quantities typically found in the output of RA can be obtained by simple manipulations (rescalings) of $\tilde{\mathbf{U}}$, $\tilde{\mathbf{D}}$, and $\tilde{\mathbf{V}}'$. The matrix of weights applied to $\mathbf{X}$ to obtain redundancy components is $\mathbf{W} = n^{1/2}\tilde{\mathbf{U}}$. The matrix of redundancy components is $\mathbf{F} = \mathbf{X}\mathbf{W} = n^{1/2}\mathbf{X}\tilde{\mathbf{U}}$. The matrix of predictor loadings (correlations between predictor variables and redundancy components) is obtained by $\mathbf{C} = n^{-1}\mathbf{X}'\mathbf{F} = n^{-1/2}\mathbf{X}'\mathbf{X}\tilde{\mathbf{U}}$. The matrix of cross loadings (correlations between criterion variables and redundancy components) is $\mathbf{A} = n^{-1}\mathbf{Y}'\mathbf{F} = n^{-1/2}\tilde{\mathbf{V}}\tilde{\mathbf{D}}$.

A ridge LS (RLS) estimate of B can be obtained similarly. Let

$$\phi_\lambda(\mathbf{B}) = \mathrm{SS}(\mathbf{Y} - \mathbf{X}\mathbf{B}) + \lambda \mathrm{SS}(\mathbf{B})_{P_{X'}} \tag{4}$$

denote the RLS criterion, where λ is the ridge parameter, and $\mathbf{P}_{X'}$ is the orthogonal projector onto the row space of $\mathbf{X}$. Let

$$\hat{\mathbf{B}}(\lambda) = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{P}_{X'})^{-}\mathbf{X}'\mathbf{Y} \tag{5}$$

be a rank-free estimate of $\mathbf{B}$ that minimizes the above criterion. In a manner analogous to (3), we can rewrite (4) as (Takane and Hwang, 2007)

$$\phi_\lambda(\mathbf{B}) = \mathrm{SS}(\mathbf{Y})_{Q_X(\lambda)} + \mathrm{SS}(\hat{\mathbf{B}}(\lambda) - \mathbf{B})_{X'X+\lambda P_{X'}}, \tag{6}$$

where

$$\mathbf{Q}_X(\lambda) = \mathbf{I} - \mathbf{P}_X(\lambda), \tag{7}$$

and where

$$\mathbf{P}_X(\lambda) = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{P}_{X'})^{-}\mathbf{X}' \tag{8}$$

is called a ridge operator (Takane and Yanai, 2008). Again, since the first term on the right hand side of (6) is unrelated to $\mathbf{B}$, a reduced rank RLS estimate of $\mathbf{B}$ is obtained by minimizing the second term. This is achieved by $\mathrm{GSVD}(\hat{\mathbf{B}}(\lambda))_{X'X+\lambda P_{X'}, I}$. The rest of the procedure remains essentially the same as in the LS estimation. For later use, we introduce the following ridge metric matrix,

$$\mathbf{M}(\lambda) = \mathbf{P}_X + \lambda(\mathbf{X}\mathbf{X}')^{+}, \tag{9}$$

where $\mathbf{P}_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ is the orthogonal projector onto the column space of $\mathbf{X}$, and $(\mathbf{X}\mathbf{X}')^{+}$ is the Moore-Penrose inverse of $\mathbf{X}\mathbf{X}'$. Then, $\mathbf{X}'\mathbf{X} + \lambda\mathbf{P}_{X'}$ can be rewritten as

$$\mathbf{X}'\mathbf{X} + \lambda\mathbf{P}_{X'} = \mathbf{X}'\mathbf{M}(\lambda)\mathbf{X}. \tag{10}$$

Partial Redundancy Analysis (PRA)

We extend the above procedures to partial RA (PRA). Again, we first discuss the LS estimation, and then extend it to the ridge LS (RLS) estimation. Suppose $\mathbf{X}$ consists of two subsets of variables, i.e., $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$, where $\mathbf{X}_1$ is n by q_1, and $\mathbf{X}_2$ is n by q_2. We assume $\mathbf{X}_1$ and $\mathbf{X}_2$ are disjoint in the sense that $\text{rank}(\mathbf{X}) = \text{rank}(\mathbf{X}_1) + \text{rank}(\mathbf{X}_2)$. We also partition $\mathbf{B}$ accordingly, i.e., $\mathbf{B} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix}$. The model for PRA can be written as

$$\mathbf{Y} = \mathbf{X}_1\mathbf{B}_1 + \mathbf{X}_2\mathbf{B}_2 + \mathbf{E}. \tag{11}$$

Let

$$\phi(\mathbf{B}) = \text{SS}(\mathbf{Y} - \mathbf{X}\mathbf{B}) = \text{SS}(\mathbf{Y} - \mathbf{X}_1\mathbf{B}_1 - \mathbf{X}_2\mathbf{B}_2) = \phi(\mathbf{B}_1, \mathbf{B}_2) \tag{12}$$

be the LS criterion. We minimize this criterion with respect to $\mathbf{B}_1$ and $\mathbf{B}_2$ subject to a rank restriction $\text{rank}(\mathbf{B}_1) = r$. For this purpose, we first rewrite model (11) as (Reinsel and Velu, 1998)

$$\mathbf{Y} = \mathbf{Q}_{X_2}\mathbf{X}_1\mathbf{B}_1 + \mathbf{X}_2\mathbf{B}_2^* + \mathbf{E}, \tag{13}$$

where

$$\mathbf{Q}_{X_2} = \mathbf{I} - \mathbf{P}_{X_2} = \mathbf{I} - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^-\mathbf{X}_2', \tag{14}$$

and

$$\mathbf{B}_2^* = \mathbf{B}_2 + (\mathbf{X}_2'\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{X}_1\mathbf{B}_1. \tag{15}$$

Note that $\mathbf{X}_2'\mathbf{Q}_{X_2} = \mathbf{0}$, so that the first two terms in this model are mutually orthogonal. We also rewrite the LS criterion (12) as

$$\phi(\mathbf{B}_1, \mathbf{B}_2^*) = \text{SS}(\mathbf{Y} - \mathbf{Q}_{X_2}\mathbf{X}_1\mathbf{B}_1 - \mathbf{X}_2\mathbf{B}_2^*), \tag{16}$$

which, due to the orthogonality, can further be rewritten as

$$\phi(\mathbf{B}_1, \mathbf{B}_2^*) = \text{SS}(\mathbf{Q}_X\mathbf{Y}) + \text{SS}(\hat{\mathbf{B}}_1 - \mathbf{B}_1)_{X_1'Q_{X_2}X_1} + \text{SS}(\hat{\mathbf{B}}_2^* - \mathbf{B}_2^*)_{X_2'X_2}, \tag{17}$$

where

$$\hat{\mathbf{B}}_1 = (\mathbf{X}_1'\mathbf{Q}_{X_2}\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{Q}_{X_2}\mathbf{Y}, \tag{18}$$

and

$$\hat{\mathbf{B}}_2^* = (\mathbf{X}_2'\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{Y}. \tag{19}$$

Since the first and the third terms on the right hand side of (17) are unrelated to $\mathbf{B}_1$ (the third term can always be made equal to zero by taking $\mathbf{B}_2^* = \hat{\mathbf{B}}_2^*$), (17) can be minimized by minimizing the second term subject to $\text{rank}(\mathbf{B}_1) = r$. This amounts to $\text{GSVD}(\hat{\mathbf{B}}_1)_{X_1'Q_{X_2}X_1, I}$, where $\mathbf{X}_1'\mathbf{Q}_{X_2}\mathbf{X}_1$ and $\mathbf{I}$ are metric matrices. The rest of the procedure remains essentially the same as in ordinary RA. A LS estimate of $\mathbf{B}_2$ is obtained by

$$\tilde{\mathbf{B}}_2 = (\mathbf{X}_2'\mathbf{X}_2)^-\mathbf{X}_2'(\mathbf{Y} - \mathbf{X}_1\tilde{\mathbf{B}}_1) = \hat{\mathbf{B}}_2^* - (\mathbf{X}_2'\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{X}_1\tilde{\mathbf{B}}_1, \tag{20}$$

where $\tilde{\mathbf{B}}_1$ is a reduced rank estimate of $\mathbf{B}_1$ obtained above.

We now extend the LS estimation to the RLS estimation. We minimize

$$\phi_\lambda(\mathbf{B}_1, \mathbf{B}_2) = \text{SS}(\mathbf{Y} - \mathbf{X}_1\mathbf{B}_1 - \mathbf{X}_2\mathbf{B}_2) + \lambda\text{SS}(\begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix})_{P_{X'}} \tag{21}$$

with respect to $\mathbf{B}_1$ and $\mathbf{B}_2$ subject to $\text{rank}(\mathbf{B}_1) = r$. We rewrite the model (11) as

$$\mathbf{Y} = \mathbf{Q}_{X_2}(\lambda)\mathbf{X}_1\mathbf{B}_1 + \mathbf{X}_2\mathbf{B}_2^* + \mathbf{E}, \tag{22}$$

where

$$\mathbf{Q}_{X_2}(\lambda) = \mathbf{I} - \mathbf{P}_{X_2}(\lambda) = \mathbf{I} - \mathbf{X}_2(\mathbf{X}_2'\mathbf{M}(\lambda)\mathbf{X}_2)^-\mathbf{X}_2', \tag{23}$$

and

$$\mathbf{B}_2^* = \mathbf{B}_2 + (\mathbf{X}_2'\mathbf{M}(\lambda)\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{X}_1\mathbf{B}_1. \tag{24}$$

Note that $\mathbf{X}_2'\mathbf{M}(\lambda)\mathbf{Q}_{X_2}(\lambda)\mathbf{X}_1 = \mathbf{0}$ (see Takane and Yanai (2008) for a proof), so that the first two terms in this model are mutually orthogonal with respect to the metric matrix $\mathbf{M}(\lambda)$. Because of the orthogonality, criterion (21) can also be rewritten as (see Appendix for detail)

$$\begin{aligned}\phi_\lambda(\mathbf{B}_1, \mathbf{B}_2^*) &= \text{SS}(\mathbf{Y})_{Q_X(\lambda)} \\ &+\text{SS}(\hat{\mathbf{B}}_1(\lambda) - \mathbf{B}_1)_{X_1'Q_{X_2}(\lambda)\mathbf{X}_1 + \lambda\mathbf{P}_{X_1'}} + \text{SS}(\hat{\mathbf{B}}_2^*(\lambda) - \mathbf{B}_2^*)_{X_2'M(\lambda)X_2},\end{aligned} \tag{25}$$

where

$$\begin{aligned}\hat{\mathbf{B}}_1(\lambda) &= (\mathbf{X}_1'\mathbf{Q}_{X_2}(\lambda)\mathbf{X}_1 + \lambda\mathbf{P}_{X_1'})^-\mathbf{X}_1'\mathbf{Q}_{X_2}(\lambda)\mathbf{Y} \\ &= (\mathbf{X}_1'\mathbf{Q}_{X_2}(\lambda)\mathbf{M}(\lambda)\mathbf{Q}_{X_2}(\lambda)\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{Q}_{X_2}(\lambda)\mathbf{Y},\end{aligned} \tag{26}$$

and

$$\hat{\mathbf{B}}_2^*(\lambda) = (\mathbf{X}_2'\mathbf{M}(\lambda)\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{Y}. \tag{27}$$

Again, the first and the third terms on the right hand side of (25) are unrelated to $\mathbf{B}_1$ (the third term can always be made zero by taking $\mathbf{B}_2^* = \hat{\mathbf{B}}_2^*(\lambda)$), (25) can be minimized by minimizing the second term subject to $\text{rank}(\mathbf{B}_1) = r$, which is achieved by $\text{GSVD}(\hat{\mathbf{B}}_1(\lambda))_{X_1'Q_{X_2}(\lambda)M(\lambda)Q_{X_2}(\lambda)X_1, I}$. The rest of the procedure remains essentially the same as in the LS case. An estimate of $\mathbf{B}_2$ is obtained by

$$\begin{aligned}\hat{\mathbf{B}}_2(\lambda) &= (\mathbf{X}_2'\mathbf{M}(\lambda)\mathbf{X}_2)^-\mathbf{X}_2'(\mathbf{Y} - \mathbf{X}_1\tilde{\mathbf{B}}_1(\lambda)) \\ &= \hat{\mathbf{B}}_2^*(\lambda) - (\mathbf{X}_2'\mathbf{M}(\lambda)\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{X}_1\tilde{\mathbf{B}}_1(\lambda), \end{aligned} \tag{28}$$

where $\tilde{\mathbf{B}}_1(\lambda)$ is a reduced rank estimate of $\mathbf{B}_1$ obtained above.

Constrained Redundancy Analysis (CRA)

The model for constrained RA (CRA) remains the same as (1), but the matrix of regression coefficients $\mathbf{B}$ is subject to two kinds of restrictions: $\mathbf{R}'\mathbf{B} = \mathbf{0}$ (the linear restriction) for a given constraint matrix $\mathbf{R}$, and $\text{rank}(\mathbf{B}) = r$ (the rank restriction). Without loss of generality we assume that $\text{Sp}(\mathbf{R}) \subset \text{Sp}(\mathbf{X}')$ (i.e., the range space of $\mathbf{R}$ is in the row space of $\mathbf{X}$). In the LS estimation, we minimize

$$\phi(\mathbf{B}) = \text{SS}(\mathbf{Y} - \mathbf{X}\mathbf{B}) \tag{29}$$

with respect to $\mathbf{B}$ subject to these constraints. Let $\mathbf{T}$ be a matrix such that $\mathbf{P}_{X'} - \mathbf{R}(\mathbf{R}'\mathbf{R})^-\mathbf{R}' = \mathbf{T}\mathbf{T}'$ and $\mathbf{T}'\mathbf{T} = \mathbf{I}$. Then, the constraint $\mathbf{R}'\mathbf{B} = \mathbf{0}$ can be reparameterized as

$$\mathbf{B} = \mathbf{T}\mathbf{B}^* \tag{30}$$

for some $\mathbf{B}^*$, and $\phi(\mathbf{B})$ in (29) can be rewritten as

$$\phi(\mathbf{B}) = \text{SS}(\mathbf{Q}_{XT}\mathbf{Y}) + \text{SS}(\hat{\mathbf{B}}^* - \mathbf{B}^*)_{T'X'XT}, \tag{31}$$

where

$$\mathbf{Q}_{XT} = \mathbf{I} - \mathbf{X}\mathbf{T}(\mathbf{T}'\mathbf{X}'\mathbf{X}\mathbf{T})^{-1}\mathbf{T}'\mathbf{X}', \tag{32}$$

and

$$\hat{\mathbf{B}}^* = (\mathbf{T}'\mathbf{X}'\mathbf{X}\mathbf{T})^{-1}\mathbf{T}'\mathbf{X}'\mathbf{Y} \tag{33}$$

is the rank-free LS estimate of $\mathbf{B}^*$. Since the first term on the right hand side of (31) is unrelated to $\mathbf{B}^*$, (31) can be minimized by minimizing the second term subject to the rank restriction on $\mathbf{B}^*$. Note that $\text{rank}(\mathbf{B}^*) = \text{rank}(\mathbf{B})$. The rest of the procedure remains

essentially the same as in ORA. Note that $\hat{\mathbf{B}}^{(c)} = \mathbf{T}\hat{\mathbf{B}}^*$ is the rank-free LS estimate of $\mathbf{B}$, and since

$$\mathrm{SS}(\hat{\mathbf{B}}^* - \mathbf{B}^*)_{T'X'XT} = \mathrm{SS}(\hat{\mathbf{B}}^{(c)} - \mathbf{B})_{X'X}, \tag{34}$$

(31) can also be minimized by $\mathrm{GSVD}(\hat{\mathbf{B}}^{(c)})_{X'X,I}$.

In some cases, constraints on $\mathbf{B}$ may be supplied in the form of $\mathbf{B} = \mathbf{HA}$, where $\mathbf{H}$ is a known design matrix, and $\mathbf{A}$ is the matrix of regression weights to be estimated. This is similar to (30), but unlike $\mathbf{T}$, $\mathbf{H}$ may not satisfy the required orthogonality. Such an $\mathbf{H}$, however, can always be turned into $\mathbf{T}$ with the required property. Let the singular value decomposition of $\mathbf{H}$ be denoted by $\mathbf{H} = \mathbf{U}_H\mathbf{D}_H\mathbf{V}'_H$. Then, $\mathbf{T}$ can be set equal to $\mathbf{U}_H$, and $\mathbf{B}^* = \mathbf{D}_H\mathbf{V}'_H\mathbf{A}$.

In the RLS estimation, we minimize (4) with respect to $\mathbf{B}$ subject to $\mathbf{R}'\mathbf{B} = \mathbf{0}$ and $\mathrm{rank}(\mathbf{B}) = r$. In a manner analogous to (6), the RLS criterion can be rewritten as

$$\phi_\lambda(\mathbf{B}) = \mathrm{SS}(\mathbf{Y})_{Q_{XT}(\lambda)} + \mathrm{SS}(\hat{\mathbf{B}}^*(\lambda) - \mathbf{B}^*)_{T'X'XT+\lambda I}, \tag{35}$$

where

$$\mathbf{Q}_{XT}(\lambda) = \mathbf{I} - \mathbf{XT}(\mathbf{T}'\mathbf{X}'\mathbf{XT} + \lambda\mathbf{I})^{-1}\mathbf{T}'\mathbf{X}', \tag{36}$$

and

$$\hat{\mathbf{B}}^*(\lambda) = (\mathbf{T}'\mathbf{X}'\mathbf{XT} + \lambda\mathbf{I})^{-1}\mathbf{T}'\mathbf{X}'\mathbf{Y} \tag{37}$$

is the RLS estimate of $\mathbf{B}^*$. Since the first term on the right hand side of (35) is unrelated to $\mathbf{B}^*$, (35) can be minimized by minimizing the second term subject to $\mathrm{rank}(\mathbf{B}) = \mathrm{rank}(\mathbf{B}^*) = r$. This is achieved by $\mathrm{GSVD}(\hat{\mathbf{B}}^*(\lambda))_{T'X'XT+\lambda I,I}$. Again, the rest of the procedure remains essentially the same as in ORA. Note that $\hat{\mathbf{B}}^{(c)}(\lambda) = \mathbf{T}\hat{\mathbf{B}}^*(\lambda)$ is the rank-free RLS estimate of $\mathbf{B}$, and since

$$\mathrm{SS}(\hat{\mathbf{B}}^*(\lambda) - \mathbf{B}^*)_{T'X'XT+\lambda I} = \mathrm{SS}(\hat{\mathbf{B}}^{(c)}(\lambda) - \mathbf{B})_{X'X+\lambda P_{X'}}, \tag{38}$$

(35) can also be minimized by $\mathrm{GSVD}(\hat{\mathbf{B}}^{(c)}(\lambda))_{X'X+\lambda P_{X'},I}$. Note also that since $\mathbf{T}'\mathbf{P}_{X'}\mathbf{T} = \mathbf{T}'\mathbf{T} = \mathbf{I}$, $\mathbf{T}'\mathbf{X}'\mathbf{XT} + \lambda\mathbf{I}$ can be rewritten as $\mathbf{T}'(\mathbf{X}'\mathbf{X} + \lambda\mathbf{P}_{X'})\mathbf{T} = \mathbf{T}'\mathbf{X}'\mathbf{M}(\lambda)\mathbf{XT}$.

Partial and Constrained Redundancy Analysis (PCRA)

In PCRA, we combine the preceding two methods. The model is the same as (11). In the LS estimation, we minimize (12) with respect to $\mathbf{B}_1$ and $\mathbf{B}_2$ subject to the restriction that $\mathbf{R}'\mathbf{B}_1 = \mathbf{0}$ and $\mathrm{rank}(\mathbf{B}_1) = r$. We first rewrite the model as in (13) and reparameterize $\mathbf{B}_1$ as

$$\mathbf{B}_1 = \mathbf{TB}_1^* \tag{39}$$

for some $\mathbf{B}_1^*$. Let

$$\hat{\mathbf{B}}_1^* = (\mathbf{T}'\mathbf{X}_1'\mathbf{Q}_{X_2}\mathbf{X}_1\mathbf{T})^{-1}\mathbf{T}'\mathbf{X}_1'\mathbf{Q}_{X_2}\mathbf{Y} \tag{40}$$

be the rank-free LS estimate of $\mathbf{B}_1^*$, and let

$$\hat{\mathbf{B}}_2^* = (\mathbf{X}_2'\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{Y} \tag{41}$$

be a LS estimate of $\mathbf{B}_2^*$. Then, a reduced rank estimate of $\mathbf{B}_1^*$ is obtained by $\mathrm{GSVD}(\hat{\mathbf{B}}_1^*)_{T'X_1'Q_{X_2}X_1T,\,I}$, and that of $\mathbf{B}_1$ by $\mathrm{GSVD}(\hat{\mathbf{B}}_1^{(c)})_{X_1'Q_{X_2}X_1,\,I}$, where $\hat{\mathbf{B}}_1^{(c)} = \mathbf{T}\hat{\mathbf{B}}_1^*$ is the rank-free LS estimate of $\mathbf{B}_1$. A LS estimate of $\mathbf{B}_2$ is obtained by

$$\tilde{\mathbf{B}}_2 = (\mathbf{X}_2'\mathbf{X}_2)^-\mathbf{X}_2'(\mathbf{Y} - \mathbf{X}_1\tilde{\mathbf{B}}_1^{(c)}) = \hat{\mathbf{B}}_2^* - (\mathbf{X}_2'\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{X}_1\tilde{\mathbf{B}}_1^{(c)}, \tag{42}$$

where $\tilde{\mathbf{B}}_1^{(c)}$ is a reduced rank estimate of $\mathbf{B}_1$ obtained above.

In the RLS estimation, we minimize (21) with respect to $\mathbf{B}_1$ and $\mathbf{B}_2$ subject to $\mathbf{R}'\mathbf{B}_1 = \mathbf{0}$ and $\mathrm{rank}(\mathbf{B}_1) = r$. We first rewrite the model (11) as (22). We reparameterize $\mathbf{B}_1$ as in (39). Let

$$\hat{\mathbf{B}}_1^*(\lambda) = (\mathbf{T}'\mathbf{X}_1'\mathbf{Q}_{X_2}(\lambda)\mathbf{X}_1\mathbf{T} + \lambda\mathbf{I})^{-1}\mathbf{T}'\mathbf{X}_1'\mathbf{Q}_{X_2}(\lambda)\mathbf{Y} \tag{43}$$

be the rank-free RLS estimate of $\mathbf{B}_1^*$, and let

$$\hat{\mathbf{B}}_2^*(\lambda) = (\mathbf{X}_2'\mathbf{M}(\lambda)\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{Y} \tag{44}$$

be an RLS estimate of $\mathbf{B}_2^*$. Then, a reduced rank estimate of $\mathbf{B}_1^*$ is obtained by $\mathrm{GSVD}(\hat{\mathbf{B}}_1^*(\lambda))_{T'X_1'Q_{X_2}(\lambda)X_1T+\lambda I,\,I}$, and that of $\mathbf{B}_1$ by $\mathrm{GSVD}(\hat{\mathbf{B}}_1^{(c)}(\lambda))_{X_1'Q_{X_2}(\lambda)X_1+\lambda P_{X_1'}}$, where $\hat{\mathbf{B}}_1^{(c)}(\lambda) = \mathbf{T}\hat{\mathbf{B}}_1^*(\lambda)$ is the rank-free RLS estimate of $\mathbf{B}_1$. An RLS estimate of $\mathbf{B}_2$ is obtained by

$$\begin{aligned}\tilde{\mathbf{B}}_2(\lambda) &= (\mathbf{X}_2'\mathbf{M}(\lambda)\mathbf{X}_2)^-\mathbf{X}_2'(\mathbf{Y} - \mathbf{X}_1\tilde{\mathbf{B}}_1^{(c)}(\lambda)) \\ &= \hat{\mathbf{B}}_2^*(\lambda) - (\mathbf{X}_2'\mathbf{M}(\lambda)\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{X}_1\tilde{\mathbf{B}}_1^{(c)}(\lambda),\end{aligned} \tag{45}$$

where $\tilde{\mathbf{B}}_1^{(c)}(\lambda)$ is a reduced rank estimate of $\mathbf{B}_1$ obtained above.

Cross Validation, Permutation Tests, and the Bootstrap

The ridge parameter λ regulates how much of a shrinkage effect we would like to incorporate in the estimation. So far in the "Method" section, we have assumed that the value of λ is known. However, it is usually unknown and must be estimated in some way. An "optimal" value of λ may be chosen through cross validation. In G-fold cross validation, the data set is randomly divided into G subsets, one of which is used as a test sample, while the remaining $G-1$ subsets are used to estimate model parameters. These estimates

are then applied to the test sample to calculate the prediction error. This is repeated G times with each one of the G subsets serving as the test sample in turn. The total amount of prediction error is accumulated over the G test samples. The index is then normalized by SS($\mathbf{Y}$) to obtain the normalized prediction error. These steps are repeated for different values of λ (e.g., 0, 1, 5, 10, 20, 50), and the value of λ that gives the smallest prediction error is chosen. (In the event that the prediction error monotonically decreases within the range of λ examined, a larger value of λ must be tried.) When $G = n$ (the number of cases in the original data), this procedure is equivalent to the leaving-one-out (LOO) method.

An optimal value of r (dimensionality of solutions) can also be determined by a similar method. In this case, we systematically vary r ($1 \leq r \leq \text{rank}(\mathbf{X}'\mathbf{Y})$), evaluate the sum of squared prediction errors in the same way as above, and choose the value that gives the smallest prediction error. There is one drawback in this procedure, however. Zero dimensional solutions (the hypothesis of complete independence between $\mathbf{X}$ and $\mathbf{Y}$) cannot be tested against other dimensionalities using the cross validation technique. (This is because it is impossible to cross validate the zero dimensional solutions.) This is especially problematic when the optimal dimensionality found by the cross validation is one. In this case, we may use permutation tests (Takane and Hwang, 2002) for testing the significance of the first dimension. For general discussions of permutation tests in similar contexts, the reader is referred to Legendre and Legendre (1998), and ter Braak and Šmilauer (1998).

A Bootstrap method (Efron and Tibshirani, 1993) is used to assess the stability of the estimates. In the Bootstrap method we repeatedly draw random samples of size n (called bootstrap samples) from the original data set with replacement. We apply the method of analysis to each of the bootstrap samples to obtain parameter estimates. We then calculate means and variances of the estimates, from which we estimate biases and standard errors. The Bootstrap method may also be used to test whether the estimated parameters are significantly positive or negative. Suppose an estimate with the original data turns out to be positive. We count the number of times the estimate of the same parameter comes out to be negative in bootstrap samples. If the relative frequency of the bootstrap estimates crossing over zero is less than a prescribed significance level (e.g., $\alpha = .05$ or .01), we conclude that it is significantly positive.

Examples of Application

We now apply the proposed methods to two example data sets. The first data set pertains to the situation in which both criterion and predictor variables are continuous, while the second concerns the case in which criterion variables are discrete and predictor variables are a mixture of continuous and discrete variables. The second example represents applications of PRA, CRA, and PCRA to discriminant analysis.

Example 1: Food and Cancer Data

In the first example, we predict cancer mortality rates based on the amount of various foods people eat. The average daily intakes of the following food categories in 34 countries are used as the predictor variables: (x_1) alcohol, (x_2) meat, (x_3) fish, (x_4) cereal, (x_5) vegetable, (x_6) milk products, and (x_7) the total number of calories per capita. The criterion variables are mortality rates by four types of cancer: (y_1) esophagus, (y_2) stomach, (y_3) pancreas, and (y_4) liver. Prior to all the analyses, both $\mathbf{X}$ and $\mathbf{Y}$ were columnwise standardized.

We first applied the regularized ordinary RA (ORA) to obtain some benchmark results. Thirty four-fold cross validation indicated that the optimal value of λ was 5, and the best dimensionality was one. However, the difference between $\lambda = 5$ and $\lambda = 10$ was very small with a normalized prediction error of .702 vs .703. The RLS estimation with the optimal value of λ yielded a smaller squared prediction error than the LS estimation (.716), although the improvement was rather modest. The condition number for the matrix of predictor variables in this analysis was 4.60, which raised no serious concern about multicollinearity. Since the unidimensional model was found to be the best solution by cross validation, permutation tests were applied to ensure that this dimension was statistically significant against the 0-dimensional model. The permutation tests indicated that the first component was highly significant ($s_1^2 = 48.58, p < .000$, where s_1^2 indicates the sum of squares of Y that can be explained by the first redundancy component, and the p-value indicates the empirical significance level), while the second component was not ($s_2^2 = 9.11, p > .188$).

Table 1 compares the LS and the RLS estimates of component weights, predictor loadings, and cross loadings along with their standard error estimates (in parentheses) obtained by the Bootstrap method. The estimates of parameters are similar across the two estimation methods. However, as expected, their standard errors are almost uniformly smaller for the RLS estimates. Predictor loadings indicate that the first component is significantly negatively correlated with alcohol, meat, and the total number of calories, and significantly positively with cereal. The first redundancy component may thus be interpreted as representing low-fat and low-cholesterol diet. This component is also significantly negatively correlated with mortality rates for esophagus, pancreas, and liver cancers.

According to the signs of the predictor loadings obtained by ORA, food variables could be grouped into two groups: fish, cereal, and vegetable, on one hand, and alcohol, meat, milk products, and the total number of calories, on the other. The two groups of variables are expressed in the reparameterized form as:

TABLE 1.
The LS and the RLS estimates of component weights (Weight), predictor loadings (P. Load.), and cross loadings (C. Load.) by ORA. (Bootstrap standard error estimates are given in parentheses. "*" indicates a significance at the 5% level, and "**" at the 1% level.)

	LS		RLS	
Pred.	Weight	P. Load.	Weight	P. Load.
x_1 (alcohol)	-.123	**-.740	-.204	**-.750
	(.256)	(.143)	(.142)	(.135)
x_2 (meat)	-.045	**-.773	-.178	**-.785
	(.248)	(.123)	(.125)	(.115)
x_3 (fish)	.154	.086	.087	.077
	(.167)	(.225)	(.121)	(.225)
x_4 (cereal)	**.465	**.698	**.381	**.676
	(.166)	(.167)	(.111)	(.167)
x_5 (vegetable)	*.358	.211	**.264	.182
	(.188)	(.192)	(.111)	(.180)
x_6 (milk products)	-.016	-.228	-.059	-.230
	(.177)	(.238)	(.119)	(.230)
x_7 (calories)	**-.706	**-.648	**-.495	**-.609
	(.175)	(.168)	(.104)	(.208)
Crit.	C. Load.		C. Load.	
y_1 (esophagus)	**-.661		**-.636	
	(.156)		(.142)	
y_2 (stomach)	.316		.289	
	(.313)		(.272)	
y_3 (pancreas)	**-.594		**-.554	
	(.215)		(.205)	
y_4 (liver)	**-.829		**-.797	
	(.127)		(.126)	

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \end{bmatrix}'. \tag{46}$$

We applied CRA with the $\mathbf{H}$ matrix given above. Matrix $\mathbf{H}$ can readily be transformed into $\mathbf{T}$ by simply normalizing each column of $\mathbf{H}$. Cross validation indicates that the optimal value of λ is 10, and the dimensionality is one. Again, the difference between $\lambda = 10$ and $\lambda = 5$ is rather small with a normalized prediction error of .747 vs .748. Although the RLS estimation with the optimal value of λ yields a smaller prediction error compared to the LS estimation (.754), the improvement is small. The condition number for this analysis was 1.40, which was even lower than that of ORA. An important thing is that we still see some improvement in prediction error by regularization even in such a case. Permutation tests confirm that the first component is highly significant ($s_1^2 = 39.51, p < .000$), while the second is not ($s_2^2 = .91, p > .597$).

Table 2 provides the LS and the RLS estimates of component weights, predictor loadings, and cross loadings obtained by CRA along with their standard errors (in parentheses) estimated by the bootstrap method. Estimates of parameters are again similar across the two estimation methods, but the standard errors of the RLS estimates are consistently smaller than those of LS estimates.

The prediction errors are somewhat larger for CRA (.749) than ORA (.705) due to the additional constraint imposed on $\mathbf{B}$. However, the constraint expressed by matrix $\mathbf{H}$ has the effect of producing more stable estimates than the unconstrained case. However, it is important to note the *a posteriori* nature of the constraint. Note also that more stable estimates are obtained at a cost of slightly increased bias, which is found by the Bootstrap method to be approximately 1/10 of the variance.

In predicting cancer mortality rates from food variables, it is important to consider other factors such as overall wealth of the countries, which may have some extraneous effects on the relationships. After all, what food people can afford to eat and how accessible their health care system is depend on how wealthy they are on average. It is of interest, therefore, to include variables indicating wealth status as covariates, whose effects are eliminated in evaluating the predictability of mortality rates by the food variables. This highlights more intrinsic aspects of the relationships between the two sets of variables. We use Gross Domestic Product (GDP) as an indicator of the overall wealth status.

PRA was applied with the seven food variables as $\mathbf{X}_1$, and GDP as $\mathbf{X}_2$. Cross validation indicated that the optimal value of λ was 20 (the normalized prediction error = .717), and the dimensionality was one. A sizable improvement in predictability is obtained by the RLS estimation relative to the LS estimation (.790). The condition number for PRA was at about the same level (3.59) as that in ORA. Permutation tests indicate that the

TABLE 2.

The LS and the RLS estimates of component weights (Weight), predictor loadings (P. Load.), and cross loadings (C. Load.) by CRA with the constraint matrix given in (46). (Bootstrap standard error estimates are given in parentheses. "*" indicates a significance at the 5% level, and "**" at the 1% level.)

	LS		RLS	
Pred.	Weight	P. Load.	Weight	P. Load.
x_1 (alcohol)	**-.305	**-.766	**-.287	**-.718
	(.043)	(.109)	(.034)	(.110)
x_2 (fish)	**-.305	**-.807	**-.287	**-.758
	(.043)	(.121)	(.034)	(.120)
x_3 (meat)	**.187	*.316	**.172	*.293
	(.062)	(.193)	(.055)	(.176)
x_4 (cereal)	**.187	**.548	**.172	**.511
	(.062)	(.132)	(.055)	(.122)
x_5 (vegetable)	**.187	.223	**.172	.204
	(.062)	(.225)	(.055)	(.212)
x_6 (milk products)	**-.305	**-.481	**-.287	**-.450
	(.043)	(.179)	(.034)	(.110)
x_7 (calories)	**-.305	**-.557	**-.287	**-.526
	(.043)	(.200)	(.034)	(.199)
Crit.	C. Load.		C. Load.	
y_1 (esophagus)	**-.670		**-.627	
	(.142)		(.138)	
y_2 (stomach)	.262		.245	
	(.259)		(.240)	
y_3 (pancreas)	**-.534		**-.500	
	(.220)		(.208)	
y_4 (liver)	**-.724		**-.678	
	(.123)		(.132)	

first component is highly significant ($s_1^2 = 27.19, p < .000$) despite the fact that the value of s_1^2 is reduced considerably from that obtained by ORA due to the elimination of the effect of GDP. As before, the second component is not significant ($s_2^2 = 5.05, p > .272$).

Table 3 displays the LS and the RLS estimates of component weights, predictor loadings, and cross loadings obtained by PRA along with their standard error estimates (in parentheses) obtained by the Bootstrap method. Estimates of parameters are similar across the two methods of estimation. However, the RLS estimates tend to have smaller standard errors compared to the LS estimates. The first redundancy component is significantly negatively correlated with meat, alcohol, and the total number of calories, but positively with cereal. This is similar to ORA. This component again represents low-fat and low-cholesterol diet as in ORA, but unaffected by GDP.

The fourth analysis pertains to PCRA, in which we incorporate the linear constraint. The cross validation indicates that the unidimensional solution associated with $\lambda = 20$ gives the smallest prediction error (.717). The difference between $\lambda = 10$ and $\lambda = 20$ is rather small with a normalized prediction error of .7174 vs .7171. This represents some improvement from the LS estimation (.735). The condition number for PCRA was at about the same level (1.31) as that for CRA. Permutation tests indicate that the first component is still highly significant ($s_1^2 = 25.37, p < .000$), while the second component is not ($s_2^2 = .78, p > .577$). PCRA gives a prediction error slightly larger than PRA due to the constraint imposed. However, standard errors of the estimates in PCRA tend to be much smaller (Table 4) than those in PRA. While PRA tends to produce less stable estimates than ORA (compare Table 3 with Table 1), CRA tends to produce more stable estimates than ORA, and PCRA than PRA. In particular, PCRA produced more estimates significantly different from zero than PRA. Apparently, the constraint imposed via the matrix $\mathbf{H}$ has an additional stabilizing effect on the estimates of parameters. In general, the RLS estimation yields more stable estimates of parameters across all four methods (ORA, PRA, CRA, and PCRA).

Example 2: Panel Data

In the second example, we predict male employment status from factors that influence work-family conflict. Men are being called upon to handle more family demands of taking care of the house and the family in addition to holding down a job (Duxbury, Higgins, and Lee, 1994). Studies have supported the positive relationship between this conflict and employees' job withdrawal, such as turnover intentions or behaviors (Anderson, Coffey, and Byerly, 2002). Their employment status may depend on how they cope with this conflict. They may keep working, look for another job, or leave their work life. The data come from the Panel Study of Income Dynamics for 2003. We randomly selected 88 subjects who

TABLE 3.
The LS and the RLS estimates of component weights (Weight), predictor loadings (P. Load.), and cross loadings (C. Load.) by PRA. (Bootstrap standard error estimates are given in parentheses. "*" indicates a significance at the 5% level, and "**" at the 1% level.)

	LS		RLS	
Pred.	Weight	P. Load.	Weight	P. Load.
x_1 (alcohol)	-.237	*-.713	**-.298	**-.684
	(.442)	(.198)	(.111)	(.114)
x_2 (meat)	-.022	*-.618	**-.256	**-.658
	(.425)	(.219)	(.089)	(.122)
x_3 (fish)	.329	*.563	.146	*.334
	(.257)	(.226)	(.091)	(.191)
x_4 (cereal)	.485	*.393	*.296	**.442
	(.293)	(.188)	(.101)	(.131)
x_5 (vegetable)	.364	.257	*.179	.171
	(.304)	(.234)	(.084)	(.267)
x_6 (milk products)	-.057	-.336	-.105	-.277
	(.280)	(.251)	(.112)	(.211)
x_7 (calories)	*-.771	*-.428	**-.338	*-.419
	(.335)	(.198)	(.063)	(.171)
Crit.	C. Load.	B_2'	C. Load.	B_2'
y_1 (esophagus)	**-.622	-.086	**-.550	.012
	(.180)	(.212)	(.128)	(.071)
y_2 (stomach)	.224	-.094	.206	-.080
	(.466)	(.347)	(.282)	(.141)
y_3 (pancreas)	-.368	.295	*-.363	*.214
	(.240)	(.255)	(.177)	(.094)
y_4 (liver)	**-.556	.302	**-.569	**.226
	(.165)	(.253)	(.108)	(.074)

TABLE 4.
The LS and the RLS estimates of component weights (Weight), predictor loadings (P. Load.), and cross loadings (C. Load.) by PCRA. (Bootstrap standard error estimates are given in parentheses. "*" indicates a significance at the 5% level, and "**" at the 1% level.)

	LS		RLS	
Pred.	Weight	P. Load.	Weight	P. Load.
x_1 (alcohol)	**-.308	**-.694	**-.272	**-.631
	(.066)	(.120)	(.034)	(.097)
x_2 (meat)	**-.308	**-.647	**-.272	**-.623
	(.066)	(.136)	(.034)	(.119)
x_3 (fish)	*.252	**.589	**.191	**.419
	(.090)	(.176)	(.058)	(.168)
x_4 (cereal)	*.252	*.377	**.191	**.380
	(.090)	(.156)	(.058)	(.119)
x_5 (vegetable)	*.252	.309	**.191	.230
	(.090)	(.243)	(.058)	(.219)
x_6 (milk products)	**-.308	**-.533	**-.272	**-.449
	(.066)	(.164)	(.034)	(.154)
x_7 (calories)	**-.308	-.329	**-.272	*-.373
	(.066)	(.195)	(.034)	(.185)
Crit.	C. Load.	B_2'	C. Load.	B_2'
y_1 (esophagus)	**-.605	.033	**-.550	.053
	(.151)	(.108)	(.127)	(.068)
y_2 (stomach)	.204	-.142	.195	-.098
	(.308)	(.260)	(.246)	(.127)
y_3 (pancreas)	*-.365	*.363	**-.374	**.239
	(.186)	(.168)	(.174)	(.087)
y_4 (liver)	**-.507	**.422	**-.516	**.281
	(.128)	(.134)	(.115)	(.072)

had a job in 2002 from the database of the Panel Study. There are three criterion groups: (y_1) employed (38 subjects), (y_2) turnover (36 subjects), (y_3) keeping house (14 subjects) (a baseline category), and five predictor variables: (x_1) number of children in the home under 18 years of age, (x_2) the actual age in years of the youngest child, (x_3) housework hours per week (time spent cooking, cleaning, and doing other work around the house), (x_4) total work weeks last year, (x_5) weekly work hours in the previous year. Since age may also affect the degree of work-family conflict and directly influence employment status (Sparrow, 1996), this variable was used as an extraneous variable. Again, both $\mathbf{X}$ and $\mathbf{Y}$ were columnwise standardized before analysis.

As before, we applied all four methods (ORA, PRA, CRA, and PCRA). Table 5 summarizes the results of cross validation. For all the analyses, the best dimensionality turned out to be of full rank ($r = 2$), as determined by the 44-fold cross validation method. (Which constraint matrix is used in CRA and PCRA will be explained in the next paragraph.) The optimal value of λ varies from one method to another (10 to 50) as well as over different dimensionalities. (The latter indicates the importance of applying the cross validation for all combinations of λ and r.) However, in this example, different values of λ do not result in large differences in prediction error even for the LS case ($\lambda = 0$). Overall, CRA gives the best predictions, although the difference between CRA and PCRA is fairly minor. (As will be seen later, PCRA gives the smallest classification error.) The condition numbers for predictor variables in ORA, PRA, CRA, and PCRA were 1.93, 2.00, 1.21, and 1.15, respectively. All were very small, indicating no serious multicollinearity problems in these analyses. Note that the normalized prediction errors are almost all above .9, indicating the difficulty of prediction in this data set. This is consistent with the fact that the cross validated classification error rate is nearly 45% even in the most favorable case (PCRA) in Table 6. While this is certainly smaller than the chance level, it is far from being impressive.

Figure 3 shows a plot of the RLS estimates of cross loadings (y's) and predictor loadings (x's) obtained by PRA. Note that (x_3) housework hours per week has small correlation with both the first and the second components. The first component is highly positively correlated with (x_4) total work weeks last year, and negligibly with all of the other predictor variables. This component may be called a high commitment to work. This component discriminates between (y_1) employed and (y_3) keeping house. The second component, on the other hand, is highly positively correlated with (x_2) actual age in years of the youngest child, and moderately positively with (x_1) number of children in the home under 18 years of age and (x_5) weekly work hours in the previous year. These variables indicate longer working hours linked with heavy responsibilities for children. This component may be called family-and-work conflict. The second component discriminates between

TABLE 5.
Cross validation results for ORA, PRA, CRA, and PCRA obtained from the panel data. ("*" indicates the best optimal λ within a method.)

λ	ORA		PRA		CRA		PCRA	
	r = 1	r = 2	r = 1	r = 2	r = 1	r = 2	r = 1	r = 2
0	.937	.928	.954	.954	.937	.897	.949	.907
1	.938	.926	.952	.952	.936	.896	.948	.906
5	.935	.921	.949	.945	.935	.895	.946	.904
10	.932	.916	.945	.938	.935	*.894	.944	.903
20	.930	.911	.941	.929	.934	.895	.942	*.901
50	.933	*.910	.939	*.921	.938	.902	.941	.905
100	.943	.921	.944	.926	.948	.916	.948	.917

(y_2) turnover, and (y_3) keeping house. Since (y_2) turnover closely relates to the three predictor variables associated with the second component, this component may indicate the likelihood of male employees' turnover. Male employees facing serious work-family conflict are likely to leave their jobs and look for another workplace. Therefore, organizations should encourage a healthy balance between work and personal life for their employees to decrease their turnover rates. The predictor variables were classified into two groups depending on which of the two redundancy components they were more highly correlated with, from which the following H matrix was constructed and used in CRA and PCRA:

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}'.$$

(The pattern of correlations remained essentially the same for ORA and PRA, so that the same constraint matrix was used in both CRA and PCRA.)

In this example, the criterion variables were group indicators, so cross validated classification error rates may provide a better indication of the performance for these four methods. Cases (subjects) were assigned to the group such that the predicted value is the closest to the dummy-coded group indicators among all three criterion groups. Table 6 summarizes the results of cross-validated classifications for the four methods obtained with the respective optimal values of the ridge parameter. The rows indicate the observed groups to which subjects actually belong. The columns indicate the predicted groups. The numbers in the diagonal of each subtable show the numbers of correct classifications. The overall error rates are .500 for ORA, .511 for PRA, .465 for CRA, and .454 for PCRA in the RLS estimation. While these numbers are by no means impressive, they still represent

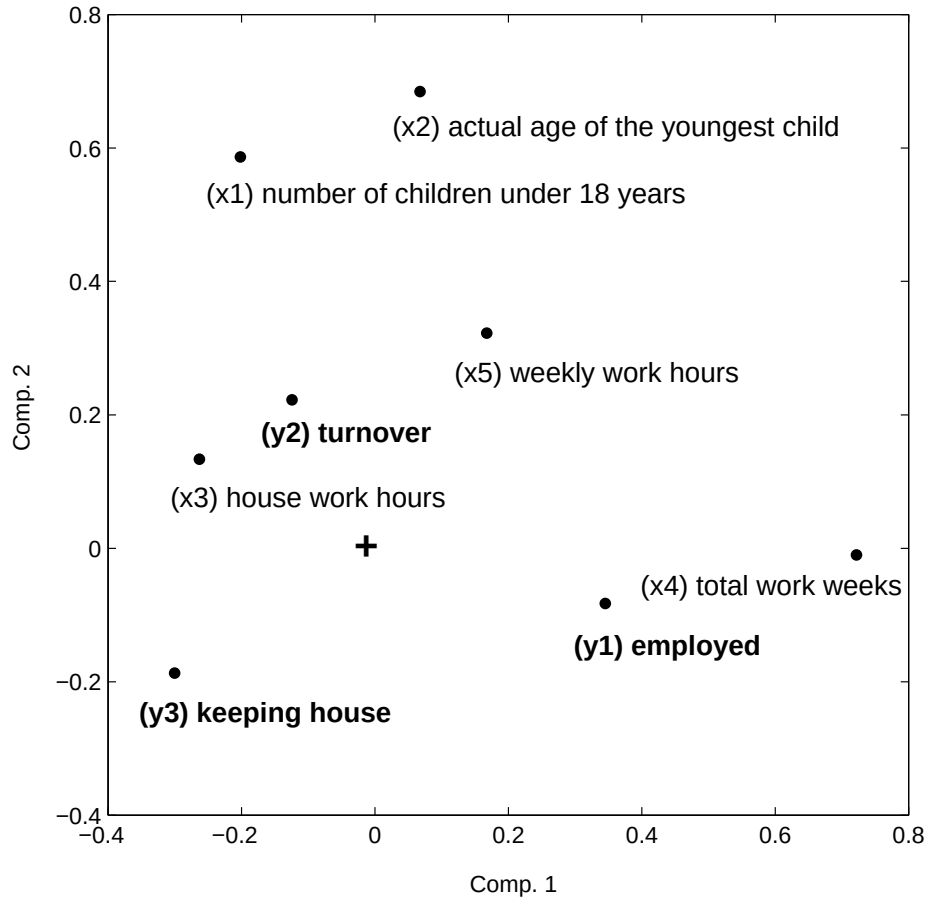

FIGURE 3.
A plot of the RLS estimates of cross loadings and predictor loadings obtained by PRA from the panel data. ("+" indicates the origin.)

substantial improvements from random assignments. These numbers also compare favorably with those obtained from the LS estimation, which are .523 for ORA, .522 for PRA, .477 for CRA, and .465 for PCRA. On the whole, the ridge PCRA gives the best results in terms of the correct classification rates. This leads to the idea that in classification problems like in this example, an optimal combination of λ and r may be selected by the cross validated classification error rate (rather than the prediction error).

Concluding Remarks

RA extracts components of predictor variables which are most predictive of criterion variables. These components thus summarize the relationships between X and Y in a concise manner. In the present article, we proposed a ridge type of regularized estimation for three variants of RA (PRA, CRA, and PCRA). PRA involves dividing predictor variables

TABLE 6.
Cross validated classification error rates obtained by regularized ORA, PRA, CRA, and PCRA.

Obs. Group	ORA			PRA			CRA			PCRA		
	Pre. Group			Pre. Group			Pre. Group			Pre. Group		
	1	2	3	1	2	3	1	2	3	1	2	3
1	25	8	5	24	10	4	25	9	4	25	10	3
2	11	13	12	12	13	11	12	15	9	11	16	9
3	5	3	6	6	2	6	5	2	7	5	2	7

into two sets, one of which has reduced rank structure, and the other of which is treated as covariates whose effects are to be eliminated. In CRA, a linear constraint is imposed on regression parameters. CRA may yield more stable estimates of parameters by reducing the number of parameters to be estimated, provided that the imposed constraint is consistent with the data. PCRA is obtained by combining PRA and CRA. We have shown that the ridge LS (RLS) estimates of regression coefficients can be obtained in closed form for all three cases above, given a fixed value of the ridge parameter λ. The optimal value of λ in turn can be determined by cross validation.

We have demonstrated the usefulness of the RLS estimation through example data sets. In the first example we reported, both criterion and predictor variables are continuous, while in the second, criterion variables are discrete, and predictor variables are mostly continuous. Other combinations of the types of criterion and predictor variables are possible. When $\mathbf{Y}$ is continuous, and $\mathbf{X}$ is discrete, RA reduces to multivariate analysis of variance (MANOVA). When both $\mathbf{X}$ and $\mathbf{Y}$ are discrete, RA reduces to nonsymmetric correspondence analysis (NSCA; D'Ambra and Lauro, 1989). Partial and/or constrained MANOVA and NSCA with the RLS estimation feature could be of interest in many data analytic situations.

A number of regularization procedures have been proposed for multivariate prediction in the literature. (See Reinsel and Velu (1998, Chapter 9) for an excellent overview of some of these procedures.) Among these, Breiman and Friedman's (1997) curds and whey (CW) method is particularly interesting. It transforms the multivariate multiple regression problem into canonical coordinates, shrinks canonical correlations, and then transforms back to the original coordinate system. It is reported to work very well; it works better than reduced rank without shrinkage, or separate ridge regressions of criterion variables without reduced rank. It is of interest to find out whether it works better than reduced rank and shrinkage combined, which is essentially our regularized RA.

There have also been attempts to use a penalty term defined in norms other than the

SS norm. Tibshirani (1996) proposed lasso which uses the L_1 norm. The lasso is known to be very effective in subset selection in multiple regression situations. More recently, it has been extended to the L_p norm by Yuan, Ekici, Lu, and Monteiro (2007). Since the SS norm is a special case in which $p = 2$, systematic comparison over different values of p is important from an empirical perspective.

There are a number of possible extensions of the proposed methods, of which we point out only one. We may consider the RLS estimation for generalized MANOVA (GMANOVA; Reinsel and Velu, 1998) and a mixture of GMANOVA and MANOVA (Chinchilli and Elswick, 1985). The former involves repeated measurements of criterion variables at several time points, constituting "growth" curves, which are represented by reduced rank polynomial regression models in time. The latter concerns a mixture of GMANOVA and MANOVA, with the GMANOVA part subject to the reduced-rank regression model, and the latter viewed as covariates. Since the basic models of GMANOVA and GMANOVA-MANOVA are similar to those of RA, the proposed estimation procedures for PRA, CRA, and PCRA can, in a straightforward manner, be extended to GMANOVA and GMANOVA-MANOVA.

Appendix

We show the equivalence of (21) and (25). From the RLS estimation of ordinary RA, we know (21) can be rewritten as

$$\phi_\lambda(\mathbf{B}) = \mathrm{SS}(\mathbf{Y})_{Q_X(\lambda)} + \mathrm{SS}(\hat{\mathbf{B}}(\lambda) - \mathbf{B})_{X'M(\lambda)X}, \tag{47}$$

where $\mathbf{M}(\lambda)$ is as defined in (19). (The above equation is the same as (6).) The first term on the right hand side of (47) is equal to the first term on the right hand side of (25). Hence, we only need to prove

$$\begin{aligned}\mathrm{SS}(\hat{\mathbf{B}}(\lambda) - \mathbf{B})_{X'M(\lambda)X} = \mathrm{SS}(\hat{\mathbf{B}}_1(\lambda) - \mathbf{B}_1)_{X_1'Q_{X_2}(\lambda)X_1+\lambda P_{X_1'}} \\ +\mathrm{SS}(\hat{\mathbf{B}}_2^*(\lambda) - \mathbf{B}_2^*)_{X_2'M(\lambda)X_2}.\end{aligned} \tag{48}$$

Note that as has been assumed $\mathbf{X}_1$ and $\mathbf{X}_2$ are disjoint, and that $\mathbf{X}_1\mathbf{Q}_{X_2}(\lambda)\mathbf{X}_1 + \lambda\mathbf{P}_{X_1'} = \mathbf{X}_1'\mathbf{Q}_{X_2}(\lambda)\mathbf{M}(\lambda)\mathbf{Q}_{X_2}(\lambda)\mathbf{X}_1$. Now let

$$\mathbf{N} = \begin{bmatrix} \mathbf{I}_{q_1} & \mathbf{0} \\ -(\mathbf{X}_2'\mathbf{M}(\lambda)\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1 & \mathbf{I}_{q_2} \end{bmatrix}. \tag{49}$$

Then,

$$\begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix} = \mathbf{N} \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2^* \end{bmatrix}, \tag{50}$$

and

$$\left[\begin{array}{c} \hat{\mathbf{B}}_1(\lambda) \\ \hat{\mathbf{B}}_2(\lambda) \end{array} \right] = \mathbf{N} \left[\begin{array}{c} \hat{\mathbf{B}}_1(\lambda) \\ \hat{\mathbf{B}}_2^*(\lambda) \end{array} \right]. \tag{51}$$

Thus,

$$\mathrm{SS}(\hat{\mathbf{B}}(\lambda) - \mathbf{B})_{X'M(\lambda)X} = \mathrm{SS} \left(\left[\begin{array}{c} \hat{\mathbf{B}}_1(\lambda) \\ \hat{\mathbf{B}}_2^*(\lambda) \end{array} \right] - \left[\begin{array}{c} \mathbf{B}_1 \\ \mathbf{B}_2^* \end{array} \right] \right)_{N'X'M(\lambda)XN}. \tag{52}$$

However,

$$\begin{aligned} \mathbf{N}'\mathbf{X}'\mathbf{M}(\lambda)\mathbf{X}\mathbf{N} &= \left[\begin{array}{cc} \mathbf{X}_1'\mathbf{Q}_{X_2}(\lambda)\mathbf{X}_1 + \lambda\mathbf{P}_{X_1'} & \mathbf{A} \\ \mathbf{A}' & \mathbf{X}_2'\mathbf{M}(\lambda)\mathbf{X}_2 \end{array} \right] \\ &= \left[\begin{array}{cc} \mathbf{X}_1'\mathbf{Q}_{X_2}(\lambda)\mathbf{X}_1 + \lambda\mathbf{P}_{X_1'} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2'\mathbf{M}(\lambda)\mathbf{X}_2 \end{array} \right], \end{aligned} \tag{53}$$

where $\mathbf{A} = \mathbf{X}_1'\mathbf{Q}_{X_2}(\lambda)\mathbf{X}_2 - \lambda\mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}(\lambda)\mathbf{X}_2)^- = \mathbf{0}$. That $\mathbf{A} = \mathbf{0}$ may be seen as follows. We have

$$\mathbf{X}_1'\mathbf{Q}_{X_2}(\lambda)\mathbf{M}(\lambda)\mathbf{X}_2 = \mathbf{X}_1'\mathbf{X}_2 - \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}(\lambda)\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{M}(\lambda)\mathbf{X}_2 = \mathbf{0}. \tag{54}$$

We also have

$$\begin{aligned} &\mathbf{X}_1'\mathbf{Q}_{X_2}(\lambda)\mathbf{M}(\lambda)\mathbf{X}_2 \\ &\quad = \mathbf{X}_1'\mathbf{Q}_{X_2}(\lambda)\mathbf{X}_2 + \lambda\mathbf{X}_1'\mathbf{Q}_{X_2}(\lambda)(\mathbf{X}\mathbf{X}')^+\mathbf{X}_2 \\ &\quad = \mathbf{X}_1'\mathbf{Q}_{X_2}(\lambda)\mathbf{X}_2 + \lambda\mathbf{X}_1'(\mathbf{X}\mathbf{X}')^+\mathbf{X}_2 - \lambda\mathbf{X}_1'\mathbf{P}_{X_2}(\lambda)(\mathbf{X}\mathbf{X}')^+\mathbf{X}_2 \\ &\quad = \mathbf{X}_1'\mathbf{Q}_{X_2}(\lambda)\mathbf{X}_2 - \lambda\mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}(\lambda)\mathbf{X}_2)^- = \mathbf{A}. \end{aligned} \tag{55}$$

We thus have $\mathbf{A} = \mathbf{0}$. Using (53), the right hand side of (52) can be rewritten as

$$\begin{aligned} &\mathrm{SS} \left(\left[\begin{array}{c} \hat{\mathbf{B}}_1(\lambda) \\ \hat{\mathbf{B}}_2^*(\lambda) \end{array} \right] - \left[\begin{array}{c} \mathbf{B}_1 \\ \mathbf{B}_2^* \end{array} \right] \right)_{N'X'M(\lambda)XN} = \\ &\qquad \mathrm{SS}(\hat{\mathbf{B}}_1(\lambda) - \mathbf{B}_1)_{X_1'Q_{X_2}(\lambda)X_1 + \lambda P_{X_1'}} + \mathrm{SS}(\hat{\mathbf{B}}_2^*(\lambda) - \mathbf{B}_2^*)_{X_2'M(\lambda)X_2}. \end{aligned} \tag{56}$$

References

Anderson, S. E., Coffey, B. S., and Byerly, R. (2002). Formal organizational initiatives and informal workplace practices: Links to workfamily conflict and job-related outcomes. *Journal of Management*, *28*, 787-810.

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

Breiman, L., and Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society, Series B*, *59*, 3-54.

Chinchilli, V. M., and Elswick, R. K. (1985). A mixture of the MANOVA and GMANOVA models. *Communications in Statistics, Theory and Methods*, *14*, 3075-3089.

D'Ambra, L., and Lauro, N. (1989). Non symmetrical analysis of three-way contingency tables. In R. Coppi and S. Bolasco (Eds.), *Multiway data analysis*. Amsterdam: North Holland.

Duxbury, L., Higgins, C., and Lee, C. (1994). Work-family conflict: A comparison by gender, family type, and perceived control. *Journal of Family Issues*, *25*, 449-466.

Efron, B., and Tibshirani, R. J. (1993). *An introduction to the Bootstrap*. Boca Raton, Florida: CRC Press.

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

Groß, J. (2003). *Linear regression*. Berlin: Springer.

Legendre, P., and Legendre, L. (1998). *Numerical Ecology*. The Second English Edition. Oxford: Elsevier.

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

Reinsel, G. C., and Velu, R. P. (1998). *Multivariate Reduced-Rank Regression: Theory and Applications*. New York: Springer-Verlag.

Sparrow, P. R. (1996). Transitions in the psychological contract: Some evidence from the banking sector. *Human Resource Management Journal*, *6*, 75-92.

Takane, Y., and Hunter, M. A. (2001). Constrained principal component analysis: A comprehensive theory, *Applicable Algebra in Engineering, Communication and Computing*, *12*, 391-419.

Takane, Y., and Hwang, H. (2002). Generalized constrained canonical correlation analysis. *Multivariate Behavioral Research*, *37*, 163-195.

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

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

Takane, Y., and Shibayama, T. (1991). Principal component analysis with external information on both subjects and variables. *Psychometrika*, *56*, 97-120.

Takane, Y., and Yanai, H. (2008) On ridge operators. *Linear Algebra and Its Applications*, *428*, 1778-1790.

ten Berge, J. M. F. (1993). Least squares optimization in multivariate analysis. Leiden, The Netherlands: DSWO Press.

ter Braak, C. J. F., and Šmilauer, P. (1998). *CANOCO Reference Manual and User's Guide to Canoco for Windows*. Ithaka, N.Y.: Microcomputer Power.

Tibshirani, R. (1996). regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Sireis B*, *58*, 267-288.

Van den Wollenberg, A. L. (1977). Redundancy analysis: alternative for canonical analysis. *Psychometrika*, *42*, 207-219.

Velu, R. P. (1991). Reduced rank models with two sets of regressors. *Applied Statistics*, *40*, 159-170.

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society, Series B*, *69*, 329-346.