REGULARIZED MULTIPLE-SET CANONICAL CORRELATION ANALYSIS

Yoshio Takane and Heungsun Hwang

mcgill university

Hervé Abdi

university of texas at dallas

The work reported in this paper is supported by Grants 10630 and 290439 from the Natural Sciences and Engineering Research Council of Canada to the first and the second authors, respectively. The authors would like to thank the two editors (old and new), the associate editor, and four anonymous reviewers for their insightful comments on earlier versions of this paper. Correspondence regarding this article 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 the paper are available upon request.
E-Mail: takane@psych.mcgill.ca
Phone: 514-398-6125
Fax: 514-398-4896
Website: http://takane.brinkster.net/Yoshio/

REGULARIZED MULTIPLE-SET CANONICAL CORRELATION ANALYSIS

Abstract

Multiple-set canonical correlation analysis (Generalized CANO or GCANO for short) is an important technique because it subsumes a number of interesting multivariate data analysis techniques as special cases. More recently, it has also been recognized as an important technique for integrating information from multiple sources. In this paper we present a simple regularization technique for GCANO and demonstrate its usefulness. Regularization is deemed important as a way of supplementing insufficient data by prior knowledge, and/or of incorporating certain desirable properties in the estimates of parameters in the model. Implications of regularized GCANO for multiple correspondence analysis are also discussed. Examples are given to illustrate the use of the proposed technique.

Key words: Information integration, Prior information, Ridge regression, Generalized singular value decomposition (GSVD), G-fold cross validation, Permutation tests, the Bootstrap method, Multiple correspondence analysis (MCA).

1. Introduction

Multiple-set canonical correlation analysis (GCANO) subsumes a number of representative techniques of multivariate data analysis as special cases (e.g., Gifi, 1990). Perhaps for this reason it has attracted attention of so many researchers (e.g., Gardner, et al., 2006; Takane and Oshima-Takane, 2002; van de Velden and Bijmolt, 2006; van der Burg, 1988). When the number of data sets K is equal to two, GCANO reduces to the usual (2-set) canonical correlation analysis (CANO), which in turn specializes into canonical discriminant analysis or MANOVA, when one of the two sets of variables consists of indicator variables, and into correspondence analysis (CA) of two-way contingency tables when both sets consist of indicator variables. GCANO also specializes into multiple correspondence analysis (MCA) when all K data sets consist of indicator variables representing patterns of responses to multiple-choice items, and into principal component analysis (PCA) when each of the K data sets consists of a single continuous variable. Thus, introducing some useful modification to GCANO has far reaching implications beyond what is normally referred to as GCANO.

GCANO analyzes the relationships among K sets of variables. It can also be viewed as a method for information integration from K distinct sources (Takane and Oshima-Takane, 2002; see also Dahl and Næs (2006), Devaux, et al. (1998), Fischer, et al. (2007), and Sun, et al. (2005).). For example, information regarding an object comes through multi-modal sensory channels, e.g., visual, auditory, and tactile. The information coming through multiple pathways must be integrated in some way before an identification judgment is made about this object. GCANO mimics this information integration mechanism. As another example, let us look at Table 1. This is a small data set from Abdi and Valentin (2007) in which three expert judges evaluated six brands of wine according to several criteria. As in this example, these criteria are not necessarily identical across different judges. In a situation like this, one may be tempted to ask: 1) What are the most discriminating factors among the six brands of wine that are commonly used by the three judges? 2) Where are those wines positioned in terms of those factors? These questions can best be answered by applying GCANO, which extracts a set of attributes (called canonical variates or components) most representative of all three judges in characterizing the wines. In the application section of this paper, a GCANO analysis of this data set will be given in some detail.

In this paper we discuss a simple regularization technique for GCANO and demonstrate its usefulness in data analysis. Regularization can broadly be construed as a process for incorporating prior knowledge in data analysis for better understanding of data, and as such, it includes all such processes that are variously called penalizing, smoothing, shrinking, soft-constraining, etc. Regularization has proven useful as a means of identi-

Table 1: Wine tasting data from Abdi and Valentin (2007).

		Expert 1			Expert 2				Expert 3		
wine	Oak-type	fruity	woody	coffee	red fruit	roasted	vanillin	woody	fruity	butter	woody
1	1	1	6	7	2	5	7	6	3	6	7
2	2	5	3	2	4	4	4	2	4	4	3
3	2	6	1	1	5	2	1	1	7	1	1
4	2	7	1	2	7	2	1	2	2	2	2
5	1	2	5	4	3	5	6	5	2	6	6
6	1	3	4	4	3	5	4	5	1	7	5

fying an over-parameterized model (e.g., Tikhonov and Arsenin, 1977), of supplementing insufficient data by prior knowledge (e.g., Poggio and Girosi, 1990), of incorporating certain desirable properties (e.g., smoothness) in the estimates of parameters (e.g., Ramsay and Silverman, 2005), and of obtaining estimates of parameters with better statistical properties (e.g., Hoerl and Kennard, 1970).

There are a variety of regularization techniques that have been developed. In this paper, however, we focus on a ridge type of shrinkage estimation initially developed in the context of regression analysis. In ridge regression (Hoerl and Kennard, 1970), the vector of regression coefficients b is estimated by

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

where $\mathbf{X}$ is the matrix of predictor variables (assumed columnwise nonsingular), $\mathbf{y}$ is the vector of observations on the criterion variable, $\mathbf{I}$ is the identity matrix of appropriate size, and λ is called the ridge parameter. A small positive value of λ in (1) often provides an estimate of $\mathbf{b}$ which is on average closer to the true parameter value than the least squares (LS) estimator (Hoerl and Kennard, 1970). Let $\boldsymbol{\theta}$ represent a generic parameter vector, and let $\hat{\boldsymbol{\theta}}$ represent its estimator. One way to measure the average closeness of an estimator to the population value is provided by the mean square error (MSE) defined by

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}) = \mathrm{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})], \tag{2}$$

where E indicates an expectation operation. The $\mathrm{MSE}(\hat{\boldsymbol{\theta}})$ can be decomposed into two parts,

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}) = (\boldsymbol{\theta} - \mathrm{E}(\hat{\boldsymbol{\theta}}))'(\boldsymbol{\theta} - \mathrm{E}(\hat{\boldsymbol{\theta}})) + \mathrm{E}[(\hat{\boldsymbol{\theta}} - \mathrm{E}(\hat{\boldsymbol{\theta}}))'(\hat{\boldsymbol{\theta}} - \mathrm{E}(\hat{\boldsymbol{\theta}}))], \tag{3}$$

where the first term on the right hand side is called "squared bias" and the second term "variance". The LS estimator is usually unbiased, but tends to have a large variance. The ridge estimator, on the other hand, is usually biased (albeit often slightly), but is associated with a much smaller variance. As a result, the latter tends to have a smaller MSE than its LS counterpart. The ridge estimation has been found particularly useful when there are a large number of predictor variables (compared to the number of cases), and/or when they are highly correlated (e.g., Hoerl and Kennard, 1970). It can also be easily adapted to the estimation of parameters in many multivariate data analysis techniques (Takane and Hwang, 2006, 2007; Takane and Jung, 2006; Takane and Yanai, 2008). In this paper we demonstrate the usefulness of the ridge regularization in GCANO through the analysis of both Monte Carlo data sets and actual data sets. In Takane and Hwang (2006), a special case of regularized GCANO (RGCANO), regularized multiple correspondence analysis (RMCA), was discussed. However, it was assumed in that paper that K sets of variables

were disjoint. In this paper, this assumption is lifted, and RGCANO is developed under a general condition.

This paper is organized as follows. In the next section, we present the proposed method of RGCANO in some detail. We first (section 2.1) briefly discuss ordinary (non-regularized) GCANO. This is for preparation to introduce regularization in the following subsection (section 2.2). We then discuss how to choose an optimal value of the regularization parameter (section 2.3). In section 2.4, we discuss some additional considerations necessary to deal with multiple-choice categorical data by RGCANO (RMCA). In section 3, we first (section 3.1) give a simple demonstration of the effect of regularization using a Monte Carlo method. We then (sections 3.2 through 3.5) illustrate practical applications of RGCANO using actual data sets. In none of these examples, the disjointness condition holds. We conclude the paper by a few remarks about the method. The appendix provides further technical information.

2. The Methods

A number of procedures have been proposed so far for relating multiple sets of variables. See Gifi (1990, section 5.1) and Smilde, Bro, and Geladi (2004) for a concise summary of these procedures. We consider only one of them in this paper, developed by Carroll (1968; see also Horst (1961), and Meredith (1964)). This approach is most attractive because the solution can be obtained non-iteratively (Kroonenberg, 2008).

2.1. Multiple-set Canonical Correlation Analysis (GCANO)

Let $\mathbf{X}_k$ $(k = 1, \cdots, K)$ denote the n-case by p_k-variable $(n > p_k)$ matrix of the k^{th} data set. Unless otherwise stated, we assume that $\mathbf{X}_k$ is column-wise standardized. Let $\mathbf{X}$ denote an n by p $(= \sum_k p_k)$ row block matrix, $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_K]$. Let $\mathbf{W}$ denote a p by t matrix of weights applied to $\mathbf{X}$ to obtain canonical (variate) scores, where t is the dimensionality of the solution (the number of canonical variates to be extracted). Let $\mathbf{W}$ be partitioned conformably with the partition of X, that is, $\mathbf{W} = [\mathbf{W}_1', \ldots, \mathbf{W}_K']'$, where $\mathbf{W}_k$ is a p_k by t matrix. In GCANO, we obtain $\mathbf{W}$ which maximizes

$$\phi(\mathbf{W}) = \mathrm{tr}(\mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W}), \tag{4}$$

subject to the restriction that $\mathbf{W}'\mathbf{D}\mathbf{W} = \mathbf{I}_t$, where $\mathbf{D}$ is a block diagonal matrix formed from $\mathbf{D}_k = \mathbf{X}_k'\mathbf{X}_k$ as the k^{th} diagonal block. This leads to the generalized eigen equation of the form,

$$\mathbf{X}'\mathbf{X}\mathbf{W} = \mathbf{D}\mathbf{W}\mathbf{\Delta}^2, \tag{5}$$

where $\mathbf{\Delta}^2$ is the diagonal matrix of the t largest generalized eigenvalues of $\mathbf{X}'\mathbf{X}$ with respect to $\mathbf{D}$ (arranged in descending order of magnitude), and $\mathbf{W}$ is the matrix of the corresponding generalized eigenvectors. Matrix of canonical scores (variates) $\mathbf{F}$ can be obtained by $\mathbf{F} = \mathbf{XW\Delta}^{-1}$. In the above generalized eigen problem, $\mathbf{D}$ is not necessarily of full rank. A way to avoid the null space of $\mathbf{D}$ in the solution has been given by de Leeuw (1982).

Essentially the same results as above can also be obtained by the generalized singular value decomposition (GSVD) of matrix $\mathbf{XD}^-$ with column metric $\mathbf{D}$, where $\mathbf{D}^-$ is a g-inverse of $\mathbf{D}$. This is written as

$$\text{GSVD}(\mathbf{XD}^-)_{I_n,\,D}. \tag{6}$$

In this decomposition, we obtain a matrix of left singular vectors $\mathbf{F}^*$ such that $\mathbf{F}^{*\prime}\mathbf{F}^* = \mathbf{I}_r$ (where r is the rank of $\mathbf{X}$), a matrix of right generalized singular vectors $\mathbf{W}^*$ such that $\mathbf{W}^{*\prime}\mathbf{D}\mathbf{W}^* = \mathbf{I}_r$, and a *pd* (positive definite) diagonal matrix of generalized singular values $\mathbf{\Delta}^*$ (in descending order of magnitude) such that $\mathbf{XD}^- = \mathbf{F}^*\mathbf{\Delta}^*\mathbf{W}^{*\prime}$ (e.g., Abdi, 2007; Greenacre, 1984; Takane and Hunter, 2001). To obtain $\text{GSVD}(\mathbf{XD}^-)_{I_n,\,D}$, we first calculate the ordinary SVD of $\mathbf{XD}^{-1/2}$, denoted by $\mathbf{XD}^{-1/2} = \tilde{\mathbf{F}}\tilde{\mathbf{\Delta}}\tilde{\mathbf{W}}'$, and then obtain $\mathbf{XD}^- = \tilde{\mathbf{F}}\tilde{\mathbf{\Delta}}\tilde{\mathbf{W}}'\mathbf{D}^{-1/2} = \mathbf{F}^*\mathbf{\Delta}^*\mathbf{W}^{*\prime}$, where $\mathbf{F}^* = \tilde{\mathbf{F}}$, $\mathbf{\Delta}^* = \tilde{\mathbf{\Delta}}$, and $\mathbf{W}^* = \mathbf{D}^{-1/2}\tilde{\mathbf{W}}$. The matrix $\mathbf{W}$ that maximizes (4) is obtained by retaining only the first t columns of $\mathbf{W}^*$ corresponding to the t largest generalized singular values (assuming that $t \leq r$), and matrix $\mathbf{\Delta}$ in (5) is obtained by retaining only the leading t by t block of $\mathbf{\Delta}^*$. The matrix of canonical scores $\mathbf{F}$ is obtained directly by retaining only the first t columns of $\mathbf{F}^*$.

We choose $\mathbf{D}^-\mathbf{D}^{1/2}$ for $\mathbf{D}^{-1/2}$ above, where $\mathbf{D}^-$ is an arbitrary g-inverse and $\mathbf{D}^{1/2}$ is the symmetric square root factor of $\mathbf{D}$. This choice of $\mathbf{D}^{-1/2}$ is convenient since $\mathbf{X}(\mathbf{D}^-\mathbf{D}^{1/2})^2 = \mathbf{XD}^+$, where $\mathbf{D}^+$ is the Moore-Penrose g-inverse of $\mathbf{D}$, uniquely determines the solution to the above GSVD problem. (Note, however, a different choice of $\mathbf{D}^{-1/2}$ is typically made in MCA as will be explained in section 2.4.)

There is another popular criterion for GCANO called the homogeneity criterion (Gifi, 1990). It is defined as

$$\psi(\mathbf{F}, \mathbf{B}) = \sum_{k=1}^{K} \text{SS}(\mathbf{F} - \mathbf{X}_k\mathbf{B}_k), \tag{7}$$

where $\text{SS}(\mathbf{Y}) = \text{tr}(\mathbf{Y}'\mathbf{Y})$, and $\mathbf{B}_k$ is the p_k by t matrix of weights. Let $\mathbf{B}$ denote a column block matrix $\mathbf{B} = [\mathbf{B}_1', \cdots, \mathbf{B}_K']'$. We minimize (7) with respect to $\mathbf{B}$ and $\mathbf{F}$ under the restriction that $\mathbf{F}'\mathbf{F} = \mathbf{I}_t$. Minimizing (7) with respect to $\mathbf{B}$ for fixed $\mathbf{F}$ leads to $\hat{\mathbf{B}} = \mathbf{D}^-\mathbf{X}'\mathbf{F}$. By putting this estimate of $\mathbf{B}$ in (7), we obtain $\psi^*(\mathbf{F}) = \psi(\mathbf{F}, \hat{\mathbf{B}})$. Minimizing $\psi^*(\mathbf{F})$ with respect to $\mathbf{F}$ under the restriction that $\mathbf{F}'\mathbf{F} = \mathbf{I}_t$ leads to the following eigen

equation:

$$\mathbf{X}\mathbf{D}^{-}\mathbf{X}'\mathbf{F} = \mathbf{F}\boldsymbol{\Delta}^2. \tag{8}$$

Matrix $\mathbf{B}$ is related to $\mathbf{W}$ in the previous formulation by $\mathbf{B} = \mathbf{W}\boldsymbol{\Delta}^{-1}$. Note that $\mathbf{X}\mathbf{D}^{-}\mathbf{X}'$ is invariant over the choice of a g-inverse $\mathbf{D}^{-}$ because $\mathrm{Sp}(\mathbf{X}') \subset \mathrm{Sp}(\mathbf{D})$ (Rao and Mitra, 1970, Lemma 2.2.4(iii)), where Sp indicates a range space. Let $\mathbf{D}^{-(*)}$ denote a block diagonal matrix with $\mathbf{D}_k^{-}$ as its k^{th} diagonal block. Clearly, $\mathbf{D}^{-(*)} \in \{\mathbf{D}^{-}\}$ (i.e., $\mathbf{D}^{-(*)}$ is a g-inverse of $\mathbf{D}$). Thus, $\mathbf{X}\mathbf{D}^{-}\mathbf{X}' = \mathbf{X}\mathbf{D}^{-(*)}\mathbf{X}' = \sum_{k=1}^{K}\mathbf{P}_k$, where $\mathbf{P}_k = \mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k)^{-}\mathbf{X}_k'$ is the orthogonal projector onto $\mathrm{Sp}(\mathbf{X}_k)$. Note also $\mathbf{X}\mathbf{D}^{-}\mathbf{D}\mathbf{D}^{-}\mathbf{X}' = \mathbf{X}\mathbf{D}^{-}\mathbf{X}'$, which explains the relationship between (8) and (6), where we noted a similar relationship between (5) and (6).

Among the three approaches, the first approach (solving (5)) has a computational advantage when the sample size n is greater than the total number of variables p, while the homogeneity approach (solving (8)) has the advantage when $p > n$. (These methods obtain eigenvalues and vectors of a p by p and an n by n matrix, respectively, and the smaller the size of the matrix, the more quickly the solution can be obtained.) The GSVD approach (solving (6)) is numerically most stable (least prone to rounding errors because it avoids calculating the matrix of sums of squares and products) and provides a theoretical bridge between the first two.

The following notations, though not essential in non-regularized GCANO, will be extremely useful in regularized GCANO to be described in the following section. Let $\mathbf{D}_X$ denote the block diagonal matrix with $\mathbf{X}_k$ as the k^{th} diagonal block, and let $\mathbf{N} = \mathbf{1}_K \otimes \mathbf{I}_n$, where $\mathbf{1}_K$ is the K-component vector of ones, and $\otimes$ indicates a Kronecker product. (We define $\mathbf{A} \otimes \mathbf{B} = [a_{ij}\mathbf{B}]$.) Then, $\mathbf{X} = \mathbf{N}'\mathbf{D}_X$, and $\mathbf{D} = \mathbf{D}_X'\mathbf{D}_X$. The homogeneity criterion (7) can also be rewritten as

$$\psi(\mathbf{F}, \mathbf{B}) = \mathrm{SS}(\mathbf{N}\mathbf{F} - \mathbf{D}_X\mathbf{B}). \tag{9}$$

An estimate of $\mathbf{B}$ that minimizes (9) for fixed $\mathbf{F}$ can then be written as $\hat{\mathbf{B}} = \mathbf{D}^{-}\mathbf{D}_X'\mathbf{N}\mathbf{F} = \mathbf{D}^{-}\mathbf{X}'\mathbf{F}$.

2.2. Regularized GCANO (RGCANO)

We now incorporate a ridge type of regularization into GCANO. In RGCANO, we maximize

$$\phi_\lambda(\mathbf{W}) = \mathrm{tr}(\mathbf{W}'(\mathbf{X}'\mathbf{X} + \lambda\mathbf{J}_p)\mathbf{W}) \tag{10}$$

subject to the ortho-normalization restriction $\mathbf{W}'\mathbf{D}(\lambda)\mathbf{W} = \mathbf{I}_t$, where λ is the regularization parameter (a shrinkage factor), $\mathbf{D}(\lambda) = \mathbf{D} + \lambda\mathbf{J}_p$, and $\mathbf{J}_p$ is the block diagonal

matrix with $\mathbf{J}_{p_k} = \mathbf{X}_k'(\mathbf{X}_k\mathbf{X}_k')^-\mathbf{X}_k$ as the k^{th} diagonal block. (Matrix $\mathbf{J}_{p_k}$ is the orthogonal projector onto the row space of $\mathbf{X}_k$. It reduces to an identity matrix of order p_k, if $\mathbf{X}_k$ is columnwise nonsingular.) An optimal value of λ is determined by a cross validation procedure (to be explained in the next section). It usually takes a small positive value and has the effect of shrinking the estimates of canonical weights $\mathbf{W}$. As has been alluded to earlier, this tends to produce estimates with a smaller mean square error (Hoerl and Kennard, 1970) than in the non-regularized case ($\lambda = 0$).

The above criterion leads to the following generalized eigen equation to be solved:

$$(\mathbf{X}'\mathbf{X} + \lambda\mathbf{J}_p)\mathbf{W} = \mathbf{D}(\lambda)\mathbf{W}\boldsymbol{\Delta}^2. \tag{11}$$

As before, essentially the same results can also be obtained by GSVD. Using the notations introduced at the end of the previous section, let

$$\mathbf{T} = \begin{bmatrix} \mathbf{N} & \lambda^{1/2}\mathbf{D}_X\mathbf{D}^+ \end{bmatrix} \tag{12}$$

be a row block matrix, and define

$$\mathbf{M}_{D_x}(\lambda) = \mathbf{T}\mathbf{T}' = \mathbf{N}\mathbf{N}' + \lambda(\mathbf{D}_X\mathbf{D}_X')^+. \tag{13}$$

(Note that $(\mathbf{D}_X\mathbf{D}_X')^+ = \mathbf{D}_X(\mathbf{D}_X'\mathbf{D}_X)^{+2}\mathbf{D}_X' = \mathbf{D}_X(\mathbf{D}^+)^2\mathbf{D}_X'$.) Then, $\mathbf{X}'\mathbf{X} + \lambda\mathbf{J}_p$ can be rewritten as

$$\mathbf{X}'\mathbf{X} + \lambda\mathbf{J}_p = \mathbf{D}_X'\mathbf{M}_{D_x}(\lambda)\mathbf{D}_X. \tag{14}$$

This suggests that we obtain

$$\mathrm{GSVD}(\mathbf{T}'\mathbf{D}_X\mathbf{D}(\lambda)^-)_{I_{n+p},\, D(\lambda)}, \tag{15}$$

where the matrix in parentheses reduces to

$$\mathbf{T}'\mathbf{D}_X\mathbf{D}(\lambda)^- = \begin{bmatrix} \mathbf{X} \\ \lambda^{1/2}\mathbf{J}_p \end{bmatrix}\mathbf{D}(\lambda)^-. \tag{16}$$

Let this GSVD be represented by

$$\mathbf{T}'\mathbf{D}_X\mathbf{D}(\lambda)^- = \mathbf{F}^*\boldsymbol{\Delta}^*\mathbf{W}^{*\prime} = \begin{bmatrix} \mathbf{F}_1^* \\ \mathbf{F}_2^* \end{bmatrix}\boldsymbol{\Delta}^*\mathbf{W}^{*\prime}, \tag{17}$$

where $\mathbf{F}^{*\prime}\mathbf{F}^* = \mathbf{I}$. We split the $\mathbf{F}^*$ matrix into two parts, one ($\mathbf{F}_1^*$) corresponding to the $\mathbf{X}$ part, and the other ($\mathbf{F}_2^*$) corresponding to the $\lambda^{1/2}\mathbf{J}_p$ part of $\mathbf{T}'\mathbf{D}_X$ in (16). We are typically only interested in the first part. As before, $\mathbf{W}$ in (11) can be obtained by retaining the only t leading columns of $\mathbf{W}^*$.

There is an interesting relationship between $\mathbf{F}_1^*$ and $\mathbf{F}_2^*$, namely

$$\mathbf{F}_1^* = \lambda^{-1/2}\mathbf{X}\mathbf{F}_2^* \tag{18}$$

that allows further reduction of the above GSVD problem. From (16) and (17), it is obvious that $\mathrm{Sp}\left(\begin{bmatrix}\mathbf{F}_1^* \\ \mathbf{F}_2^*\end{bmatrix}\right) \subset \mathrm{Sp}\left(\begin{bmatrix}\mathbf{X} \\ \lambda^{1/2}\mathbf{J}_p\end{bmatrix}\right)$, which implies $\begin{bmatrix}\mathbf{F}_1^* \\ \mathbf{F}_2^*\end{bmatrix} = \begin{bmatrix}\mathbf{X} \\ \lambda^{1/2}\mathbf{J}_p\end{bmatrix}\mathbf{G}$ for some $\mathbf{G}$. From the bottom portion of this relationship, we obtain $\mathbf{J}_p\mathbf{G} = \lambda^{-1/2}\mathbf{F}_2^*$. Note that $\mathbf{X}\mathbf{J}_p = \mathbf{X}$. Note also that the restriction $\mathbf{F}^{*\prime}\mathbf{F}^* = \mathbf{I}$ is turned into

$$\lambda^{-1}\mathbf{F}_2^{*\prime}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{J}_p)\mathbf{F}_2^* = \mathbf{I}. \tag{19}$$

By premultiplying (17) by $\lambda^{-1/2}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{J}_p)^{+1/2}[\lambda^{-1/2}\mathbf{X}' \ \mathbf{J}_p]$, where $(\mathbf{X}'\mathbf{X} + \lambda\mathbf{J}_p)^{+1/2}$ is the symmetric square root factor of $(\mathbf{X}'\mathbf{X} + \lambda\mathbf{J}_p)^+$, and also taking into account (16), we obtain

$$(\mathbf{X}'\mathbf{X} + \lambda\mathbf{J}_p)^{1/2}\mathbf{D}(\lambda)^- = \lambda^{-1/2}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{J}_p)^{1/2}\mathbf{F}_2^*\mathbf{\Delta}^*\mathbf{W}^{*\prime}. \tag{20}$$

By setting

$$\tilde{\mathbf{F}}_2^* = \lambda^{-1/2}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{J}_p)^{1/2}\mathbf{F}_2^*, \tag{21}$$

we obtain

$$(\mathbf{X}'\mathbf{X} + \lambda\mathbf{J}_p)^{1/2}\mathbf{D}(\lambda)^- = \tilde{\mathbf{F}}_2^*\mathbf{\Delta}^*\mathbf{W}^{*\prime}. \tag{22}$$

This is the GSVD$((\mathbf{X}'\mathbf{X} + \lambda\mathbf{J}_p)^{1/2}\mathbf{D}(\lambda)^-)_{I,\,D(\lambda)}$, since $\tilde{\mathbf{F}}_2^{*\prime}\tilde{\mathbf{F}}_2^* = \mathbf{I}$ from (19). The matrix whose GSVD is obtained in (22) is usually much smaller in size than the one in (17). Once $\mathbf{F}_2^*$ is obtained, $\mathbf{F}_1^*$ can easily be calculated by (18).

The homogeneity criterion (7) can also be adapted for regularization. Let

$$\begin{aligned}\psi_\lambda(\mathbf{F}, \mathbf{B}) &= \mathrm{SS}(\mathbf{N}\mathbf{F}_1 - \mathbf{D}_X\mathbf{B}) + \lambda\mathrm{SS}(\bar{\mathbf{F}}_2 - \mathbf{B})_{J_p} + \lambda(K-1)\mathrm{SS}(\bar{\mathbf{F}}_2)_{J_p} \\ &= \mathrm{SS}\left(\begin{bmatrix}\mathbf{N}\mathbf{F}_1 \\ \mathbf{F}_2 \\ (K-1)^{1/2}\mathbf{F}_2\end{bmatrix} - \begin{bmatrix}\mathbf{D}_X \\ \lambda^{1/2}\mathbf{J}_p \\ \mathbf{0}\end{bmatrix} B\right)\end{aligned} \tag{23}$$

be the regularized homogeneity criterion, where in general $\mathrm{SS}(\mathbf{A})_M = \mathrm{tr}(\mathbf{A}'\mathbf{M}\mathbf{A})$, and $\mathbf{F}_2 = \lambda^{1/2}\mathbf{J}_p\bar{\mathbf{F}}_2$. This criterion is minimized with respect to $\mathbf{B}$ and $\mathbf{F} = [\mathbf{F}_1', \mathbf{F}_2']'$ under the restriction that $\mathbf{F}'\mathbf{F} = \mathbf{I}_t$. For fixed $\mathbf{F}$, an estimate of $\mathbf{B}$ that minimizes (23) is given by

$$\hat{\mathbf{B}} = \mathbf{D}(\lambda)^- \begin{bmatrix}\mathbf{D}_X' & \lambda^{1/2}\mathbf{J}_p & \mathbf{0}\end{bmatrix} \begin{bmatrix}\mathbf{N}\mathbf{F}_1 \\ \mathbf{F}_2 \\ (K-1)^{1/2}\mathbf{F}_2\end{bmatrix}, \tag{24}$$

where $\mathbf{D}(\lambda) = \mathbf{D}_X'\mathbf{D}_X + \lambda\mathbf{J}_p$. By putting this estimate of $\mathbf{B}$ in (23), we obtain

$$\psi_\lambda^*(\mathbf{F}) \equiv \psi_\lambda(\mathbf{F}, \hat{\mathbf{B}}) = \mathrm{SS}\left(\begin{bmatrix}\mathbf{N}\mathbf{F}_1 \\ \mathbf{F}_2 \\ (K-1)^{1/2}\mathbf{F}_2\end{bmatrix}\right)_{I-R}$$

$$= \text{Const.} - \text{SS}\left(\left[\begin{array}{c} \mathbf{NF}_1 \\ \mathbf{F}_2 \\ (K-1)^{1/2}\mathbf{F}_2 \end{array}\right]\right)_R, \tag{25}$$

where

$$\mathbf{R} = \left[\begin{array}{c} \mathbf{D}_X \\ \lambda^{1/2}\mathbf{J}_p \\ \mathbf{0} \end{array}\right] \mathbf{D}(\lambda)^- \left[\begin{array}{ccc} \mathbf{D}_X' & \lambda^{1/2}\mathbf{J}_p & \mathbf{0} \end{array}\right]. \tag{26}$$

It can be easily verified that

$$\text{SS}\left(\left[\begin{array}{c} \mathbf{NF}_1 \\ \mathbf{F}_2 \\ (K-1)^{1/2}\mathbf{F}_2 \end{array}\right]\right)_I = K(\mathbf{F}_1'\mathbf{F}_1 + \mathbf{F}_2'\mathbf{F}_2) = K\mathbf{I} \tag{27}$$

is a constant. The second term on the right hand side of (25) can be further rewritten as

$$\begin{aligned} &\text{tr}\left(\left[\begin{array}{c} \mathbf{NF}_1 \\ \mathbf{F}_2 \\ \tilde{K}\mathbf{F}_2 \end{array}\right]' \mathbf{R} \left[\begin{array}{c} \mathbf{NF}_1 \\ \mathbf{F}_2 \\ \tilde{K}\mathbf{F}_2 \end{array}\right]\right) \\ &\quad = \text{tr}\left(\left[\begin{array}{ccc} \mathbf{F}_1'\mathbf{N}' & \mathbf{F}_2' & \tilde{K}\mathbf{F}_2' \end{array}\right] \left[\begin{array}{c} \mathbf{D}_X \\ \lambda^{1/2}\mathbf{J}_p \\ \mathbf{0} \end{array}\right] \mathbf{D}(\lambda)^- \left[\begin{array}{ccc} \mathbf{D}_X' & \lambda^{1/2}\mathbf{J}_p & \mathbf{0} \end{array}\right] \left[\begin{array}{c} \mathbf{NF}_1 \\ \mathbf{F}_2 \\ \tilde{K}\mathbf{F}_2 \end{array}\right]\right) \\ &\quad = \text{tr}\left(\left[\begin{array}{cc} \mathbf{F}_1' & \mathbf{F}_2' \end{array}\right] \left[\begin{array}{cc} \mathbf{XD}(\lambda)^-\mathbf{X}' & \lambda^{1/2}\mathbf{XD}(\lambda)^-\mathbf{J}_p \\ \lambda^{1/2}\mathbf{J}_p\mathbf{D}(\lambda)^-\mathbf{X}' & \lambda\mathbf{J}_p\mathbf{D}(\lambda)^-\mathbf{J}_p \end{array}\right] \left[\begin{array}{c} \mathbf{F}_1 \\ \mathbf{F}_2 \end{array}\right]\right), \end{aligned} \tag{28}$$

where $\tilde{K} = (K-1)^{1/2}$. Minimizing (25) with respect to $\mathbf{F}$ subject to $\mathbf{F}'\mathbf{F} = \mathbf{I}_t$ is equivalent to maximizing (28) under the same normalization restriction on $\mathbf{F}$, which leads to the following eigen equation:

$$\left[\begin{array}{cc} \mathbf{XD}(\lambda)^-\mathbf{X}' & \lambda^{1/2}\mathbf{XD}(\lambda)^-\mathbf{J}_p \\ \lambda^{1/2}\mathbf{J}_p\mathbf{D}(\lambda)^-\mathbf{X}' & \lambda\mathbf{J}_p\mathbf{D}(\lambda)^-\mathbf{J}_p \end{array}\right] \left[\begin{array}{c} \mathbf{F}_1 \\ \mathbf{F}_2 \end{array}\right] = \left[\begin{array}{c} \mathbf{F}_1 \\ \mathbf{F}_2 \end{array}\right] \mathbf{\Delta}^2, \tag{29}$$

where as in the non-regularized case, $\mathbf{XD}(\lambda)^-\mathbf{X}'$ is invariant over the choice of a g-inverse $\mathbf{D}(\lambda)^-$ since $\text{Sp}(\mathbf{X}') \subset \text{Sp}(\mathbf{D}(\lambda))$. Let $\mathbf{D}(\lambda)^{-(*)}$ be a block diagonal matrix with $\mathbf{D}_k(\lambda)^-$ as the k^{th} diagonal block. Clearly, $\mathbf{D}(\lambda)^{-(*)} \in \{\mathbf{D}(\lambda)^-\}$, so that $\mathbf{XD}(\lambda)^-\mathbf{X}' = \mathbf{XD}(\lambda)^{-(*)}\mathbf{X}' = \sum_{k=1}^K \mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k + \lambda\mathbf{J}_{p_k})^-\mathbf{X}_k'$. Again, we are only interested in $\mathbf{F}_1$. This $\mathbf{F}_1$ is equal to the t leading columns of $\mathbf{F}_1^*$ in (17).

Takane and Hwang (2006) developed regularized MCA, a special case of RGCANO for multiple-choice data, under the disjointness condition on $\mathbf{X}_k$'s. Matrices $\mathbf{X}_k$'s are said

to be disjoint if the following rank additivity condition holds (Takane and Yanai, 2008):

$$\text{rank}(\mathbf{X}) = \sum_{k=1}^{K} \text{rank}(\mathbf{X}_k). \tag{30}$$

The above development is substantially more general in that no such condition is required. That the present formulation is indeed more general than that by Takane and Hwang (2006) is explicitly shown in Appendix (A).

Criterion (10) can also be readily generalized into the maximization of

$$\phi_\lambda^{(L)}(\mathbf{W}) = \text{tr}(\mathbf{W}'(\mathbf{X}'\mathbf{X} + \lambda\mathbf{L})\mathbf{W}) \tag{31}$$

subject to the restriction that $\mathbf{W}'(\mathbf{D} + \lambda\mathbf{L})\mathbf{W} = \mathbf{I}_t$, where $\mathbf{L}$ is a block diagonal matrix with $\mathbf{L}_k$ as the k^{th} diagonal block. Matrix $\mathbf{L}_k$ could be any symmetric *nnd* (non-negative definite) matrix such that $\text{Sp}(\mathbf{L}_k) = \text{Sp}(\mathbf{X}_k')$. This generalization is often useful when we need a regularization term more complicated than $\mathbf{J}_p$. Such cases arise, for example, when we wish to incorporate certain degrees of smoothness in curves to be approximated by way of regularization (Adachi, 2002; Ramsay and Silverman, 2005). In this case, we define $\mathbf{M}_{D_X}^{(L)}(\lambda) = \mathbf{T}\mathbf{T}'$, where

$$\mathbf{T} = \begin{bmatrix} \mathbf{N} & (\lambda\mathbf{L})^{1/2} \end{bmatrix}. \tag{32}$$

2.3. The Choice of λ

We use the G-fold cross validation method to choose an "optimal" value of λ. In this method, the entire data set is partitioned into G subsets, one of which is set aside in turn as the test sample. Estimates of parameters are obtained from the remaining $G-1$ subsets, which are then used to predict the test sample to assess the amount of prediction error. This process is repeated G times with the test sample changed systematically. Let $\mathbf{X}^{(-g)}$ denote the data matrix with the g^{th} subset, $\mathbf{X}^{(g)}$, removed from $\mathbf{X}$. We apply RGCANO to $\mathbf{X}^{(-g)}$ to obtain $\mathbf{W}^{(-g)}$, from which we calculate $\mathbf{X}^{(g)}\mathbf{W}^{(-g)}\mathbf{W}^{(-g)\prime}$. This gives the cross validatory prediction of $\mathbf{X}^{(g)}\mathbf{D}(\lambda)^-$. We repeat this for all g's ($g = 1, \cdots, G$), and collect all $\mathbf{X}^{(g)}\mathbf{W}^{(-g)}\mathbf{W}^{(-g)\prime}$ in the matrix $\widehat{\mathbf{X}\mathbf{D}}(\lambda)^-$. We then calculate

$$\varepsilon(\lambda) = \text{SS}(\mathbf{X}\mathbf{D}(\lambda)^- - \widehat{\mathbf{X}\mathbf{D}}(\lambda)^-)_{I_n,\, D(\lambda)} \tag{33}$$

as an index of cross validated prediction error, where $\text{SS}(\mathbf{Y})_{I_n,D(\lambda)} = \text{tr}(\mathbf{Y}'\mathbf{Y}\mathbf{D}(\lambda))$. We evaluate $\varepsilon(\lambda)$ for different values of λ (e.g., $\lambda = 0$, 5, 10, 20, 50, 100), and choose the value of λ associated with the smallest value of the prediction error.

The above procedure is based on the following rationale. Let $\mathbf{F}_1^*\mathbf{\Delta}^*\mathbf{W}^{*\prime}$ denote $\text{GSVD}(\mathbf{X}\mathbf{D}^-(\lambda))_{I,\, D(\lambda)}$, and let $\mathbf{F}_1$, $\mathbf{\Delta}$, and $\mathbf{W}$ represent the reduced rank matrices obtained from $\mathbf{F}_1^*$, $\mathbf{\Delta}^*$, and $\mathbf{W}^*$. Then, $\mathbf{X}\mathbf{W}\mathbf{W}' = \mathbf{F}_1^*\mathbf{\Delta}^*\mathbf{W}^{*\prime}\mathbf{D}(\lambda)\mathbf{W}\mathbf{W}' = \mathbf{F}_1\mathbf{\Delta}\mathbf{W}'$ (denoted

by $\widehat{\mathbf{XD}}(\lambda)^-$) gives the best reduced rank approximation to $\mathbf{XD}(\lambda)^-$. In cross validation we use $\mathbf{X}^{(g)}\mathbf{W}^{(-g)}\mathbf{W}^{(-g)\prime}$ to obtain the best prediction to $\mathbf{X}^{(g)}\mathbf{D}(\lambda)^-$, which is accumulated over g.

A similar procedure can be used for selecting an optimal number of canonical variates to be extracted. It is, however, rather time-consuming to vary both the value of λ and the number of canonical variates simultaneously in the G-fold cross validation procedure. It is more economical to choose the number of canonical variates by some other means, and then apply the cross validation method to find an optimal value of λ. We use permutation tests for dimensionality selection. This procedure is similar to the one used by Takane and Hwang (2002) in generalized constrained CANO. General descriptions of the permutation tests for dimensionality selection can be found in Legendre and Legendre (1998), and ter Braak (1990).

We also use a bootstrap method (Efron, 1979) to assess the reliability of parameter estimates derived by RGCANO. In this procedure, random samples (called bootstrap samples) of the same size as the original data are repeatedly sampled with replacement from the original data. RGCANO is applied to each bootstrap sample to obtain estimates of parameters each time. We then calculate the mean and the variance-covariance of the estimates (after reflecting and permuting dimensions if necessary) across the bootstrap samples, from which we calculate estimates of standard errors of the parameter estimates, or draw confidence regions to indicate how reliably parameters are estimated. The latter is done under the assumption of asymptotic normality of the parameter estimates. When the asymptotic normality assumption is not likely to hold, we may simply plot the empirical distribution of parameter estimates. In most cases, this is sufficient to get a rough idea of how reliably parameters are estimated.

Significance tests of canonical weights and structure vectors (correlations between canonical scores and observed variables) may also be performed as byproducts of the bootstrap method described above. We count the number of times bootstrap estimates "cross" the value of zero (i.e., if the original estimate is positive, we count how many times the corresponding bootstrap estimate comes out to be negative, and vice versa.) If the relative frequency (the p-value) of "crossing" zero is smaller than a prescribed α level, we conclude that the corresponding parameter is significantly positive (or negative).

2.4. Regularized Multiple Correspondence Analysis (RMCA)

GCANO reduces to MCA when each of the K data sets consists of indicator variables (e.g., Gifi, 1990). However, there is a subtle "difference" between them that needs to be addressed. In MCA, the data are usually only columnwise centered (but not standardized). The columnwise centering, however, reduces the rank of an indicator matrix by one.

Consequently, $\mathbf{D}$ and $\mathbf{D}(\lambda)$ defined in sections 2.1 and 2.2, respectively, are bound to be rank deficient, and the choice of g-inverses of $\mathbf{D}$ and $\mathbf{D}(\lambda)$ is crucial from a computational point of view. Both MCA and RMCA use special g-inverses (Takane and Hwang, 2006).

Let us begin with the non-regularized case. Let $\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_K]$ denote a matrix of raw indicator matrices, and let $\mathbf{D}_Z$ denote a block diagonal matrix with $\mathbf{Z}_k$ as the k^{th} diagonal block. To allow missing data (zero rows in $\mathbf{Z}_k$) in our formulation, let $\mathbf{D}_{w_k}$ $(k = 1, \ldots, K)$ denote a diagonal matrix with its i^{th} diagonal element equal to 1 if the i^{th} subject has responded to item k, and 0 otherwise. (This approach for handling missing data is similar to that of van de Velden and Bijmolt (2006).) Let $\mathbf{Q}_{D_{w_k}} = \mathbf{I}_n - \mathbf{1}_n(\mathbf{1}_n'\mathbf{D}_{w_k}\mathbf{1}_n)^{-1}\mathbf{1}_n'\mathbf{D}_{w_k}$ (the orthogonal projector onto $\mathrm{Sp}(\mathbf{1}_n)$ in the metric $\mathbf{D}_{w_k}$), where $\mathbf{1}_n$ is the n-component vector of ones. Then,

$$\mathbf{X}_k = \mathbf{Q}_{D_{w_k}}'\mathbf{Z}_k = \mathbf{Z}_k\mathbf{Q}_{1_{p_k}/\tilde{D}_k}, \tag{34}$$

for $k = 1, \ldots, K$, where $\tilde{\mathbf{D}}_k = \mathbf{Z}_k'\mathbf{Z}_k$,

$$\mathbf{Q}_{1_{p_k}/\tilde{D}_k} = \mathbf{I}_{p_k} - \mathbf{1}_{p_k}(\mathbf{1}_{p_k}'\tilde{\mathbf{D}}_k\mathbf{1}_{p_k})^{-1}\mathbf{1}_{p_k}'\tilde{\mathbf{D}}_k, \tag{35}$$

and $\mathbf{1}_{p_k}$ is the p_k-component vector of ones. Matrix $\mathbf{Q}_{1_{p_k}/\tilde{D}_k}$ is the orthogonal projector onto $\mathbf{1}_{p_k}$ in the metric $\tilde{\mathbf{D}}_k$. To show (34), we simply note that $\mathbf{Z}_k\mathbf{1}_{p_k} = \mathbf{D}_{w_k}\mathbf{1}_n$, and $\mathbf{1}_n'\mathbf{D}_{w_k}\mathbf{1}_n = \mathbf{1}_{p_k}'\tilde{\mathbf{D}}_k\mathbf{1}_{p_k}$. Let $\mathbf{Q}_{D_w}'$ denote a supermatrix formed from $\mathbf{Q}_{D_{w_k}}'$ arranged side by side (i.e., $\mathbf{Q}_{D_w}' = [\mathbf{Q}_{D_{w_1}}', \ldots, \mathbf{Q}_{D_{w_K}}']$). Then, the columnwise centered data matrix is obtained by

$$\mathbf{X} = \mathbf{Q}_{D_w}'\mathbf{D}_Z = \mathbf{Z}\mathbf{Q}_{1_p/\tilde{D}}, \tag{36}$$

where $\mathbf{Q}_{1_p/\tilde{D}}$ is a block diagonal matrix with $\mathbf{Q}_{1_{p_k}/\tilde{D}_k}$ as the k^{th} diagonal block.

Let $\tilde{\mathbf{D}} = \mathbf{D}_Z'\mathbf{D}_Z$ denote the diagonal matrix with $\tilde{\mathbf{D}}_k$ as the k^{th} diagonal block, and let $\mathbf{D}$ denote the block diagonal matrix with $\mathbf{D}_k = \mathbf{X}_k'\mathbf{X}_k$ as the k^{th} diagonal block. Then,

$$\mathbf{D}_k = \tilde{\mathbf{D}}_k\mathbf{Q}_{1_{p_k}/\tilde{D}_k} = \mathbf{Q}_{1_{p_k}/\tilde{D}_k}'\tilde{\mathbf{D}}_k, \tag{37}$$

and

$$\mathbf{D} = \tilde{\mathbf{D}}\mathbf{Q}_{1_p/\tilde{D}} = \mathbf{Q}_{1_p/\tilde{D}}'\tilde{\mathbf{D}}. \tag{38}$$

To see (37), we note that $\mathbf{D}_k = \mathbf{Q}_{1_{p_k}/\tilde{D}_k}'\tilde{\mathbf{D}}_k\mathbf{Q}_{1_{p_k}/\tilde{D}_k}$, but $\tilde{\mathbf{D}}_k\mathbf{Q}_{1_{p_k}/\tilde{D}_k} = \mathbf{Q}_{1_{p_k}/\tilde{D}_k}'\tilde{\mathbf{D}}_k$ in general, and $\mathbf{Q}_{1_{p_k}/\tilde{D}_k}$ (and $\mathbf{Q}_{1_{p_k}/\tilde{D}_k}'$) are idempotent.

For the sake of generality, we do not assume that $\tilde{\mathbf{D}}$ is always nonsingular in the following discussion. That is, the existence of response categories with zero marginal frequencies is allowed. When the original data set includes such categories, they can be removed *a*

priori from all subsequent analyses. However, it is important to be able to handle such categories, since they may occur quite frequently in applications of the bootstrap methods.

A straightforward application of GCANO with columnwise centered data requires

$$\text{GSVD}(\mathbf{ZQ}_{1_p/\tilde{D}}\mathbf{D}^{-})_{I_n,\,D}, \tag{39}$$

whereas in MCA we typically obtain

$$\text{GSVD}(\mathbf{ZQ}_{1_p/\tilde{D}}\tilde{\mathbf{D}}^{+})_{I_n,\,\tilde{D}}, \tag{40}$$

where $\tilde{\mathbf{D}}^{+}$ indicates the Moore-Penrose g-inverse of $\tilde{\mathbf{D}}$. The latter is solved by first post-multiplying $\mathbf{ZQ}_{1_p/\tilde{D}}\tilde{\mathbf{D}}^{+}$ by $\tilde{\mathbf{D}}^{1/2}$, that is, $\mathbf{ZQ}_{1_p/\tilde{D}}\tilde{\mathbf{D}}^{+}\tilde{\mathbf{D}}^{1/2} = \mathbf{ZQ}_{1_p/\tilde{D}}(\tilde{\mathbf{D}}^{+})^{1/2}$, whose ordinary SVD is then obtained. Let this SVD be denoted by $\mathbf{ZQ}_{1_p/\tilde{D}}(\tilde{\mathbf{D}}^{+})^{1/2} = \tilde{\mathbf{F}}\tilde{\mathbf{\Delta}}\tilde{\mathbf{W}}'$. Then $\text{GSVD}(\mathbf{ZQ}_{1_p/\tilde{D}}\tilde{\mathbf{D}}^{+})_{I_n,\,\tilde{D}}$, denoted by $\mathbf{F}^{*}\mathbf{\Delta}^{*}\mathbf{W}^{*\prime}$, is obtained by $\mathbf{F}^{*} = \tilde{\mathbf{F}}$, $\mathbf{\Delta}^{*} = \tilde{\mathbf{\Delta}}$, and $\mathbf{W}^{*} = \tilde{\mathbf{W}}(\tilde{\mathbf{D}}^{+})^{1/2}$. This is equivalent to making the following sequence of choices in solving (39):

(a) Take $\tilde{\mathbf{D}}^{+}$ as a g-inverse of $\mathbf{D}$ and obtain $\text{GSVD}(\mathbf{ZQ}_{1_p/\tilde{\mathbf{D}}}\mathbf{D}^{+})_{I_n,\,D}$. (That $\tilde{\mathbf{D}}^{+}$ is a g-inverse of $\mathbf{D}$ can easily be shown by $\mathbf{D}\tilde{\mathbf{D}}^{+}\mathbf{D} = \mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}\tilde{\mathbf{D}}^{+}\tilde{\mathbf{D}}\mathbf{Q}_{1_p/\tilde{D}} = \mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}\mathbf{Q}_{1_p/\tilde{D}}$ $= \tilde{\mathbf{D}}\mathbf{Q}_{1_p/\tilde{D}} = \mathbf{D}$ due to (38).)

(b) Take $\mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}^{1/2}$ as a square root factor of the metric matrix $\mathbf{D}$, where $\tilde{\mathbf{D}}^{1/2}$ is the symmetric square root factor of $\tilde{\mathbf{D}}$. By Theorem 1 in Appendix (B), postmultiplying $\mathbf{ZQ}_{1_p/\tilde{D}}\tilde{\mathbf{D}}^{+}$ by $\mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}^{1/2}$ leads to $\mathbf{ZQ}_{1_p/\tilde{D}}(\tilde{\mathbf{D}}^{+})^{1/2}$, whose SVD we obtain. Note that postmultiplying $\mathbf{ZQ}_{1_p/\tilde{D}}\tilde{\mathbf{D}}^{+}$ by $\mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}^{1/2}$ has the same effect as postmulitplying the former by merely $\tilde{\mathbf{D}}^{1/2}$. (That $\mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}^{1/2}$ is a square root factor of $\mathbf{D}$ can easily be shown by $\mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}^{1/2}\tilde{\mathbf{D}}^{1/2}\mathbf{Q}_{1_p/\tilde{D}} = \mathbf{D}$ due to (38).)

(c) Take $(\tilde{\mathbf{D}}^{+})^{1/2}$ as a g-inverse of $\mathbf{D}^{1/2} = \mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}^{1/2}$. The SVD of $\mathbf{ZQ}_{1_p/\tilde{D}}(\tilde{\mathbf{D}}^{+})^{1/2}$ obtained above is postmultiplied by $(\tilde{\mathbf{D}}^{+})^{1/2}$ to obtain $\text{GSVD}(\mathbf{ZQ}_{1_p/\tilde{D}}\tilde{\mathbf{D}}^{+})_{I,\,\tilde{D}}$. (That $(\tilde{\mathbf{D}}^{+})^{1/2}$ is a g-inverse of $\mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}^{1/2}$ can easily be shown by $\mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}^{1/2}(\tilde{\mathbf{D}}^{+})^{1/2}\mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}^{1/2} = \mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}^{1/2}$.)

Note that the choices of solutions in the above steps are not necessarily unique, but they are chosen to make the solution to (39) equivalent to that of (40).

An important question is if an analogous relationship holds for regularized MCA. The answer is "yes", as shown below. Let $\tilde{\mathbf{D}}(\lambda) = \tilde{\mathbf{D}} + \lambda\mathbf{J}_p$, where $\mathbf{J}_p$ is, as defined earlier, a block diagonal matrix with $\mathbf{J}_{p_k} = \mathbf{X}'_k(\mathbf{X}_k\mathbf{X}'_k)^{-}\mathbf{X}_k$ as the k^{th} diagonal block. Then,

$$\mathbf{D}(\lambda) = \tilde{\mathbf{D}}(\lambda)\mathbf{Q}_{1_p/\tilde{D}} = \mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}(\lambda), \tag{41}$$

since $\mathbf{Q}'_{1_p/\tilde{D}}\mathbf{J}_p = \mathbf{J}_p\mathbf{Q}_{1_p/\tilde{D}} = \mathbf{J}_p$. A straightforward application of (15) with columnwise centered data requires

$$\text{GSVD}(\mathbf{T}'\mathbf{D}_Z\mathbf{Q}_{1_p/\tilde{D}}\mathbf{D}(\lambda)^-)_{I_n,\, D(\lambda)}, \tag{42}$$

whereas RMCA typically obtains (Takane and Hwang, 2006)

$$\text{GSVD}(\mathbf{T}'\mathbf{D}_Z\mathbf{Q}_{1_p/\tilde{D}}\tilde{\mathbf{D}}(\lambda)^+)_{I_n,\, \tilde{D}(\lambda)}, \tag{43}$$

where $\tilde{\mathbf{D}}(\lambda)^+$ is the Moore-Penrose g-inverse of $\tilde{\mathbf{D}}(\lambda)$. The two GSVD problems can be made equivalent by making the following sequence of choices in solving (42):

(a) Take $\tilde{\mathbf{D}}(\lambda)^+$ as a g-inverse of $\mathbf{D}(\lambda)$ and obtain GSVD($\mathbf{T}'\mathbf{D}_Z\mathbf{Q}_{1_p/\tilde{\mathbf{D}}}\mathbf{D}(\lambda)^+$). (That $\tilde{\mathbf{D}}(\lambda)^+$ is a g-inverse of $\mathbf{D}(\lambda)$ can be shown by $\mathbf{D}(\lambda)\tilde{\mathbf{D}}(\lambda)^+\mathbf{D}(\lambda) = \mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}(\lambda)\tilde{\mathbf{D}}(\lambda)^+\tilde{\mathbf{D}}(\lambda)\mathbf{Q}_{1_p/\tilde{D}} = \mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}(\lambda)\mathbf{Q}_{1_p/\tilde{D}} = \tilde{\mathbf{D}}(\lambda)\mathbf{Q}_{1_p/\tilde{D}} = \mathbf{D}(\lambda)$ due to (41).)

(b) Take $\mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}(\lambda)^{1/2}$ as a square root factor of $\mathbf{D}(\lambda)$. By Theorem 2 in Appendix (B), postmultiplying $\mathbf{T}'\mathbf{D}_Z\mathbf{Q}_{1_p/\tilde{D}}\tilde{\mathbf{D}}(\lambda)^+$ by $\mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}(\lambda)^{1/2}$ leads to $\mathbf{T}'\mathbf{D}_Z\mathbf{Q}_{1_p/\tilde{D}}(\tilde{\mathbf{D}}(\lambda)^+)^{1/2}$, whose SVD we obtain. Note that postmultiplying $\mathbf{T}'\mathbf{D}_Z\mathbf{Q}_{1_p/\tilde{D}}\tilde{\mathbf{D}}(\lambda)^+$ by $\mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}(\lambda)^{1/2}$ has the same effect as postmultiplying the former by merely $\tilde{\mathbf{D}}(\lambda)^{1/2}$. (That $\mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}(\lambda)^{1/2}$ is a square root factor of $\mathbf{D}(\lambda)$ can be shown by $\mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}(\lambda)^{1/2}\tilde{\mathbf{D}}(\lambda)^{1/2}\mathbf{Q}_{1_p/\tilde{D}} = \mathbf{D}(\lambda)$ due to (41).)

(c) Take $(\tilde{\mathbf{D}}(\lambda)^+)^{1/2}$ as a g-inverse of $\mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}(\lambda)^{1/2}$. The SVD of $\mathbf{T}'\mathbf{D}_Z\mathbf{Q}_{1_p/\tilde{D}}(\tilde{\mathbf{D}}(\lambda)^+)^{1/2}$ is postmultiplied by $(\tilde{\mathbf{D}}(\lambda)^+)^{1/2}$ to obtain GSVD($\mathbf{T}'\mathbf{D}_Z\mathbf{Q}_{1_p/\tilde{D}}\tilde{\mathbf{D}}(\lambda)^+)_{I,\, \tilde{D}(\lambda)}$. (That $(\tilde{\mathbf{D}}(\lambda)^+)^{1/2}$ is a g-inverse of $\mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}(\lambda)^{1/2}$ can be shown by $\mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}(\lambda)^{1/2}(\tilde{\mathbf{D}}(\lambda)^+)^{1/2}\mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}(\lambda)^{1/2} = \mathbf{Q}'_{1_p/\tilde{D}}\tilde{\mathbf{D}}(\lambda)^{1/2}$.)

Again, the choices in the above steps make the solution to (42) equivalent to that of (43).

3. Some Numerical Examples

In this section we report some results on applications of RGCANO. We first present a simple demonstration of the effect of regularization using a Monte Carlo technique. We then discuss four applications of RGCANO to illustrate practical uses of the method. The first two of these analyze continuous data (applications of RGCANO proper), while the last two analyze multiple-choice data (applications of GCANO for RMCA). In none of these examples the rank additivity condition (30) holds, requiring the generalized formulation developed in this paper.

3.1. A Monte Carlo Study

In this demonstration a number of data sets were generated from a population GCANO model. (This is somewhat different from the Monte Carlo study conducted for RMCA by Takane and Hwang (2006), where no population model existed according to which the data could be generated. Instead, multiple-choice data with a very large sample size was taken as the population data from which a number of data sets were sampled.) They were then analyzed by RGCANO to examine the effect of regularization on the estimates of parameters. Since "true" population values are known in this case, MSE can be directly calculated as a function of the regularization parameter.

The population model used was as follows. First, it was assumed that there were three sets of variables, and that the first set had 3 variables, the second set 4 variables and the third set 5 variables ($p = 12$). A model with two canonical variates was then postulated by assuming

$$\mathbf{A}' = \left[\begin{array}{ccc|cccc|ccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} & -\frac{3}{\sqrt{20}} & -\frac{1}{\sqrt{20}} & \frac{1}{\sqrt{20}} & \frac{3}{\sqrt{20}} & -\frac{2}{\sqrt{10}} & -\frac{1}{\sqrt{10}} & 0 & \frac{1}{\sqrt{10}} & \frac{2}{\sqrt{10}} \end{array}\right],$$

from which a population covariance matrix was generated by

$$\mathbf{\Sigma} = a\mathbf{A}\mathbf{A}' + b\mathbf{I},$$

where both a and b were further assumed to be unity. (Note that columns of $\mathbf{A}$ consist of a constant vector and a vector of linear trend coefficients within each subset of variables. There was no strong reason to postulate this particular $\mathbf{A}$, except that we wanted a population covariance matrix with two canonical variates.) The population parameters ($\mathbf{W}$) in GCANO can be derived from the generalized eigen decomposition of $\mathbf{\Sigma}$ with respect to $\mathbf{D}_\Sigma$, where $\mathbf{D}_\Sigma$ is a block diagonal matrix formed from diagonal blocks of $\mathbf{\Sigma}$ of order 3, 4, and 5. (This decomposition has exactly 2 eigenvalues larger than one.) A number of data matrices (100 within a particular sample size), each row following the p-variate normal distribution with mean 0 and covariance matrix $\mathbf{\Sigma}$, were generated and subjected to GCANO with the value of λ systematically varied ($\lambda = 0$, 10, 50, 100, 200, 400). This was repeated for different sample sizes ($n = 50$, 100, 200, and 400).

Figure 1 depicts MSEs as a function of the sample size and the value of the ridge parameter. The MSE was calculated according to (1). As can be seen, the MSE decreases as soon as λ departs from 0, but begins to increase after a while. The minimum value of MSE occurs somewhere in the middle. This tendency is clearer for small sample sizes, but can still be observed for a sample size as large as 400. Figure 2 breaks down the MSE function for $n = 200$ into its two constituents, squared bias and variance. The squared bias consistently increases as the value of λ increases, while the variance decreases. The sum

of the two (MSE) takes its smallest value in the mid range. Figures 1 and 2 are similar to those derived theoretically by Hoerl and Kennard (1970) in multiple regression analysis, and it is reassuring to find that essentially the same holds for GCANO as well. (See also Takane and Hwang, 2006.) We also tried a number of parameter combinations within the two-component model, and the one-component model to assess the generality of the above results. We obtained essentially the same results.

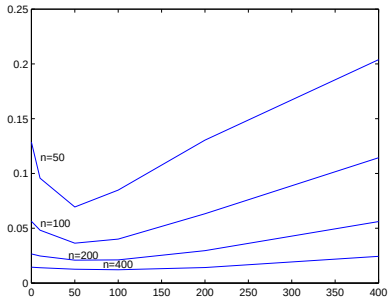

FIGURE 1.
MSE as a function of the regularization parameter (λ) and sample size (n).

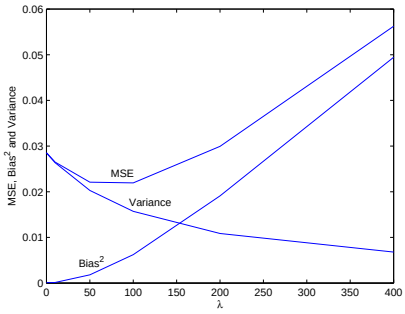

FIGURE 2.
MSE, Squared Bias and Variance as a function of the regularization parameter (λ) for $n = 200$.

3.2. The First Example of Application

We first analyze the data set given in Table 1. We refer the reader to the introduction section for a description of the data and motivations for data analysis. Permutation tests indicated that there was only one significant component ($p < .0005$ for the first component, and $p > .15$ for the second component, both with $\lambda = 10$.) The six-fold cross validation found that an "optimal" value of λ was 10. The cross validation index (ϵ) was .755 for $\lambda = 10$, and .781 for $\lambda = 0$.

The top portion of Table 2 provides the estimates of weights (applied to the observed variables to derive the canonical component) and the correlations between the canonical component and the observed variables for non-regularized GCANO (columns 3 and 4) and RGCANO (columns 5 and 6). It can be observed that the weights (and the correlations) obtained by RGCANO tend to be closer to zero than their non-regularized counter-parts, indicating the shrinkage effect of regularization. The weights as well as the correlations indicate that this component represents the type of material used to store the wine. It is highly positively correlated with "woodiness" for all three judges, and negatively with "fruitiness". The bottom portion of the table give canonical scores for the six wines on the derived component. The wines are grouped into two groups, wines 1, 5, and 6 on

the positive side, and wines 2, 3, and 4 on the negative side, which coincide with the two different types of oak to store the wines indicated in the second column of Table 1. Patterns of canonical scores are similar in both non-regularized and regularized cases. However, the standard errors given in parentheses indicate that the estimates of canonical scores in the latter are much more reliable than in the former. (The former are at least 10 times as large.) The overall size of the canonical scores was equated across the two estimation methods to ensure that the smaller standard errors in RGCANO are not merely due to smaller sizes of the estimates. (This roughly has the effect of equating the degree of bias across the two methods.)

TABLE 2.
Analysis of wine tasting data in Table 1 by non-regularized and regularized GCANO.

Column	Scale	Non-regularized		Regularized	
		Weight	Correlation	Weight	Correlation
Expert 1	fruity	-.415	-.993	-.264	-.890
	woody	.500	.996	.275	.902
	coffee	.097	.926	.225	.837
Expert 2	red fruit	-.298	-.928	-.174	-.824
	roasted	-.187	.928	.205	.872
	vanillin	.490	.974	.186	.873
	woody	.443	.947	.222	.884
Expert 3	fruity	.267	-.447	-.067	-.511
	butter	.218	.884	.270	.856
	woody	.943	.982	.324	.903
Row		Score	Std. Error	Score	Std. Error
Wine 1		.597	(.418)	.539	(.024)
Wine 2		-.142	(.366)	-.133	(.009)
Wine 3		-.457	(.523)	-.540	(.040)
Wine 4		-.514	(.616)	-.466	(.032)
Wine 5		.351	(.292)	.341	(.016)
Wine 6		.165	(.313)	.258	(.025)

3.3. The Second Example

The data to be analyzed in this section are similar to the wine tasting data in the previous section, but on a much larger scale (Tillman, Dowling, and Abdi, in preparation). Twenty six judges rated twenty one music pieces on six rating scales. The 21

music pieces were sampled from works by three composers (seven pieces each): 1. Bach, 2. Beethoven, and 3. Mozart. (See Table 3 for more detailed descriptions.) After hearing each piece for ten seconds, each judge rated the piece on the following six bipolar scales: 1. simple/complex, 2. fierce/gentle, 3. positive/negative, 4. elegant/plain, 5. regular/irregular, and 6. jagged/smooth. (The underlined symbols are used as plotting symbols in Figure 3.) Each rating scale had eight response categories numbered from 1 to 8. A larger number indicated a more complex, more gentle, more negative, plainer, more regular, and smoother piece.

TABLE 3.
Description of the 21 music pieces.

No.	Composer	Symbol	Description
1	Bach 1	11	A† - English Suite No. 1, Guigue 806
2	Bach 2	12	Bb - Partitas No. 1, BWV 825
3	Bach 3	13	C - Three-Part Invention, BWV 787
4	Bach 4	14	C minor - French Suite No. 2, BWV 813
5	Bach 5	15	D - Prelude No. 5, BWV 850 (Well-tempered Piano I)
6	Bach 6	16	F - Little Fugue, BWV 556
7	Bach 7	17	G - French Suite No. 5, BWV 816
8	Beethoven 1	21	A - Sonata K331, Allegro
9	Beethoven 2	22	Bb - Sonata K281, Allegro
10	Beethoven 3	23	C - Sonata K545, Allegro
11	Beethoven 4	24	C minor - Sonata K457, Allegro assai
12	Beethoven 5	25	D - Sonata K576, Allegro
13	Beethoven 6	26	F - Sonata K280, Allegro
14	Beethoven 7	27	G - Sonata K283, Allegro
15	Mozart 1	31	A - Sonata No. 2, Op. 2, Allegro
16	Mozart 2	32	Bb - Sonata No. 11, Op. 22
17	Mozart 3	33	C - Sonata in C, Op. 21, Allegro con brio
18	Mozart 4	34	C minor - Sonata No. 5, Op. 10 No. 1, Allegro
19	Mozart 5	35	D - Sonata No. 7, Op. 10, Presto
20	Mozart 6	36	F - Sonata No. 6, Op. 10 No. 2
21	Mozart 7	37	G - Sonata No. 10, Op. 14, Allegro.

†A capital letter indicates key, and a lower case "b" a bemol (flat).
All are major except those explicitly designated as minor.

Our interest in applying GCANO in this case is to find a representation of the 21

music pieces most representative of the 26 judges. We therefore mainly focus on canonical scores ($\mathbf{F}_1$) that indicate the spatial locations of the pieces in relation to each other. Permutation tests found that the first two components were clearly significant (both $p <$ .0005), while the third one was on the borderline. The third component was not significant for the non-regularized case with a p-value of .120, while it was marginally significant ($p = .021$) with the value of $\lambda = 10$. For ease of comparison, however, we adopted a two-dimensional solution for both cases. (The third component was also rather difficult to interpret, because none of the six scales was correlated highly with this component.) The 21-fold cross validation found that an optimal value of the ridge parameter was 10. The cross validation index (ϵ) was .926 for $\lambda = 10$ compared to .931 for $\lambda = 0$. The difference is rather small, but this is due to the fact that this data set is fairly large (21 by 6 by 26).

Figure 3 displays the two-dimensional configuration of the 21 music pieces from RGCANO (the plot of $\mathbf{F}_1$). Music pieces are indicated by number pairs, the first one indicating the composer number followed by the piece number within a composer. For example, 23 indicates Beethoven's piece number 3. Works by the same composer loosely form clusters; those composed by Bach tend to be located toward the upper left corner, those by Beethoven toward the right, and those by Mozart toward the lower left corner. This is seen from the convex hulls (shown by connected line segments) drawn to enclose the works by the three composers separately.

Mean ratings of the 26 judges on the six scales were mapped into the configuration as vectors. (The mean ratings can be taken in this example because all judges used the same six scales. This is in contrast to the previous example in which different sets of attributes were used by different judges.) Six arrows indicate the directions with which the mean ratings on the six rating scales are most highly correlated. Bach's compositions were rated plainer and more negative, Beethoven's works more gentle and smoother, and Mozart's pieces more irregular and complex. The numbers in parentheses indicate the multiple correlations, which are fairly large in all cases.

The bootstrap method was used to assess the reliability of the estimated locations of the music pieces. (The number of bootstrap samples was set to 1000.) Figure 4 displays the two-dimensional configuration of music pieces obtained by the non-regularized GCANO along with the 95% confidence regions. The variabilities of point locations are fairly large in almost all cases. This is partly due to the fact that the 26 judges varied considerably in their ratings. Figure 5, on the other hand, displays the same (as Figure 4) but that derived by RGCANO. The confidence regions derived by RGCANO are almost uniformly smaller than those derived by non-regularized GCANO. Again, the configuration in Figure 5 has been resized to match the size of Figure 4 to ensure that the tighter confidence regions in the former were not merely due to the shrunken configuration size. (As before, this

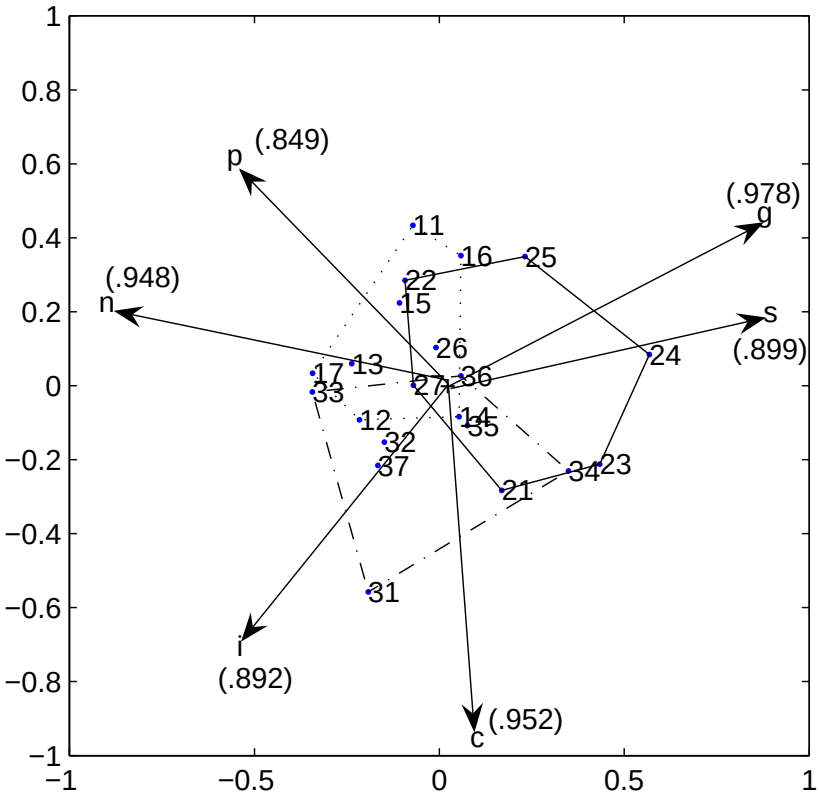

FIGURE 3.
Regularized GCANO with the musical piece data along with vectors representing the directions with which average ratings on six scales are most highly correlated

adjustment may be seen as a bias correction.)

3.4 The Third Example

The third example pertains to a small multiple-choice data set used by Maraun, Slaney, and Jalava (2005) to illustrate the use of MCA. Two groups of subjects (5 depressed inpatients, and 5 university undergraduates) responded to four items of the Beck Depression Inventory (BDI; Beck, 1996). The items and the response categories used were: 1. Item 1 – Sadness (1: I do not feel sad, 2: I feel sad much of the time, 3: I am sad all the time, 4: I am so sad or unhappy that I can't stand it). 2. Item 4 – Loss of pleasure (1: I get as much pleasure as I ever did from the things I enjoy, 2: I don't enjoy things as much as I used to, 3: I get very little pleasure from the things I used to enjoy, 4: I can't get any pleasure from the things I used to enjoy). 3. Item 7 – Self-dislike (1: I feel the same about

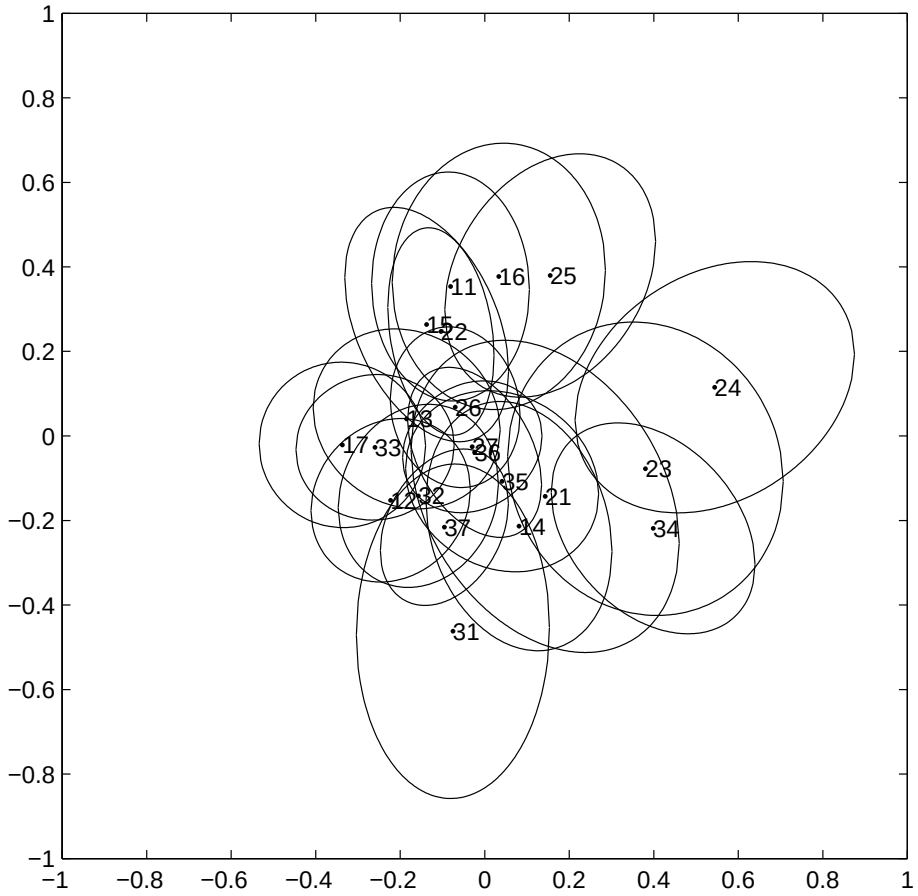

FIGURE 4.
Two-dimensional stimulus configuration from the musical piece data obtained by non-regularized GCANO along with 95% confidence regions

myself as ever, 2: I have lost confidence about myself, 3: I am disappointed about myself, 4: I dislike myself). 4. Item 21 – Loss of interest in sex (1: I have not noticed any recent change in my interest in sex, 2: I am less interested in sex than I used to be, 3: I am much less interested in sex now, 4: I have lost interest in sex completely). Although the response options are roughly ordinal, they were treated as nominal for the purpose of MCA.

Permutation tests indicated that there was only one significant component ($p < .0005$ for the first component and $p > .15$ for the second component with $\lambda = 1$). The 10-fold cross validation indicated that the optimal value of λ was 1. The value of ϵ was .240 for $\lambda = 1$, and .468 for $\lambda = 0$. Table 4 shows that the overall patterns of estimates remain the same across the non-regularized and the regularized cases. The derived component represents the degree of depression with the negative side indicating more serious depression. (For category 3 of item 7 the weight estimate is 0 with 0 variance. This is because no respondents

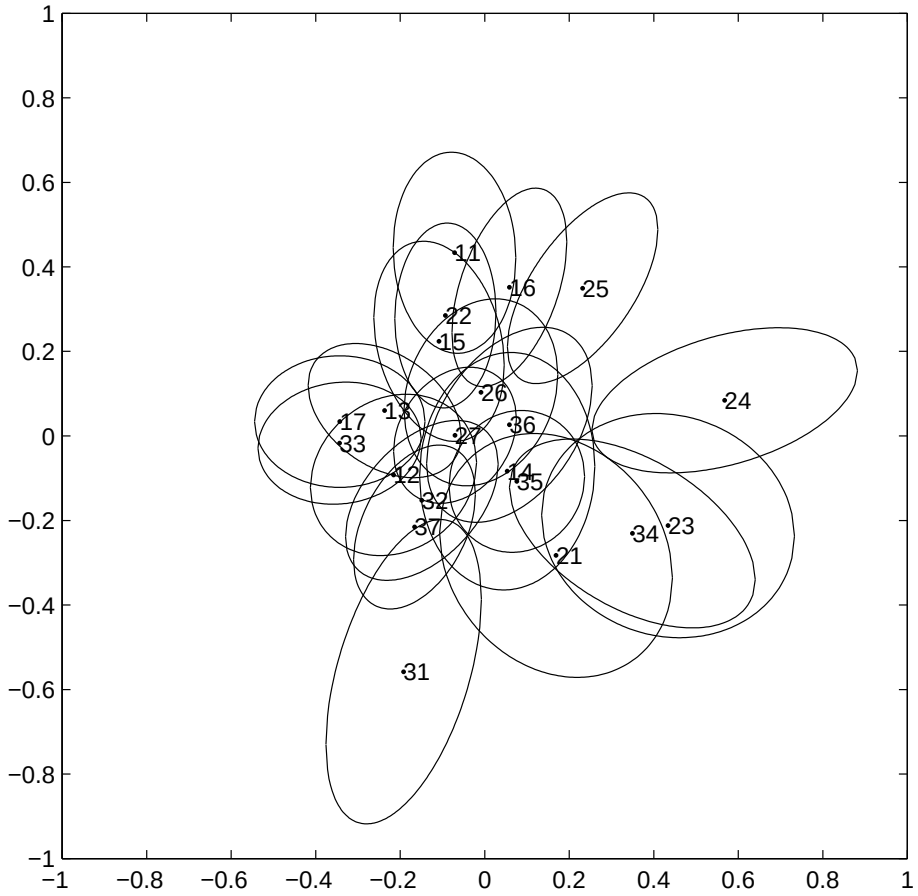

FIGURE 5.
Two-dimensional stimulus configuration from the musical piece data obtained by regularized GCANO along with 95% confidence regions

chose this category.) Scores of depressed inpatients tend to be on the negative side (they are also more variable), while those of university undergrads on the positive side. Standard errors were almost uniformly smaller in the regularized case. (The regularized estimates were again scaled up to match the size of the non-regularized estimates.)

3.5 The Fourth Example

The fourth and final example concerns the analysis of sorting data by MCA (Takane, 1980). Ten university students sorted 29 "Have" words into as many groups as they liked in terms of the similarity in meaning. The 29 Have words were: 1. Accept, 2. Beg, 3. Belong, 4. Borrow, 5. Bring, 6. Buy, 7. Earn, 8. Find, 9. Gain, 10. Get, 11. Get rid of, 12. Give, 13. Have, 14. Hold, 15. Keep, 16. Lack, 17. Lend, 18. Lose, 19. Need, 20. Offer, 21. Own, 22. Receive, 23. Return, 24. Save, 25. Sell, 26. Steal, 27. Take, 28. Use, and 29. Want. The

TABLE 4.
Analysis of Maraun et al.,'s (2005) data by regularized GCANO (MCA).

BDI	Category	Non-regularized		Regularized	
		Weight	Std. Error	Weight	Std. Error
Item 1	1	.905	(.298)	.827	(.223)
	2	.540	(.259)	.321	(.219)
	3	.671	(.370)	.390	(.243)
	4	-1.149	(.285)	-1.537	(.223)
Item 4	1	.813	(.399)	1.086	(.313)
	2	.983	(.660)	.732	(.452)
	3	-1.586	(.804)	-1.128	(.663)
	4	-.663	(.537)	-.691	(.463)
Item 7	1	.992	(.447)	.943	(.361)
	2	.449	(.360)	.580	(.322)
	3	.000	(.000)	.000	(.000)
	4	-1.410	(.385)	-1.523	(.341)
Item 21	1	.797	(.274)	.843	(.236)
	2	.866	(.472)	.682	(.417)
	3	.088	(.387)	.065	(.330)
	4	-1.410	(.411)	-1.589	(.349)
	Subj.	Score	Std. Error	Score	Std. Error
Depressed	1	-1.322	(.575)	-1.374	(.499)
Inpatients	2	.358	(.320)	.345	(.322)
	3	-1.322	(.575)	-1.374	(.499)
	4	-1.586	(.558)	-1.490	(.476)
	5	-.364	(.414)	-.374	(.407)
University	1	.983	(.495)	.820	(.378)
Undergrads	2	.846	(.360)	.934	(.352)
	3	.540	(.323)	.592	(.341)
	4	1.001	(.434)	1.030	(.388)
	5	.866	(.385)	.891	(.346)

sorting data are a special case of multiple-choice data with rows of the table representing stimuli, while columns representing sorting clusters elicited by the subjects.

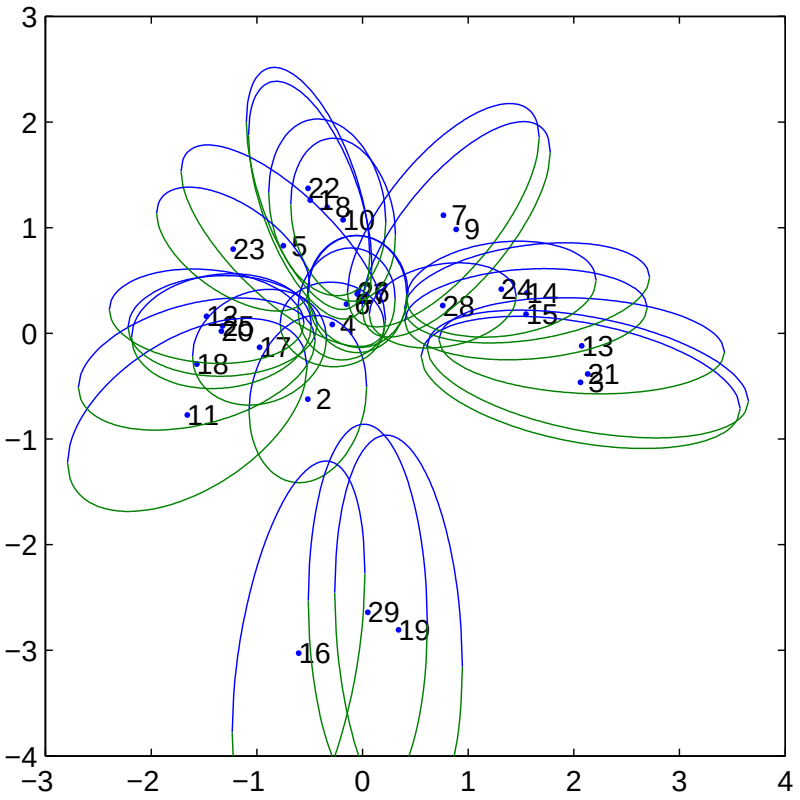

FIGURE 6.
Two-dimensional stimulus configuration for 29 have words obtained by non-regularized MCA along with 95% confidence regions

Permutation tests indicated that there were seven significant components ($p < .0005$ up to the seventh component, and $p > .85$ for the eighth component with $\lambda = 1$). However, for ease of presentation we adopt a two-dimensional solution defined by the two most dominant components. The 29-fold cross validation found the optimal value of λ was 1. The value of ϵ was .593 for $\lambda = 1$, and .906 for $\lambda = 0$. Figures 6 and 7 display the two-dimensional stimulus configurations by non-regularized and regularized MCA, respectively, along with 95% confidence regions for the point locations. The configurations are similar in both cases. We find verbs such as 13. Have, 21. Own, 3. Belong, 15. Keep, etc. on the left, while 12. Give, 25. Sell, 11. Get rid of, 18. Lose, etc. on the right. At the bottom, we see 16. Lack, 19. Need, and 29. Want, while at the top, 22. Receive, 10. Get, 7. Earn,

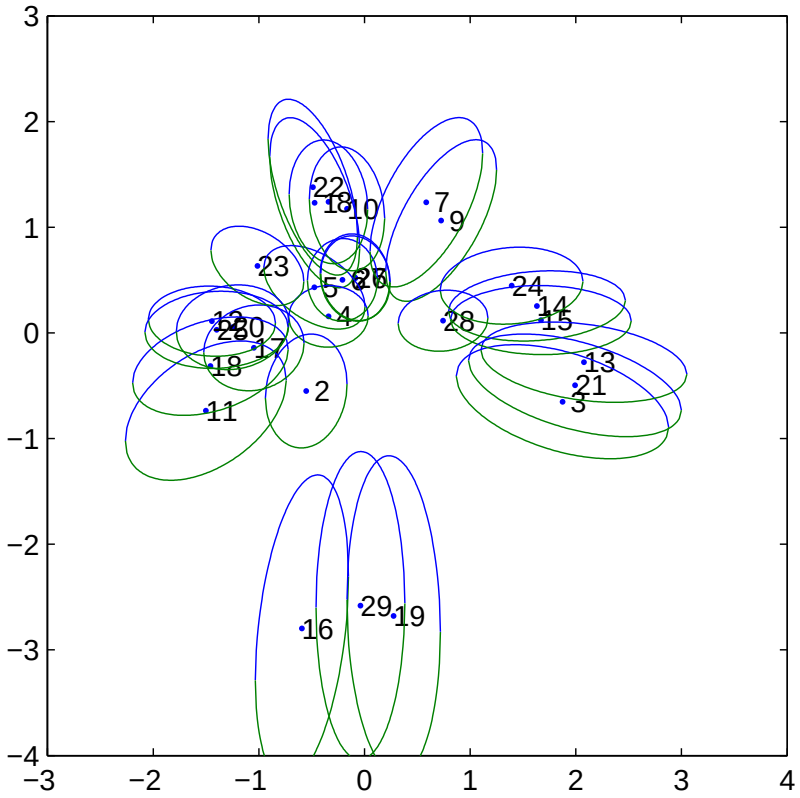

FIGURE 7.
Two-dimensional stimulus configuration for 29 have words obtained by regularized GCANO along with 95% confidence regions

8. Find, etc. We may interpret dimension 1 (the horizontal direction) as contrasting two states of possession, stable on the left and unstable on the right, while dimension 2 (the vertical direction) contrasting the two states of nonpossession, stable at the bottom and unstable at the top. Although the configurations are similar, confidence regions are almost uniformly smaller in the regularized case.

Concluding Remarks

We presented a simple regularization technique for multiple-set canonical correlation analysis (RGCANO). This technique is a straightforward extension of the ridge estimation in regression analysis (Hoerl and Kennard, 1970) to multiple-set canonical correlation analysis (GCANO). We outlined the theory for regularized GCANO and extended it to multiple correspondence analysis (RMCA: Takane and Hwang, 2006). We demonstrated

the usefulness of RGCANO using both a Monte Carlo method and actual data sets. The regularization technique similar to the one presented here may be incorporated in many other multivariate data analysis methods. Redundancy analysis (Takane and Hwang, 2007; Takane and Jung, 2006), canonical discriminant analysis (DiPillo, 1976; Friedman, 1989), PCA, hierarchical linear models (HLM), logistic discrimination, generalized linear models, log-linear models, and structural equation models (SEM) are but a few examples of the techniques in which regularization might be useful.

GCANO (Multiple-set CANO) has not been used extensively in data analysis so far. Hardly any use of it has been reported in the literature, except in food sciences (Dahl and Næs, 2006; see, however, Devaux, et al. (1998), Fischer, et al. (2007), Sun, et al. (2005).), and in the special cases of GCANO, such as MCA and an optimal scaling approach to nonlinear multivariate analysis (Gifi, 1990). The point of view given in the introduction section that it can be viewed as a method of information integration from multiple sources may broaden the scope of GCANO and generate more applications in the future.

Incorporating prior knowledge is essential in many data analyses. Information obtained from the data is never sufficient and must be supplemented by prior information. In regression analysis, for example, the regression curve (the conditional expectation of y on X) is estimated for the entire range of X based on a finite number of observations. In linear regression analysis, this is made possible by the assumption that the regression curve is linear at least within the range of X of our interest. Regularization provides one way of incorporating prior knowledge in data analysis.

Appendix

(A): RGCANO when the rank additivity condition (30) holds

We show that the formulation of RGCANO presented in section 2.2 reduces to that of Takane and Hwang (2006) developed under the assumption of rank additivity (30). Under this condition, a solution of RGCANO was obtained by

$$\text{GSVD}(\mathbf{XD}(\lambda)^{-})_{M(\lambda),\,D(\lambda)}, \tag{44}$$

where

$$\mathbf{M}(\lambda) = \mathbf{J}_n + \lambda(\mathbf{XX}')^{+} \tag{45}$$

is called the ridge metric matrix (Takane and Yanai, 2008). Here, $\mathbf{J}_n$ is any matrix such that $\mathbf{X}'\mathbf{J}_n\mathbf{X} = \mathbf{X}'\mathbf{X}$ (e.g., $\mathbf{J}_n = \mathbf{I}_n$, $\mathbf{J}_n = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$, etc.). In this appendix, it will be shown that (15) reduces to (44) if (30) holds.

Let (17) be the solution to (15), where $\mathbf{F}^{*\prime}\mathbf{F}^{*} = \begin{bmatrix} \mathbf{F}_1^{*} \\ \mathbf{F}_2^{*} \end{bmatrix}' \begin{bmatrix} \mathbf{F}_1^{*} \\ \mathbf{F}_2^{*} \end{bmatrix} = \mathbf{I}$. Then, obviously

$$\mathbf{X}\mathbf{D}(\lambda)^{-} = \mathbf{F}_1^{*}\mathbf{\Delta}^{*}\mathbf{W}^{*\prime} \tag{46}$$

holds. We will show that under (30) this is the solution to (44). That is,

$$\mathbf{F}_1^{*\prime}\mathbf{M}(\lambda)\mathbf{F}_1^{*} = \mathbf{I}. \tag{47}$$

Let

$$\mathbf{M}(\lambda) = \mathbf{C}\mathbf{C}', \tag{48}$$

where

$$\mathbf{C} = \begin{bmatrix} \mathbf{J}_n & \lambda^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{X})^{+} \end{bmatrix}. \tag{49}$$

Note that $(\mathbf{X}\mathbf{X}')^{+} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{+2}\mathbf{X}'$. When (30) holds,

$$\mathbf{C}'\mathbf{X}\mathbf{D}(\lambda)^{-} = \begin{bmatrix} \mathbf{X} \\ \lambda^{1/2}(\mathbf{X}'\mathbf{X})^{+}\mathbf{X}'\mathbf{X} \end{bmatrix}\mathbf{D}(\lambda)^{-} = \begin{bmatrix} \mathbf{X} \\ \lambda^{1/2}\mathbf{J}_p \end{bmatrix}\mathbf{D}(\lambda)^{-} = \mathbf{T}'\mathbf{D}_X\mathbf{D}(\lambda)^{-}. \tag{50}$$

(Note that $(\mathbf{X}'\mathbf{X})^{+}\mathbf{X}'\mathbf{X} = \mathbf{P}_{X'}$ which in turn is equal to $\mathbf{J}_p$ if and only if (30) holds (Takane and Yanai, 2008).) This leads to

$$\mathbf{M}(\lambda)\mathbf{F}_1^{*} = \mathbf{C}\begin{bmatrix} \mathbf{F}_1^{*} \\ \mathbf{F}_2^{*} \end{bmatrix}. \tag{51}$$

We thus have

$$\mathbf{F}_1^{*\prime}\mathbf{M}(\lambda)\mathbf{F}_1^{*} = \begin{bmatrix} \mathbf{F}_1^{*} \\ \mathbf{F}_2^{*} \end{bmatrix}'\mathbf{C}'\mathbf{M}(\lambda)^{+}\mathbf{C}\begin{bmatrix} \mathbf{F}_1^{*} \\ \mathbf{F}_2^{*} \end{bmatrix} = \mathbf{I}, \tag{52}$$

since $\mathrm{Sp}(\mathbf{F}^{*}) \subset \mathrm{Sp}(\mathbf{C}')$.

From (51), $\mathbf{F}_1^{*} + \lambda(\mathbf{X}\mathbf{X}')^{+}\mathbf{F}_1^{*} = \mathbf{F}_1^{*} + \lambda^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{X})^{+}\mathbf{F}_2^{*}$, from which it follows that

$$\mathbf{F}_2^{*} = \lambda^{1/2}(\mathbf{X}'\mathbf{X})^{+}\mathbf{X}'\mathbf{F}_1^{*}, \tag{53}$$

indicating $\mathrm{Sp}(\mathbf{F}_2^{*}) \subset \mathrm{Sp}(\mathbf{X}')$, or

$$\mathbf{F}_1^{*} = \lambda^{-1/2}\mathbf{X}\mathbf{F}_2^{*}, \tag{54}$$

indicating $\mathrm{Sp}(\mathbf{F}_1^{*}) \subset \mathrm{Sp}(\mathbf{X})$. (The latter always holds as given in (18), while the former holds only when the rank additivity holds.)

By (53) we obtain

$$\mathbf{X}\mathbf{D}(\lambda)^{-}\mathbf{X}'\mathbf{F}_1^{*} + \lambda\mathbf{X}\mathbf{D}(\lambda)^{-}(\mathbf{X}'\mathbf{X})^{+}\mathbf{X}'\mathbf{F}_1^{*} = \mathbf{X}\mathbf{D}(\lambda)^{-}\mathbf{X}'\mathbf{M}(\lambda)\mathbf{F}_1^{*} = \mathbf{F}_1^{*}\mathbf{\Delta}^{*2} \tag{55}$$

from (29), where $(\mathbf{X}'\mathbf{X})^+\mathbf{X}' = \mathbf{X}'(\mathbf{X}\mathbf{X}')^+$ was used to establish the first equality. (Note that $\mathbf{F}_1$ is the t leading columns of $\mathbf{F}_1^*$.) Premultiplying the equation by $\mathbf{M}(\lambda)^{1/2}$, where $\mathbf{M}(\lambda)^{1/2}$ is the symmetric square root factor of $\mathbf{M}(\lambda)$, leads to

$$\mathbf{M}(\lambda)^{1/2}\mathbf{X}\mathbf{D}(\lambda)^-\mathbf{X}'\mathbf{M}(\lambda)^{1/2}\tilde{\mathbf{F}}_1 = \tilde{\mathbf{F}}_1\mathbf{\Delta}^2, \tag{56}$$

where $\tilde{\mathbf{F}}_1 = \mathbf{M}(\lambda)^{1/2}\mathbf{F}_1^*$, and $\tilde{\mathbf{F}}_1'\tilde{\mathbf{F}}_1 = \mathbf{I}$. As in the non-regularized case, $\mathbf{X}\mathbf{D}(\lambda)^-\mathbf{X}'$ is invariant over the choice of a g-inverse $\mathbf{D}(\lambda)^-$ since $\mathrm{Sp}(\mathbf{X}') \subset \mathrm{Sp}(\mathbf{D}(\lambda))$. Let $\mathbf{D}(\lambda)^{-(*)}$ be the block diagonal matrix with $\mathbf{D}_k(\lambda)^-$ as the k^{th} diagonal block. Clearly, $\mathbf{D}(\lambda)^{-(*)} \subset \{\mathbf{D}(\lambda)^-\}$, so that $\mathbf{M}(\lambda)^{1/2}\mathbf{X}\mathbf{D}(\lambda)^-\mathbf{X}'\mathbf{M}(\lambda)^{1/2} = \mathbf{M}(\lambda)^{1/2}\mathbf{X}\mathbf{D}(\lambda)^{-(*)}\mathbf{X}'\mathbf{M}(\lambda)^{1/2} = \sum_{k=1}^K \mathbf{P}_{M(\lambda)^{1/2}X_k}(\lambda)$, where $\mathbf{P}_{M(\lambda)^{1/2}X_k}(\lambda) = \mathbf{M}(\lambda)^{1/2}\mathbf{X}_k(\mathbf{X}_k'\mathbf{M}(\lambda)\mathbf{X}_k)^-\mathbf{X}_k'\mathbf{M}(\lambda)^{1/2}$ is the orthogonal projector onto $\mathrm{Sp}(\mathbf{M}(\lambda)^{1/2}\mathbf{X}_k)$.

When $K = 2$, the above procedure leads to the canonical ridge "regression" proposed by Vinod (1976). Matrix $\mathbf{M}(\lambda)^{1/2}\mathbf{X}\mathbf{D}(\lambda)^-\mathbf{X}'\mathbf{M}(\lambda)^{1/2}$ in (56) reduces to the sum of two orthogonal projectors when $K = 2$. In the standard CANO, on the other hand, the eigenvalues and vectors of the product of the two orthogonal projectors are typically obtained. However, the dominant eigenvalues and the corresponding eigenvectors of the sum of two projectors are related in a simple manner to those of the products of the two projectors (ten Berge, 1979).

(B): Two theorems bridging (R)GCANO and (R)MCA

In Appendix (B), we give some nontrivial results used in section 2.4. We refer the reader to that section for definitions of various symbols. However, to avoid notational clutters, we use $\mathbf{Q}$ for $\mathbf{Q}_{1_p/\tilde{D}}$, $\mathbf{1}$ for $\mathbf{1}_p$, and $\mathbf{J}$ for $\mathbf{J}_p$. (Remember that $\mathbf{J}_p$ is defined shortly after Equation (10), and that $\mathbf{Q}_{1_p/\tilde{D}}$ is a block diagonal matrix with $\mathbf{Q}_{1_{p_k}/\tilde{D}_k}$ defined in (35) as the k^{th} diagonal block, so that $\mathbf{Q}_{1_p/\tilde{D}}$ has the following expression: $\mathbf{Q}_{1_p/\tilde{D}} = \mathbf{I}_n - \mathbf{A}\tilde{\mathbf{D}}$, where $\mathbf{A}$ is the block diagonal matrix with $\mathbf{1}_{p_k}(\mathbf{1}_{p_k}'\tilde{\mathbf{D}}_k\mathbf{1}_{p_k})^{-1}\mathbf{1}_{p_k}'$ as the k^{th} diagonal block.)
Theorem 1. $\mathbf{Z}\mathbf{Q}\tilde{\mathbf{D}}^+\mathbf{Q}' = \mathbf{Z}\mathbf{Q}\tilde{\mathbf{D}}^+$.

Proof. The left hand side of the above equation can be expanded as $\mathbf{Z}\{\tilde{\mathbf{D}}^+ - \mathbf{1}(\mathbf{1}'\tilde{\mathbf{D}}\mathbf{1})^{-1}\mathbf{1}'\tilde{\mathbf{D}}\tilde{\mathbf{D}}^+ - \tilde{\mathbf{D}}^+\tilde{\mathbf{D}}\mathbf{1}(\mathbf{1}'\tilde{\mathbf{D}}\mathbf{1})^{-1}\mathbf{1}' + \mathbf{1}(\mathbf{1}'\tilde{\mathbf{D}}\mathbf{1})^{-1}\mathbf{1}'\tilde{\mathbf{D}}\tilde{\mathbf{D}}^+\tilde{\mathbf{D}}\mathbf{1}(\mathbf{1}'\tilde{\mathbf{D}}\mathbf{1})^{-1}\mathbf{1}'\}$, the third and the fourth terms of which cancel out because $\mathbf{Z}\tilde{\mathbf{D}}^+\tilde{\mathbf{D}}\mathbf{1} = \mathbf{Z}\mathbf{1}$. QED.

A stronger result holds when $\tilde{\mathbf{D}}$ is nonsingular, namely $\mathbf{Q}\tilde{\mathbf{D}}^{-1} = \tilde{\mathbf{D}}^{-1}\mathbf{Q}'$, but this is rather trivial.

The following lemma is necessary to prove Theorem 2.
Lemma. $\tilde{\mathbf{D}}(\lambda)^+ = \tilde{\mathbf{D}}^+ - \tilde{\mathbf{D}}^+\mathbf{J}\mathbf{S}^{-1}\mathbf{J}\tilde{\mathbf{D}}^+$, where $\mathbf{S} = \mathbf{J}\tilde{\mathbf{D}}^+\mathbf{J} + \lambda^{-1}\mathbf{I}$.

Proof. We first prove $\tilde{\mathbf{D}}(\lambda)\tilde{\mathbf{D}}(\lambda)^+ = \tilde{\mathbf{D}}\tilde{\mathbf{D}}^+$ (symmetric). By expanding the left

hand side we obtain $\tilde{\mathbf{D}}(\lambda)\tilde{\mathbf{D}}(\lambda)^+ = \tilde{\mathbf{D}}\tilde{\mathbf{D}}^+ - \tilde{\mathbf{D}}\tilde{\mathbf{D}}^+\mathbf{J}(\mathbf{J}\tilde{\mathbf{D}}^+\mathbf{J} + \lambda^{-1}\mathbf{I})^{-1}\mathbf{J}\tilde{\mathbf{D}}^+ + \lambda\mathbf{J}\tilde{\mathbf{D}}^+ - \lambda\mathbf{J}\tilde{\mathbf{D}}^+\mathbf{J}(\mathbf{J}\tilde{\mathbf{D}}^+\mathbf{J} + \lambda^{-1}\mathbf{I})^{-1}\mathbf{J}\tilde{\mathbf{D}}^+$. The second and the fourth terms on the right hand side of this equation can be rewritten as $-\lambda\mathbf{J}\lambda^{-1}\mathbf{I}(\mathbf{J}\tilde{\mathbf{D}}^+\mathbf{J}+\lambda^{-1}\mathbf{I})^{-1}\mathbf{J}\tilde{\mathbf{D}}^+$, and $-\lambda\mathbf{J}\mathbf{J}\tilde{\mathbf{D}}^+\mathbf{J}(\mathbf{J}\tilde{\mathbf{D}}^+\mathbf{J}+\lambda^{-1}\mathbf{I})^{-1}\mathbf{J}\tilde{\mathbf{D}}^+$, respectively, and these two terms add up to $-\lambda\mathbf{J}\tilde{\mathbf{D}}^+$, establishing $\tilde{\mathbf{D}}(\lambda)\tilde{\mathbf{D}}(\lambda)^+ = \tilde{\mathbf{D}}\tilde{\mathbf{D}}^+$. This in turn implies the other Penrose conditions: $\tilde{\mathbf{D}}(\lambda)^+\tilde{\mathbf{D}}(\lambda) = \tilde{\mathbf{D}}^+\tilde{\mathbf{D}}$ (symmetric), $\tilde{\mathbf{D}}(\lambda)\tilde{\mathbf{D}}(\lambda)^+\tilde{\mathbf{D}}(\lambda) = \tilde{\mathbf{D}}(\lambda)$, and $\tilde{\mathbf{D}}(\lambda)^+\tilde{\mathbf{D}}(\lambda)\tilde{\mathbf{D}}(\lambda)^+ = \tilde{\mathbf{D}}(\lambda)^+$. QED.

Theorem 2. $\mathbf{ZQ}\tilde{\mathbf{D}}(\lambda)^+\mathbf{Q}' = \mathbf{ZQ}\tilde{\mathbf{D}}(\lambda)^+$.

Proof. Using the expression of $\tilde{\mathbf{D}}(\lambda)^+$ in the previous lemma, the left hand side of the above equation can be expanded as $\mathbf{ZQ}\tilde{\mathbf{D}}^+\mathbf{Q}' - \mathbf{ZQ}\tilde{\mathbf{D}}^+\mathbf{J}\mathbf{S}^{-1}\tilde{\mathbf{D}}^+\mathbf{Q}'$, the second term of which can further be expanded as $\mathbf{Z}\{\tilde{\mathbf{D}}^+\mathbf{J}\mathbf{S}^{-1}\mathbf{J}\tilde{\mathbf{D}}^+ - \mathbf{1}(\mathbf{1}'\tilde{\mathbf{D}}\mathbf{1})^{-1}\mathbf{1}'\tilde{\mathbf{D}}\tilde{\mathbf{D}}^+\mathbf{J}\mathbf{S}^{-1}\mathbf{J}\tilde{\mathbf{D}}^+ - \tilde{\mathbf{D}}^+\mathbf{J}\mathbf{S}^{-1}\mathbf{J}\tilde{\mathbf{D}}^+\tilde{\mathbf{D}}\mathbf{1}(\mathbf{1}'\tilde{\mathbf{D}}\mathbf{1})^{-1}\mathbf{1}' + \mathbf{1}(\mathbf{1}'\tilde{\mathbf{D}}\mathbf{1})^{-1}\mathbf{1}'\tilde{\mathbf{D}}\tilde{\mathbf{D}}^+\mathbf{J}\mathbf{S}^{-1}\mathbf{J}\tilde{\mathbf{D}}^+\tilde{\mathbf{D}}\mathbf{1}(\mathbf{1}'\tilde{\mathbf{D}}\mathbf{1})^{-1}\mathbf{1}'\}$. Each of the last three terms of this expression vanishes since $\mathbf{J}\tilde{\mathbf{D}}^+\tilde{\mathbf{D}}\mathbf{1} = \mathbf{0}$, leaving $\mathbf{ZQ}\tilde{\mathbf{D}}^+ - \mathbf{ZQ}\tilde{\mathbf{D}}^+\mathbf{J}\mathbf{S}^{-1}\mathbf{J}\tilde{\mathbf{D}}^+ = \mathbf{ZQ}\tilde{\mathbf{D}}(\lambda)^+$. QED.

Again, a stronger result holds when $\tilde{\mathbf{D}}(\lambda)$ (and consequently, $\tilde{\mathbf{D}}$) is nonsingular, namely $\mathbf{Q}\tilde{\mathbf{D}}(\lambda)^{-1} = \tilde{\mathbf{D}}(\lambda)^{-1}\mathbf{Q}'$. This can be shown as follows: When $\tilde{\mathbf{D}}(\lambda)$ is nonsingular, its inverse takes the form of $\tilde{\mathbf{D}}(\lambda)^{-1} = \tilde{\mathbf{D}}^{-1} - \tilde{\mathbf{D}}^{-1}\mathbf{J}\mathbf{S}^{-1}\mathbf{J}\tilde{\mathbf{D}}^{-1}$, where $\mathbf{S} = \mathbf{J}\tilde{\mathbf{D}}^{-1}\mathbf{J} + \lambda^{-1}\mathbf{I}$ (see the lemma above). Thus, $\tilde{\mathbf{D}}(\lambda)^{-1}\mathbf{Q}' = \tilde{\mathbf{D}}^{-1}\mathbf{Q}' - \tilde{\mathbf{D}}^{-1}\mathbf{J}\mathbf{S}^{-1}\mathbf{J}\tilde{\mathbf{D}}^{-1}\mathbf{Q}' = \tilde{\mathbf{D}}^{-1}\mathbf{Q}' - \tilde{\mathbf{D}}^{-1}\mathbf{Q}'\mathbf{J}\mathbf{S}^{-1}\mathbf{J}\mathbf{Q}\tilde{\mathbf{D}}^{-1} = \mathbf{Q}\tilde{\mathbf{D}}^{-1} - \mathbf{Q}\tilde{\mathbf{D}}^{-1}\mathbf{J}\mathbf{S}^{-1}\mathbf{J}\tilde{\mathbf{D}}^{-1} = \mathbf{Q}\tilde{\mathbf{D}}(\lambda)^{-1}$. Note that $\mathbf{Q}'\mathbf{J} = \mathbf{J}\mathbf{Q} = \mathbf{J}$.

References

Abdi, H., and Valentin, D. (2007). The STATIS method. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics*, (pp. 955-962). Thousand Oaks (CA): Sage.

Abdi, H. (2007). Singular value decomposition (SVD) and generalized singular decomposition (GSVD). In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics*, (pp. 907-912). Thousand Oak, CA: Sage.

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

Beck, A. (1996). *BDI-II*. San Antonio, TX: Psychological Corporation.

Carroll, J. D. (1968). A generalization of canonical correlation analysis to three or more sets of variables. Proceedings of the 76^{th} Annual Convention of the American Psychological Association, 227-228.

Dahl, T., and Næs, T. (2006). A bridge between Tucker-1 and Carroll's generalized canonical analysis. *Computational Statistics and Data Analysis*, **50**, 3086-3098.

de Leeuw, J. (1982). Generalized eigenvalue problems with posivite semi-definite matrices. *Psychometrika*, **47**, 87-93.

Devaux, M.-F., Courcoux, P., Vigneau, E., and Novales, B. (1998). Generalized canonical correlation analysis for the interpretation of fluorescence spectral data. *Analusis*, **26**, 310-316.

DiPillo, P. J. (1976). The application of bias to discriminant analysis. *Communications in Statistics - Theory and Methods*, **5**, 843-859.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, **7**, 1-26.

Efron, B., and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

Fischer, B., Ross, V., and Buhmann, J. M. (2007). Time-series alignment by non-negative multiple generalized canonical correlation analysis. In F. Massuli, S. Mitra, and G. Pasi (Eds.), *Applications of fuzzy set theory*, (pp. 505-511). Berlin: Springer-Verlag.

Friedman, J. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, **84**, 165-175.

Gardner, S., Gower, J. C., and le Roux, N. J. (2006). A synthesis of canonical variate analysis, generalized canonical correlation and Procrustes analysis. *Computational Statistics and Data Analysis*, **50**, 107-134.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester: Wiley.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Hoerl, A. F., and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthgonal problems. *Technometrics*, **12**, 55-67.

Horst, P. (1961). Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology*, **17**, 331-347.

Kroonenberg, P. M. (2008). *Applied multiway data analysis*. New York: Wiley.

Legendre, P., and Legendre, L. (1998). *Numerical ecology*. Amsterdam: North Holland.

Maraun, M., Slaney, K., and Jalava, J. (2005). Dual scaling for the analysis of categorical data. *Journal of Personality Assessment*, **85**, 209-217.

Meredith, W. (1964). Rotation to achieve factorial invariance. *Psychometrika*, **29**, 187-206.

Poggio, T., and Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, **247**, 978-982.

Ramsay, J. O., and Silverman, B. W. (2005). *Functional data analysis*, (Second Edition). New York: Springer.

Smilde, A., Bro, R., and Geladi, P. (2004). *Multi-way analysis: Applications in the chemical sciences*. New York: Wiley.

Sun, Q.-S., Heng, P.-A., Jin, Z., and Xia, D.-S. (2005). Face recognition based on generalized canonical correlation analysis. In D. S. Huang, X.-P. Zhang, G.-B. Huang (Eds.), *Advances in intelligent computing*, (pp. 958-967). Berlin: Springer-Verlag.

Takane, Y. (1980). Analysis of categorizing behavior by a quantification method. *Behaviormetrika*, **8**, 75-86.

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 correspondence analysis. In M. J. Greenacre, and J. Blasius, (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**, 392-405.

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

Takane, Y., and Oshima-Takane, Y. (2002). Nonlinear generalized canonical correlation analysis by neural network models. In S. Nishisato, Y. Baba, H. Bozdogan, and K. Kanefuji (Eds.), *Measurement and multivariate analysis* (pp. 183-190). Tokyo: Springer Verlag.

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

ten Berge, J. M. F. (1979). On the equivalence of two oblique congruence rotation methods, and orthogonal approximations. *Psychometrika*, **44**, 359-364.

ter Braak, C. J. F. (1990). *Update notes: CANOCO Version 3.10*. Wageningen, The Netherlands: Agricultural Mathematics Group.

Tikhonov, A. N., and Arsenin, V. Y. (1977). *Solutions of ill-posed problems*. Washington, DC: Winston.

Tillman, B., Dowling, J., and Abdi, H. (in preparation). Bach, Mozart or Beethoven? Indirect investigations of musical style perception with subjective judgments and sorting tasks.

van de Velden, M., and Bijmolt, T. H. A. (2006). Generalized canonical correlation analysis of matrices with missing rows: A simulation study. *Psychometrika*, **71**, 323-331.

van der Burg, E. (1988). *Nonlinear canonical correlation and some related techniques*. Leiden: DSWO Press.

Vinod, H. D. (1976). Canonical ridge and econometrics of joint production. *Journal of Econometrics*, **4**, 47-166.