PSYCHOMETRIKA—VOL. 47, NO. 4.
DECEMBER, 1982

CONSTRAINED CANONICAL CORRELATION

WAYNE S. DESARBO, ROBERT E. HAUSMAN, SHEN LIN, AND WESLEY THOMPSON

BELL LABORATORIES

This paper explores some of the problems associated with traditional canonical correlation. A response surface methodology is developed to examine the stability of the derived linear functions, where one wishes to investigate how much the coefficients can change and still be in an ε-neighborhood of the globally optimum canonical correlation value. In addition, a discrete (or constrained) canonical correlation method is formulated where the derived coefficients of these linear functions are constrained to be in some small set, e.g., $\{1, 0, -1\}$, to aid in the interpretation of the results. An example concerning the psychographic responses of Wharton MBA students of the University of Pennsylvania regarding driving preferences and life-style considerations is provided.

Key words: canonical correlation, constrained multivariate analysis, response surface analysis.

Introduction

Brief Review of Canonical Correlation

Hotelling [1936] first formulated canonical correlation analysis as a method to examine the relationships between two sets of measurements made on the same subjects. Assume that one has two sets of variables, $\mathbf{X}$ (p standardized variables which we shall refer to as the predictor set) and $\mathbf{Y}$ (q standardized variables, which we shall refer to as the criterion set), all measured on the same N experimental units. One can then derive a correlation matrix:

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{xx} & \mathbf{R}_{xy} \\ \mathbf{R}_{yx} & \mathbf{R}_{yy} \end{bmatrix} \tag{1}$$

partitioned as in (1). In canonical correlation analysis, one defines a pair of canonical components (variates) to be linear combinations of the form $\{\mathbf{w}'\mathbf{X}, \mathbf{v}'\mathbf{Y}\}$, where these linear combinations are normally standardized to zero mean and unit variance. We shall refer to each such pair of components as a factor.

The first such factor is determined by maximizing the correlation [cf. Anderson, 1958] between its two components $\mathbf{w}_1'\mathbf{X}$ and $\mathbf{v}_1'\mathbf{Y}$. This canonical correlation can be written as

$$z_1 = \frac{\mathbf{w}_1'\mathbf{R}_{xy}\mathbf{v}_1}{(\mathbf{w}_1'\mathbf{R}_{xx}\mathbf{w}_1)^{1/2}(\mathbf{v}_1'\mathbf{R}_{yy}\mathbf{v}_1)^{1/2}}\,. \tag{2}$$

After extraction of the first pair of canonical variates ($\mathbf{w}_1'\mathbf{X}$, $\mathbf{v}_1'\mathbf{Y}$), a second pair ($\mathbf{w}_2'\mathbf{X}$, $\mathbf{v}_2'\mathbf{Y}$) can be determined having maximum correlation, with the restriction that the derived canonical variates are uncorrelated with all other canonical variates except with their counterparts in the other set. Note that the derived vectors of coefficients, $\mathbf{w}$ and $\mathbf{v}$, will

Wayne S. DeSarbo, Robert Jausman, Shen Lin, and Wesley Thompson are all Members of Technical Staff at Bell Laboratories. We wish to express our gratitude to the editor and reviewers of this paper for their insightful remarks.

Requests for reprints should be addressed to: Wayne S. DeSarbo, Bell Labs., Room 2C-479, 600 Mountain Ave., Murray Hill, N.J. 07974.

0033-3123/82/1200-3110$75/0

not be orthogonal. This process of obtaining canonical variates can continue up to min $\{p, q\} = r$ factors, assuming $r < N$.

Chambers [1977] and Kettenring [Note 3] discuss efficient algorithms for deriving such canonical coefficients employing general orthogonalization procedures.

Some Problems as Cited in the Literature

Much has been written concerning some problems associated with traditional canonical correlation analysis. Kendall and Stuart [1966] and Kendall [1968] cite the problem of difficulty of interpretation:

> "The difficulties of interpretation are such that not many examples of convincing applications of canonical correlation analysis appear in the literature" [Kendall, p. 69].

Clearly, whenever one attempts to optimize some mathematical criterion, there is no guarantee of interpretability of the results. And this problem seems to be especially prevalent with canonical correlation analysis (for reasons we shall examine shortly).

The obvious first suggestion on how to interpret the derived canonical variates is to examine what went into their composition, i.e., examining the elements of the matrices of coefficients or weights **W** and **V** which represent the direct contribution of each of the original variables to the composites. While this suggestion appears quite simple and obvious, and indeed is made in several textbooks presentations, it is one which can be quite misleading. For example, a large degree of multicollinearity would typically result in wide confidence intervals around these coefficients, so that one variable may thus hide or suppress the importance of another variable highly correlated with it. According to Levine [1977], this suppression issue is probably the most crucial one in hindering the ability to interpret the canonical variates. What typically happens is that if two variables in the same set are closely correlated with each other, once one of the two has made its contribution to the composite, the other has no additional autonomous contribution to make. The first variable's coefficient may be high, the second's near zero, that is, suppressed by the first. Alternatively, both coefficients may be depressed.

Several authors have proposed additional analyses to be performed to enhance interpretation. Cooley and Lohnes [1962] and Meredith [1964] propose examining "structure correlations," the correlations of the original variables with the associated canonical variate(s). Cliff and Krus [1976] recommend a procedure for an orthogonal rotation of the derived canonical variates to "simple structure" which would preserve some of the original canonical variates' optimal properties. Weinberg and Darlington [1976] propose "integrizing" or rounding the canonical coefficients to $+1$, 0, or -1 based on the observation of subsequent matrices of partial correlations (more will be said about this later). Vinod [1977] proposes canonical ridge regression, to derive reliable coefficient estimates when multicollinearity is present. Gnanadesikan [1977] recommends plotting the derived canonical coefficients in a variety of ways to reval possible idiosyncracies in the data. Kettenring [Note 3] suggests possibly deleting "unusual" variables and observations and assessing their impact.

Another very related problem is one of coefficient instability, or what Cooley [1965] calls "bouncing betas". This problem may result from excessive multicollinearity in the data. However, as we shall see shortly, the coefficient instability problem is inherent in the estimation technique itself. Even without serious levels of multicollinearity, one can tamper with the derived coefficients and not affect the value of the objective function all that much, which suggests that the response surface can be quite flat for many types of dependence structures. Many authors have realized this and have suggested sub-sampling procedures to assess coefficient instability. Thorndike and Weiss [1973] have recommended cross-validation for all canonical correlation applications to obtain some insight into how stable the coefficients are. Dempster [1966] suggest jack-knifing to assess coefficient stability (one could also employ the bootstrap method).

Finally, there is the problem of variable importance in canonical correlation analysis, that is, which variables in each vector **X** and **Y** are most important in obtaining the designated canonical correlation. There are few reliable and widely used tests of significance one can employ, like the t-tests in regression, to delete insignificant variables from the analysis. Likelihood ratio tests can be applied if the underlying distributional assumptions are met. Green, Carroll, and DeSarbo [1979] propose a δ^2-importance measure based on a singular value decomposition procedure.

Many of these issues have been addressed in the context of the basic multiple linear regression model. For example, there has been a trend in psychometric research suggesting that criterion functions for linear composites remain surprisingly stable despite wide variations in the component weights. Green [1977] has investigated parameter sensitivity in multiple regression and principal components. He has derived indifference regions for these parameters which indicate the potential for wide variation in estimates for these parameters. In turn, this has provoked considerable attention to the use of weights other than those obtained by least squares methods (cf. Bibby & Toutenberg, 1977; Darlington, 1978; Green, 1977; Pruzek & Frederick, 1978; Winer, 1978; Wainer, 1976, 1978).

Proposed Research

Given the problems associated with canonical correlation analysis previously mentioned, we propose a two-phase research plan which investigates and/or solves some of these problems:

(i) *Response Surface Methodology*—to explore the response surface of the canonical correlation objective function (z) and examine how sensitive z is to changes in $\mathbf{w}_i$ and/or $\mathbf{v}_i$.
(ii) "*Constrained*" *Canonical Correlation Analysis*—to constrain "optimally" the canonical correlation coefficients to $+1$, 0 or -1 (or some other small set), enhance interpretation, and possibly provide insight into variable importance.

Response Surface Methodology

Motivation

Given the fact that empirical studies performed [Thorndike & Weiss, 1973 & Thorndike, et al. 1968] have demonstrated the instability in the elements of $\mathbf{w}$ and $\mathbf{v}$, the logical question which follows is "how much?" That is, how sensitive is the objective function to perturbations in the elements of $\mathbf{w}$ and/or $\mathbf{v}$? If in fact the response surface of the particular canonical correlation, as a function of $\mathbf{w}$ and $\mathbf{v}$, is generally flat, then one could perhaps perturb some subset of the coefficients in $\mathbf{w}$ and $\mathbf{v}$ and derive a new solution for which the corresponding value z of the objective function was not much below the global optimum z^*. On the other hand, if the response surface were quite "peaked", then such changes might drastically affect the resulting value of z. In addition, there is the question of *which* coefficient(s), when perturbed, would affect the resulting value of z the most. If one were to perturb all coefficients, which direction would result in the largest (or smallest) reduction in z? The answers to these questions may provide some insight into the problem of variable importance. One may find that one can perturb a particular coefficient quite a bit and not affect the resulting z value or that very slight perturbations in a coefficient affect the z value substantially.

An alternative methodology is to examine simultaneous confidence intervals and respective ellipsoids suggested by Bartlett [1933], Roy [1957], and Scheffé [1959] as applied to canonical correlation. These procedures utilize an objective function which is proportional to the one used for parameter sensitivity analysis, but they focus in upon a slightly different issue—sampling variation—while we choose to examine parameter sensitivity. As Green [1977] notes:

> "The difference between sampling variability and parameter sensitivity is important. Sampling variability depends mainly on the number of cases. Parameter sensitivity is, by contrast, independent of *n*. With the entire population at hand, sampling variability disappears, but parameter sensitivity is still a problem. The model will still not fit perfectly; further, two quite different sets of parameters may still yield nearly the same fit of the model to the data. In a sense, the issue of sensitivity, or tolerable parameter variation, is more a question about the nature of the model than about the data being described. (p. 264)"

Thus, we are primarily interested in investigating parameter sensitivity as opposed to sampling variation. We will be interested in how the value of the canonical correlation changes (both direction and magnitude) as values of the parameters (canonical coefficients) change. We need not make any distributional assumptions about the nature of the data to do this.

A Simple Example

Professor D. F. Morrison of the University of Pennsylvania [personal communication] provided a simple four variable problem to illustrate the problem of coefficient instability *without multicollinearity*. Assume two predictor variables, X_1 and X_2, and two criterion variables, Y_1 and Y_2, with correlation matrix:

$$\mathbf{R} = \begin{array}{c} \\ X_1 \\ X_2 \\ Y_1 \\ Y_2 \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & Y_1 & Y_2 \end{array} \\ \left| \begin{array}{cccc} 1 & 0 & c_1 & 0 \\ 0 & 1 & 0 & c_2 \\ c_1 & 0 & 1 & 0 \\ 0 & c_2 & 0 & 1 \end{array} \right| \end{array}, \tag{3}$$

where $c_1 > c_2$. Then, letting $\mathbf{X} = (X_1, X_2)'$ and $\mathbf{Y} = (Y_1, Y_2)'$, the first set of canonical variates is:

$$\begin{aligned} \mathbf{w}_1'\mathbf{X} &= X_1 \\ \mathbf{v}_1'\mathbf{Y} &= Y_1, \end{aligned} \tag{4}$$

and the maximum canonical correlation is therefore c_1. The second set of canonical variates is:

$$\begin{aligned} \mathbf{w}_2'\mathbf{X} &= X_2 \\ \mathbf{v}_2'\mathbf{Y} &= Y_2 \end{aligned} \tag{5}$$

with canonical correlation c_2. Now, consider arbitrary linear functions:

$$U_1 = X_1 + aX_2$$

and

$$T_1 = Y_1 + bY_2, \tag{6}$$

where we introduce "contamination" in the first pair of canonical variates. We wish to examine how non-null values of a and b affect the correlation:

$$\text{Corr}\,(U_1, T_1) = \frac{c_1 + abc_2}{(1 + a^2)^{1/2}(1 + b^2)^{1/2}} = z_1, \tag{7}$$

where:

$$c_1 - \varepsilon \leq z_1 \leq c_1$$
$$\varepsilon = \text{some small constant} > 0. \tag{8}$$

That is, we wish to investigate how large one can make a and b and still remain in some ε-neighborhood of c_1. Thus

$$\frac{c_1 + abc_2}{(1 + a^2)^{1/2}(1 + b^2)^{1/2}} \geq c_1 - \varepsilon, \tag{9}$$

for a prespecified ε. We denote the area defined by (9) as the "ε-indifference region". The boundary of this region is an oval, ellipse-like figure, symmetric about the origin, whose major and minor axes are the 45° and 135° lines through the origin.

If we let $c_1 = .8$ and $c_2 = .5$, then Table 1 displays the amount of "contamination" (measured by the values of a and b from (11)) allowed for ε values of .10, .05, and .01. Even for small values of ε, the values of a and b are quite substantial, indicating that even in the absence of multicollinearity, the derived coefficients may be perturbed substantially from their optimal values without greatly affecting the value of the objective function, z_1.

General Case

Formulation. Let

$$\Phi(\boldsymbol{\theta}) = \frac{\mathbf{w}'\mathbf{R}_{xy}\mathbf{v}}{(\mathbf{w}'\mathbf{R}_{xx}\mathbf{w})^{1/2}(\mathbf{v}'\mathbf{R}_{yy}\mathbf{v})^{1/2}}, \qquad \text{where } \boldsymbol{\theta} = (\mathbf{w}, \mathbf{v})$$

and Φ^* be the maximum canonical correlation for, say, the first factor. One can then define an ε-indifference (or tolerance) region as:

$$\{\boldsymbol{\theta} \mid \Phi^* - \Phi(\boldsymbol{\theta}) \leq \varepsilon\}, \tag{10}$$

where ε is the largest difference or reduction in the objective function that one is indifferent towards. Thus,

$$\{\boldsymbol{\theta} \mid \Phi(\boldsymbol{\theta}) = \Phi^* - \varepsilon\} \tag{11}$$

is the boundary of the "indifference region". Taking a Taylor series approximation of $\Phi(\boldsymbol{\theta})$ near $\boldsymbol{\theta}^*$, i.e., in the neighborhood defined by (11), and examining the first three terms, one

Table 1

Contamination for Simple Four Variable Example

For: $c_1 = .8$

$c_2 = .5$

ε	z_1	a	b
.10	.70	−.707	.707
.05	.75	−.447	.447
.01	.79	−.186	.186

obtains:

$$\Phi(\boldsymbol{\theta}) \approx \Phi^* + \mathbf{g}^{*\prime}\mathbf{r} + \frac{\mathbf{r}'\mathbf{H}^*\mathbf{r}}{2}, \tag{12}$$

where:

$$\begin{aligned} \mathbf{r} &= \boldsymbol{\theta} - \boldsymbol{\theta}^* \\ \mathbf{g}^* &= \text{the gradient of } \Phi(\boldsymbol{\theta}) \text{ evaluated at } \boldsymbol{\theta} = \boldsymbol{\theta}^* \\ \mathbf{H}^* &= \text{the Hessian matrix of } \Phi(\boldsymbol{\theta}) \text{ evaluated at } \boldsymbol{\theta} = \boldsymbol{\theta}^* \\ &\quad \text{(necessarily negative definite or negative semidefinite).} \end{aligned}$$

At the maximum, $\mathbf{g}^*$ is zero giving:

$$\Phi(\boldsymbol{\theta}) \approx \Phi^* + \frac{\mathbf{r}'\mathbf{H}^*\mathbf{r}}{2}, \tag{13}$$

and the ε-indifference region can be *approximated* by [Bard, 1974]:

$$|\mathbf{r}'\mathbf{H}^*\mathbf{r}| \leq 2\varepsilon. \tag{14}$$

If we let $\mathbf{A} = -\mathbf{H}^*$ so that $\mathbf{A}$ is at least positive semi-definite, then $\mathbf{r}'\mathbf{A}\mathbf{r} = 2\varepsilon$ is the equation of an n-dimensional ellipsoid where $n \leq p + q$. The ellipsoids corresponding to different values of ε are concentric and similar in shape and orientation, so that much information can be gained from the analysis of the matrix $\mathbf{A}$, without regard to the actual value of ε.

Single Parameter Changes. Given the previous formulation, we can now answer the question of how much a single parameter θ_i can be varied from the optimal value θ_i^* so as to remain in an ε-indifference region. That is, how much can we perturb θ_i^* so as not to reduce the optimal canonical correlation by more than ε.

Let:

$$r_j = 0 \quad \forall \quad j \neq i; \tag{15}$$

then

$$A_{ii} r_i^2 \leq 2\varepsilon, \tag{16}$$

where A_{ii} is the scalar in the i-th row and i-th column of $\mathbf{A}$. Expanding r_i in (16), one obtains:

$$\theta_i^* - \left(\frac{2\varepsilon}{A_{ii}}\right)^{1/2} \leq \theta_i \leq \theta_i^* + \left(\frac{2\varepsilon}{A_{ii}}\right)^{1/2}, \tag{17}$$

or

$$|\theta_i - \theta_i^*| \leq \left(\frac{2\varepsilon}{A_{ii}}\right)^{1/2} \tag{18}$$

for any single parameter change. Bard [1974] considers θ_i to be "well-determined" if the quantity $(2\varepsilon/A_{ii})^{1/2}$ is small on the scale by which θ_i is measured, and θ_i is "ill-determined" if $(2\varepsilon/A_{ii})^{1/2}$ is large. However, since the objective function

$$z = \frac{\mathbf{w}'\mathbf{R}_{xy}\mathbf{v}}{(\mathbf{w}'\mathbf{R}_{xx}\mathbf{w})^{1/2}(\mathbf{v}'\mathbf{R}_{yy}\mathbf{v})^{1/2}} \tag{19}$$

is not quadratic, we are dealing with an approximation whose accuracy must be examined by evaluating the objective function at this boundary.

All-Parameter Changes. Here, we wish to generalize the procedure for single parameter changes to the case where all parameters can be altered simultaneously. Figure 1 presents a simple two-variable example used solely for illustrative purposes. Given the elliptical structure of the ε-indifference region pictured, one could alter all the coefficients infinitely many ways and still be on the boundary of the ellipse. What appears to be a good synopsis of these all-parameter changes is the examination of the "best" and "worst" cases. By "best" we mean the direction and length of a vector from the origin to the boundary corresponding to the minor axis of the ellipse in Figure 1 (minimum length). This direction corresponds to the linear combination of the parameters which is "best" determined. By "worst" we, correspondingly, refer to the direction and length of the major axis of that ellipse (maximum length). Clearly, the point on the boundary furthest away from the origin renders information concerning the most that one can change all the parameters and still obtain an objective function value less than or equal to the global optimum minus ε.

Mathematically, we wish to find the vector terminus with maximum distance [Bard, 1974] from the origin ("worst" case) on the ε-boundary, i.e.:

$$\text{Max } \mathbf{r}'\mathbf{r} \quad \text{(distance from origin)}$$

such that

$$\mathbf{r}'\mathbf{A}\mathbf{r} = 2\varepsilon \quad \text{(on boundary).} \tag{20}$$

After forming the Lagrangian,

$$L = \mathbf{r}'\mathbf{r} - \lambda(\mathbf{r}'\mathbf{A}\mathbf{r} - 2\varepsilon), \tag{21}$$

we take partial derivatives and set them equal to zero:

$$\frac{\partial L}{\partial \mathbf{r}} = 2\mathbf{r} - 2\lambda\mathbf{A}\mathbf{r} = 0 \tag{22}$$

$$\frac{\partial L}{\partial \lambda} = \mathbf{r}'\mathbf{A}\mathbf{r} - 2\varepsilon = 0. \tag{23}$$

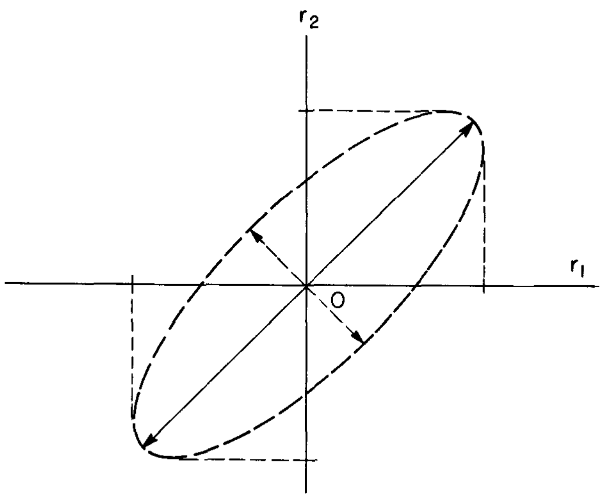

FIGURE 1.
Effects of all-parameter changes: A simple two variable example.

Simplifying, we obtain an eigenstructure problem of the form

$$(\mathbf{A}^{-1} - \lambda\mathbf{I})\mathbf{r} = 0. \tag{24}$$

Here the length of the axes is related to the size of the eigenvalues of $\mathbf{A}^{-1}$ (or the reciprocals of the eigenvalues of $\mathbf{A}$). The direction of the "worst" case vector then is given by the eigenvector associated with the maximum eigenvalue of $\mathbf{A}^{-1}$. Given the direction, one can then find the coordinate of the vector terminus by scaling the eigenvector to satisfy the constraint. Similarly, the eigenvector associated with the smallest eigenvalue of $\mathbf{A}^{-1}$ determines the direction of the "best" case vector (or the most "sensitive" direction to move). This eigenvector can be properly scaled to lie on the constraint boundary. This procedure, then, provides linear combinations of the parameters which are best determined ("best" case) and those which are worst determined ("worst" case).

A word of caution is in order at this point. In the canonical correlation problem, one can multiply the predictor and criterion set variables by scalars and still obtain the same canonical correlation. This scale invariance for both sets implies that $\mathbf{A}$ is always positive semi-definite (at least two zero eigenvalues), and therefore always singular, so that $\mathbf{A}^{-1}$ does not exist. Thus, one must use the eigenstructure of $\mathbf{A}$ to obtain these "best" and "worst" cases. Here, the worst determined linear combination of the parameters corresponds to the eigenvector associated with the smallest nonzero eigenvalue of $\mathbf{A}$, while the best determined linear combination corresponds to the eigenvector associated with the largest eigenvalue of $\mathbf{A}$.

Subsets of Parameter Changes. One can now adapt the previous case of altering all $n = p + q$ parameters simultaneously to the case of changing some subset, say t of them, simultaneously. Now, we hold $n - t$ parameters fixed at their optimal values. By reordering, if necessary, we may assume that the last $n - t$ coordinates are held fixed at their optimal values. Thus, we set

$$\mathbf{r}' = (\mathbf{a}\ \mathbf{0}), \tag{25}$$

where $\mathbf{a}$ represents the change in the nonfixed parameters. As before, we wish to find the "best" and "worst" cases, i.e.,

$$\text{Max } \mathbf{r}'\mathbf{r}$$

such that:

$$\mathbf{r}'\mathbf{A}\mathbf{r} = 2\varepsilon. \tag{26}$$

However, the rows and columns of $\mathbf{A}$ must be permuted to agree with the order specified by r. Let us rearrange and partition $\mathbf{A}$ as follows:

$$\mathbf{A} = \begin{array}{c c} & \begin{array}{cc} t & n-t \end{array} \\ \begin{array}{c} t \\ n-t \end{array} & \begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{D} & \mathbf{E} \end{bmatrix} \end{array}, \tag{27}$$

where:

$\mathbf{B} = t \times t$ sub-matrix of $\mathbf{A}$ which corresponds to elements being altered,

$\mathbf{E} = (n - t) \times (n - t)$ sub-matrix of $\mathbf{A}$ which corresponds to elements *not* being altered,

$\mathbf{C} = \mathbf{D}'$ sub-matrix of cross partial derivatures between those elements being altered and these not being altered.

The boundary constraint, $\mathbf{r}'\mathbf{A}\mathbf{r} = 2\varepsilon$, becomes

$$(\mathbf{a}'\ \mathbf{0}')\begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{D} & \mathbf{E} \end{bmatrix}\begin{pmatrix} \mathbf{a} \\ \mathbf{0} \end{pmatrix} = 2\varepsilon \tag{28}$$

or $\mathbf{a}'\mathbf{B}\mathbf{a} = 2\varepsilon$. Now, one can form the "reduced" Lagrangian (for "worst" case analysis), letting

$$L = \mathbf{a}'\mathbf{a} - \lambda(\mathbf{a}'\mathbf{B}\mathbf{a} - 2\varepsilon). \tag{29}$$

We perform the same analysis as in the case of all parameters changes, except that now the eigenstructure of $\mathbf{B}^{-1}$, not $\mathbf{A}^{-1}$, is relevant. Thus, the "worst" case is represented by properly scaling the eigenvector associated with the largest eigenvalue of $\mathbf{B}^{-1}$ to satisfy the boundary constraint. Similarly, the "best" case is represented by properly scaling the eigenvector associated with the smallest eigenvalue of $\mathbf{B}^{-1}$ to satisfy the boundary constraint.

A potential singularity problem also exists here if, say, one were to select $\mathbf{w}$ (or $\mathbf{v}$) as fixed and examined changes in $\mathbf{v}$ (or $\mathbf{w}$). In this case, $\mathbf{B}$ is singular (one zero root), and the adjustments discussed in the previous section apply. Since $\mathbf{B}^{-1}$ does not exist, we use the eigenstructure of $\mathbf{B}$ in place of $\mathbf{B}^{-1}$. That is, the worst determined linear combination of the parameters corresponds to the eigenvector associated with the smallest nonzero eigenvalue of $\mathbf{B}$, while the best determined linear combination corresponds to the eigenvector associated with the largest eigenvalue of $\mathbf{B}$.

Constrained Canonical Correlation Analysis

Motivation

Given the interpretation and coefficient instability problems alluded to earlier, we propose an alternative formulation to canonical correlation analysis where we constrain the elements of $\mathbf{w}$ and $\mathbf{v}$ to be $+1$, 0, or -1, i.e.,:

$$\underset{\mathbf{w},\,\mathbf{v}}{\text{Max}}\left[z = \frac{\mathbf{w}'\mathbf{R}_{xy}\mathbf{v}}{(\mathbf{w}'\mathbf{R}_{xx}\mathbf{w})^{1/2}(\mathbf{v}'\mathbf{R}_{yy}\mathbf{v})^{1/2}}\right]$$

such that

$$\begin{aligned} w_i &\in \{-1, 0, +1\}, \qquad i = 1, \ldots, p \\ v_i &\in \{-1, 0, +1\}, \qquad j = 1, \ldots, q. \end{aligned} \tag{30}$$

This procedure attempts to alleviate the interpretation problem in that the constrained coefficients imply that a variable has a positive $(+1)$, negative (-1), or null (0) effect or contribution to a derived canonical variate. This may well be easier to interpret than the traditional rational number form of $\mathbf{w}$ and $\mathbf{v}$, often extended to six or so decimal places. In addition, it attempts to simplify the analysis in that variables with zero coefficients can be ignored. These zero coefficients appear to be related to the question of variable importance.

A similar formulation has been proposed by Weinberg and Darlington [1976] to alleviate coefficient instability and "shrinkage" of the canonical correlation associated with cross-validating canonical correlation results. There are, however, some serious problems with their procedure:

(i) There is no mention of optimality or objective function maximization. It is not clear to us what properties these derived coefficients have, in general—they could be quite "sub-optimal".

(ii) The vectors extracted after the first pair are, as the authors point out, *not* in terms of the original variables, but rather in terms of residual scores. As such, it is not evident how interpretation is enhanced by use of such a procedure when one is now dealing with complex functions of the original variables. In addition, these subsequent vectors are not integral $\{+1, 0, -1\}$ with respect to the original variables (only with respect to the residuals).

We therefore propose an alternative solution which attempts to enhance interpretation and stability (as was the intention of Weinberg and Darlington), but also achieves optimality and restricts the interpretation to the original set of variables.

Algorithms and Method Comparisons

Three methods were applied to solve the constrained canonical correlation problem stated in (30):

(i) Complete Enumeration
(ii) Combinatorial Optimization
(iii) Branch and Bound.

The mathematical details of each method and comparisons between methods follow.

Complete Enumeration. This procedure entails evaluating all 3^{p+q} possible solutions and selecting the one with the highest z value. Obviously, for even moderately sized problems, the computer cost of performing such calculations is quite high. However, the procedure does guarantee finding the globally optimum solution.

Taking advantage of the structure of the problem, the authors have developed an algorithm excluding approximately 75% of the total calculations using synthetic division (see Appendix I). Even so, we do not recommend such a procedure because of cost considerations since 25% of 3^{p+q} can still be quite large for moderately sized p and q. In addition, the algorithm, unless modified, is not conducive to examining a number of possible solutions. One may, for example, be willing to trade-off some canonical correlation value for parsimony if many more coefficients can be set to zero.

For the extraction of more than one factor, we will require that the vectors within each set be orthogonal to the preceding vectors extracted. A discussion of this will follow later. While constraints to check orthogonality can be embedded in this complete enumeration framework, a more efficient procedure has been developed which utilizes a tree traversal search algorithm. Note that all the solutions to the constrained canonical correlation problem can be exhibited via a tree structure as in Figure 2. Three branches $(-1, 0, +1)$ emanate from each node and each level in the tree corresponds to a coefficient in $(\mathbf{w}, \mathbf{v})$. Assume that previous factors $\mathbf{W}$, $\mathbf{V}$ have been entered row-wise into matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ respectively. Then for $\mathbf{w}^*$ and $\mathbf{v}^*$ to be orthogonal to previous derived factors, we require:

$$\mathbf{A}_1\mathbf{w}^* = 0, \qquad \mathbf{A}_2\mathbf{v}^* = 0. \tag{31}$$

A depth-first order [Reingold, et al. [1977]] search procedure is conducted to evaluate some of the nodes of the tree and to derive the set of all possible orthogonal solutions which can then be evaluated to select the best one(s). At each node of the tree, an evaluation is made to see whether the vector $\mathbf{w}^*$ (similarly for $\mathbf{v}^*$) can be completed to be orthogonal with the concatenated vectors in $\mathbf{A}_1$ (similarly for $\mathbf{A}_2$). That is, at a given node, one checks to see that for that subset of the elements of $\mathbf{w}^*$, say $\mathbf{x}_1$, whose elements have already been assigned values of $+1$, 0, or -1, there exists some $\mathbf{x}_2$ with similar elements such that

$$\mathbf{A}_1 \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{0} \end{pmatrix} + \mathbf{A}_1 \begin{pmatrix} \mathbf{0} \\ x_2 \end{pmatrix} = \mathbf{0}. \tag{32}$$

This amounts to checking to see whether particular "partial sums" in the first term on the left hand side of (32) can be offset by some $\mathbf{x}_2$ via the second term to give a zero vector for all vectors in $\mathbf{A}_1$. This quite efficient method is related to the Branch and Bound algorithm discussed shortly.

Combinatorial Optimization. An alternative to complete enumeration for solving the constrained canonical correlation problem is a modification of the Lin and Kernighan [1973] combinatiorial optimization algorithm. The algorithm attempts to generate a number of local optima by examining changes in a total of m elements in $(\mathbf{w}, \mathbf{v})$ at a time. The steps of the modified algorithm are as follows:

(i) Set $K = 0$; select m from $\{1, 2, \ldots, p + q\}$;
Set maximum number of iterations (MAXIT);
(ii) Generate random feasible $\mathbf{w}$ and $\mathbf{v}$ vectors;
(iii) Set $K = K + 1$;
(iv) Evaluate objective function (z) and let $z^* = z$;
(v) Generate a random map, i.e., a random permutation of the first $p + q$ positive integers. This map indicates (randomly) the order in which elements will be changed m at a time;
(vi) Try to improve—one attempts to evaluate changes in $\mathbf{w}$ and $\mathbf{v}$ m at a time according to the random map until either one improves $(z > z^*)$ or one has evaluated all possible m changes according to the map without improvement. If there is improvement, then set $z^* = z$, store the $\mathbf{w}$ and $\mathbf{v}$ solution that resulted in that z, and go to step 5. If no improvement results $(z \leq z^*)$, store $\mathbf{w}$, $\mathbf{v}$, and z-value. If $K <$ MAXIT, go to step 2, otherwise output results and stop.

The algorithm is much cheaper to run than the complete enumeration routine previously discussed. Lin and Kernighan [1973] estimate computer costs for the algorithm as

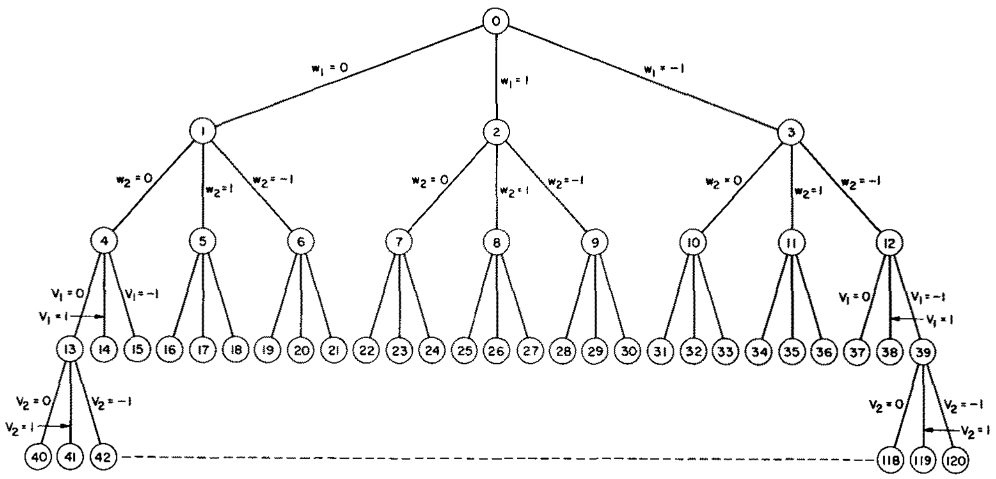

FIGURE 2.
The branch and bound tree for $p = q = 2$.

a function of the number of parameters and describe some if its interesting mathematical properties for the traveling salesman problem. While the procedure does not guarantee a globally optimum result, setting MAXIT ≥ 40 in our problem has always resulted in obtaining the globally optimum solution at least once for $m = 1$ or 2. Obviously, the larger one makes m and MAXIT, the higher the probability of finding the globally optimal result, but at a much higher computer cost.

Another desirable feature of the algorithm is that it finds MAXIT local optimum solutions. This gives the analyst the opportunity of examining a number of "good" solutions enabling him to perhaps make his own trade-off between a z-value and say coefficients set equal to zero. Or, perhaps one solution may be much more interpretable than the others, given the nature of the application at hand. At any rate, examining a number of "good" solutions is one of the distinct advantages of this method.

For the extraction of more than one factor, one can either embed the orthogonality constraints in the algorithm or use the algorithm discussed in the next section.

Branch and Bound. Here, we present a modified branch and bound method to solve the constrained canonical correlation problem in (30) similar to that used in integer programming [see Garfinkel & Nemhauser, 1972]. In order to understand the algorithm, consider the tree illustrated in Figure 2 which considers the case where $p = q = 2$. Each node in this tree corresponds to a problem of the form:

$$\max_{\mathbf{w}, \mathbf{v}} z(\mathbf{w}, \mathbf{v}) \tag{33}$$

subject to all the constraints listed above that node. For example, the problem at node seven is:

$$\max_{\mathbf{w}, \mathbf{v}} z(\mathbf{w}, \mathbf{v}),$$

such that:

$$w_1 = 1,$$

$$\text{and} \quad w_2 = 0. \tag{34}$$

Note that the leaf (or terminating) nodes (nodes 40 through 120) correspond to the feasible solutions of (30).

Let the maximal value of $z(\mathbf{w}, \mathbf{v})$ at node i be z_i. Then if i is an ancestor of j (i.e., above node j in the tree), $z_i \geq z_j$. We use this relationship to limit the number of nodes which we must evaluate. Suppose we have evaluated the leaf node t and the value is z_t. Suppose further, that node i has value $z_i < z_t$. Then we know that leaf t is better than any node, and specifically any leaf, which is a descendant of node i. Thus, none of those descendants need to be evaluated.

Also, since in our problem $\mathbf{w}$ is equivalent to $-\mathbf{w}$, we can eliminate nodes 3 and 6, along with their descendants, due to redundancy. Similar pruning is possible lower in the tree due to redundancy among the $\mathbf{v}$'s.

The branch and bound algorithm essentially consists of working our way down the tree, visiting nodes with high values first. We stop when we have found a leaf node with a value which we can guarantee is larger than that of any other leaf node.

The remaining problem is the evaluation of z_i. Note first that $z(\mathbf{w}, \mathbf{v})$ is invariant under scale transformations to $\mathbf{w}$ and $\mathbf{v}$. Thus, in place of the constraints, say

$$\mathbf{w}' = (1, -1)$$

$$\mathbf{v}' = (1, 1) \tag{35}$$

we can use

$$w_1 = -w_2$$
$$v_1 = v_2 \tag{36}$$

or

$$\mathbf{w} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} w_1, \qquad \mathbf{v} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \mathbf{v}_1 \tag{37}$$

for some nonzero w_1 and v_1.

We may therefore restate the problem for node i as

$$\operatorname*{Max}_{\mathbf{a},\mathbf{b}} z_i,$$

where

$$z_i = \frac{\mathbf{w}'\mathbf{R}_{xy}\mathbf{v}}{(\mathbf{w}'\mathbf{R}_{xx}\mathbf{w})^{1/2}(\mathbf{v}'\mathbf{R}_{yy}\mathbf{v})^{1/2}} \tag{38}$$

such that

$$\mathbf{w} = \mathbf{T}_1\mathbf{a},$$
$$\mathbf{v} = \mathbf{T}_2\mathbf{b},$$

and $\mathbf{T}_i$ has the form:

$$\mathbf{T}_i = \begin{bmatrix} \mathbf{t}_i & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \qquad i = 1, 2. \tag{39}$$

Here, $\mathbf{t}_1$ and $\mathbf{t}_2$ handle the constraints for that node. If there are c_1 constraints on $\mathbf{w}$ and c_2 constraints on $\mathbf{v}$ at this node, then $\mathbf{t}_1$ is $c_1 \times 1$, $\mathbf{T}_1$ is $p \times (p - c_1 + 1)$, $\mathbf{a}$ is $(p - c_1 + 1) \times 1$, $\mathbf{t}_2$ is $c_2 \times 1$, $\mathbf{T}_2$ is $q \times (q - c_2 + 1)$, and $\mathbf{b}$ is $(q - c_2 + 1) \times 1$. $\mathbf{T}_1$ maps the first element of $\mathbf{a}$ into the first c_1 elements of $\mathbf{w}$ and the rest of the elements of $\mathbf{a}$ into the remaining elements of $\mathbf{w}$. Similarly, $\mathbf{T}_2$ maps the first element of $\mathbf{b}$ into the first c_2 elements of $\mathbf{v}$ and the rest of the elements of $\mathbf{b}$ into the remaining elements of $\mathbf{v}$.

If we replace $\mathbf{w}$ and $\mathbf{v}$ by $\mathbf{T}_1\mathbf{a}$ and $\mathbf{T}_2\mathbf{b}$ respectively, then we can rewrite (38) as

$$\operatorname*{Max}_{\mathbf{a},\mathbf{b}} z_i,$$

where

$$z_i = \frac{\mathbf{a}'\mathbf{T}_1'\mathbf{R}_{xy}\mathbf{T}_2\mathbf{b}}{(\mathbf{a}'\mathbf{T}_1'\mathbf{R}_{xx}\mathbf{T}_1\mathbf{a})^{1/2}(\mathbf{b}'\mathbf{T}_2'\mathbf{R}_{yy}\mathbf{T}_2\mathbf{b})^{1/2}}, \tag{40}$$

which can be solved by standard canonical correlation techniques.

This branch and bound algorithm will provide the first constrained canonical correlation and the corresponding constrained canonical variates.

In continuous (unconstrained) canonical correlation analysis, further canonical variates are defined by a similar maximization, but with the additional constraints that they be uncorrelated with the prior canonical variates. If we attempt to add these constraints to our formulation as well, then a second constrained canonical correlation may not exist in general. This will be discussed in further detail later. We can, however, require that further discrete canonical weight vectors be orthogonal to the prior discrete canonical weight vectors in the same set. This is handled in the algorithm by checking for feasibility before evaluating a node. Specifically, a node is considered feasible if and only if at least one of its descendants is a leaf which satisfies the orthogonality constraints. The complete method for evaluating a node under the enforced orthogonality is covered in Appendix II.

In some cases, we must stop before extracting all canonical factors. Consider, for example, a case with three predictor variables where $\mathbf{w}_1 = [1 \quad 1 \quad 1]'$ and $\mathbf{w}_2 = [1 \quad 0 \quad -1]'$. Due to orthogonality, $\mathbf{w}_3$ would have to be a scalar multiple of $[1 \quad -2 \quad 1]'$. But this conflicts withour requirement that

$$w_i \in \{0, \pm 1\}, \qquad \forall i = 1, \ldots, p. \tag{41}$$

In most canonical correlation analyses, however, we are only interested in the first few factors so that this problem is not a major consideration.

Algorithm Comparisons. One can compare the three algorithms presented in terms of the following criteria:

a. *Cost*—cost of computation (e.g., CPU Time);
b. *Accuracy*—whether algorithm can always provide the globally optimum solution; and,
c. *Flexibility*—whether the algorithm can present a number of "good" solutions (e.g., solutions where many coefficients are set equal to zero).

The complete enumeration procedure is clearly the most expensive to run. It does, however, provide the globally optimum solution. It lacks flexibility in terms of not being able to readily provide alternative "good" solutions.

The combinatorial optimization procedure is the cheapest to run, especially for large $p + q$. However, there is no guarantee that the globally optimum solution will be found. It does have the flexibility of being able to provide a number of locally opitmum solutions as previously mentioned.

The branch and bound procedure is much cheaper to run than complete enumeration, but is more expensive to run than the combinatorial optimization algorithm, especially for large $p + q$. It does have the advantage of guaranteeing a globally optimum result. However, because of the branching process, it is difficult to generate a range of "good" suboptimal solutions. One could circumvent this latter difficulty by using a modified objective function which associated a positive reward with each coefficient which is set to zero. The resulting solutions will then have the property that they have maximum canonical correlation for solution vectors with at least that many zeros.

An Example—Wharton MBA 5 × 5 Case

Study Description

Thirty-four Wharton MBA graduate students in Marketing from the University of Pennsylvania completed a questionnaire which examined their driving attitudes, self image, demographic profile, automobile brand preferences, and automobile usage. From this larger study, we choose to investigate the relationship, if any, between five psychographic questions, **Y**, concerning driving attitudes and five other psychographic questions, **X**, measuring self image. The items, listed in Table 2, are all measured on the same seven-point scale from -3 to $+3$ where, for example, "-3" denotes completely disagree, "0" denotes neutral, and "$+3$" denotes completely agree.

The Unconstrained Solution

Table 2 also shows the correlation matrix **R** for these ten variables, and Table 3 presents the results of the unconstrained canonical correlation analysis. Only the first set of factors is significant ($p < .05$), and is listed in Table 4 with the associated structure correlations, i.e., the correlations between the factors and the variables.

Table 2

Psychographic Questions and Correlations for Example

Question	Variable	Y_1	Y_2	Y_3	Y_4	Y_5	X_1	X_2	X_3	X_4	X_5
I spend a lot of time driving.	Y_1	1.000	0.614	0.494	0.291	−0.289	−0.189	0.386	0.277	0.022	0.119
I need a car because of my work.	Y_2		1.000	0.415	0.116	−0.099	0.121	0.137	0.328	−0.252	0.069
I enjoy driving.	Y_3			1.000	0.396	−0.295	−0.134	0.318	0.440	0.197	0.031
I'd rather drive on the turnpike than in the city.	Y_4				1.000	−0.027	−0.037	0.107	0.461	0.101	0.086
Given a choice I'd prefer public transportation over driving.	Y_5					1.000	0.306	0.039	−0.195	−0.011	−0.251
Those who know me consider me thrifty.	X_1						1.000	−0.092	0.021	−0.008	0.150
I'm influential in my circle of friends.	X_2							1.000	−0.056	0.040	−0.241
I like to know what tomorrow has in store for me.	X_3								1.000	0.097	0.279
I admit I dress to please others.	X_4									1.000	0.449
I try to keep up with the Jones'.	X_5										1.000

A question that immediately arises is how one should interpret these results in Table 4. If one were to examine the canonical coefficients together with the associated structure correlations, it is not obviously clear what the appropriate interpretation should be. What should the decision rule be for claiming that a canonical coefficient or structure correlation is large enough for that variable to be included in the interpretation of the results? What if conflicts arise? For example, Table 5 shows that v_4 is small yet the associated structure correlation is quite large. There appears to be some ambiguity here with respect to interpretation. There is also the question of how stable these coefficients are—how well are they determined? These issues will now be examined by applying our response surface methods and constrained canonical correlation procedure to this example.

Table 3

Canonical Correlation Results for Example

Number	Eigenvalue	Canonical Correlation	Wilks Lambda	Chi-Square	D.F.	Significance
1	0.443	0.666	0.245	38.588	25	0.040
2	0.344	0.587	0.442	22.448	16	0.129
3	0.235	0.484	0.674	10.829	9	0.288
4	0.117	0.343	0.881	3.459	4	0.484
5	0.000	0.024	0.999	0.016	1	0.897

Table 4

Canonical Coefficients and Structure Correlation for Example

Canonical Coefficients

Coefficients for Variables of the Criterion Set		Coefficients for Variables of the Predictor Set		Structure Correlations			
	$\mathbf{v}_1$		$\mathbf{w}_1$		$\mathbf{v}_1'\mathbf{Y}$		$\mathbf{w}_1'\mathbf{X}$
v_1	0.546	w_1	−0.310	Y_1	0.707	X_1	−0.357
v_2	−0.370	w_2	0.588	Y_2	0.267	X_2	0.598
v_3	0.670	w_3	0.311	Y_3	0.867	X_3	0.638
v_4	0.229	w_4	0.311	Y_4	0.610	X_4	0.390
v_5	0.029	w_5	−0.024	y_5	−0.296	X_5	0.109

Response Surface Application

Table 5 presents the results for single parameter changes at ε values of .001, .005, .01, and .05. Clearly, as ε grows, so does the indifference region for each parameter. Even for very small ε, one notes some degree of coefficient instability. In particular, for ε as small as .005, mostly all of the coefficients vary so as to render ambiguous interpretations depending upon whether one examines the upper limit (U) or lower limit (L) value. Specifically, the w_5 and v_5 estimates can change signs, while the remainder of the estimates, in general, can vary between relatively small values ($|L| \leq .5$) to relatively high values ($|U| \geq .5$). Interpretation can thus become a problem considering these individual coefficient instabilities. The problem obviously becomes worse when one considers $\varepsilon = .05$ where there are substantial fluctuations as Table 5 demonstrates. These approximations were evaluated to check their accuracy by substituting each upper and lower bound in the optimal solution one at a time and calculating the value of the objective function to see if it were

Table 5

Single Parameter Changes for Example

U = upper limit value
L = lower limit value

	The First Factor Coefficients for Global Optimum	ϵ .001		.005		.01		.05	
		L	U	L	U	L	U	L	U
w_1	−.311	−.362	−.261	−.425	−.198	−.472	−.151	−.671	.048
w_2	.588	.529	.647	.456	.720	.401	.775	.169	1.007
w_3	.654	.593	.716	.517	.792	.460	.850	.219	1.091
w_4	.312	.260	.363	.196	.427	.148	.475	−.053	.676
w_5	−.025	−.072	.023	−.131	.082	−.176	.126	−.362	.313
v_1	.546	.478	.613	.395	.696	.333	.758	.071	1.020
v_2	−.371	−.420	−.321	−.481	−.260	−.526	−.215	−.719	−.022
v_3	.670	.574	.765	.456	.883	.368	.972	−.006	1.345
v_4	.229	.170	.290	.096	.364	.040	.419	−.194	.653
v_5	.030	−.021	.078	−.083	.140	−.129	.186	−.323	.380

within the ε-value from .666, the globally optimum value. While the obtained value was never *exactly* on the boundary, the approximation was quite close for all four ε values checked.

Table 6 contains the results for the case of all parameters changing at once. Here, the results are expressed in terms of the amount that is to be added or subtracted from the global optimum solution so as to lie in the specified ε-indifference region. Although the results here are not nearly so drastic as that presented in Table 5 (as to be expected), one still notes a substantial amount of "indeterminacy" for higher ε values, especially for coefficient estimates for the "worst case" for w_4, v_2, and v_4. Again, in all cases, for all values of ε specified, these approximations were checked by adjusting the globally optimal solution by these values and then evaluating the objective function. Here too, while these calculated values rarely lay *exactly* on the ε-indifference region boundary, they were very close to it.

Constrained Canonical Correlation Results

All three algorithms were applied to the correlation matrix presented in Table 2, and exactly the same solution was obtained. Table 7 presents the globally optimum constrained solution for the first factor and the associated canonical correlation value. As one can readily note, the value of the objective function (.649) is not reduced all that much (.071) by the constraints. In addition, interpretation of the results is much easier. At one extreme, those who score highly on the first criterion factor tend to agree with:

—I spend a lot of time driving,
—I do not need car for work or business,
—I enjoy driving.

Those who score highly on the first predictor factor tend to agree with:

—I am not considered thrifty,
—I am influential with friends,
—I like to know what tomorrow brings,
—I dress to please others.

Table 6

All Parameter Changes For Example

B = Best Case
W = Worst Case

Coefficients for the First Factor	Global Optimum	$\epsilon = .001$ W	$\epsilon = .001$ B	$\epsilon = .005$ W	$\epsilon = .005$ B	$\epsilon = .01$ W	$\epsilon = .01$ B	$\epsilon = .05$ W	$\epsilon = .05$ B
		Distance along the Axis							
w_1	−.311	+.011	+.000	+.024	+.001	+.034	+.001	+.076	+.003
w_2	.588	−.007	+.015	−.015	+.033	−.022	+.047	−.049	+.105
w_3	.654	+.051	−.005	+.115	−.011	+.163	−.016	+.364	−.035
w_4	.312	−.081	−.019	−.182	−.042	−.257	−.059	−.574	−.132
w_5	−.025	+.048	−.019	+.107	−.043	+.152	−.061	+.339	−.136
v_1	.546	−.002	−.008	−.004	−.018	−.006	−.026	−.014	−.058
v_2	−.371	+.074	−.013	+.166	−.029	+.234	−.040	+.524	−.090
v_3	.670	+.018	−.001	+.041	−.003	+.057	−.005	+.128	+.010
v_4	.229	+.077	+.004	+.172	+.009	+.243	+.013	+.542	+.029
v_5	.030	−.045	−.009	−.102	−.020	−.144	−.028	−.321	−.062

Table 7

Discrete Canonical Correlation Solution

	Continuous Solution	Discrete Solution
w_1	−.311	−1
w_2	.588	1
w_3	.654	1
w_4	.312	1
w_5	−.025	0
v_1	.546	1
v_2	−.371	−1
v_3	.670	1
v_4	.229	0
v_5	.030	0
Correlation	.666	.649

In this simple illustration, driving frequency and enjoyment appear to be positively related to sociability-related items.

Consideration of Criteria for the Extraction of More Than One Factor

Recall that the unconstrained solution provides vectors that render uncorrelated canonical variates. In general, however, the corresponding eigenvectors within each set are *not* orthogonal.

The constrained case discussed presents some difficulties along this line when one extracts more than one factor. The problem lies in which criteria to select as a constraint:

(i) uncorrelatedness of canonical variates,
(ii) orthogonality of coefficient vectors,
(iii) maximization of "partial" canonical correlation, or
(iv) some combination of the above.

Surely, one wishes to derive additional factors which present a different view of the data. Without requiring any additional constraints for extracting other factors, one would derive factors which are identical to the first.

In most applications, the uncorrelatedness of subsequent canonical factors (within sets at least) is perhaps the most relevant characteristic that one would want the results to reflect. However, as was indicated earlier, it may not always be possible to derive subsequent uncorrelated factors. For example, for most low-dimensional applications, say $p = q = 2$, uncorrelatedness under discrete canonical correlation is possible only under very extraordinary circumstances such as in the Morrison example. Even for problems of rather large dimensionality, such a "tight" constraint might remove such a large portion of possible discrete solutions that those feasible ones remaining may result in a low canonical correlation.

The second approach, and one which is currently utilized as was mentioned earlier, is to constrain the derived vectors to be orthogonal, i.e.,

$$\mathbf{w}_i'\mathbf{w}_j = 0$$

and

$$\mathbf{v}_i'\mathbf{v}_j = 0, \qquad \forall i \neq j;\ i, j = 1, \ldots, \min\{p, q\}. \tag{42}$$

While this does not, in general, ensure uncorrelated factors, it does provide a method of obtaining additional information concerning the structure of the data. This procedure should, for many cases, provide low intercorrelations between canonical variates of the same set. In addition, the above constraints are easily embedded in the algorithms.

Returning to the Wharton MBA example, we have presented the first three factors in the discrete analysis in Table 8. Table 8 also presents their intercorrelations. Ignoring the interpretation of these factors (since the last two factors are not significant) for now, one notes that there is a more pronounced drop in canonical correlation for the last two factors as compared to their continuous counterparts [from Tables 4 and 10, factor two: .587 (continuous) vs. .463 (discrete); factor three; .484 (continuous) vs. .395 (discrete)], which indicates that the orthogonality constraint does appear to be quite tight. Unfortunately, the orthogonality constraint does not provide pure uncorrelatedness as demonstrated by the one or two relatively high inter-factor correlations in Table 8. Naturally, the dimensionality of the problem whould have some effect on the degree of interset correlation among factors as well as the original correlation structure **R**.

The third method for extracting additional factors involves updating the correlation matrix after each factor is extracted. This technique eliminates the need for the tight orthogonality or uncorrelatedness constraints discussed above. Weinberg and Darlington [1976] regress each of the variables in each of the two sets on the first canonical variate from the corresponding set and then calculate an updated correlation matrix from the residuals. We prefer to update the correlation matrix by partialling out the effects of both of the canonical variates on all of the variables. The maximum canonical correlation

Table 8

First Three Discrete Factors of Wharton MBA Example

	FACTOR				Correlation Matrix for Three Factors in Discrete Analysis					
	I	II	III		$\mathbf{w}_1'X$	$\mathbf{w}_2'X$	$\mathbf{w}_3'X$	$\mathbf{v}_1'Y$	$\mathbf{v}_2'Y$	$\mathbf{v}_3'Y$
w_1	−1	0	−1	$\mathbf{w}_1'X$	1.000	+.065	−.129	+.649	+.394	+.196
w_2	1	0	0	$\mathbf{w}_2'X$		1.000	+.229	.000	+.464	+.179
w_3	1	−1	−1	$\mathbf{w}_3'X$			1.000	+.067	+.084	.395
w_4	1	1	0							
w_5	0	−1	1							
v_1	1	1	0	$\mathbf{v}_1'Y$				1.000	+.539	+.258
v_2	−1	1	1	$\mathbf{v}_2'Y$					1.000	−.441
v_3	1	0	1	$\mathbf{v}_3'Y$						1.000
v_4	0	1	0							
v_5	0	−1	1							
Correlation	.649	.464	.395							

associated with this matrix is then the maximum partial canonical correlation given the first two canonical variates.

Let the original variables be denoted by

$$\mathbf{T} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix}, \tag{43}$$

where $\mathbf{X}$ is a p-dimensional vector and $\mathbf{Y}$ is q-dimensional. Furthermore, let $\mathbf{R}$ denote the correlation matrix for $\mathbf{T}$ where

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{xx} & \mathbf{R}_{xy} \\ \mathbf{R}_{yx} & \mathbf{R}_{yy} \end{bmatrix}. \tag{44}$$

Let the first pair of discrete canonical variates be

$$A_1 = \mathbf{w}_1'\mathbf{X} \qquad \text{and} \qquad B_1 = \mathbf{v}_1'\mathbf{Y}, \tag{45}$$

and the "proposed" second pair be

$$A_2 = \mathbf{w}_2'\mathbf{X} \qquad \text{and} \qquad B_2 = \mathbf{v}_2'\mathbf{Y}. \tag{46}$$

The correlation matrix for $[A_1, B_1, A_2, B_2]'$ is thus:

$$\mathbf{R}_{A_1B_1A_2B_2} = \begin{vmatrix} \mathbf{w}_1'\mathbf{R}_{xx}\mathbf{w}_1 & \mathbf{w}_1'\mathbf{R}_{xy}\mathbf{v}_1 & \mathbf{w}_1'\mathbf{R}_{xx}\mathbf{w}_2 & \mathbf{w}_1'\mathbf{R}_{xy}\mathbf{v}_2 \\ \mathbf{v}_1'\mathbf{R}_{yx}\mathbf{w}_1 & \mathbf{v}_1'\mathbf{R}_{yy}\mathbf{v}_1 & \mathbf{v}_1'\mathbf{R}_{yx}\mathbf{w}_2 & \mathbf{v}_1'\mathbf{R}_{yy}\mathbf{R}_{yy}\mathbf{v}_2 \\ w_2'\mathbf{R}_{xx}\mathbf{w}_1 & \mathbf{w}_2'\mathbf{R}_{xy}\mathbf{v}_1 & \mathbf{w}_2'\mathbf{R}_{xx}\mathbf{w}_2 & \mathbf{w}_2'\mathbf{R}_{xy}\mathbf{v}_2 \\ \mathbf{v}_2'\mathbf{R}_{yx}\mathbf{w}_1 & \mathbf{v}_2'\mathbf{R}_{yy}\mathbf{v}_1 & \mathbf{v}_2'\mathbf{R}_{yx}\mathbf{w}_2 & \mathbf{v}_2'\mathbf{R}_{yy}\mathbf{v}_2 \end{vmatrix}. \tag{47}$$

This matrix may be rewritten as

$$\mathbf{R}_{A_1B_1A_2B_2} = \begin{bmatrix} \mathbf{c}_1'\mathbf{R}\mathbf{c}_1 & \mathbf{c}_1'\mathbf{R}\mathbf{c}_2 \\ \mathbf{c}_2'\mathbf{R}\mathbf{c}_1 & \mathbf{c}_2'\mathbf{R}\mathbf{c}_2 \end{bmatrix}, \tag{48}$$

where

$$\mathbf{c}_1 = \begin{bmatrix} \mathbf{w}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{v}_1 \end{bmatrix} \qquad \text{and} \qquad \mathbf{c}_2 = \begin{bmatrix} \mathbf{w}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{v}_2 \end{bmatrix}. \tag{49}$$

We may now write the conditional covariance matrix of A_2, B_2 given A_1, B_1, as

$$\begin{aligned} \mathbf{R}_{A_2B_2\cdot A_1B_1} &= \mathbf{c}_2'\mathbf{R}\mathbf{c}_2 - \mathbf{c}_2'\mathbf{R}\mathbf{c}_1(\mathbf{c}_1'\mathbf{R}\mathbf{c}_1)^{-1}\mathbf{c}_1'\mathbf{R}\mathbf{c}_2 \\ &= \mathbf{c}_2'[\mathbf{R} - \mathbf{R}\mathbf{c}_1(\mathbf{c}_1'\mathbf{R}\mathbf{c}_1)^{-1}\mathbf{c}_1'\mathbf{R}]\mathbf{c}_2 \\ &= \mathbf{c}_2'\mathbf{R}_1\mathbf{c}_2, \end{aligned} \tag{50}$$

where

$$\mathbf{R}_1 = \mathbf{R} - \mathbf{R}\mathbf{c}_1(\mathbf{c}_1'\mathbf{R}\mathbf{c}_1)^{-1}\mathbf{c}_1'\mathbf{R}. \tag{51}$$

$\mathbf{R}_1$ then represents the covariance matrix for $\mathbf{T}$ after partialling out the effects of A_1 and B_1.

We partition $\mathbf{R}_1$ as

$$\mathbf{R}_1 = \begin{bmatrix} \mathbf{R}_{1xx} & \mathbf{R}_{1xy} \\ \mathbf{R}_{1yx} & \mathbf{R}_{1yy} \end{bmatrix}, \tag{52}$$

and the partial correlation between A_2 and B_2 can now be seen to be

$$\rho_{A_2B_2\cdot A_1B_1} = \frac{\mathbf{w}_2'\mathbf{R}_{1xy}\mathbf{v}_2}{[(\mathbf{w}_2'\mathbf{R}_{1xx}\mathbf{w}_2)(\mathbf{v}_2'\mathbf{R}_{1yy}\mathbf{v}_2)]^{1/2}}. \tag{53}$$

Table 9

Extraction of Additional Factors from Wharton MBA Example by Maximizing Partial Correlation

	FACTOR I	II	III
w_1	-1	0	-1
w_2	1	0	0
w_3	1	1	-1
w_4	1	-1	0
w_5	0	1	1
v_1	1	0	1
v_2	-1	1	0
v_3	1	1	1
v_4	0	1	0
v_5	0	-1	-1
Partial Correlation:	.649	.555	.432

We may now maximize $\rho_{A_2 B_2 \cdot A_1 B_1}$ subject to the constraint that components of $\mathbf{w}_2$ and $\mathbf{v}_2$ must be in $\{-1, 0, 1\}$, thereby obtaining the optimal choice of A_2 and B_2. It is clear that this process may be extended to extract additional factors.

This optimization may be carried out using any of the discrete optimization algorithms mentioned earlier. We only need replace $\mathbf{R}$ by $\mathbf{R}_1$. We feel that this is an improvement over the Weinberg and Darlington [1976] heuristic in that our procedures attempt to optimize a meaningful objective function. In addition, our procedure guarantees that the coefficients of the original variables which determine the canonical variates will all be in $\{-1, 0, 1\}$. This is true only for the first factor if one uses the Weinberg and Darlington procedure.

This procedure can also be coupled with orthogonality constraints. This we have done for the Wharton MBA example. Our findings are reported in Table 9.

Discussion and Needs for Future Research

Relationships Between Continuous and Discrete Solutions

A naive approach to the discrete canonical correlation problem is to suggest somehow rounding off the continuous solution to obtain the constrained solution. For example, one could arbitrarily devise the following rule for a particular factor:

If: w_i or $v_i < -c_1$ then assign it the value -1;

w_i or $v_i > +c_2$ then assign it the value $+1$;

otherwise, assign it the value 0. (53)

Using this ad hoc decision rule for the first continuous factor in our Wharton MBA example with $c_1 = c_2 = .5$, one obtains

$$\mathbf{w}_1 = (0, 1, 1, 0, 0)$$
$$\mathbf{v}_1 = (1, 0, 1, 0, 0), \tag{55}$$

rendering a suboptimal canonical correlation of $z = .599$ (vs. the .649 discrete canonical correlation value).

If one set $c_1 = c_2 = .3$, one could recover the constrained optimum solution. This is true for this example, but not in general. It is clear that if such a rounding rule is to be useful, a wise choice of c_1 and c_2 must be made. In addition, the results to be presented in the next section make it clear that such a rounding rule cannot be successful in all cases. The conditions under which such a rule leads to the optimal discrete canonical correlation value are therefore of interest.

Perhaps a more profitable avenue of research in canonical correlation analysis would be to investigate the effects of rounding optimal estimates to the first decimal place in canonical correlation similar to what Bibby [1980] has done for regression and principal components analyses.

Effects of Multicollinearity

What are the effects of multicollinearity upon both continuous and discrete solutions? Levine [1977] has found that high levels of multicollinearity contribute to the instability of the continuous canonical correlation coefficients. What happens to the discrete solution?

We adjusted the correlation matrix for the Wharton MBA example by altering only one entry in $\mathbf{R}_{yy}$ in order to introduce additional multicollinearity. We have changed the correlation between Y_1 and Y_2 from .615 to .790. Table 10 presents the continuous solutions for the first factor for both before and after the adjustments. Most of the coefficients change quite drastically:

(i) w_3 changes in sign and quite substantially in magnitude;
(ii) Coefficients w_4, v_1, v_2 and v_5 all change in magnitude;
(iii) v_4 changes in sign.

Note the increase in the canonical correlation coefficient obtained merely by altering one entry (and its symmetric counterpart) in $\mathbf{R}_{yy}$.

Effects of Multicollinearity on Continuous and Discrete Solutions

	Effects of Multicollinearity on Continuous Solution			Effects of Multicollinearity on Discrete Solution	
	FACTOR 1			FACTOR 1	
	Before	After		Before	After
w_1	−.310	−.473	w_1	−1	−1
w_2	.588	.536	w_2	+1	+1
w_3	.654	−.120	w_3	+1	+1
w_4	.311	.634	w_4	+1	+1
w_5	.024	−.080	w_5	0	0
v_1	.546	1.515	v_1	+1	+1
v_2	−.370	−1.572	v_2	−1	−1
v_3	.670	.517	v_3	+1	+1
v_4	.229	−.358	v_4	0	0
v_5	.029	.315	v_5	0	0
Correlation	.666	.804	Correlation	.649	.717

Table 10 also depicts the constrained solution. Here *none* of the elements change and a canonical correlation of .717 results. Clearly, for this particular example, this solution has not been affected by the level of multicollinearity artifically introduced.

More research is needed to examine the effect of multicollinearity and other types of dependence structure "abnormalities" on both continuous and discrete solutions. In particular, it would be desirable to perform a comparative study of the robustness of the constrained vs. continuous solutions against various changes in **R**.

Extension to Other Models

These same constrained methodologies could be applied similarly to other models such as regression, multiple discriminant analysis, MANOVA, principal components/factor analysis, and multidimensional scaling. In fact, since regression, MANOVA and discriminant analysis can be expressed in terms of "special" canonical correlation problems, one could adapt the current algorithms and apply them directly [see Hausman, Note 1].

In fact, some literature exists on the effects of imposing such constraints $(+1, 0, -1)$ in multiple regression. Many authors [e.g., Wainer, 1976] have proposed that such directional weights for predictors may yield almost as accurate prediction of a criterion as the more optimal least-squares coefficients, and that they may cross-validate better. Bentler and Woodward [1979] present a heuristic algorithm for imposing such constraints in multiple regression.

Hausman [Note 2] has successfully developed a branch and bound algorithm for principal components which restricts the coefficients of the components to $+1$, 0, or -1. Hausman and DeSarbo are currently investigating the response surface for principal components analysis. Certainly, the same can be done for many multidimensional scaling methods.

Expansion of Feasible Values/Scaling Considerations

The elements of $\mathbf{w}$ and $\mathbf{v}$ obtained by the above methods are, of course, coefficients of the standardized variables. Coefficients for the original variables may be obtained by dividing the coefficients of the standardized variables by the standard deviations of the original variables. Note that these coefficients of the original variables may not all be of the same order of magnitude or even in the same units.

If, however, all the variables in the X-set are measured on commensurate scales and all the variables in the Y-set are measured on commensurate scales, we may be more interested in constraining the coefficients of the original variables. This may be accomplished by using the covariance matrix in place of the correlation matrix to obtain the desired coefficients directly. Note that both of these methods for obtaining coefficients of the original variables lead to the same coefficients in a standard canonical correlation analysis. In the constrained analysis, however they lead to different solutions since the constraints on the coefficients of the standardized variables are not equivalent to the constraints on the coefficients of the original variables.

For applications where the variables within a set are scaled or measured quite differently, one can also relax the constraint somewhat to include numerous other discrete integer and/or noninteger values as previously mentioned. Each of the algorithms can be adjusted to accommodate such a change.

Cross-Validation Comparisons

As mentioned earlier, many authors have recommended cross-validating the results of canonical correlation whenever possible because of the coefficient instability problem.

This usually entails performing the same analysis on either a "hold-out" sample or two randomly split halves of the original sample, and comparing results by computing the canonical correlation in one sample corresponding to the vectors derived in the other sample. Thorndike and Weiss [1973] report some inherent difficulties with this procedure. In practice, "variation in sample-specific error from one sample to another causes shrinkage in cross-validation or 'bouncing betas' in replication of many studies using multiple or canonical correlation" (p. 125). Thorndike, et al. [1968] have performed cross-validation studies with real data finding that such problems do indeed arise.

As discussed earlier, Weinberg and Darlington [1976] propose a $\{+1, 0, -1\}$ constrained canonical correlation procedure which attempts to alleviate some of these problems. In 56 empirical comparisions between their technique and ordinary canonical correlation, they find that their constrained approach is superior in terms of the cross-validation criteria tested (higher cross validities and lower validity shrinkage). Due to our small sample size ($n = 34$) with our MBA example, we feel that such split-half cross-validation would not be very meaningful.

More research should be performed in this area, especially concerning comparison in validity/cross-validity between continuous and our discrete $\{-1, 0, +1\}$ canonical correlation analyses. In addition, jackknifing and bootstrapping techniques could be employed to examine the stability of solutions over subjects and/or variables.

Appendix I

Complete Enumeration Algorithm

In this Appendix we wish to develop an efficient complete enumeration algorithm for the discrete canonical correlation problem with p and q variables in the two sets.

Problem: List all leaves of a tree with depth $n = p + q$ where each nonleaf node has j children. Thus, we want a function f such that:

(i) $f(t) = (x_n(t), \ldots, x_1(t))$,
(ii) The range of x_i is $\{0, \ldots, j-1\}$,
(iii) as t varies between 1 and j^n, we reach each point in E, the n-fold Cartesian product of $\{0, 1, \ldots, j-1\}$.

Solution: Associate with each t its "unique" representation in base j. As t ranges between 1 and j^n, we enumerate all outcomes.

Example:

$$n = 10$$

$$p = 5$$

$$q = 5$$

$$j = 3$$

To convert t to its trinary representation, use "synthetic division", e.g.:

$$\text{for} \quad t = 17 = 122_3,$$

$$x_1 = 2 \qquad x_1' = -1$$

$$x_2 = 2 \qquad x_2' = -1$$

$$x_3 = 1 \qquad x_3' = 1$$

$$
\begin{array}{cc}
x_4 = 0 & x_4' = 0 \\
\cdot\cdot & \cdot\cdot \\
\cdot\cdot & \cdot\cdot \\
\cdot\cdot & \cdot\cdot \\
x_{10} = 0 & x_{10}' = 0,
\end{array}
$$

where the conversion rule is:

$$
\begin{array}{cc}
x_1 & \Rightarrow x_i' \\
0 & 0 \\
1 & +1 \\
2 & -1.
\end{array}
$$

Computational Savings. In our example, we can skip those outcomes that render solutions in sets A, B, C, and D below:

(i) A: $\{x: 0\ 0\ 0\ 0\ 0\ x_5\, x_4\, x_3\, x_2\, x_1\}$

Here, the first canonical variate would have zero variance, so we can ignore the first j^p or 243 calculations and start with $t = 244$;

(ii) B: $\{x: x_{10}\, x_9\, x_3\, x_7\, x_6\ 0\ 0\ 0\ 0\ 0\}$

We have 3^5 of these which correspond to those t divisible by 3^5:

$$\#\ \{B\}\backslash\{\mathbf{0}\} = 3^5 - 1;$$

(iii) C: $\{x: x_i = 2 \text{ and } x_{i+1} = x_{i+2} = \ldots x_{10} = 0 \text{ for some } i > 5\}$
(iv) D: $\{x: x_i = 2 \text{ and } x_{i+1} = x_{i+2} = \ldots x_5 = 0 \text{ for some } i \leq 5\}$

Outcomes in C and D may be ignored since canonical correlation is not affected by reflection of the signs of either set of canonical variates.

As a result, the total number of outcomes which we must consider is slightly less than $3^{p+q}/4$.

Appendix II

Node Evaluation Under Enforced Orthogonality

In this Appendix, we develop the mathematics for node evaluation in the branch and bound algorithm under enfored orthogonality.

Let $\mathbf{C}$ be the $k \times p$ matrix whose columns are the first k factors in the first set. Let $\mathbf{D}$ be the $k \times q$ matrix defined similarly for the second set. In solving for the k + 1st factor, our objective at any single node is to solve a problem of the form:

$$\max_{\mathbf{t},\,\mathbf{u},\,\mathbf{v},\,\mathbf{w}} \ z = \mathbf{w}'\mathbf{R}_{xy}\mathbf{v}$$

such that

$$\mathbf{w}'\mathbf{R}_{xx}\mathbf{w} = 1,$$

$$\mathbf{v}'\mathbf{R}_{yy}\mathbf{v} = 1,$$

$$\mathbf{C}\mathbf{w} = 0,$$

$$\mathbf{D}\mathbf{v} = 0,$$

$$\mathbf{w} = \mathbf{A}\mathbf{t},$$

and

$$\mathbf{v} = \mathbf{B}\mathbf{u}, \tag{AII.1}$$

where **A** and **B** depend on the particular node being evaluated.

By substituting **At** for **w** and **Bu** for **v**, we can rewrite (AII.1) as

$$\operatorname*{Max}_{\mathbf{t},\,\mathbf{u}} z = \mathbf{t}'\mathbf{A}'\mathbf{R}_{xy}\mathbf{B}\mathbf{u}$$

such that

$$\mathbf{t}'\mathbf{A}'\mathbf{R}_{xx}\mathbf{A}\mathbf{t} = 1,$$

$$\mathbf{u}'\mathbf{B}'\mathbf{R}_{yy}\mathbf{B}\mathbf{u} = 1,$$

$$\mathbf{C}\mathbf{A}\mathbf{t} = 0,$$

and

$$\mathbf{D}\mathbf{B}\mathbf{u} = 0. \tag{AII.2}$$

Suppose **A** is $p \times n$ ($n \leq p$). Then **CA** is $k \times n$. Let the rank of **CA** be r. If $r \geq n$, then t must be the null vector and the value of the node is zero. Otherwise, let $(\mathbf{CA})^*$ be an $n \times (n-r)$ matrix of full column rank $(n-r)$, such that

$$(\mathbf{CA})(\mathbf{CA})^* = \mathbf{0}. \tag{AII.3}$$

Then the statement

$$\mathbf{C}\mathbf{A}\mathbf{t} = \mathbf{0} \tag{AII.4}$$

and

$$\exists \mathbf{x} \quad \text{such that} \quad \mathbf{t} = (\mathbf{CA})^*\mathbf{x} \tag{AII.5}$$

both define $(n-r)$-dimensional linear subspaces for **t**. Further,

$$\mathbf{t} = (\mathbf{CA})^*\mathbf{x} \tag{AII.6}$$

implies

$$\mathbf{C}\mathbf{A}\mathbf{t} = \mathbf{C}\mathbf{A}(\mathbf{CA})^*\mathbf{x} = \mathbf{0}, \tag{AII.7}$$

so that they both define the same subspace. Thus, we can rewrite (AII.2) as

$$\operatorname*{Max}_{\mathbf{t},\,\mathbf{u},\,\mathbf{x},\,\mathbf{y}} z = \mathbf{t}'\mathbf{A}'\mathbf{R}_{xy}\mathbf{B}\mathbf{u}$$

such that

$$\mathbf{t}'\mathbf{A}'\mathbf{R}_{xx}\mathbf{A}\mathbf{t} = 1,$$

$$\mathbf{u}'\mathbf{B}'\mathbf{R}_{yy}\mathbf{B}\mathbf{u} = 1,$$

$$t = (\mathbf{CA})^*\mathbf{x},$$

and

$$\mathbf{u} = (\mathbf{DB})^*\mathbf{y}. \tag{AII.8}$$

Finally, we substitute $(\mathbf{CA})^*\mathbf{x}$ for **t** and $(\mathbf{DB})^*\mathbf{y}$ for **u**, yielding

$$\operatorname*{Max}_{\mathbf{x},\,\mathbf{y}} z = \mathbf{x}'(\mathbf{CA})^{*\prime}\mathbf{A}'\mathbf{R}_{xy}\mathbf{B}(\mathbf{DB})^*\mathbf{y}$$

such that

$$\mathbf{x}'(\mathbf{CA})^{*\prime}\mathbf{A}'\mathbf{R}_{xx}\mathbf{A}(\mathbf{CA})^*\mathbf{x} = 1$$

and

$$\mathbf{y}'(\mathbf{DB})^{*\prime}\mathbf{B}'\mathbf{R}_{yy}\mathbf{B}(\mathbf{DB})^{*}\mathbf{y} = 1. \tag{AII.9}$$

Thus, the value of the node is the largest root of

$$|\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx} - \lambda\mathbf{S}_{xx}| = \mathbf{0}, \tag{AII.10}$$

where

$$\mathbf{S}_{xy} = (\mathbf{CA})^{*\prime}\mathbf{A}'\mathbf{R}_{xy}\mathbf{B}(\mathbf{DB})^{*}, \tag{AII.11}$$

$$\mathbf{S}_{xx} = (\mathbf{CA})^{*\prime}\mathbf{A}'\mathbf{R}_{xx}\mathbf{A}(\mathbf{CA})^{*}, \tag{AII.12}$$

and

$$\mathbf{S}_{yy} = (\mathbf{DB})^{*\prime}\mathbf{B}'\mathbf{R}_{yy}\mathbf{B}(\mathbf{DB})^{*}. \tag{AII.13}$$

Thus, we can embed the orthogonality constraints directly into the objective function and solve for subsequent factors as before.

REFERENCE NOTES

1. Hausman, R., Constrained multivariate analysis, *Unpublished Memorandum*, Bell Laboratories, Holmdel, N.J., 1980a.
2. Hausman, R., Constrained principal components, *Unpublished Memorandum*, Bell Laboratories, Holmdel, N.J., 1980b.
3. Kettenring, J. R., Canonical analysis, *Unpublished Memorandum*, Bell Laboratories, Murray Hill, N.J., 1979.

REFERENCES

Anderson, T. W., *Introduction to Multivariate Statistical Analysis*, New York: John Wiley & Sons, 1958.

Bard, Y., *Nonlinear Parameter Estimation*, New York: Academic Press, 1974.

Bartlett, M. S., On the theory of statistical regression, *Proceedings of the Royal Society of Edinburgh*, 1933, *53*, pp. 260–283.

Bentler, P. M., & Woodwards, J. A., Regression on linear composites: Statistical theory and applications. *Multivariate Behavioral Research Monographs*, 1979, No. 79-1.

Bibby, J., & Toutenberg, H., *Prediction and improved estimation in linear models*. New York: Wiley, 1977.

Bibby, J., Some Effects of Rounding Optimal Estimates, *Working Paper*, The Open University, England, 1980.

Chambers, J., *Computational Methods for Data Analysis*, New York: John Wiley & Sons, 1977.

Cliff, N. & Krus, D. J., Interpretation of canonical analysis: rotated vs. unrotated solutions, *Psychometrika*, 1976, *4*, 35–42.

Cooley, W. W., Canonical correlation, Paper presented at the 74-th Annual Convention, *American Psychological Association* (mimeograph), 1965.

Cooley, W. W. & Lohnes, P. R., *Multivariate Procedures for the Behavioral Sciences*, New York: John Wiley & Sons, 1962.

Darlington, R. B. & Bishop, C. H., Increasing test validity by considering interitem correlations, *Journal of Applied Psychology*, 1966, *50*, 322–330.

Darlington, R. B., Reduced-variance regression. *Psychological Bulletin*, 1978, *85*, 1238–1255.

Dempster, A. P., Estimation in multivariate analysis, in *Multivariate Analysis* (P. R. Krishnaiah, ed.), New York: Academic Press, 1966, 315–334.

Garfinkel, R. S. and Nemhauser, G. L., *Integer Programming*, New York: John Wiley & Sons, 1972.

Gnanadesikan, R., *Methods for Statistical Data Analysis of Multivariate Observations*, New York: John Wiley & Sons, 1977.

Green, B. R. Jr., Parameter sensitivity in multivariate methods. *Multivariate Behavioral Research*, 1977, *12*, 263–287.

Green, P. E., Carroll, J. D., & DeSarbo, W. S., An extended application of the δ^2 measure of predictor-variable importance, *Proceedings of the American Marketing Association Educator's Conference*, Chicago: A.M.A., 1979.

Hotelling, H., Relations between two sets of variates, *Biometrika*, 1936, *28*, 321–377.

Kendall, M. G., and Stuart, A. S., *The Advanced Theory of Statistics*, Vol. 3, New York: Hafner Publishing Co., 1966.

Kendall, M. G., *A Course in Multivariate Analysis*, 4-th impression, New York: Hafner Publishing Co., 1968.

Levine, M. S., *Canonical Analysis and Factor Comparisons*, Beverly Hills: Sage Publications 1977.

Lin, S. and Kernighan, M., An effective heuristic algorithm for the traveling salesman problem, *Operations Research*, 1973, *21*, 498–516.

Meredith, W. Canonical correlations with fallible data, *Psychometrika*, 1964, *29*, 55–65.

Pruzek, R. M., & Frederick, R. C., Weighting predictors in linear models: Alternatives to least squares and limitations of equal weights. *Psychological Bulletin*, 1978, *85*, 254–266.

Reingold, E. M., Nievergelt, J., & Deo, N., *Combinatorial Algorithms*, Englewood Cliffs, New Jersey: Prentice Hall, 1977.

Roy, S. N., *Some Aspects of Multivariate Analysis*, New York: John Wiley and Sons, Inc., 1957.

Scheffé, H., *The Analysis of Variance*, New York: John Wiley and Sons, Inc., 1959.

Thorndike, R. M., Weiss, D. J., & Davis, R. V., Canonical correlation of vocational interests and vocational needs, *Journal of Counseling Psychology*, 1968, *15*, 101–106.

Thorndike, R. M. & Weiss, D. J., A Study of the stability of canonical correlations and canonical components, *Educational and Psychological Measurement*, 1973, *33*, 123–134.

Wainer, H., Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, 1976, *83*, 213–217.

Wainer, H., On the sensitivity of regression and regressors. *Psychological Bulletin*, 1978, *85*, 267–273.

Weinberg, S. L. & Darlington, R. B., Canonical analysis when number of variables is large relative to sample size, *Journal of Educational Statistics*, 1976, *1*, 313–332.

Winer, B. J., Statistics and data analysis: Trading bias for reduced mean square error. *Annual Review of Psychology*, 1978, *29*, 647–6811.

Vinod, H. D., Canonical ridge and econometrics of joint production, *Journal of Econometrics*, 1977, *4*, 147–166.

Manuscript received 6/15/81
First revision received 2/1/82
Final revision received 6/28/82