

**Correspondence Analysis, Multiple Correspondence Analysis and Recent
Developments**

Heungsun Hwang

McGill University

Marc A. Tomiuk

HEC Montreal

Yoshio Takane

McGill University

February 10, 2008

The work reported in this paper was supported by Grant 290439 and Grant 10630 from the Natural Sciences and Engineering Research Council of Canada to the first and third authors, respectively. Data from the 2000 Canadian Election Survey were provided by the Institute for Social Research, York University. The survey was funded by the Social Sciences and Humanities Research Council of Canada, and was completed for the 2000 Canadian Election Team of André Blais (Université de Montréal), Elisabeth Gidengil (McGill University), Richard Nadeau (Université de Montréal) and Neil Nevitte (University of Toronto). Neither the Institute for Social Research, the SSHRC, nor the Canadian Election Survey Team are responsible for the analyses and interpretations presented here.

1. Introduction

The use of multiple-choice response formats is common in psychology and other fields of inquiry. This format offers several advantages: Firstly, it provides respondents with a faster and less tedious response format in comparison to rating or rank-order question formats. Secondly, its use leads to higher survey completion rates while enabling the inclusion of a greater number of questions and/or response categories in a survey (Arimond & Elfessi, 2001; Dolničar & Leisch, 2001). Thirdly, the use of multiple-choice question formats represents a simpler means of data collection/management thus reducing data entry costs (Javalgi, Whipple, McManamon & Edick, 1992). Finally, multiple-choice response formats are highly flexible in the sense that other types of categorical data such as binary, frequency table and sorting data can be regarded as special cases of this general format (e.g., Nishisato, 1994; Takane, 1980).

Correspondence analysis (CA) and multiple correspondence analysis (MCA) represent *descriptive multivariate* techniques for exploring the associations inherent to multiple-choice questions (Benzécri, 1973; Gifi, 1990; Greenacre, 1984; Lebart, Morineau, & Warwick, 1984; Nishisato, 1980). The distinction between CA and MCA rests in the former's focus on interrelationships between two multiple-choice questions whereas the latter emphasizes interrelationships among more than two multiple-choice questions. The reader is referred to Nishisato (2007) for an extensive historical overview of CA and MCA.

Technically, CA and MCA are closely related to *canonical correlation analysis* (CCA) (Hotelling, 1936) and *multiple-set canonical correlation analysis* (MCCA) (Carroll, 1968; Horst, 1961; Meredith, 1964), respectively. CCA is used to describe

interrelationships between two sets of ‘continuous’ variables whereas MCCA captures those among more than two sets of continuous variables. In CCA and MCCA, a series of linear combinations or weighted composites of each set of variables, called the *canonical variates*, are obtained in such a way that they are mutually orthogonal to each other within the same set of linear combinations while remaining maximally correlated with different set(s) of linear combinations. These correlations between the variates are termed *canonical correlations*.

CA and MCA aim to construct linear combinations of the ‘response categories’ of multiple-choice questions in the same way as in CCA and MCCA, respectively. Thus they treat a single response category of each multiple-choice question as one variable in each set of variables in CCA and MCCA. CA and MCA typically display the weights for the linear combinations of response categories jointly in a low-dimensional graphical map. By representing interrelationships among the response categories of multiple-choice questions in the map, CA and MCA have proved useful to both practitioners and academics alike (Hoffman, de Leeuw, & Arjunji, 1994). Moreover, they are nonparametric approaches and therefore do not require the *a priori* and correct specification of the distribution underlying multiple-choice data. Thus, CA and MCA are popular mapping methods that describe the association structures in multiple-choice data without recourse to stringent distribution assumptions (Green, Krieger, & Carroll, 1987).

The purpose of this chapter is to provide an account of the technical underpinnings and applications of CA and MCA. As stated earlier, when data are in the form of multiple-choice questions, CA and MCA may be regarded as special cases of CCA and MCCA, respectively. Hence, we will begin with descriptions of CCA and

MCCA so as to facilitate understanding of CA and MCA. Subsequently, we shall discuss two latest extensions of MCA – regularized MCA and a combined approach to MCA and a hard-clustering technique (*c*-means) for accommodating cluster-level respondent heterogeneity.

2. Correspondence Analysis

2.1. Canonical Correlation Analysis

Canonical correlation analysis (CCA) aims to extract linear combinations from each of two sets of continuous variables which are simultaneously: (1) correlated as highly as possible with a different set of linear combinations and (2) uncorrelated within the same set. Let \mathbf{X}_1 and \mathbf{X}_2 denote n by p and n by q matrices of variables, respectively, where n is the number of respondents, and p and q are the numbers of variables. Assume that \mathbf{X}_1 and \mathbf{X}_2 are mean-centered, indicating that each column mean is eliminated from the individual cases of the column as follows: Let \mathbf{Z}_1 and \mathbf{Z}_2 denote the original, uncentered data matrices. Then, $\mathbf{X}_1 = \mathbf{QZ}_1$ and $\mathbf{X}_2 = \mathbf{QZ}_2$, where $\mathbf{Q} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}'$ and \mathbf{I} is an identity matrix and $\mathbf{1}$ is an n by 1 vector of ones.

Let $\mathbf{J} = \mathbf{X}_1'\mathbf{X}_2$. Note that $\mathbf{Q}'\mathbf{Q} = \mathbf{Q}$, and we have

$$\mathbf{J} = \mathbf{X}_1'\mathbf{X}_2 = \mathbf{Z}_1'\mathbf{QZ}_2 = \mathbf{Z}_1'(\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}')\mathbf{Z}_2 = \mathbf{Z}_1'\mathbf{Z}_2 - n^{-1}\mathbf{Z}_1'\mathbf{1}\mathbf{1}'\mathbf{Z}_2.$$

Let \mathbf{W}_1 and \mathbf{W}_2 denote p by d and q by d matrices consisting of canonical weights assigned to the variables (= columns) of \mathbf{X}_1 and \mathbf{X}_2 , respectively, where $d \leq \min(p, q)$. Then, $\mathbf{F}_1 = \mathbf{X}_1\mathbf{W}_1$ and $\mathbf{F}_2 = \mathbf{X}_2\mathbf{W}_2$ indicate the linear combinations or canonical variates of \mathbf{X}_1 and \mathbf{X}_2 , respectively.

The objective of CCA is to determine \mathbf{W}_1 and \mathbf{W}_2 in such a way that the resultant canonical variates, \mathbf{F}_1 and \mathbf{F}_2 , are maximally correlated between them and are uncorrelated within each. This problem is equivalent to maximizing the following criterion:

$$\phi_1(\mathbf{W}_1, \mathbf{W}_2) = \text{trace}(\mathbf{W}_1' \mathbf{X}_1' \mathbf{X}_2 \mathbf{W}_2) = \text{trace}(\mathbf{W}_1' \mathbf{J} \mathbf{W}_2), \quad (1)$$

with respect to \mathbf{W}_1 and \mathbf{W}_2 , subject to the within-set orthonormality constraints

$$\mathbf{W}_1' \mathbf{X}_1' \mathbf{X}_1 \mathbf{W}_1 = \mathbf{F}_1' \mathbf{F}_1 = \mathbf{I} \text{ and } \mathbf{W}_2' \mathbf{X}_2' \mathbf{X}_2 \mathbf{W}_2 = \mathbf{F}_2' \mathbf{F}_2 = \mathbf{I} \text{ (e.g., ten Berge, 1993, p. 53).}$$

This maximization criterion can be re-expressed as

$$\phi_1(\mathbf{W}_1, \mathbf{W}_2) = \text{trace}(\mathbf{M}_1' (\mathbf{X}_1' \mathbf{X}_1)^{-1/2} \mathbf{J} (\mathbf{X}_2' \mathbf{X}_2)^{-1/2} \mathbf{M}_2), \quad (2)$$

where $\mathbf{M}_1 = (\mathbf{X}_1' \mathbf{X}_1)^{1/2} \mathbf{W}_1$ and $\mathbf{M}_2 = (\mathbf{X}_2' \mathbf{X}_2)^{1/2} \mathbf{W}_2$, subject to the constraints

$$\mathbf{M}_1' \mathbf{M}_1 = \mathbf{M}_2' \mathbf{M}_2 = \mathbf{I} \text{ (ten Berge, 1993, p.53). Thus, maximizing (2) with respect to } \mathbf{M}_1$$

and \mathbf{M}_2 is equivalent to solving the following singular value decomposition (SVD)

problem:

$$\text{SVD}\left((\mathbf{X}_1' \mathbf{X}_1)^{-1/2} \mathbf{J} (\mathbf{X}_2' \mathbf{X}_2)^{-1/2}\right) = \mathbf{\Gamma} \mathbf{\Lambda} \mathbf{\Sigma}', \quad (3)$$

where $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ are matrices of row and column singular vectors, respectively, with the orthonormality property $\mathbf{\Gamma}' \mathbf{\Gamma} = \mathbf{\Sigma}' \mathbf{\Sigma} = \mathbf{I}$, and $\mathbf{\Lambda}$ is a diagonal matrix consisting of singular values (λ 's) as elements in descending order. Then, $\mathbf{M}_1 = \mathbf{\Gamma}$ and $\mathbf{M}_2 = \mathbf{\Sigma}$. In turn, the canonical weights for CCA can be obtained by:

$$\mathbf{W}_1 = (\mathbf{X}_1' \mathbf{X}_1)^{-1/2} \mathbf{\Gamma} \text{ and } \mathbf{W}_2 = (\mathbf{X}_2' \mathbf{X}_2)^{-1/2} \mathbf{\Sigma}. \quad (4)$$

Moreover, each singular value in $\mathbf{\Lambda}$ is equivalent to the canonical correlation between a pair of the canonical variates from each of the two sets of variables.

This approach to CCA involving (2), (3), and (4) is also known to be equivalent to the generalized singular value decomposition (GSVD) of the following matrix:


$$\mathbf{C} = (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{J} (\mathbf{X}_2' \mathbf{X}_2)^{-1} \quad (5)$$

with $\mathbf{X}_1' \mathbf{X}_1$ and $\mathbf{X}_2' \mathbf{X}_2$ as row and column metric matrices, respectively (see Greenacre, 1984; Takane & Hwang, 2002).

2.2. Correspondence Analysis

CA can be viewed as a special case of CCA where \mathbf{Z}_1 and \mathbf{Z}_2 are n by p and n by q ‘indicator’ matrices of two multiple-choice questions, respectively, where p and q indicate the numbers of the response categories to the two questions. Here, an indicator matrix represents a data format where 1 is assigned to the response category chosen by a respondent and 0 to the other response categories of non-choice for each question. To illustrate, consider that five respondents are measured on two multiple-choice questions (Q1 and Q2) with three response categories each, as displayed in the left-hand table below. This table simply presents which category is chosen by each respondent. The two multiple-choice questions in this condensed format can be transformed into two indicator matrices (\mathbf{Z}_1 and \mathbf{Z}_2), as shown in the right-hand table below.

Q1	Q2
1	2
2	3
2	1
3	1
1	3



\mathbf{Z}_1	\mathbf{Z}_2
1 0 0	0 1 0
0 1 0	0 0 1
0 1 0	1 0 0
0 0 1	1 0 0
1 0 0	0 0 1

Let $\mathbf{D}_1 = \mathbf{Z}_1' \mathbf{Z}_1$ and $\mathbf{D}_2 = \mathbf{Z}_2' \mathbf{Z}_2$ denote diagonal matrices of the column sums of \mathbf{Z}_1 and \mathbf{Z}_2 , respectively. Again, \mathbf{X}_1 and \mathbf{X}_2 represent the mean-centered matrices of \mathbf{Z}_1

and \mathbf{Z}_2 , respectively. CA aims to choose weights, \mathbf{W}_1 and \mathbf{W}_2 , assigned to the response categories (columns) of \mathbf{X}_1 and \mathbf{X}_2 in the same way as in CCA. This in turn involves the calculation of the GSVD of \mathbf{C} in (5) with metric matrices $\mathbf{X}_1'\mathbf{X}_1$ and $\mathbf{X}_2'\mathbf{X}_2$. Note that in CA, $\mathbf{X}_1'\mathbf{X}_1$ and $\mathbf{X}_2'\mathbf{X}_2$ in \mathbf{C} can be replaced by \mathbf{D}_1 and \mathbf{D}_2 , respectively, because the data are presented in the form of indicator matrices (refer to Takane & Hwang, 2002).

Thus, CA is equivalent to calculating the GSVD of the following matrix:

$$\mathbf{C} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{J}(\mathbf{X}_2'\mathbf{X}_2)^{-1} = \mathbf{D}_1^{-1}\mathbf{J}\mathbf{D}_2^{-1}, \quad (6)$$

with \mathbf{D}_1 and \mathbf{D}_2 as row and column metric matrices, respectively. As described earlier, this GSVD involves solving the following SVD problem:

$$\text{SVD}(\mathbf{D}_1^{-1/2}\mathbf{J}\mathbf{D}_2^{-1/2}) = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Sigma}'. \quad (7)$$

Then, $\mathbf{W}_1 = \mathbf{D}_1^{-1/2}\mathbf{\Gamma}$ and $\mathbf{W}_2 = \mathbf{D}_2^{-1/2}\mathbf{\Sigma}$. In CA, these canonical weights are called the *standard coordinates* of the response categories of each multiple-choice question. Again, $\mathbf{\Lambda}$ contains singular values in descending order.

If the matrix \mathbf{C} in (6) is divided by n , a more familiar formulation of CA in the literature is obtained as follows:

$$\begin{aligned} n^{-1}\mathbf{C} &= n^{-1}\mathbf{D}_1^{-1}\mathbf{J}\mathbf{D}_2^{-1} \\ &= \mathbf{D}_1^{-1}n^{-1}(\mathbf{Z}_1'\mathbf{Z}_2 - n^{-1}\mathbf{Z}_1'\mathbf{1}\mathbf{1}'\mathbf{Z}_2)\mathbf{D}_2^{-1} \\ &= \mathbf{D}_1^{-1}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_2^{-1}, \end{aligned} \quad (8)$$

where $\mathbf{P} = n^{-1}\mathbf{Z}_1'\mathbf{Z}_1$ is the so-called p by q *correspondence* matrix (= the frequency table of two multiple-choice questions/ n), $\mathbf{r} = n^{-1}\mathbf{Z}_1'\mathbf{1}$ is a p by 1 vector of row masses (row totals of the frequency table / n), $\mathbf{c}' = n^{-1}\mathbf{1}'\mathbf{Z}_2$ is a 1 by q vector of column masses (column

totals of the frequency table/ n) (e.g., Blasius & Greenacre, 1994). The GSVD of (8) result in the same standard coordinates as those from the GSVD of (6).

Dual scaling (Nishisato, 1980) provides essentially the same solutions as those in CA, although it optimizes a different criterion to obtain \mathbf{W}_1 and \mathbf{W}_2 . It aims to determine each column of \mathbf{W}_1 and \mathbf{W}_2 successively by maximizing the corresponding squared correlation ratio η , i.e., the between-subject sum of squares divided by the total sum of squares in ANOVA. As an example, the first column of \mathbf{W}_2 , say \mathbf{w}_2 , is obtained by maximizing:

$$\phi_2(\mathbf{w}_2) = \eta = \frac{\mathbf{w}_2' \mathbf{J}' \mathbf{D}_r^{-1} \mathbf{J} \mathbf{w}_2}{\mathbf{w}_2' \mathbf{D}_c \mathbf{w}_2}. \quad (9)$$

By setting the derivative of (9) with respect to \mathbf{w}_2 (divided by 2) equal to zeros, we have

$$\begin{aligned} \frac{1}{2} \frac{\partial \phi_2(\mathbf{w}_2)}{\partial \mathbf{w}_2} &= \mathbf{J}' \mathbf{D}_r^{-1} \mathbf{J} \mathbf{w}_2 - \eta \mathbf{D}_c \mathbf{w}_2 \\ &= (\mathbf{J}' \mathbf{D}_r^{-1} \mathbf{J} - \eta \mathbf{D}_c) \mathbf{w}_2 \\ &= (\mathbf{J}' \mathbf{D}_r^{-1} \mathbf{J} \mathbf{D}_c^{-1/2} - \eta \mathbf{D}_c^{1/2}) \mathbf{D}_c^{1/2} \mathbf{w}_2 \\ &= (\mathbf{D}_c^{-1/2} \mathbf{J}' \mathbf{D}_r^{-1} \mathbf{J} \mathbf{D}_c^{-1/2} - \eta \mathbf{I}) \mathbf{D}_c^{1/2} \mathbf{w}_2 \\ &= (\mathbf{A}' \mathbf{A} - \eta \mathbf{I}) \mathbf{m}_2 = \mathbf{0}, \end{aligned} \quad (10)$$

where $\mathbf{m}_2 = \mathbf{D}_c^{1/2} \mathbf{w}_2$ and $\mathbf{A} = \mathbf{D}_r^{-1/2} \mathbf{J} \mathbf{D}_c^{-1/2}$. Solving (10) comes down to calculating the eigenvalue decomposition (EVD) of $\mathbf{A}' \mathbf{A}$ as in principal components analysis (PCA). The first eigenvector of $\mathbf{A}' \mathbf{A}$ equals to \mathbf{m}_2 . Then, $\mathbf{w}_2 = \mathbf{D}_c^{-1/2} \mathbf{m}_2$, which is equivalent to the first column of \mathbf{W}_2 in CA. By a similar procedure, the first column of \mathbf{W}_1 , say \mathbf{w}_1 , is obtained by solving $(\mathbf{A} \mathbf{A}' - \eta \mathbf{I}) \mathbf{m}_1 = \mathbf{0}$, where $\mathbf{m}_1 = \mathbf{D}_r^{1/2} \mathbf{w}_1$. The first eigenvector of $\mathbf{A} \mathbf{A}'$ equals to \mathbf{m}_1 . Then, $\mathbf{w}_1 = \mathbf{D}_r^{-1/2} \mathbf{m}_1$, which is equivalent to the first column of \mathbf{W}_1 in

CA. The next columns of \mathbf{W}_1 and \mathbf{W}_2 are successively obtained by eliminating the effects of the previous solutions from $\mathbf{A}'\mathbf{A}$ and $\mathbf{A}\mathbf{A}'$, respectively (see Nishisato, 1994, p. 105).

In CA, the so-called *principal coordinates* (Greenacre, 1984, p.90) are obtained by post-multiplying the standard coordinates by $\mathbf{\Lambda}$:

$$\tilde{\mathbf{W}}_1 = \mathbf{D}_1^{-1/2}\mathbf{\Gamma}\mathbf{\Lambda} \text{ and } \tilde{\mathbf{W}}_2 = \mathbf{D}_2^{-1/2}\mathbf{\Sigma}\mathbf{\Lambda} . \quad (11)$$

Thus, these principal coordinates are simply the standard coordinates rescaled by singular values. Note that they can be re-expressed as

$$\tilde{\mathbf{W}}_1 = \mathbf{D}_1^{-1}\mathbf{X}_1'\mathbf{F}_2 \text{ and } \tilde{\mathbf{W}}_2 = \mathbf{D}_2^{-1}\mathbf{X}_2'\mathbf{F}_1 . \quad (12)$$

Equation (12) is derived from

$\tilde{\mathbf{W}}_1 = \mathbf{D}_1^{-1/2}\mathbf{\Gamma}\mathbf{\Lambda} = \mathbf{D}_1^{-1/2}\mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Sigma}\mathbf{\Sigma} = \mathbf{D}_1^{-1/2}\mathbf{D}_1^{-1/2}\mathbf{X}_1'\mathbf{X}_2\mathbf{D}_2^{-1/2}\mathbf{\Sigma} = \mathbf{D}_1^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{W}_2 = \mathbf{D}_1^{-1}\mathbf{X}_1'\mathbf{F}_2$ ($\tilde{\mathbf{W}}_2$ is also derived in a similar way). This indicates that the principal coordinates for one multiple-choice question are obtained by regressing the canonical variates of the other question onto the question in a way similar to estimating regression coefficients in linear regression analysis (Hirschfeld, 1935). This is called the *barycentric* principle or *dual relations* (Nishisato, 1980) in CA. Roughly speaking, this principle holds that the principal coordinates for one multiple-choice question depend on the canonical variates (and in turn the standard coordinates) of the other question.

The results of CA are graphically displayed in a low-dimensional space. In practice, the principal coordinates for two multiple-choice questions are jointly displayed in a low-dimensional space. This is called the *symmetric* map. The principal coordinates in this map are comparable to each other given that they are expressed in the same unit. Also, as shown above, the principal coordinates for one multiple-choice question rely on those for the other multiple-choice questions, i.e., the *barycentric* principle. More

precisely, each of them is a weighted average of the canonical variates of the other multiple-choice. Thus, the principal coordinates may be interpreted in terms of closeness, e.g., response categories positioned close together are similar to each other. However, it is noteworthy that no distance-based interpretations are feasible between the principal coordinates for different multiple-choice questions because \mathbf{W}_1 and \mathbf{W}_2 are involved in different data sets, \mathbf{X}_1 and \mathbf{X}_2 , respectively, so that the computation of the distance between them is not justifiable (e.g., Greenacre, 1994; Lebart, Morineau, & Warwick, 1984, Nishisato, 2007).

As in CCA and other data-reduction techniques, CA also invites a focus on the first few dimensions for interpretation. The number of dimensions may be determined in various ways. For example, as in PCA, we may select the dimensions whose eigenvalues (= squared singular values) explain a majority of the total sum of eigenvalues. In CA, the eigenvalues are often called *inertias*. Also, a scree plot of inertias against dimensions may be examined to identify an elbow point in the trajectory of eigenvalues. Furthermore, other criteria such as graphical and/or substantive interpretability may also be considered for dimensionality selection. For instance, in practice, a two-dimensional solution is usually displayed for facilitating interpretation.

Other than these heuristics for dimensionality selection, the permutation test may be employed for directly testing the significance of canonical correlations (Takane & Hwang, 2002). The permutation test is beneficial because it does not rely on any distributional assumptions on the data. In principle, this test is applied only for testing the significance of the largest canonical correlation. However, the significance of subsequent canonical correlations can also be examined by eliminating the effects of previous

canonical correlations from the data sets through the procedure discussed below (see Legendre & Legendre, 1998; ter Braak, 1990).

The permutation test based on Manley's (1997) procedure for testing the significance of the largest canonical correlation can be carried out as follows:

Step 1: Apply CA to \mathbf{X}_1 and \mathbf{X}_2 , and compute the observed value of Bartlett's (1938) statistic φ_o , given by

$$\varphi_o = - \left[(N-1) - \frac{1}{2}(p+q+1) \right] \sum_{j=1}^J \log(1-\lambda_j^2), \quad (13)$$

where $J = \min(p,q)$, and λ_j is a sample canonical correlation obtained from CA.

Step 2: Randomly permute the cases (or randomly select one case at a time without replacement) of one data matrix, say \mathbf{X}_2 , so as to create a 'permuted' sample of the data matrix, denoted by \mathbf{X}_2^* .

Step 3: Apply CA to \mathbf{X}_1 and \mathbf{X}_2^* , and calculate the permuted Bartlett's statistic, denoted by φ_p .

Step 4: Repeat Steps 2 and 3 B times (e.g., $B = 1,000$). This results in the null distribution of φ_p , i.e., the distribution of φ_p under the independence assumption between two data sets.

Step 5: Compute the so-called Permutation Achieved Significance Level (PASL) which is equal to the probability that $\varphi_p \geq \varphi_o$.

If the PASL is less than .05, we may reject the null hypothesis of independence at a 5% level, indicating that the largest canonical correlation is significantly different from zero.

To test the second largest canonical correlation, we remove the effect of the largest canonical correlation from \mathbf{X}_1 and \mathbf{X}_2 . Specifically, the effect of the largest

canonical correlation can be eliminated from \mathbf{X}_1 and \mathbf{X}_2 by $\mathbf{X}_1 = \mathbf{X}_1\boldsymbol{\Omega}_1$ and $\mathbf{X}_2 = \mathbf{X}_2\boldsymbol{\Omega}_2$, respectively, where $\boldsymbol{\Omega}_1 = \mathbf{I} - \mathbf{w}_1\mathbf{w}_1'\mathbf{X}_1'\mathbf{X}_1$ and $\boldsymbol{\Omega}_2 = \mathbf{I} - \mathbf{w}_2\mathbf{w}_2'\mathbf{X}_2'\mathbf{X}_2$. As a result, the second largest canonical correlation now becomes the largest one because the effect of the latter disappears in the data. Thus, the same permutation procedure described above can be carried out to test the significance of the second largest canonical correlation. The same strategy is utilized for testing the significance of subsequent canonical correlations. This approach is essentially the same as that of Legendre and Legendre (1998). Note that although the above procedure employs Bartlett's statistic, other statistics such as Roy's max lambda ($n\lambda^2$) can also be used for the permutation test.

The bootstrap method (Efron, 1979) can be used for assessing the reliability of the weight estimates of CA. In this method, a number of random samples (bootstrap samples) of \mathbf{X}_1 and \mathbf{X}_2 are repeatedly sampled from the original data matrices with replacement. CA is applied to each bootstrap sample so as to obtain the estimates of weights. Then, the mean and the variance-covariance of the estimates are calculated across entire bootstrap samples. They are used for the computation of the standard errors or the construction of the confidence regions (Ramsay, 1978) of the estimates, which indicate how reliable the estimates are.

2.3. Application: The 2000 Canadian Federal Election Data

The present example is part of the Canadian Election Survey (CES) conducted by the Institute for Social Research at York University to investigate political opinions or preferences of Canadians during the 2000 federal election campaign. Telephone interviews were given to randomly chosen Canadian citizens of voting age (18 years of age or older), which began on October 24, 2000 and terminated at the last day of the

campaign - November 26, 2000.

Two items were selected from the CES data for this example. Item 1 asked the province where respondents live, and item 2 asked which party respondents would vote for in the upcoming election. We removed the respondents who refused to answer the second question from the original data. Item 1 consisted of 10 Canadian provinces from East to West: 1 = Newfoundland (NF), 2 = Prince Edward Island (PE), 3 = Nova Scotia (NS), 4 = New Brunswick (NB), 5 = Quebec (QC), 6 = Ontario (ON), 7 = Manitoba (MA), 8 = Saskatchewan (SK), 9 = Alberta (AB), 10 = British Columbia (BC). Item 2 comprised 10 response categories: 1 = Other, 2 = Liberal Party, 3 = Alliance Party, 4 = Conservative Party, 5 = New Democratic Party, 6 = Bloc Quebecois Party, 7 = Green Party, 8 = Will not vote, 9 = None, 10 = Don't know/undecided. The sample size was 3185.

Table 1 provides the inertias (squared canonical correlations) estimated from CA and their percentages of the total inertia. It was found from the permutation test with 1000 permuted samples that the first four canonical correlations turned out to be significant, although the last two significant ones appear quite small. This may be due to the large sample size. In fact, the first two inertias accounted for about 87% of the total inertia. This suggests that the two-dimensional solution is likely to capture a majority of the associations among the response categories of the two items.

Insert Table 1 about here

Figure 1 displays the two-dimensional symmetric plot of the principal coordinates of the response categories of the two items. To make the figure concise, the coordinates of the ten provinces in item 1 are labelled NF, PE, NS, NB, QC, ON, MA, SK, AB, and BC. The order of the labels is equivalent to that of the categories in item 1. The coordinates of the ten response categories in item 2 are represented by their category numbers from 1 to 10. The symbol ‘+’ indicates the origin of the two-dimensional plot.

Insert Figure 1 about here

In Figure 1, ‘QC’ is closely located with ‘1 (Other)’, ‘6 (Bloc Quebecois)’, ‘8 (Will not vote)’, and ‘9 (None)’. This suggests that Quebec residents were more likely to vote for Bloc Quebecois among federal parties in the upcoming election. Moreover, they seemed to show less preference to current federal parties or were more likely to give up voting in the election, compared to those in other provinces. In addition, they were more likely to choose other parties than extant federal parties compared to other provinces’ residents. On the other hand, ‘AB’ is very close to ‘3 (Alliance)’, indicating that the residents of Alberta were more likely to support the Alliance Party. ‘BC’ and ‘SK’ appear to close to ‘7 (Green Party)’, suggesting that the major supporters for the party resided in the two provinces. Furthermore, ‘ON’ and ‘MA’ seem to be closely located with ‘2 (Liberal Party)’, ‘5 (New Democratic Party)’, and ‘10 (Don’t know/undecided)’. Thus, the residents of the two provinces were more likely to vote for a centrist (Liberal) or centrist-left (New Democratic) party. Also, the two provinces were likely to entail more swing voters. Finally, the provinces on the east coast of Canada, including ‘PE’, ‘NS’,

‘NB’, and ‘NF’, appear to be close to ‘4 (Conservative)’ as well as ‘5 (New Democratic)’. Thus, the residents in these provinces were more inclined towards the Conservative Party or the New Democratic Party than those in other provinces.

3. Multiple Correspondence Analysis

Multiple correspondence analysis (MCA) is used to describe interrelationships among more than two multiple-choice questions. As stated earlier, MCA may be regarded as a special case of multiple-set canonical correlation analysis (MCCA) (Carroll, 1968; Horst, 1961; Meredith, 1964). Thus, we begin with the description of MCCA.

3.1. Multiple-set Canonical Correlation Analysis (MCCA)

Let \mathbf{X}_k denote n by p_k matrix of variables, where p_k is the number of variables ($k = 1, \dots, K$). Assume that \mathbf{X}_k is mean-centered. Let \mathbf{W}_k denote a p_k by d matrix of canonical weights assigned to the variables of \mathbf{X}_k . Then, $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_K]$ is an n by

p row block matrix consisting of \mathbf{X}_k side by side, where $p = \sum_{k=1}^K p_k$,

$\mathbf{W} = [\mathbf{W}_1', \mathbf{W}_2', \dots, \mathbf{W}_K']'$ is a p by d column block matrix stacking \mathbf{W}_k one below another,

and $\mathbf{\Phi} = \text{diag}[\mathbf{X}_1' \mathbf{X}_1, \mathbf{X}_2' \mathbf{X}_2, \dots, \mathbf{X}_K' \mathbf{X}_K]$ is a block diagonal matrix consisting of

$\mathbf{X}_k' \mathbf{X}_k$ as the k -th diagonal block.

The objective of MCCA is to determine \mathbf{W}_k in such a way that the resultant canonical variates are maximally correlated among different sets of canonical variates while uncorrelated within the same set. This problem is equivalent to maximizing the following criterion:

$$\phi_3(\mathbf{W}) = \text{trace}(\mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W}), \quad (14)$$

with respect to \mathbf{W} , subject to the within-set orthogonality constraints $\mathbf{W}'\mathbf{\Phi}\mathbf{W} = \mathbf{I}$

(Carroll, 1968). This maximization criterion can be re-expressed as:

$$\phi_2(\mathbf{W}) = \text{trace}(\mathbf{M}'\mathbf{\Phi}^{-1/2}\mathbf{X}'\mathbf{X}\mathbf{\Phi}^{-1/2}\mathbf{M}), \quad (15)$$

where $\mathbf{M} = \mathbf{\Phi}^{1/2}\mathbf{W}$, subject to the constraint $\mathbf{M}'\mathbf{M} = \mathbf{I}$. Thus, maximizing (15) with respect to \mathbf{M} is equivalent to obtaining the following eigenvalue decomposition (EVD):

$$\text{EVD}(\mathbf{\Phi}^{-1/2}\mathbf{X}'\mathbf{X}\mathbf{\Phi}^{-1/2}) = \mathbf{\Sigma}\mathbf{\Lambda}^2\mathbf{\Sigma}', \quad (16)$$

where $\mathbf{\Sigma}'\mathbf{\Sigma} = \mathbf{I}$ and $\mathbf{\Lambda}^2$ is a diagonal matrix consisting of eigenvalues (squared singular values) as elements. The EVD in (16) is equivalent to the singular value decomposition (SVD) of $\mathbf{\Phi}^{-1/2}\mathbf{X}'\mathbf{X}\mathbf{\Phi}^{-1/2}$, whose singular values become equal to the eigenvalues from (16). Then, $\mathbf{M} = \mathbf{\Sigma}$. In turn, \mathbf{W} is obtained by

$$\mathbf{W} = \mathbf{\Phi}^{-1/2}\mathbf{\Sigma}. \quad (17)$$

This approach to MCCA involving (15), (16) and (17) is known to be equivalent to the generalized eigenvalue decomposition (GEVD) of the following matrix:

$$\mathbf{G} = \mathbf{\Phi}^{-1}\mathbf{X}'\mathbf{X}\mathbf{\Phi}^{-1} \quad (18)$$

with $\mathbf{\Phi}$ as both row and column metric matrices (Greenacre, 1984; Takane & Hwang, 2002).

MCCA can be alternatively formulated through the criterion for *homogeneity analysis* or *K-set canonical correlation* (Gifi, 1990; Yanai, 1998). This is equivalent to minimizing the following criterion:

$$\phi_4(\mathbf{F}, \mathbf{B}_k) = \sum_{k=1}^K \text{SS}(\mathbf{F} - \mathbf{X}_k\mathbf{B}_k), \quad (19)$$

with respect to \mathbf{F} and \mathbf{B}_k , subject to $\mathbf{F}'\mathbf{F} = \mathbf{I}$, where $SS(\mathbf{H}) = \text{trace}(\mathbf{H}'\mathbf{H})$, \mathbf{F} is an n by d matrix of canonical variates, and \mathbf{B}_k is a p_k by d matrix of canonical weights. Let $\mathbf{B} = [\mathbf{B}_1', \mathbf{B}_2', \dots, \mathbf{B}_K']$ denote a column block matrix stacking \mathbf{B}_k one below another. If \mathbf{F} is considered to be fixed, minimizing (19) reduces to solving the least-squares estimation problem with respect to \mathbf{B}_k as in linear regression analysis. Thus, we obtain

$$\hat{\mathbf{B}}_k = (\mathbf{X}_k' \mathbf{X}_k)^{-1} \mathbf{X}_k' \mathbf{F}, \text{ or collectively, } \hat{\mathbf{B}} = \mathbf{\Phi}^{-1} \mathbf{X}' \mathbf{F}. \quad (20)$$

By inserting (20) to (19), we obtain

$$\phi_5(\mathbf{F}) = \sum_{k=1}^K SS(\mathbf{F} - \mathbf{X}_k (\mathbf{X}_k' \mathbf{X}_k)^{-1} \mathbf{X}_k' \mathbf{F}). \quad (21)$$

Let $\mathbf{\Omega}_k = \mathbf{X}_k (\mathbf{X}_k' \mathbf{X}_k)^{-1} \mathbf{X}_k'$. Note that $\mathbf{\Omega}_k' \mathbf{\Omega}_k = \mathbf{\Omega}_k$ and $\mathbf{\Omega}_k' = \mathbf{\Omega}_k$. This criterion can be re-expressed as:

$$\begin{aligned} \phi_5(\mathbf{F}) &= \sum_{k=1}^K SS(\mathbf{F} - \mathbf{\Omega}_k \mathbf{F}) \\ &= \sum_{k=1}^K \text{trace}(\mathbf{F}' \mathbf{F} - \mathbf{F}' \mathbf{\Omega}_k \mathbf{F}) \\ &= Kd - \text{trace} \left(\mathbf{F}' \left[\sum_{p=1}^P \mathbf{\Omega}_k \right] \mathbf{F} \right). \end{aligned} \quad (22)$$

Thus, minimization of (22) with respect to \mathbf{F} is equivalent to maximizing

$$\text{trace} \left(\mathbf{F}' \left[\sum_{k=1}^K \mathbf{\Omega}_k \right] \mathbf{F} \right). \quad (23)$$

This problem reduces to calculating the eigenvalue decomposition of $\sum_{k=1}^K \mathbf{\Omega}_k$ whose eigenvectors are equal to \mathbf{F} (Yanai, 1998). The matrix \mathbf{B} in (20) is related to \mathbf{W} in (17) by $\mathbf{B} = \mathbf{W}\mathbf{\Lambda}$.

CCA may be viewed as a special case of MCCA when there are only two sets of variables ($K = 2$). Specifically, let $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$, and $\mathbf{W} = [\mathbf{W}_1', \mathbf{W}_2']$. Then, (16) can be expressed as

$$\begin{aligned}
\Phi^{-1/2} \mathbf{X}' \mathbf{X} \Phi^{-1/2} &= \begin{bmatrix} (\mathbf{X}_1' \mathbf{X}_1)^{-1/2} & \mathbf{0} \\ \mathbf{0} & (\mathbf{X}_2' \mathbf{X}_2)^{-1/2} \end{bmatrix} \begin{bmatrix} \mathbf{X}_1' \mathbf{X}_1 & \mathbf{X}_1' \mathbf{X}_2 \\ \mathbf{X}_2' \mathbf{X}_1 & \mathbf{X}_2' \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} (\mathbf{X}_1' \mathbf{X}_1)^{-1/2} & \mathbf{0} \\ \mathbf{0} & (\mathbf{X}_2' \mathbf{X}_2)^{-1/2} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{I} & (\mathbf{X}_1' \mathbf{X}_1)^{-1/2} \mathbf{J} (\mathbf{X}_2' \mathbf{X}_2)^{-1/2} \\ (\mathbf{X}_2' \mathbf{X}_2)^{-1/2} \mathbf{J}' (\mathbf{X}_1' \mathbf{X}_1)^{-1/2} & \mathbf{I} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{I} & \Gamma \Lambda \Sigma \\ \Sigma \Lambda \Gamma & \mathbf{I} \end{bmatrix}.
\end{aligned} \tag{24}$$

From (24), it follows that $\mathbf{W}_1 = \Phi_1^{-1/2} \Gamma$ and $\mathbf{W}_2 = \Phi_2^{-1/2} \Sigma$, and $\Lambda^2 = \mathbf{I} + \Lambda$ (ten Berge, 1979; also see Gifi, 1990, p. 273). This indicates that the canonical weights from CCA are equivalent to those from MCCA when $K = 2$.

3.2. Multiple Correspondence Analysis (MCA)

MCA can be viewed as a special case of MCCA where \mathbf{X}_k is an n by p_k ‘indicator’ matrix of a multiple-choice question, where p_k indicates the number of response categories of the question. In MCA, the metric matrix for MCCA, i.e., Φ , can be replaced by $\mathbf{D} = \text{diag}[\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_K]$ which is a block diagonal matrix consisting of $\mathbf{D}_k = \mathbf{Z}_k' \mathbf{Z}_k$ as the k -th diagonal block, similarly to the CA case.

Thus, MCA is equivalent to calculating the generalized eigenvalue decomposition (GEVD) of the following matrix:

$$\mathbf{G} = \mathbf{D}^{-1} \mathbf{X}' \mathbf{X} \mathbf{D}^{-1} \tag{25}$$

with \mathbf{D} as both row and column metric matrices. In MCA, $\mathbf{X}'\mathbf{X}$ is called the centered Burt table: $\mathbf{X}'\mathbf{X} = \mathbf{Z}'\mathbf{Q}\mathbf{Z} = \mathbf{Z}'\mathbf{Z} - n^{-1}\mathbf{Z}'\mathbf{1}\mathbf{1}'\mathbf{Z}$, where $\mathbf{Z}'\mathbf{Z}$ is called the (uncentered) Burt table.

As described earlier, this GEVD involves solving the following SVD problem:

$$\text{SVD}(\mathbf{D}^{-1/2}\mathbf{X}'\mathbf{X}\mathbf{D}^{-1/2}) = \mathbf{\Sigma}\mathbf{\Lambda}^2\mathbf{\Sigma}'. \quad (26)$$

Then, the standard coordinates \mathbf{W} are obtained by

$$\mathbf{W} = \mathbf{D}^{-1/2}\mathbf{\Sigma}. \quad (27)$$

The principal coordinates can be obtained by

$$\tilde{\mathbf{W}} = \mathbf{D}^{-1/2}\mathbf{\Sigma}\mathbf{\Lambda}. \quad (28)$$

Thus, \mathbf{B} in (20) is equivalent to $\tilde{\mathbf{W}}$ in (28) in MCA, i.e., the principal coordinates of response categories.

As shown above, CCA can be viewed as a special case of MCCA when $K = 2$. Similarly, CA is also a special case of MCA when there are only two multiple-choice questions. The only difference between the two approaches is in the eigenvalue value matrix, thus rendering principal coordinates scaled differently from each other, i.e.,

$$\tilde{\mathbf{W}}_1 = \mathbf{B}_1 = \mathbf{D}_1^{-1/2}\mathbf{\Gamma}(\mathbf{I} + \mathbf{\Lambda})^{1/2} \text{ and } \tilde{\mathbf{W}}_2 = \mathbf{B}_2 = \mathbf{D}_2^{-1/2}\mathbf{\Sigma}(\mathbf{I} + \mathbf{\Lambda})^{1/2} \text{ in MCA.}$$

In MCA, the proportions of the total inertia (squared singular values) accounted for by the inertias tend to be underestimated because the total inertia is inflated due to fitting both diagonal and off-diagonal blocks of the Burt table (Greenacre, 1984). One way of dealing with this problem is to adjust the inertias greater than $1/K$ using Benzécri's (1979) formula, quoted in Greenacre (1984, p.145). Let $\tilde{\gamma}_j$ denote the adjusted inertia for the j -th inertia, γ_j . Then, the formula is given by

$$\tilde{\gamma}_j = \left(\frac{K}{K-1} \right)^2 \left(\gamma_j - \frac{1}{K} \right)^2. \quad (29)$$

Then, the adjusted inertias are expressed as percentages of the following average off-diagonal inertia (Greenacre, 1993):

$$\left(\frac{K}{K-1} \right) \left(\sum_{j=1}^J \gamma_j^2 - \frac{p-K}{K^2} \right). \quad (30)$$

In MCA, the same heuristics as those for CA described in Section 2.2 may also be used for dimensionality selection. In particular, a similar permutation procedure may be applied to MCA, in which one data matrix is fixed while the other data matrices are separately permuted at random. Then, the Permutation Achieved Significance Level (PASL) can be calculated based on Roy's max lambda ($n\lambda^2$) in order to test the significance of the largest inertia. As in CA, the bootstrap method can be adopted for examining the reliability of the weight estimates of MCA.

3.3. Application: The 2000 Canadian Federal Election Data

The present example consists of three items from the 2000 Canadian Election Survey (CES). The first two items are the same ones used in Section 2.3 for the illustration of correspondence analysis, i.e., the province of residence and the party respondents are likely to vote for in the upcoming election in 2000. The third item asked which party respondents actually voted for in the previous federal election in 1996. We selected only the respondents who recalled if they voted in the 1996 federal election and also answered the second and third questions. In the example, the first item (province of residence) involved the same 10 provinces. The second item consisted of 9 response categories: 1 = Liberal Party, 2 = Alliance Party, 3 = Conservative Party, 4 = New Democratic Party, 5 =

Bloc Quebecois Party, 6 = Green Party, 7 = Will not vote, 8 = None, 9 = Don't know/undecided. The last item consisted of 8 response categories: 1 = Liberal Party, 2 = Conservative Party, 3 = New Democratic Party, 4 = Reform Party, 5 = Bloc Quebecois Party, 6 = Annulled votes, 7 = Green Party, 8 = Other. The sample size was 2213.

Table 2 shows the adjusted inertias and their percentage of the total adjusted inertia. It is shown that the adjusted inertias appeared to gradually decrease after the first three inertias. On the other hand, the first seven inertias turned out to be significant according to the permutation test with 1000 permuted samples. This large number of significant inertias may be due to the large sample size. Here, the two dimensional solutions of the response categories are only provided so as to facilitate the interpretation of the association among the categories, although it seems to be adequate to look into higher dimensional solutions as well.

Insert Table 2 about here

Figure 2 displays the two-dimensional symmetric plot of the principal coordinates of the response categories of the three items. Again, the estimated coordinates of the ten provinces in item 1 are labelled NF, PE, NS, NB, QC, ON, MA, SK, AB, and BC. The order of the labels is equivalent to that of the categories in the item. The coordinates of the nine response categories in item 2 are represented by two digit numbers from 21 to 29. The estimated coordinates of the eight response categories in item 3 represented by two digit numbers from 31 to 38. The order of the two-digit labels is consistent to that of

the categories in items 2 and 3. The symbol '+' indicates the origin of the two-dimensional plot.

In Figure 2, 'QC' is closely located with such categories as '25 (Bloc Quebecois - 2000)', '35 (Bloc Quebecois -1997)', '27 (Will not vote in 2000)', '36 (Annulled vote in 1997)', and '28 (None in 2000)'. This suggests that Quebec residents were more supportive for Bloc Quebecois than other federal parties in both 1997 and 2000 elections. Moreover, they showed or were more inclined to no voting in both elections than those in the other provinces. Additionally, they seemed to show less preference to the extant federal parties than those in other provinces. On the other hand, 'AB' is close to '22 (Alliance Party in 2000)' and '34 (Reform Party in 1997)'. This indicates that the residents of Alberta were more likely to vote for the Reform Party in 1997 and tended to be more supportive for the Alliance Party in 2000, which was the successor to the Reform Party.

'BC' and 'SK' appear to close to '26 (Green Party in 2000)', '37 (Green Party in 1997)', '32 (Conservative Party in 1997)', '23 (Conservative Party in 2000)'. This suggests that residents in the two provinces were more supportive for the Green Party and the Conservative Party in both elections. Furthermore, 'ON', 'MA', 'PE', 'NS', 'NF', and 'NS' seem to be closely located with '31 (Liberal Party in 1997)', '21 (Liberal Party in 2000)', '24 (New Democratic Party in 2000)', '33 (New Democratic Party in 1997), and '29 (Don't know/undecided in 2000)', '38 (Other in 1997)', and '23 (Conservative Party in 2000)'. Thus, the residents of these provinces were likely to show preferences for other parties besides the parties Reform/Alliance and Bloc Quebecois in both elections.

Insert Figure 2 about here

4. Recent Developments

In this section, we introduce two latest extensions of multiple correspondence analysis – regularized MCA and a combined use of MCA and c -means for capturing cluster-level respondent heterogeneity.

4.1. Regularized multiple correspondence analysis

A regularized version of MCA has recently been proposed that often renders the estimates of MCA closer to the population parameters on average, compared to ordinary or non-regularized MCA (Takane & Hwang, 2006). This regularized MCA is easy to apply and also computationally simple as will be seen shortly.

The basic motivation of regularized MCA comes from ridge regression (Hoerl & Kennard, 1970). Ridge regression is an efficient tool for dealing with the problem of multicollinearity in multiple regression analysis, i.e., high correlations among predictor variables. Ridge regression may be described as follows: let \mathbf{X} and \mathbf{y} denote a matrix of predictor variables and a vector of dependent variable, respectively. Let \mathbf{b} denote a vector of regression coefficients. Then, the ordinary least squares estimates of regression coefficients are given by

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \quad (31)$$

In ridge regression, on the other hand, regression coefficients are estimated by

$$\hat{\mathbf{b}}_r = (\mathbf{X}'\mathbf{X} + \omega\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}, \quad (32)$$

where the additional scalar, ω , is called the ridge parameter. The ridge parameter typically takes a small positive value. The least squares estimator is known to be the best (minimum variance) unbiased estimates under mild distributional assumptions on errors. However, it may turn out to be poor estimates of regression coefficients (associated with large variances) when the matrix $\mathbf{X}'\mathbf{X}$ in (31) is ill-conditioned (nearly singular) due to multicollinearity. The ridge estimator, on the other hand, is biased but is more robust against multicollinearity. A small positive number added to the diagonals of $\mathbf{X}'\mathbf{X}$ tend to provide more stable estimates than the ordinary least squares counterparts.

The quality of parameter estimates is measured by the squared Euclidean distance between the estimates and parameters. If we take the expected value of the squared distance over data, we obtain the mean squared error (MSE). The MSE can be decomposed into two distinct components. One is the *squared bias* (the squared distance between the population parameters and the means of the estimates), and the other is the *variance* (the average distance between individual estimates and the means of the estimates). The least squares estimates involve no bias, but they may have large variances particularly in the presence of multicollinearity. On the other hand, the ridge estimates are biased but are usually associated with a smaller variance. If the variance is small enough, the ridge estimates are likely to have a smaller MSE than their least squares counterparts. In spite of their bias, therefore, the ridge estimates are on average closer to the population parameters. Indeed, for a certain range of values of ω , it is known that ridge estimators always have a smaller MSE than the ordinary least squares estimates, regardless of the existence of the multicollinearity problem (Hoerl & Kennard, 1970). Regularized MCA applies this idea of ridge regression to MCA so as to obtain better

estimates.

Let $\mathbf{\Omega}$ denote a block diagonal matrix consisting of $\mathbf{\Omega}_k$ in (22) as the k -th diagonal block. Let us define

$$\mathbf{D}(\omega) = \mathbf{D} + \omega\mathbf{\Omega}. \quad (33)$$

In (33), the value of ω is assumed to be prescribed by some cross validation method, as will be discussed later.

In regularized MCA, the following criterion is maximized

$$\phi_g(\mathbf{W}) = \text{trace}(\mathbf{W}'(\mathbf{X}'\mathbf{X} + \omega\mathbf{\Omega})\mathbf{W}), \quad (34)$$

with respect to \mathbf{W} , subject to $\mathbf{W}'\mathbf{D}(\omega)\mathbf{W} = \mathbf{I}$. Similarly to the case of ordinary MCA, maximizing (19) reduces to calculating the generalized eigenvalue decomposition of the following matrix:

$$\mathbf{D}(\omega)^{-1}(\mathbf{X}'\mathbf{X} + \omega\mathbf{\Omega})\mathbf{D}(\omega)^{-1}, \quad (35)$$

with $\mathbf{D}(\omega)$ as both row and column metric matrices (Takane & Hwang, 2006).

Once the value of the ridge parameter ω is chosen, therefore, the computation of regularized MCA is as simple as ordinary MCA. In regularized MCA, the G -fold cross-validation method (Hastie, Tibshirani, & Friedman, 2001) may be used for selecting an optimal value of the ridge parameter. In this cross-validation method, the data set at hand are randomly divided into G sub-samples. One of the sub-samples is set aside, and the estimates of parameters are obtained from the remaining sub-samples. These estimates are then used to predict the cases in the sample set aside to assess the amount of prediction error. These steps are repeated G times, setting aside one of the G sub-samples at a time.

More specifically, let $\mathbf{X}^{(g)}$ denote the g -th sample selected from \mathbf{X} and

$\mathbf{X}^{(-g)}$ denote the remaining data after $\mathbf{X}^{(g)}$ is eliminated from \mathbf{X} ($g = 1, \dots, G$).

Regularized MCA is applied to $\mathbf{X}^{(-g)}$ so as to obtain $\mathbf{W}^{(-g)}$. Then, $\mathbf{X}^{(g)}\mathbf{W}^{(-g)}\mathbf{W}^{(-g)T}$ is calculated. This procedure is repeated for all G sub-samples, and all cross validated predictions $\mathbf{X}^{(g)}\mathbf{W}^{(-g)}\mathbf{W}^{(-g)T}$ are collected in matrix $\overline{\mathbf{XD}}(\omega)^{-1}$. We then calculate

$$\varepsilon(\omega) = \text{SS}(\mathbf{XD}(\omega)^{-1} - \overline{\mathbf{XD}}(\omega)^{-1})_{I,D(\omega)}, \quad (36)$$

as an index of prediction error, where $\text{SS}(\mathbf{H})_{I,D(\omega)} = \text{trace}(\mathbf{H}'\mathbf{H}\mathbf{D}(\omega))$. We compare the values of $\varepsilon(\omega)$ for different values of ω (e.g., $\omega = 0, 1, 2, 5, 10, 20, 30$), and choose the value of ω associated with the smallest value of $\varepsilon(\omega)$.

Note that the above cross-validation procedure for determining an optimal value of ω is applied under the condition that the number of dimensions is already known in the regularized MCA solution. The permutation test may be used for dimensionality selection. The permutation test may be applied initially with $\omega = 0$, i.e., ordinary MCA, by which a tentative dimensionality is determined, and subsequently the G -fold cross validation method is applied to select an optimal value of ω .

To illustrate regularized MCA, we analyzed the data from Green and Krieger (1998). In the data, 25 consumers responded to three multiple choice items. The first item asked consumers to indicate which of four soft drinks they prefer: (1) Coke, (2) 7-up, (3) Dr. Pepper, and (4) Nehi Grape. The second item asked how much they spend on soft drinks per week: (1) Under \$2.00, (2) \$2.00 - \$ 3.99, and (3) \$4.00 and over. The last item asked consumers to indicate which snacks they prefer to eat with soft drinks: (1) Pretzels, (2) Peanuts, (3) M&M's, (4) Fritos, and (5) Dried Fruits.

We first applied ordinary/non-regularized MCA (i.e., $\omega = 0$) to the data for comparative purposes. The permutation test with 1000 permuted samples was applied to

the data under this non-regularized case. According to the permutation test, the first three inertias turned out to be significant. On the other hand, the adjusted inertias tended to decrease gradually after the first two. Moreover, the first two adjusted inertias explained about 91% of the total adjusted inertia, indicating that a two-dimensional solution accounted for a majority of the total variations among item categories. Thus, we chose dimensionality = 2. This in turn helped facilitate the interpretation of the solution.

Figure 3 displays the two-dimensional plot of the estimated principal coordinates of the response categories of the three items from non-regularized MCA. Figure 3 also provides the 95% confidence regions of the estimated category points obtained by the bootstrap method with 1000 bootstrap samples.

Insert Figure 3 about here

In the figure, the estimated coordinates for item 1 are labelled ‘d1’, ‘d2’, ‘d3’, and ‘d4’, those of item 2 are labelled ‘m1’, ‘m2’, and ‘m3’, and those of item 3 are labelled ‘s1’, ‘s2’, ‘s3’, ‘s4’, and ‘s5’. The order of the labels is equivalent to that of the categories in each item. The symbol ‘+’ indicates the origin of the two-dimensional space.

The bottom right portion of the display contains the category point of ‘m3 (\$4.00 and over)’. This point seems closer to such item categories as ‘d4 (Nehi Grape)’, and ‘s4 (Fritos)’. This suggests that heavier soft drinkers are more likely to consume Nehi Grape along with Fritos. On the other hand, the upper right portion of the display comprises the category point of ‘m1 (under \$2.00)’. This point is closely located to such item categories as ‘d2 (7-Up)’ and ‘Dried Fruits (s5)’. This indicates that light soft drink users were more

likely to consume 7-up along with dried fruits. Finally, the middle left portion of the display appears associated with moderate users of soft drinks because it embraces the category point of ‘m2 (\$2.00 - \$3.99)’. This point is positioned closer to such item categories as ‘d1 (Coke)’, ‘d3 (Dr. Pepper)’, ‘s1 (Pretzels)’, ‘s2 (Peanuts)’, and ‘s3 (M&M’s)’. This suggests that moderate soft drinkers appeared to prefer Coke and Dr. Pepper to other non-cola products, along with such snacks as Pretzels, Peanuts, and M&M’s.

Given the predetermined dimensionality, regularized MCA was subsequently applied to the same data. The G -fold cross validation method was applied to find an optimal value of the ridge parameter. In particular, in this example, we set $G = n$. This procedure is called leaving-one-out method. The leaving-one-out method was used here because the sample size was small.

The estimate of prediction error (ε) was found to be .2927 for $\omega = 0$, .2914 for $\omega = .1$, .2904 for $\omega = .2$, .2898 for $\omega = .3$, .2894 for $\omega = .4$, .2893 for $\omega = .5$, .2895 for $\omega = .6$, .2898 for $\omega = .7$, .2903 for $\omega = .8$, .2918 for $\omega = 1.0$, and .3053 for $\omega = 2.0$. Thus, the optimal value of ω was chosen as .5.

Figure 4 displays the two-dimensional plot of the estimated principal coordinates of the same response categories obtained from regularized MCA under $\omega = .5$. It also exhibits the 95% confidence regions of the estimated category points obtained by the bootstrap method with 1000 bootstrap samples. As shown in Figure 4, the confidence regions appear almost uniformly smaller for the parameter estimates obtained from regularized MCA than those from the non-regularized counterpart, indicating that the parameters were more reliably estimated in the former.

Insert Figure 4 about here

4.2. An extension of MCA for capturing cluster-level respondent heterogeneity

The parameters of MCA are currently estimated by pooling the data across respondents under the implicit assumption that all respondents come from a single, homogenous group. However, it often seems more realistic to assume that respondents come from heterogeneous groups, so that they are different with respect to their choices. Such cluster-level respondent heterogeneity has been discussed from several different theoretical and modeling perspectives (e.g., Arabie & Hubert, 1994; Bagozzi, 1982; Kamakura, Kim & Lee, 1996).

MCA was recently extended to explicitly account for cluster-level heterogeneity in respondents' preferences/choices (Hwang, Dillon, & Takane, 2007). Specifically, this approach combines MCA with the *c*-means algorithm (MacQueen, 1967) in a unified framework. The *c*-means algorithm is perhaps the most popular method for non-overlapping clustering (Wedel & Kamakura, 1998). It is efficient in dealing with large data (Green, Carmone, & Kim, 1990). More importantly, the *c*-means algorithm turns out to be beneficial because it is easily combined with the homogeneity criterion for MCA in a single framework.

We first discuss the technical underpinning of this unified approach in brevity. We then present an empirical application to illustrate the usefulness of the approach.

Let c denote the prescribed number of clusters. Let $\mathbf{\Pi}$ denote an n by c matrix of binary memberships, which allocates respondents into only one of c clusters (1 = member and 0 = non-member). Let $\mathbf{\Lambda}$ denote a c by d matrix of the centroids or mean values of clusters. Let α_1 and α_2 denote non-negative scalars.

The objective of the proposed unified approach is to combine MCA and c -means into a single framework. This problem is equivalent to minimizing the following:

$$\phi_7(\mathbf{F}, \mathbf{B}_k, \mathbf{\Pi}, \mathbf{\Lambda}) = \alpha_1 \sum_{k=1}^K \text{SS}(\mathbf{F} - \mathbf{X}_k \mathbf{B}_k) + \alpha_2 \text{SS}(\mathbf{F} - \mathbf{\Pi} \mathbf{\Lambda}), \quad (37)$$

with respect to \mathbf{F} , \mathbf{B}_k , $\mathbf{\Pi}$, and $\mathbf{\Lambda}$, subject to $\mathbf{F}'\mathbf{F} = \mathbf{I}$ and $\alpha_1 + \alpha_2 = 1$. When $\alpha_1 = 1$, the first term in (37) reduces to the homogeneity criterion for MCA in (19). When $\alpha_2 = 1$, the second term is equivalent to the standard criterion used in the c -means clustering algorithm. By minimizing both criteria in (37) simultaneously, \mathbf{F} is obtained in such a way that it recognizes the cluster structure that may be inherent in multiple-choice questions.

The values of α_1 and α_2 are *a priori* specified by the investigator. By specifying $\alpha_1 = \alpha_2 = .5$, the two terms for MCA and c -means are to be balanced. On the other hand, the two terms may be differently weighted for adjusting for their relative importance. For instance, we may wish to weigh the first term more heavily than the second term under the belief that data reduction is of more importance than clustering.

An alternating least squares algorithm (de Leeuw, Young, & Takane, 1976) is developed to minimize (37). In the algorithm, the unknown parameters, \mathbf{F} , \mathbf{B}_k , $\mathbf{\Pi}$, and $\mathbf{\Lambda}$, are updated alternately until convergence. The updates of one parameter matrix are obtained such that they minimize (37) in the least squares sense, while the others remain

fixed. Refer to Hwang et al. (2007) for the detailed description of the alternating least squares algorithm.

In effect, the alternating least squares algorithm monotonically decreases the value of criterion (37) which, in turn, is also bounded from below. The algorithm is therefore convergent. However, it does not guarantee that the convergence point is the global minimum. In particular, the c -means algorithm has been shown to be sensitive to local optima (Steinley, 2003). To safeguard against local minima, we repeat the alternating least squares procedure with a large number (say, 100) of random initial values for $\mathbf{\Pi}$. (The initial values for $\mathbf{\Delta}$ are obtained from $\mathbf{\Pi}$.) We then compare the obtained function values after convergence and subsequently choose the solution associated with the smallest one. Besides $\mathbf{\Pi}$ (and $\mathbf{\Delta}$), MCA is applied to the original data and the resultant low-dimensional data are used as rational starts for \mathbf{F} . The initial values for \mathbf{B}_k are obtained on the basis of \mathbf{F} .

In the proposed method, we need to decide *a priori* on the number of clusters, c , as well as the number of dimensions in the data, d . One simple approach consists in first selecting d by applying MCA to the data, and then deciding on the value of c by examining how the values of (37) change across different numbers of clusters (Wedel & Kamakura, 1998). It is recommended that the number of clusters be greater than the number of dimensions (Van Buuren & Heiser, 1989; Vichi & Kiers, 2001). In practice, non-statistical heuristics for evaluating the usefulness and relevance of clusters (e.g., cluster size, potential, interpretability, etc.) also plays an important role in deciding c (Arabie & Hubert, 1994; Wedel & Kamakura, 1998).

The example presented below was chosen for illustrative purposes. The data were part of the television program preference data presented in Adachi (2000). In this example, 100 Japanese undergraduate students (49 males and 51 females) were asked to provide their favourite TV program among six different program categories at each of three time points. The purpose of our analysis is to provide a low-dimensional representation of television viewing preferences while investigating whether groups of respondents exhibit qualitatively distinct patterns of choice responses to the different TV programs over time.

The three time points correspond to i) the first year of elementary school ($t = 1$), ii) the first year of junior high school ($t = 2$), and iii) the freshman year at university ($t = 3$). In Japan, these time points usually correspond with ages 6-7, 12-13 and 18-20, respectively. The six TV program categories are: animation (a), cinema (c), drama (d), music (m), sports (s), and variety (v). Thus, we can describe these data as consisting of three multiple-choice questions corresponding to the three time points, each of which is composed of six response categories corresponding to the six different TV programs.

At first, ordinary MCA was utilized so as to gain a basic understanding of the associations between variables and clusters. We chose $d = 2$ because the values of the adjusted inertias appeared to decrease slowly after the first two. The first two adjusted inertias explained about 84% of the adjusted total inertia. Next, with d fixed, we investigated changes in the value of (37) by varying numbers of clusters. The values of (37) appear to decrease gradually beyond three clusters, suggesting that no substantial changes in the criterion values are obtained by having more than three clusters. Thus, $c = 3$ was adopted for our analysis.

Next, given the predetermined numbers of dimensions and clusters, the proposed unified approach was applied to the same data. Figure 5 displays the two-dimensional plot for the estimated principal coordinates of the response categories of the two questions as well as the estimated centroids of three clusters obtained from the unified approach.

Insert Figure 5 about here

In this map, the estimated response categories at each time were represented by a two-digit label, in which the first digit indicates one of the six TV programs and the second corresponds to the time point number ($t = 1, 2, 3$). For example, ‘a1’ = animation at $t = 1$, ‘c2’ = cinema at $t = 2$, ‘d3’ = drama at $t = 3$, and so forth. Moreover, the three centroids were labelled ‘CL1’, ‘CL2’, and ‘CL3’. The symbol ‘+’ represents the origin of the display.

In Figure 5, the first cluster of respondents, whose centroid is represented by ‘CL1’, is located on the bottom of the map. ‘CL1’ is closer to such response categories as ‘v1’, ‘v2’, and ‘v3’. It suggests that the respondents in this cluster are likely to exclusively choose the variety-show program over time—in other words, they show strong preference for variety programming and their preferences do not change with time. Approximately 29% of the respondents were classified into this cluster.

On the other hand, the second cluster of respondents appears to be located on the middle right-hand side of the map, where its centroid (‘CL2’) is located. This centroid is closely located with such response categories as ‘s1’, ‘s2’, ‘s3’, and ‘d1’. This indicates

that the respondents in the second cluster seem to show preferences for sport and drama programming at an early age; moreover, their preference for sports programming does not change over time, their preference for drama programming does. About 9% of the respondents belong to the second cluster.

Finally, the middle left-hand side of Figure 5 is best associated with the third cluster. This centroid ('CL3') is positioned close to the other remaining response categories, i.e., animation, cinema, drama, music at $t = 1, 2,$ and 3 (except drama at $t = 1$). Thus, respondents in this cluster have the most eclectic viewing preferences and enjoy a broad range of TV programming from animation to music, rather than focusing on a particular genre of programming over time. This is the largest cluster representing about 62% of all respondents.

5. Conclusions

CA and MCA are flexible exploratory tools for studying interrelationships among multiple-choice questions. In this chapter, it was shown that technically CA and MCA represent special cases of canonical correlation analysis and multiple-set canonical correlation analysis, respectively, where a set of continuous variables are replaced by a set of response categories of a multiple-choice question. Accordingly, CA and MCA, each assign numerical values (weights) to the response categories of two or more multiple-choice questions. The numerical values (scaled by singular values) are graphically displayed in a low-dimensional display. This graphical display helps one to quickly understand data structures and permits the efficient communication of this information to practitioners and other researchers. Also, CA is essentially viewed as a

special case of MCA where there are two multiple-choice questions involved. Similarly, CCA represents a special case of MCCA.

CA and MCA are nonparametric techniques that do not require distributional assumptions underlying multiple-choice questions. In addition, as stated earlier, the data for CA and MCA – multiple-choice questions – are very flexible so that they may include many other types of categorical variables as special cases (e.g., binary, frequency table, sorting data, ranking data, etc.) (Nishisato, 1994). Furthermore, the interpretation of the results is straightforward and easy to understand to non-statistical experts.

Two recent extensions of MCA were also introduced in this chapter along with illustrative applications to survey data. In sum, these extensions render MCA more versatile in capability. For example, regularized MCA is useful in providing more accurate estimates of parameters, particularly when the number of respondents is small. Moreover, the unified approach to MCA and *c*-means is beneficial in revealing relationships as well as segmentation structures inherent to multiple-choice questions. This unified approach is quite versatile and therefore applicable to various clustering/segmentation situations which involve multiple-choice questions. In addition, the individual membership information furnished by this approach may be beneficial in profiling/describing the clusters when used together with demographic variables of respondents.

Nevertheless, CA and MCA do involve limitations as well. They are essentially *descriptive* statistical techniques. Thus, they are not suitable for hypothesis testing although certain types of hypotheses can ‘empirically’ be investigated by the use of linear constraints (e.g., Böckenholt & Takane, 1994; Hwang & Takane, 2002; Takane &

Hwang, 2002). This may render interpretations of solutions less objective. Moreover, they are distribution-free methods. Hence, they suffer from the lack of post-hoc fit indices for model selection (e.g., AIC or BIC). However, the entailed subjectivity of the interpretations may be regarded as a trade-off with respect to the graphical flexibility of the method (Hoffman & Franke, 1986). Furthermore, although the method is not well furnished with statistical fit measures for model selection, one can still depend on non-statistical considerations to alleviate this limitation.

The two extensions of MCA may also be further generalized so as to enhance its data-analytic capability. For example, regularized MCA is currently based on ridge-type regularization which involves the specification of a scalar. However, we may also consider other, more complicated types of regularization, for instance, a regularization term capturing the degree of smoothness in curves (Adachi, 2002; Ramsay & Silverman, 2005). Moreover, the unified approach to MCA and c-means may be extended by replacing the hard-clustering method by a fuzzy-clustering method such as fuzzy c-means (Bezdek, 1974, 1981; Dunn, 1974; Manton, Woodbury, & Tolley, 1994; Wedel & Steenkamp, 1989). This fuzzy-clustering extension may be more favorable than the current method because it provides a probabilistic classification of respondents (Wedel & Kamakura, 1998).

In sum, CA and MCA are useful techniques that afford a flexible and parsimonious graphical display of structures inherent in multiple-choice questions. They are versatile in data requirements and easy to use computationally. CA and MCA will remain as popular descriptive techniques which give rise to a broad range of applications in a variety of areas of inquiry.

References

- Adachi, K. (2000). Optimal scaling of a longitudinal choice variable with time-varying representation of individuals. *British Journal of Mathematical and Statistical Psychology*, 53, 233-253.
- 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.
- Arabie, P., & Hubert, L. (1994). Cluster analysis in marketing research. In R. P. Bagozzi (Ed.), *Advanced Methods of Marketing Research* (pp. 160-189). Oxford: Blackwell.
- Arimond, G., & A. Elfessi, A. (2001). A clustering method for categorical data in tourism market segmentation research. *Journal of Travel Research*, 39, 391-397.
- Bagozzi, R. P. (1982). A field investigation of causal relations among cognition, affect, intentions, and behavior. *Journal of Marketing Research*, 19, 562-584.
- Bartlett, M. S. (1938). Further aspects of the theory of multiple regression. *Proceedings of Cambridge Philosophical Society*, 34, 33-40.
- Benzécri, J. P. (1973). *L'analyse des données. Vol. 2. L'analyse des correspondances*. Paris: Dunod.
- Benzécri, J. P. (1979). Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire. Addendum et erratum à [BIN.MULT]. *Cahiers de L'analyse des Données*, 4, 377-378.
- Bezdek, J. C. (1974). Numerical taxonomy with fuzzy sets. *Journal of Mathematical Biology*, 1, 57-71.

- Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*.
New York: Plenum Press.
- Blasius, J. & Greenacre, M. (1994). Computation of correspondence analysis. In M.
Greenacre & J. Blasius (Eds.), *Correspondence Analysis in the Social Sciences*
(pp. 53-78). San Diego: Academic Press.
- Böckenholt, U., & Takane, Y. (1994). Linear constraints in correspondence analysis. In
M. J. Greenacre & J. Blasius (Eds.), *Correspondence Analysis in Social Sciences*
(pp. 112-127). London: Academic Press.
- Carroll, J. D. (1968). A generalization of canonical correlation analysis to three or more
sets of variables. Proceedings of the 76th Annual Convention of the American
Psychological Association, 227-228.
- de Leeuw, J., Young, F. W., & Takane, Y. (1976). Additive structure in qualitative data:
An alternating least squares method with optimal scaling features. *Psychometrika*,
41, 471-503.
- Dolničar, S., & Leisch, F. (2001). Behavioral market segmentation of binary guest survey
data with bagged clustering. In Dorffner, G., Bischof, H., & Hornik, K. (Eds.).
ICANN 2001 (pp. 111-118). Berlin: Springer-Verlag.
- Dunn, J. C. (1974). A fuzzy relative of the ISODATA process and its use in detecting
compact well-separated clusters. *Journal of Cybernetics*, 3, 32-57.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*,
7, 1-26.
- Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Chichester: Wiley.

- Green, P. E., Carmone, F. J., & Kim, J. (1990). A preliminary study of optimal variable weighting in k-means clustering. *Journal of Classification*, 7, 271-285.
- Green, P. E., & Krieger, A. M. (1995). Alternative approaches to cluster-based market segmentation. *Journal of the Market Research Society*, 37, 221-239.
- Green, P. E., & Krieger, A. M. (1998). *User's Guide to HIERMAPR*. The Wharton School. University of Pennsylvania.
- Green, P. E., Krieger, A. M., & Carroll, D. J. (1987). Conjoint analysis and multidimensional scaling: A complementary approach. *Journal of Advertising Research*, 27, 21-27.
- Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.
- Greenacre, M. J. (1993). *Correspondence Analysis in Practice*. London: Academic Press.
- Greenacre, M. J. (1994). Correspondence analysis and its interpretation. In M. Greenacre & J. Blasius (Eds.), *Correspondence Analysis in the Social Sciences* (pp. 3-22). San Diego: Academic Press.
- Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.
- Hirschfeld, H. O. (1935). A connection between correlation and contingency. *Cambridge Philosophical Society Proceedings*, 31, 520-524.
- Hoerl, A. F., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 55-67.

- Hoffman, D. L., & Franke, G. R. (1986). Correspondence analysis: Graphical representation of categorical data in marketing research. *Journal of Marketing Research*, 23, 213-227.
- Hoffman, D. L., de Leeuw, J., & Arjunji, R. V. (1994). Multiple correspondence analysis. In Bagozzi, R. P. (ed.), *Advanced Methods of Marketing Research* (pp. 260-294). Oxford: Blackwell.
- Horst, P. (1961). Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology*, 17, 331-347.
- Hotelling, H. (1936). Relations between two sets of variables. *Biometrika*, 28, 321-377.
- Hwang, H., & Takane, Y. (2002). Generalized constrained multiple correspondence analysis. *Psychometrika*, 67, 211-224.
- Hwang, H., Dillon, W. S., & Takane, Y. (2006). An extension of multiple correspondence analysis for identifying heterogeneous subgroups of respondents. *Psychometrika*, 71, 161-171.
- Javalgi, R., Whipple, T., McManamon, M., & Edick, V. (1992). Hospital image: A correspondence analysis approach. *Journal of Health Care Marketing*, 12, 34-41.
- Kamakura, W. A., Kim, B., & Lee, J. (1996). Modeling preference and structural heterogeneity in consumer choice. *Marketing Science*, 15, 152-172.
- Lebart, L., Morineau, A., & Warwick, K. M. (1984). *Multivariate Descriptive Statistical Analysis*. New York: Wiley.
- Legendre, P., and Legendre, L. (1998). *Numerical ecology*. Amsterdam: North Holland.

- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Le Cam, L. M. & Neyman, J. (Eds.). *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281-297).
- Manly, B. J. F. (1997). *Randomization and Monte Carlo Methods in Biology*. London: Chapman & Hall.
- Manton, K. G., Woodbury, M. A., & Tolley, H. D. (1994). *Statistical Applications Using Fuzzy Sets*. New York: John Wiley & Sons.
- Meredith, W. (1964). Rotation to achieve factorial invariance. *Psychometrika*, 29, 187-206.
- Nishisato, S. (1980). *Analysis of Categorical Data: Dual scaling and Its Applications*. Toronto: University of Toronto Press.
- Nishisato, S. (1994). *Elements of Dual Scaling: An Introduction to Practical Data Analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Nishisato, S. (2007). *Multidimensional Nonlinear Descriptive Analysis*. New York: Chapman & Hall/CRC.
- Ramsay, J. O. (1978). Confidence regions for multidimensional scaling analysis. *Psychometrika*, 43, 145-160.
- Ramsay, J. O., & Silverman, B. W. (2005). *Functional Data Analysis. 2nd Edition*. New York: Springer.
- Takane, Y. (1980). Analysis of categorizing behavior by a quantification method. *Behaviormetrika*, 8, 75-86.

- Takane, Y., & 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 & Hall/CRC.
- ten Berge, J. M. F. (1979). On the equivalence of two oblique congruence rotation methods, and orthogonal approximations. *Psychometrika*, 44, 359-364.
- ten Berge, J. M. F. (1993). *Least Squares Optimization in Multivariate Analysis*. Leiden University, The Netherlands: DSWO Press.
- ter Braak, C. J. F. (1990). *Update notes: CANOCO Version 3.10*. Wageningen, The Netherlands: Agricultural Mathematics Group.
- van Buuren, S., & Heiser, W. J. (1989). Clustering n objects into k groups under optimal scaling of variables. *Psychometrika*, 54, 699-706.
- Vichi, M., & Kiers, H. A. L. (2001). Factorial k-means analysis for two-way data. *Computational Statistics and Data Analysis*, 37, 49-64.
- Wedel, M., & Kamakura, W. A. (1998). *Market Segmentation: Conceptual and Methodological Foundations*. Boston: Kluwer Academic Publishers.
- Wedel, M., & Steenkamp, J.-B. E. M. (1989). Fuzzy clusterwise regression approach to benefit segmentation. *International Journal of Research in Marketing*, 6, 241-258.
- Yanai, H. (1998). Generalized canonical correlation analysis with linear constraints. In C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock, & Y. Baba (Eds.). *Data*

Science, Classification, and Related Methods (pp. 539-546). Tokyo: Springer-Verlag.

Table 1. Inertias and corresponding percentages of total inertia obtained from the 2000 Canadian Election Survey data.

Inertia	Percentage
0.2792	67.57
0.0803	19.44
0.0248	5.99
0.0192	4.64
0.0062	1.50
0.0027	1.00
0.0006	.00
0.0002	.00
0.0000	.00

Table 2. Adjusted inertias and corresponding percentages of the average off-diagonal inertia obtained from the 2000 Canadian Election Survey data.

Inertia	Percentage
9.0376	22.1414
5.0368	12.3397
3.9010	9.5571
3.4303	8.4040
3.2001	7.8400
1.9970	4.8926
1.4815	3.6297
1.1970	2.9326
1.0240	2.5088
1.0027	2.4565
1.0003	2.4506
0.9510	2.3298
0.8757	2.1454
0.7835	1.9194
0.6623	1.6226
0.6068	1.4866
0.3719	0.9110
0.3500	0.8575
0.1845	0.4520
0.0537	0.1315
0.0324	0.0794
0.0106	0.0260
0.0020	0.0050

Figure 1. The symmetric map of the 2000 Canadian Election Survey data obtained from correspondence analysis.

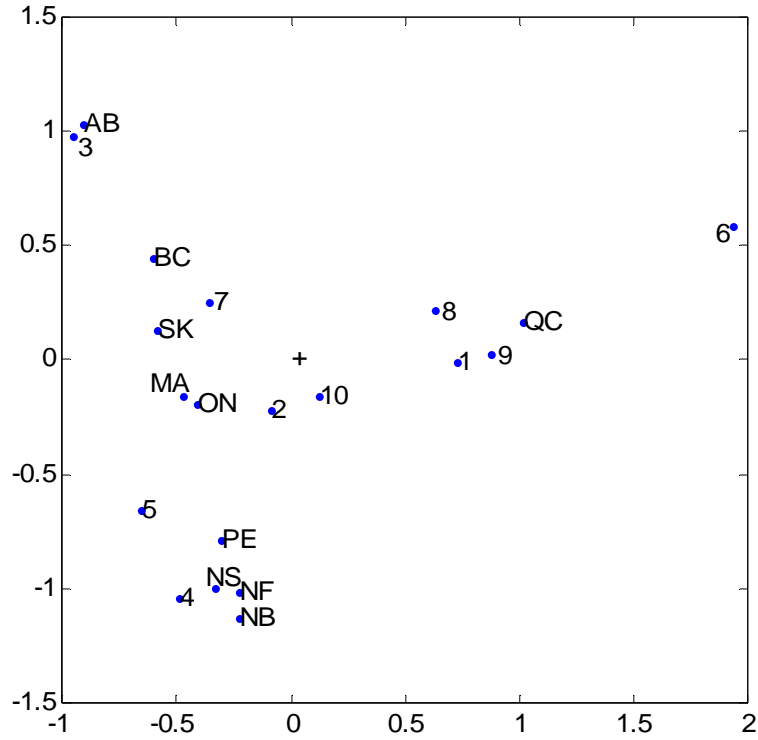


Figure 2. The symmetric map of the 2000 Canadian Election Survey data obtained from multiple correspondence analysis.

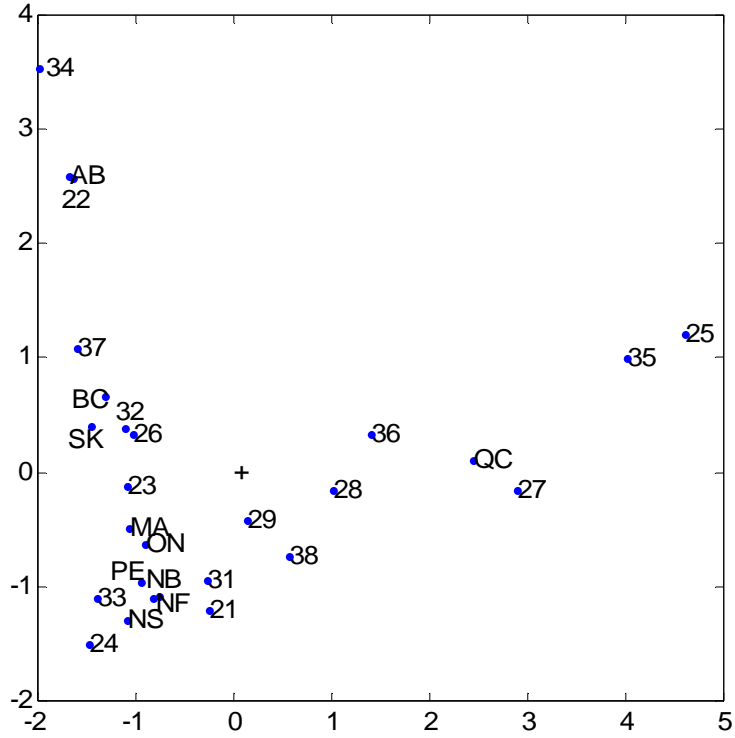


Figure 3. The symmetric map of the soft drink data obtained from ordinary, non-regularized multiple correspondence analysis, along with the 95% confidence regions of the estimated principal coordinates.

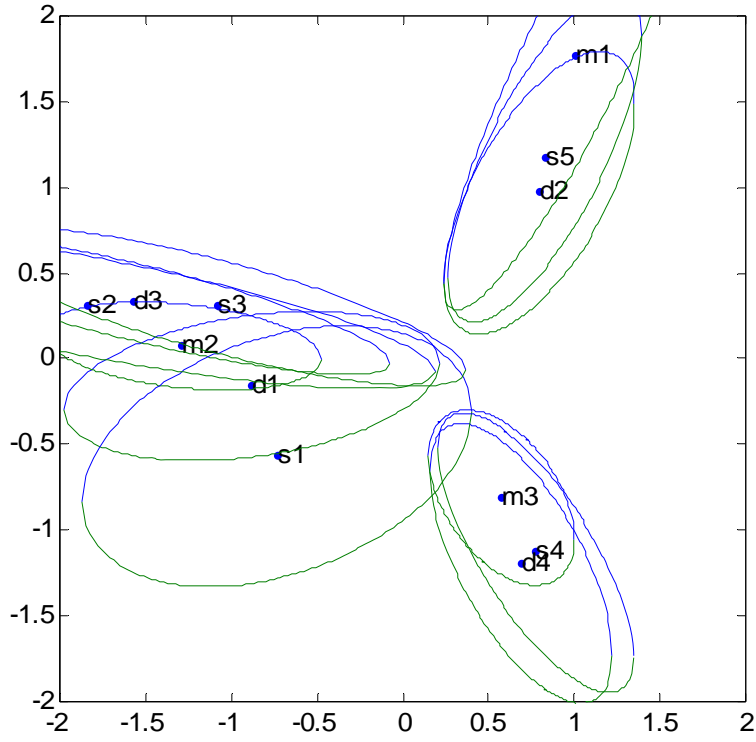


Figure 4. The symmetric map of the soft drink data obtained from regularized multiple correspondence analysis, along with the 95% confidence regions of the estimated principal coordinates.

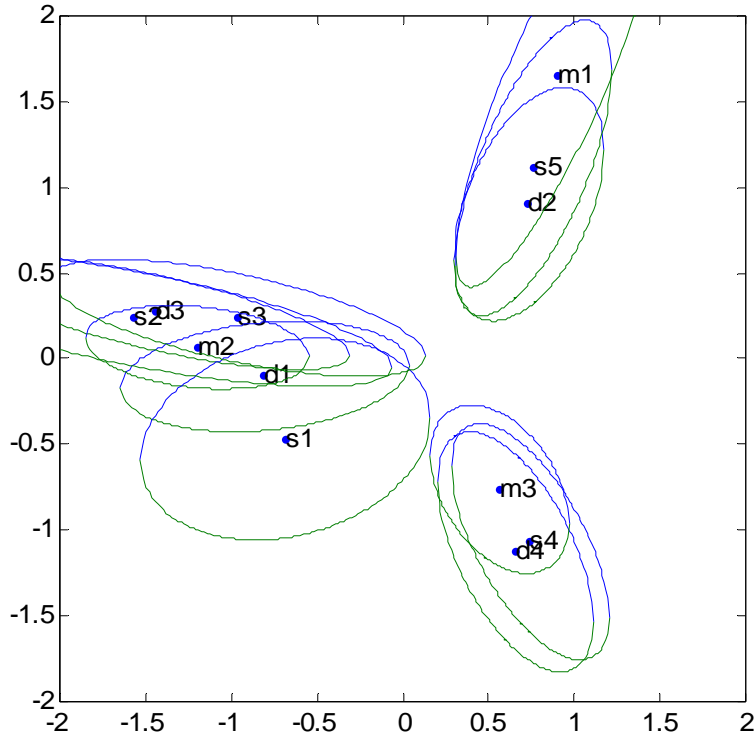


Figure 5. The symmetric map of the TV program preference data obtained from the unified approach to multiple correspondence analysis and c-means.

