GENERALIZED NONLINEAR CANONICAL CORRELATION ANALYSIS WITH ORDERED CATEGORICAL AND DICHOTOMOUS DATA

Thanoon Y. Thanoon\textsuperscript{a,c}, Robiah Adnan\textsuperscript{a}, Seyed Ehsan Saffari\textsuperscript{b}

\textsuperscript{a}Department of Mathematical Science, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia
\textsuperscript{b}Education Development Centre, Sabzevar University of Medical Sciences, 961387316-319, Sabzevar, Iran
\textsuperscript{c}Foundation of Technical Education, Technical College of Management, Mosul, Iraq

Abstract

In this paper, ordered categorical and dichotomous data are used in generalized nonlinear canonical correlation analysis to study the relationship between two or more sets of variables. Statistical analyses involving generalized nonlinear canonical correlation analysis, component loadings, and object scores are discussed in this paper. The proposed procedure is illustrated using a real data set (patients with high blood pressure). Analyses are done using SPSS program. The component loadings graph of the three sets shows the relationship between the three sets and their impact on the data set of patients with high blood pressure. The centroid graph of the categories also shows the relationship between them.

Keywords: Nonlinear canonical correlation analysis, generalized canonical correlation analysis, categorical data, dichotomous data

1.0 INTRODUCTION

Canonical correlation analysis (CCA) (see Fan & Konold,\textsuperscript{1} Timm,\textsuperscript{2} Via, et al.,\textsuperscript{3}) is a well-known technique in multivariate statistical analysis to find the maximally correlated projections between two data sets.

Generalized nonlinear canonical correlation analysis, OVERALS (see Kiers, et al.,\textsuperscript{4} Van de Velden & Takane,\textsuperscript{5} Van der Burg, et al.,\textsuperscript{6}) is one of the multivariate methods, which is the general state of the canonical correlation analysis (CCA). In this paper, the canonical correlation analysis, which represents the relationship between two linear sets (K=2) is addressed and the relationship between more than two sets (K>2) will also be discussed using the generalized nonlinear canonical correlation analysis.\textsuperscript{7}
Generalized nonlinear canonical correlation analysis corresponds to the categorical canonical correlation analysis with optimal scaling. OVERALS was developed by Gifi, Van Rijckevorsel & de Leeus which is based on generalized nonlinear canonical correlation analysis, achieving a minimum loss between object scores and canonical variables in all combined sets and optimal scales.

Standard canonical correlation analysis is an extension of multiple linear regression analysis, where instead of one it has several dependent variables.

At present, most statistical theory and computer software in the field are based on models that involve relationships among three sets of variables. More statistically sound methods for linear and nonlinear canonical correlation analysis have been proposed by Van der Burg, Van der Lans, Weenink, and Yanai & Takane.

The purpose of OVERALS is to determine how similar the sets of categorical variables are to each other. The goal of canonical correlation analysis is to clarify the difference in the relationships between a set of variables. In addition, it is used to reduce the dimensions of the variables used. The variables in each set, integrate the linear structures that have the highest correlation and give the uncorrelated subsequent linear sets to the previous sets.

Generalized nonlinear canonical correlation analysis for optimal scales expands the standard analysis in several ways.

First, you can use OVERALS to analyse data with more than two sets of variables. Second, variables in each set can be measured by nominal, ordinal or numerical scales as a result of the non-linear relationships between sets of variables. Third, instead of maximizing the correlation between variable sets as in the canonical analysis of the link between the two sets, the sets can be compared to other sets called the object scores as described in Appendix (1).

The main idea in this paper is to find the relationships between three groups of variables with ordered categorical and dichotomous data by using generalized nonlinear canonical correlation analysis. More statistically sound methods in the field are based on canonical correlation analysis which involve linear and nonlinear relationships between the multiple group of variables proposed by Hardoon, Hsieh, Lal, Thompson, and Thorndike.

The paper is organized as follows. Generalized Canonical Correlation Analysis is described in section 2. Methods and research tools are explained in Section 3. Real example to illustrate the method used is presented in Section 4. Statistical analysis and empirical results discussed in section 5. Some concluding remarks are given in section 6.

2.0 GENERALIZED CANONICAL CORRELATION ANALYSIS

In a generalized canonical correlation analysis, linear combinations are obtained in such a way that the sum of the squared correlations of the linear combinations of the variables with a so-called group configuration is a maximum.

Let \( Y \) denote the unknown group configuration. The order of \( Y \) is \( m \times k \), where \( m \) is the number of rows for each observation matrix \( X_i \) (i.e. the \( i \)th data set) and \( k \) is the dimensionality of the solution. The data matrices \( X_i \) are first centered if the variables are measured on different scales. Note that the sizes of the observation matrices \( X_i \) are \( m \times p_i \) with \( p_i \leq m -1 \) for \( i = 1, 2, \ldots, n \). The dimensionality of the solution \( k \) must be chosen by the researcher.

We can formulate as objective

\[
\text{min } \phi \left( \theta, A_i \right) = \text{min } \text{trace} \left( \sum_{i=1}^{n} \theta' - X_i A_i' (Y - X_i A_i) \right) \tag{1}
\]

subject to the restriction

\[
Y' Y = I_k
\]

It is known that for observed \( X_i \) matrices, the group configuration \( Y \) can be obtained from the Eigen equation

\[
\left( \sum_{i=1}^{n} X_i' (X_i' X_i)^{-1} X_i' \right) Y = Y_{n 	imes k} A_{n 	imes k} \tag{2}
\]

where \( \Lambda \) is a diagonal matrix with elements from the \( k \) largest eigenvalues of

\[
\sum_{i=1}^{n} X_i' (X_i' X_i)^{-1} X_i'
\]

where we have assumed that the \( X_i' \)'s are full column rank and the matrices \( A_i \) can be calculated as

\[
A_i = (X_i' X_i)^{-1} X_i' Y
\]

An interesting feature of the method is the fact that the sets of variables \( X_i \) may contain different variables. Hence, the number of variables in each set does not need to be the same.

3.0 METHODS AND RESEARCH TOOLS

We will apply OVERALS on three types of data namely, ordinal variables, single nominal variables, multiple nominal variables. The detailed description of the data is as follows:

3.1 Categorical Variables

Qualitative variables are divided into three sections:

A. Ordinal Variables: is an ordered qualitative variable and an example of this type of variable is level of education: 1: uneducated, 2: primary, 3: secondary, 4: university.

B. Single nominal variables: is an unordered qualitative variable which contains only two categories of this type of variable is gender: 1: male, 2: female.

C. Multiple nominal variables: is an unordered qualitative variable which contains more than two categories and an example of this type of variable is status on smoking habit: reduce smoking, 1: Yes, 2: No, 3: No Smoking.
3.2 Generalized Nonlinear Canonical Correlation Analysis OVERALS

In Hotelling’s canonical correlation analysis, one studies the relationship between two sets of variables after removing the linear dependencies within each of these sets. OVERALS involves comparing K sets of variables after removing the linear dependencies within each set. Various approaches suggested generalizing Hotelling’s canonical correlation procedure to K sets of variables. In a K set problem, there are K(K-1)/2 canonical correlations among the optimal set of canonical variables that can be obtained from a K × K correlation matrix R.

As noted, OVERALS maximizes the sum of correlations between columns of n×p comparison matrix and corresponding columns of n×p matrices of canonical variables. Here, n is the number of objects and p is the number of dimensions. We have K matrices of canonical variables, one matrix of canonical variables from each set. The objective of OVERALS is to minimize a loss function that can be written as follows:

\[ \sigma(X , Q , A ) = K^{-1} \sum K SSQ (X - Q_k A_k ) \]

K: is the number of sets.
X: is an n×p of comparison scores, where n is the number of objects and p is the dimensions.
Q: is an n×m partitioned matrix containing scores (to be estimated) for variables with m total number of variables.
Qk: is an n×mk matrix containing scaled variables within set k, where mk is the number of variables in set k, thus Q=[Q1 / Q2 /... /Qk /... /Qk].
A: is an m×p partitioned matrix of canonical weights.
A_k: is an mk×p matrix of canonical weights of the variables in set k, thus A=[A1 / A2 /... /Ak /... /Ak].
SSQ(): denotes the sum of squared elements of the matrix between the brackets.

4.0 EXAMPLE

4.1 Data Collection

The data was obtained from Ibn Sina hospital in Mosul-Iraq. Fourteen patients were randomly chosen and their blood pressure was taken. There are a number of variables believed to affect blood pressure. A set of doctors were consulted to identify the variables thought to be relevant in a questionnaire to be given to the patients. The variables were divided into three sets; personal variables, pathological variables and therapeutic variables. The detailed description of these variables is in Appendix (2).

4.2 Results and Discussion

Statistical software SPSS Meulman & Heiser has been used to analyse the data using generalized nonlinear canonical correlation to search for relationships and similarities between and within the three sets of variables. Table 1 shows the sets and number of categories and types of variables and categorycode in each variable, which were analysed using OVERALS. Table 2 shows the eigenvalues and the relationship described in each dimension where the maximum value for the Eigenvalues is 1 and the missing value is zero.

Clearly from the study the Eigenvalues were relatively high (0.655) and (0.623), while the real value of the fitting is (1.278), which represents the sum of Eigenvalues calculated from the differences. So we will use the two-dimensional solutions and therefore 1.278 / 2 = 63.9% of the differences accounted for in the analysis. Also:

1.278 / 0.655 from real data are calculated by fitting the first dimension.
1.278 / 0.623 of the corresponding real data are calculated by the second dimension.

Loss values representing the difference rate in each object scores in each dimension and in each set are in Table 3. The average rate of loss of the sets is actually (maximum values - real fitting values) = 2- 1.278 = 0.722, which need not be at a high level.

Sum of loss rate and fitting must be equal to the number of dimensions in the study (1.278 + 0.722 = 2). Thus the loss values indicate how small or large are the multi-correlations between the total weighted variables with optimal scales and between dimensions. Loss values, Eigen values and fit values showing the relationship between the sets are shown in Table 2.

Components loading for three sets describe the loading ratio for each variable in each set and each dimension, where the dimensions of the study were reduced to two, as shown in Table 3. Component loadings are the measures for the correlation between the objective scores and variables related to optimal scales in the absence of a loss data. The component loadings are equal to the Pearson correlation coefficient between the variables measured and the object scores. The component loadings also represent the coordinates of varying points on the chart and thus can be interpreted easily through graphical representation.

Multi nominal variables (pressure test, reduce smoking) have two component loadings, which was represented by measuring the kind of variables that are different in each dimension, as shown in Table 3. The previous variables with two-points are also shown in Table 3. The remaining variables were represented by one point.

The distance from the origin point for each point is represented by drawing a particular variable which represents the importance of that variable, so the component loadings prove that the variables (regular blood pressure check, angina, reduce weight, reduce salt, stroke, heart attack, genetic factor, age) were the most effective in the relations between sets of variables because they are far from the point of origin, which means that the variables (exercise, education, disease duration) were medium-effective, and the remaining variables (gender, reduce smoking, job,
kidney damage) do not have any impact on relations between the sets because they are close to the origin point.

### Table 1 Variables for three sets

<table>
<thead>
<tr>
<th>Sets</th>
<th>Number of category</th>
<th>Variable type</th>
<th>Category symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal var.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>3</td>
<td>Ordinal</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>2</td>
<td>Single Nominal</td>
<td></td>
</tr>
<tr>
<td>Job</td>
<td>4</td>
<td>Ordinal</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>4</td>
<td>Ordinal</td>
<td></td>
</tr>
<tr>
<td>Genetic factor</td>
<td>2</td>
<td>Single Nominal</td>
<td></td>
</tr>
<tr>
<td>Disease duration (period)</td>
<td>3</td>
<td>Ordinal</td>
<td></td>
</tr>
<tr>
<td>Pathological var.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Angina</td>
<td>2</td>
<td>Single Nominal</td>
<td></td>
</tr>
<tr>
<td>Heart attack</td>
<td>2</td>
<td>Single Nominal</td>
<td></td>
</tr>
<tr>
<td>Kidney damage</td>
<td>2</td>
<td>Single Nominal</td>
<td></td>
</tr>
<tr>
<td>Therapeutic var.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pressure Test</td>
<td>3</td>
<td>multiple Nominal</td>
<td></td>
</tr>
<tr>
<td>Reduce smoking</td>
<td>3</td>
<td>multiple Nominal</td>
<td></td>
</tr>
<tr>
<td>Reduce salt</td>
<td>2</td>
<td>Single Nominal</td>
<td></td>
</tr>
<tr>
<td>Reduce weight</td>
<td>2</td>
<td>Single Nominal</td>
<td></td>
</tr>
<tr>
<td>Exercise</td>
<td>2</td>
<td>Single Nominal</td>
<td></td>
</tr>
</tbody>
</table>

### Table 2 A summary of the two dimensional analysis

<table>
<thead>
<tr>
<th>Dimension</th>
<th>1</th>
<th>2</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loss</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Set 1</td>
<td>.175</td>
<td>.330</td>
<td>.505</td>
</tr>
<tr>
<td>Set 2</td>
<td>.540</td>
<td>.496</td>
<td>1.036</td>
</tr>
<tr>
<td>Set 3</td>
<td>.321</td>
<td>.305</td>
<td>.625</td>
</tr>
<tr>
<td>Mean</td>
<td>.345</td>
<td>.377</td>
<td>.722</td>
</tr>
<tr>
<td>Eigenvalues</td>
<td>.623</td>
<td>.655</td>
<td></td>
</tr>
<tr>
<td>Fit</td>
<td></td>
<td></td>
<td>1.278</td>
</tr>
</tbody>
</table>
5.0 STATISTICAL ANALYSIS

Figure 1 shows the correlation relationship between the three sets, “personal”, “therapeutic” and “pathological” variables. It can be seen from Figure 1 that points far away from the origin have a relationship. The categories of these variables were painted using the centroids graph as shown in Figure 2, and for the purpose of understanding the relationships between variables in all sets it is preferred to draw a circle around a set category which converged to form a cluster for the purpose of distinguishable.

Figure 2 shows the correlation relationship between four points through categories for those points (AB, F, W, L) which belong to the variables (education “secondary” damage renal “infected,” job “Manual”, regular blood pressure check “when needed”) as shown in the upper right side of Figure 2. The bottom right side of the figure has shown the importance of the following categories (O, R, U) which belong to the variables (heart attack “infected”, the duration of the disease, “more than 10 years”, genetic factor “No”). The upper left side shows the importance of category (AG) which belong to the variables (reduced salt food “no”), which have no correlation relationship with any of the categories of the rest of the variables. Finally there are correlation relationships between the categories (Z,C) which belong to the variables (age “68-84” regular blood pressure check, “irregular”) and categories (AJ, S) for the variables (angina, “infected”, exercise, “Yes”), as shown in the lower left side of Figure 2.
6.0 CONCLUSIONS

Generalized nonlinear canonical correlation analysis method (OVERALS) is very useful in graphical representation and interpretation of data by discovering the structures and similar relationships between different sets of multi-dimensional qualitative variables and categories of those variables are often used in medical data. There is a correlation relationship between three sets (personal and therapeutic and pathological variables) by drawing component loadings for three sets. First: variables regular blood pressure check, angina, reduce weight, reduce salt, heart attack, genetic factor, age were the most effective in the correlation relationships between sets of variables because they
are far from the origin. Second: variables exercise, education, disease duration were medium-affected. Third: variables gender, reduce smoking, job, kidney damage do not have any impact on relations between the sets since they are close to the origin.

There is a correlation relationship between three sets (personal, therapeutic and pathological variables) through the categories of those variables and centroid graphs for these categories shows the following

A: There is a correlation relationship between the following groups (L,W,F,AB) and the variables education: "secondary", kidney damage, "infected", "Manual", the systematic check of the pressure "when needed".

B: There is a correlation relationship between the following groups (O, R, U) and the variables heart attack "infected", disease duration "more than 10 years", genetic factor "no".

C: There is a correlation relationship between the following categories (Z, C) and the variables age "68-84", regular blood pressure check, "irregular".

D: There is a correlation relationship between the following categories (AJ, S) which belong to the variables angina "infected", exercise, "Yes".

Acknowledgements

We acknowledge the financial support from University Technologi Malaysia for the Research University Grant (Q.J130000 2526. 06H68) and the Ministry of Higher Education (MOHE) of Malaysia.

References

### Appendix 1

**Table 4** Represents object scores

<table>
<thead>
<tr>
<th></th>
<th>Dimension 1</th>
<th>Dimension 2</th>
<th>Dimension 1</th>
<th>Dimension 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-.611</td>
<td>-1.751</td>
<td>21</td>
<td>1.344</td>
</tr>
<tr>
<td>2</td>
<td>-.384</td>
<td>-.123</td>
<td>22</td>
<td>1.007</td>
</tr>
<tr>
<td>3</td>
<td>-1.605</td>
<td>.988</td>
<td>23</td>
<td>1.773</td>
</tr>
<tr>
<td>4</td>
<td>-1.188</td>
<td>1.298</td>
<td>24</td>
<td>.632</td>
</tr>
<tr>
<td>5</td>
<td>1.103</td>
<td>-.243</td>
<td>25</td>
<td>.915</td>
</tr>
<tr>
<td>6</td>
<td>.145</td>
<td>-1.725</td>
<td>26</td>
<td>1.846</td>
</tr>
<tr>
<td>7</td>
<td>-.823</td>
<td>-.999</td>
<td>27</td>
<td>-1.075</td>
</tr>
<tr>
<td>8</td>
<td>-.565</td>
<td>.403</td>
<td>28</td>
<td>.433</td>
</tr>
<tr>
<td>9</td>
<td>.558</td>
<td>.349</td>
<td>29</td>
<td>.327</td>
</tr>
<tr>
<td>10</td>
<td>.098</td>
<td>.167</td>
<td>30</td>
<td>1.138</td>
</tr>
<tr>
<td>11</td>
<td>-1.733</td>
<td>-1.357</td>
<td>31</td>
<td>.336</td>
</tr>
<tr>
<td>12</td>
<td>-.440</td>
<td>.524</td>
<td>32</td>
<td>-1.144</td>
</tr>
<tr>
<td>13</td>
<td>2.308</td>
<td>.015</td>
<td>33</td>
<td>-1.497</td>
</tr>
<tr>
<td>14</td>
<td>.566</td>
<td>1.189</td>
<td>34</td>
<td>-.710</td>
</tr>
<tr>
<td>15</td>
<td>.833</td>
<td>.450</td>
<td>35</td>
<td>-.666</td>
</tr>
<tr>
<td>16</td>
<td>-.778</td>
<td>.423</td>
<td>36</td>
<td>-.614</td>
</tr>
<tr>
<td>17</td>
<td>-.131</td>
<td>1.208</td>
<td>37</td>
<td>-1.321</td>
</tr>
<tr>
<td>18</td>
<td>-1.095</td>
<td>-1.765</td>
<td>38</td>
<td>-.330</td>
</tr>
<tr>
<td>19</td>
<td>.791</td>
<td>1.142</td>
<td>39</td>
<td>-.468</td>
</tr>
<tr>
<td>20</td>
<td>.433</td>
<td>1.200</td>
<td>40</td>
<td>.594</td>
</tr>
</tbody>
</table>

The table above represents object scores which is considered one of the outputs of the generalized nonlinear canonical correlation analysis and as shown in previous table has been reduced dimensions of the study to (two) with (forty) observations.
Appendix 2

Data Description

1- Set of Personal Variables

$X_1$: Age (1: 34-50, 2: 51-67, 3: 68-84)

$X_2$: Gender (1: Male, 2: Female)

$X_3$: Job (1: Manual, 2: Skill, 3: Professional, 4: Retired)

$X_4$: Education (1: Uneducated, 2: Primary, 3: Secondary, 4: University)

$X_5$: Genetic Factor (1: Yes, 2: No)

$X_6$: disease duration (period) (1: Less than 5 years, 2: 5-10 years, 3: (more than 10 years).

2- Set of pathological variables

$Y_1$: angina (1: infected, 2: uninfected)

$Y_2$: heart attack (1: infected, 2: uninfected)

$Y_3$: kidney damage (1: infected, 2: uninfected)

3- Set of therapeutic variables

$Z_1$: Regular blood pressure check (1: regular, 2: irregular, 3: When you need)

$Z_2$: reduce smoking (1: Yes, 2: No, 3: no smoke)

$Z_3$: reduce salt in food (1: Yes, 2: No)

$Z_4$: reducing weight (1: Yes, 2: No)

$Z_5$: exercise (1: Yes, 2: No).