Jurnal Teknologi

Full Paper

GENERALIZED NONLINEAR CANONICAL CORRELATION ANALYSIS WITH ORDERED CATEGORICAL AND DICHOTOMOUS DATA

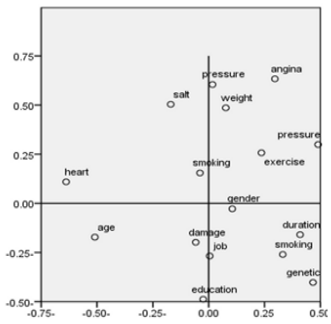
Thanoon Y. Thanoon[a,c], Robiah Adnan[a], Seyed Ehsan Saffari[b]

[a]Department of Mathematical Science, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia
[b]Education Development Centre, Sabzevar University of Medical Sciences, 9613873136-319, Sabzevar, Iran
[c]Foundation of Technical Education,Technical College of Management, Mosul, Iraq

*Corresponding author
robiaha@utm.my

Article history
Received
20 October 2014
Received in revised form
12 January 2015
Accepted
15 June 2015

Graphical abstract



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

Abstrak

Dalam kertas ini, data kategori dan data dikotomi digunakan dalam analisis korelasi berkanun taklinear teritlak untuk mengkaji hubungan antara lebih daripada dua set pembolehubah. Analisis statistik yang melibatkan analisis korelasi berkanun taklinear teritlak, komponen beban dan skor objek dibincangkan. Ilustrasi Prosedur yang dicadangkan ditunjuk melalui set data sebenar (pesakit tekanan darah tinggi). Analisis dilakukan dengan menggunakan program SPSS. Graf komponen beban daripada tiga set tersebut menunjukkan hubungan di antaranya dan juga kesannya terhadap data pesakit tekanan darah tinggi.Graf sentroid bagi set tersebut juga menunjukkan hubungan antara mereka.

Kata kunci: Tak linear analisis korelasi kanonik, umum analisis korelasi kanonik, data kategori, data sempit.

1.0 INTRODUCTION

Canonical correlation analysis (CCA) (see Fan & Konold,[1] Timm,[2] Vía, *et al*.,[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*.,[4] Van de Velden & Takane,[5] Van der Burg, *et al*.[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.[7]

Generalized nonlinear canonical correlation analysis corresponds to the categorical canonical correlation analysis with optimal scaling. OVERALS was developed by Gifi,[8] Van Rijckevorsek & de Leeus[9] 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.[10]

Standard canonical correlation analysis is an extension of multiple linear regression analysis, where instead of one it has several dependent variables.[11]

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,[6] Van der Lans,[11] Weenink,[12] and Yanai&Takane.[13]

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.[10]

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[14], Hsieh[15], Lai[16], Thompson[17], Thorndike.[18]

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 ith 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,…,n. The dimensionality of the solution k must be chosen by the researcher.

We can formulate as objective

$$\min \phi (Y, A_i) = \min \text{trace} \sum_{i=1}^{n} (Y - X_i A_i)'(Y - X_i A_i) \quad (1)$$

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

$$(\sum_{i=1}^{n} X_i (X_i' X_i)^{-1} X_i') Y = Y_{m \times k} \Lambda_{k \times k} \quad (2)$$

where Λ 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

$$A_i = (X_i' X_i)^{-1} X_i' Y \quad (3)$$

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.[5]

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:[19]

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.[11]

As noted, OVERALS maximizes the sum of correlations between columns of $n \times p$ comparison matrix and corresponding columns of $n \times p$ 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:

$$\sigma(X,Q,A)=K^{-1}\sum_{K} SSQ(X-Q_K A_K) \qquad (4)$$

K: is the number of sets.
X: is an $n \times p$ of comparison scores, where n is the number of objects and p is the dimensions.
Q: is an $n \times m$ partitioned matrix containing scores (to be estimated) for variables with m total number of variables.
Q_K: is an $n \times m_k$ matrix containing scaled variables within set k, where m_k is the number of variables in set k, thus $Q=(Q_1 / Q_2 \ldots / Q_K / \ldots Q_K)$.
A: is an $m \times p$ partitioned matrix of canonical weights.
A_k: is an $m_k \times p$ matrix of canonical weights of the variables in set k, thus $A=(A_1 / A_2 \ldots / A_K / \ldots A_K)$.
$SSQ()$: 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[20] 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 category code 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 minimum 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

Sets		Number of category	Variable type	Category symbol
Personal var.	Age	3	Ordinal	A B C
	Gender	2	Single Nominal	D E
	Job	4	Ordinal	F G H I
	Education	4	Ordinal	J K L
	Genetic factor	2	Single Nominal	M N O
	Disease duration (period)	3	Ordinal	P Q R
Pathological var.	Angina	2	Single Nominal	S T
	Heart attack	2	Single Nominal	U V
	Kidney damage	2	Single Nominal	W X Y
Therapeutic var.	Pressure Test	3	multiple Nominal	Z AB
	Reduce smoking	3	multiple Nominal	AC AD AE
	Reduce salt	2	Single Nominal	AF AG
	Reduce weight	2	Single Nominal	AH AI
	Exercise	2	Single Nominal	AJ AK

Table 2 A summary of the two dimensional analysis

Dimension				
		1	2	Sum
Loss	Set 1	.175	.330	0.505
	Set 2	.540	.496	1.036
	Set 3	.321	.305	.625
	Mean	.345	.377	.722
Eigenvalues		.623	.655	
Fit				1.278

Table 3 Component loading for three sets

				Dimension	
Sets	Variables			1	2
	Age			-.513	-.176
	Gender			.116	-.021
1	Job			.005	-.266
	Education			-.038	-.494
	Genetic factor			.468	-.405
	Disease duration (period)			.402	-.168
	Angina			.299	.627
2	Heart attack			-.629	.122
	Kidney damage			-.081	-.211
	Pressure	Dimension	1	.494	.295
	Test		2	.017	.603
3	Reduce	Dimension	1	.313	-.273
	smoking		2	-.049	.174
	Reduce salt			-.174	.499
	Reduce weight			.091	.491
	Exercise			.240	.256

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.

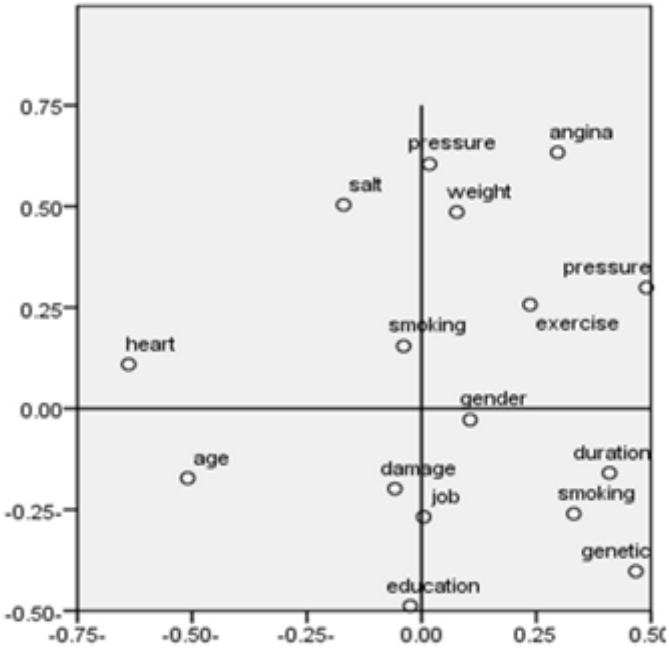


Figure 1 Component loadings for three sets

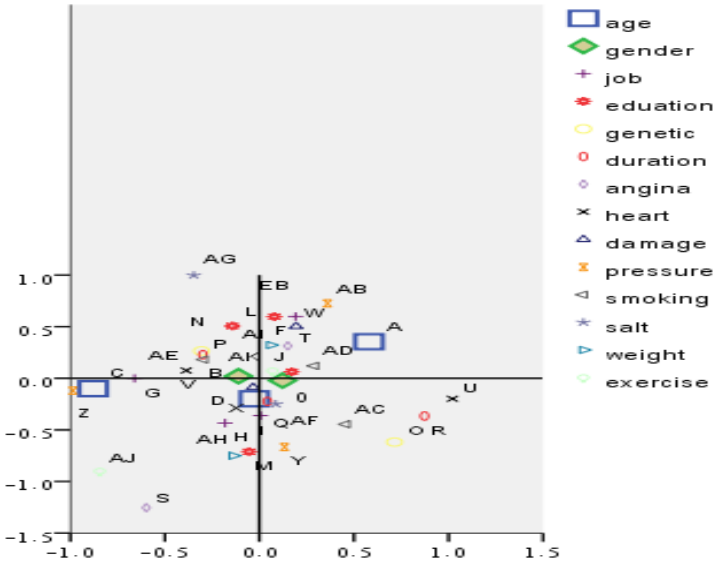


Figure 2 Centroids of all categories of variables for three sets

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

[1] Fan, Xitao and Timothy R Konold. 2010. The Reviewer's Guide to Quantitative Methods in the Social Sciences. 29-40.
[2] Albert. Gifi. 1990. *Nonlinear Multivariate Analysis*. Wiley Chichester.
[3] Hardoon, D. R., Szedmak, S. and Shawe-Taylor, J. 2004. *Neural Computation*. 16(12): 2639-2664.
[4] Hsieh, W. W. 2000. *Neural Networks*. 13(10): 1095-1105.
[5] Hsieh, William W. 2001. *Journal of Climate*. 14(1): 2528-2539.
[6] Kiers, Henk, A. L., Robert Cleroux and Jos MF Ten Berge. 1994. Computational Statistics & Data Analysis. 18(3) 331-340.
[7] Lai, P. L. and Fyfe, C. 2000. *International Journal of Neural Systems*. 10(05): 365-377.
[8] Jacqueline, J. Meulman and Willem J. Heiser. 2009. Pasw® Categories 18. Chicago: SPSS Inc.
[9] Takane, Yoshio and Yuriko Oshima-Takane. 2002. *Measurement and Multivariate Analysis*. 183-190.
[10] Bruce. Thompson. 2005. *Encyclopedia of Statistics in Behavioral Science*.
[11] Thorndike, R. M. 2000. *Applied Multivariate Statistics And Mathematical Modeling*. 237-263.
[12] Neil H. Timm. 2002. *Applied Multivariate Analysis*. Springer.
[13] Van de Velden, Michel and Yoshio Takane. 2011. *Computational Statistics*. 27(3): 551-571.
[14] Van der Burg, Eeke, Jan de Leeuw and Garmt Dijksterhuis. 1994. *Computational Statistics & Data Analysis*. 18(1): 141-163.
[15] Van der Lans, Ivo A. 1989. *Psychometrika*. 42: 207-219.
[16] Jan L. A. van Rijckevorsek and Jan de Leeus. 1988. *Component and Correspondence Analysis: Dimension Reduction by Functional Approximation*. John Wiley & Sons, Inc.
[17] Vía, Javier, Ignacio Santamaría and Jesús Pérez. 2007. *Neural Networks*. 20(1): 139-152.
[18] Weenink, David. 2003. In Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam. 25: 81-99.
[19] Yanai, Haruo and Yoshio Takane. 1992. *Linear Algebra and Its Applications*. 176: 75-89.
[20] Yazici, A, E Ögüs, Handan Ankarali and Fikret Gürbüz. 2010. *Turkish Journal Medical Science*. 40(1): 1-8.

Appendix 1

Table 4 Represents object scores

	Dimension			Dimension	
	1	2		1	2
1	-.611	-1.751	21	1.344	-1.143
2	-.384	-.123	22	1.007	.485
3	-1.605	.988	23	1.773	-1.741
4	-1.188	1.298	24	.632	1.307
5	1.103	-.243	25	.915	-.620
6	.145	-1.725	26	1.846	-.783
7	-.823	-.999	27	-1.075	-.352
8	-.565	.403	28	.433	1.200
9	.558	.349	29	.327	.825
10	.098	.167	30	1.138	.526
11	-1.733	-1.357	31	.336	.127
12	-.440	.524	32	-1.144	-.672
13	2.308	.015	33	-1.497	.244
14	.566	1.189	34	-.710	1.749
15	.833	.450	35	-.666	-.548
16	-.778	.423	36	-.614	.068
17	-.131	1.208	37	-1.321	.059
18	-1.095	-1.765	38	-.330	-1.911
19	.791	1.142	39	-.468	.703
20	.433	1.200	40	.594	-.916

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 than10 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).