Journal of Contemporary Applied Mathematics
V. 3, No 1, 2013, September
ISSN 2222-5498

The Application Of Correspondence Analysis To Logical Data Table

I. Akhisar

Abstract. Correspondence analysis is an exploratory technique designed to analyze contingence tables. All statistical techniques used a tabular representation of input data. Therefore, we can not apply it directly on a multidimensional representation. We need to convert data set to complete disjunctive table. In this study, disjunctive logic burt table obtained data based on the measurements on 220 individuals suffering skin cancer. Then correspondence analysis method applied to the data and the results obtained from the analysis are interpreted by cooperation with medical doctors.

Key Words and Phrases: correspondence Analysis, Disjunctive Logic Burt Table,Factorial Axes, Reconstitution Formula

2000 Mathematics Subject Classifications: 62Hxx, 62Pxx

1. Introduction and Statements of Results

Multiple correspondence analysis has been used as a method for the analysis of categorical data. It is possible to get inferences from the population on the basis of data [1].

A major difference between correspondence analysis and other techniques for categorical data analysis based on assumptions. In correspondence analysis, it is claimed that no model has to be hypothesized and no underlying distribution has to be assumed, but the decomposition of data can be obtained in order to study their structure [2].

let us consider a two-way contingency table of order $I \times J$ where I and J are two sets with finite number of elements as $I = \{1, 2, ..., i, ..., m\}$, $J = \{1, 2, ..., j, ..., n\}$. If $\forall (i, j) \in I \times J$ its general term $k(i, j)$ is the number of the repetition of the event (i,j), then I and J are in relation.

The set $K_{IJ} = \{k(i, j) |\ i \in I,\ j \in J\}$ is called the correspondence table and determines a statistical correspondence between the sets I and J [3-4].

By the structure of a table, we mean the relationship between the rows and the columns as expressed in correspondence analysis through graphical displays(Benzecri, 1973; Escofier and Pages, 1988-1998; Lebart et all 1998)[5-7].

 © *2011 JCAM* All rights reserved.

The comparison has to deal with both rows and columns, that is to say, the structure induced from the rows over the different set of columns, and also the structure induced over the rows by the different sets of columns.

Defining $k(i), k(j)$ and k as

$$k(i) = \sum \{k(i,j)|\ j \in J\}, k(j) = \sum \{k(i,j)|\ i \in I\} \text{ and}$$

$$k = \sum \{k(i,j)|\ i \in I,\ j \in J\}$$

respectively then we can obtain the frequency table F_{IJ} from the correspondence table K_{IJ} as follows

$$F_{IJ} = \{f_{ij}|\ i \in I,\ j \in J\};\ f_{ij} = k(i,j)/k$$

dividing every $\{f_{ij} \mid\ i \in I\}$ by the mass f_j which is the sum of j^th column, it yields

$$Z = F_I^J = \{f_I^1, f_I^2, ..., f_I^j, ..., f_I^n\}$$

where $f_I^j = f_{ij}/f_j$ In the same manner the table $T = F_J^I$ table can be obtained from F_{IJ}. On the other hand the marginal probability laws;

$$f_I = \{f_i \mid\ i \in I\};\quad f_i = \sum \{f_{ij} \mid\ j \in J\}$$

$$f_J = \{f_j \mid\ j \in I\};\quad f_j = \sum \{f_{ij} \mid\ i \in I\}$$

can also be obtained from the table F_{IJ}.

2. Burt Table

Multiple correspondence analysis(MCA) may be considered as an extension of simple correspondence analysis. Actually, we usually analyze the inner product of such matrix called Burt Table in MCA [8] . Burt table obtained from the logical table coded 0's and 1's for the answers given to some questions.

Let J_q be the answer given to the q th question and let $J = \cup\{J_q \mid\ q \in Q\}$. On the other hand we can also convert each quantitative variables to a qualitative variable by dividing variation intervals to consecutive intervals [9].

If the number of modality belonging to block of J_q is $Card(J_q)$ and $m = \sum\limits_{q \in Q} Card(J_q)$ then define f_{ij}, f_i and f_j as follows

$$\forall i\ ;\ f_{ij} = \frac{k(i,j)}{m \cdot n}\ \{k_{ij} = 0\ or\ 1\ ;\ j \in [1,m]\}$$

$$f_i = \frac{1}{n}\ (n = CardI)\ ,\ f_j = \frac{\sum\limits_i k_{ij}}{m \cdot n}$$

Using a disjoint and complete logic table with general term k_{ij}, we may define a square Burt Table with general term $\overline{k}_{jj'}$, given by

$$\overline{k}_{jj'} = Card\left\{i \mid \ i \in I \ , \ k(i,j) = k(i,j') = 1\right\}$$

The factors obtained from analyzing correspondence table are a constant multiple of the factors obtained from corresponding Burt table.

It is clear that a symmetric Burt table can be written as follows

$$\overline{k}_{jj'} = \sum \left\{k(i,j) \cdot k(i,j') \mid \ i \in I\right\} = \left(\sum \left\{k_{ij} \cdot k_{ij'}.\left(\frac{1}{k_i}\right) \mid \ i \in I\right\}\right).CardQ$$

3. The Metric Related to Table

A measure defined on a set I, denoted by $\mu_I = \{\mu_i | i \in I\}$, is a vector with n components, the set of all measures represented by R_I.

A function on a set I is denoted by $f^I = \left\{f^i | \ i \in I\right\}$ form a vector space of dimension m, represented by R^I.

We can define a metric m^{II} called as χ^2 metric with center f_i in the space R_I by the quadratic form

$$m^{II} = \{m^{ii'} \mid \ i, i' \in I\} \ ; \ \ m^{ii'} = \frac{\delta^i_{i'}}{f_i}$$

the matrix M which represents the above quadratic form defines an isometry of R_I on R^I [5]. If u^α_I is the unit vector in the space R_I according to the metric m^{II} with matrix the projection of f^j_I on the axis $\Delta(u^\alpha_I)$ represented by $G_\alpha(j)$ and defined in the tensorial form by

$$G_\alpha(j) = M(u^\alpha_I, f^j_I) = M(u^\alpha_I)(f^j_I) = \varphi^I_\alpha(f^j_I)$$

where the measure f^j_I and the function φ^I_α which is orthogonal projection on the axes $\Delta(u^\alpha_I)$ [10].

On the other, $\tau_{ii'}$ is a covariance between y^I_i and $y^J_{i'}$ elements. D_{JJ} is diagonal weights matrix and also considered a quadratic form an R_J.

$$\tau_{ii'} = \sum \{f_i y^j_i y^j_{i'} \ \mid \ j \in J\} = D(y^J_i, y^J_{i'}) = YDY'$$

and we can write $V = TDT'$, similarly.

4. The Method of Correspondence Analysis

Correspondence analysis is a exploratory method that is designed to analyze two way contingency tables containing measure between the rows and columns.

Let us consider variable set consisted of n points in R_J obtained from the data table. The metric defined it on m dimensional space R_I, is χ^2 metric with f_I center. We can not represent the set for $m > 3$, our aim is to find the best representation of the set in less

dimensional variety of R_I for $k < m$. For this reason, assigning weights to j-th variables, then we can define a quadratic form of moment of inertia [12].

It can be shown that the space H_I with k dimension is the subspace consisting the lines $\{\Delta(u_I^\alpha) \mid \ \alpha \in [1,k]\}$ M orthogonal among them [8]. These lines (directions) are the eigendirections corresponding to the largest eigenvalues of the transformation VM.

The matrices M and V are isometries of the space R_I on to R^I and R^I on to R_I respectively so the transformation VM determines an isometry of R_I on to itself.

In the space R_I by the numbers λ_α satisfying the conditions $VM(u_I^\alpha) = \lambda_\alpha u_I^\alpha$ we can define a base u_I^α M orthogonal among them.

The subspace H_I generated by the eigenvectors u_I^α corresponds the largest eigenvalues $\lambda_1, \lambda_2, ..., \lambda_k$, if $\lambda_i = \lambda_{i+1}$, it is enough to choose any three vectors which are orthogonal among them [10].

The eigendirections u_I^α ($\alpha \in [1,\ k]$) are called factorial axes, the vectors φ_α^I satisfying

$$M(f_I^J, u_I^\alpha) = M(u_I^\alpha)(f_I^j) = \varphi_\alpha^I(f_I^j) = \varphi_\alpha^I \circ f_I^j$$

is said to be α th factor. It is clear that φ_α^I ($\alpha \in A$) are the eigendirections of the transformation MV since $MV\left(\varphi_\alpha^I\right) = MVM\left(u_I^\alpha\right) = M\left(\lambda_\alpha u_I^\alpha\right) = \lambda_\alpha \varphi_\alpha^I$

5. Reconstitution of the F_{IJ} Table

If $\{u_I^\alpha |\ \alpha \in A\}$ is an orthonormal an axis system, then it can be written that

$$f_I^j - f_I = \sum \left\{ M\left(f_I^j - f_I\ ,\ u_I^\alpha\right).u_I^\alpha \mid \ \alpha \in A \right\}$$

dividing the expression f_i and using $\varphi_\alpha^i = u_i^\alpha / f_i$ we obtain that

$$\frac{f_{ij} - f_i f_j}{f_i \cdot f_j} = \sum \left\{ \lambda^{1/2} \varphi_\alpha^j \varphi_\alpha^i \mid \ \alpha \in \quad A \right\}$$

[3]. hence the reconstitution formula is obtained as

$$f_{ij} = f_i f_j \left\{ 1 + \sum \left\{ \lambda^{1/2} \varphi_\alpha^j \varphi_\alpha^i | \ \alpha \in \ A \right\} \right\}$$

the best approximation of the k th order is obtained by choosing the largest k eigenvalues. We could find a lower-dimensional space, in which retains all or almost all of the information about the differences between the rows. Then we could present all information about the similarities between the rows in a simple one, two, or more m dimensional graph.

For the zero order approximation $f_{ij} = f_i \cdot f_j$ which means that i and j are independent.

6. Application

It is well known that epidemiological studies include many variables. Within the framework of a study requested by an state hospital, the most recent complete data table

to be analyzed is based on measurements made on 220 sick individuals. The aim of tis study is to identify objective and subjective factors such as physical, environmental and interior associated with cancer causes. To try to explain the differences between health and sick individuals add the combination of factors.

The notations used for variables are listed, in appendix A, and the same variables have been grouped.

The variables from 1 to 7 refer to the physical properties, from 8 to 12 refer to environment factors and 17, 19, 21, 22, 23 and 27 refer to endocrine factors of individuals. Combination of the variables mentioned above are made up of the variables 13, 14, 15 and 16 that are affected by the variables 18, 24, 25 and 26.

This study includes 27 variables having different modalities that physical, environment and interior factors since birthmark and sun birth are both transferred by genetic and occur during the lifetime.

To obtain logical table some of the (suitable) variables are separated to groups like age intervals and gender etc.

In Appendix B, the list of the visual variables are highly correlated with hidden variables; could not be represented in graphical plane . Moreover, the graphical plane in Appendix C is examined some variables are observed to be related to each other in groups.

The contributions of the investigated parameters on the principal axes, it is also seen (Appendix C) that the first dimension include the variables FOT1, FOT2, MEL1, MEL2 and MEL3 are in the same group. We obtained the correlations (FOT1-MEL1)=.95, (FOT1-MEL2)=.99 and (FOT2-MEL3)=1.00 supporting the mention variables to be in the same group.

At the second dimension include the variables FOT3, FOT4, MEL4, MEL5 and MEL6 are seen to be in the same group. The relations (correlations)among them are given by (FOT3-MEL4)=.90, (FOT3-MEL5)=1.00 and (FOT4-MEL6)=1.00, obtained results supported by medically. On the other hand, the qualities (high moments) of these variables are as follows: FOT1=.870, FOT2=.810, FOT3=.763, FOT4=.903, MEL5=.759 and MEL6=.908.

The variables EBRO=.745, FS1=.607, FOT4=.867 and FOT3=.536 express the first axis. On the other hand FSUN, ISUN, FS1, FSB1, ISU1 and ISN1 are related to the first group variables which can be obtained from correlations and principal components.

In addition, the variables PROT is related only the variables FOT4 and MEL6 (FOT4-PROT)=.99, (MEL6-PROT)=.99 in the second group, i.e. dark skinned people. On the other hand LNA3, LNC3, LNL3 and LNB3 are seen to be in the first group.

A visualization of the results is presented in Appendix C. Also, it may be observed from correlations and principal components that relationship between EXPO and LENT is in the level of 88%. PREG and ORAL variables are together with NEVU's, however CNDP is on the other part of the axis. An other interesting result is that the variables FMEL and IMAL have been seen close to each other.

We would like to underline that the properties representing light skinned and dark skinned variables are grouped together among themselves.

7. Conclusion

The Burt table framework opens up the way for various extensions. A particulary interesting one is the possibility of comparing the structure of sub tables of different natures in the same analysis. Thus, it is possible to add various groups of quantitative and/or qualitative variables describing the same rows to set of contingency tables. The example presented in this study emphasize the possibilities together constitute a complete methodology for exploratory analysis of set of individuals, or groups of individuals, described by data of different types.

The influence of the various sub tables in a global analysis by way of a solution that is already well tested in the case of burt tables individuals variables(quantitative and/or qualitative). This solution is now extended to the case of contingency tables having different row margins. The methodology described provides an operational point of view for the analysis of several contingency tables having a common dimension.

Finally, interpreting the results from the medical doctors perspective that could find inherent relations between the variables and policies in effective way.

Appendices

A: Variables

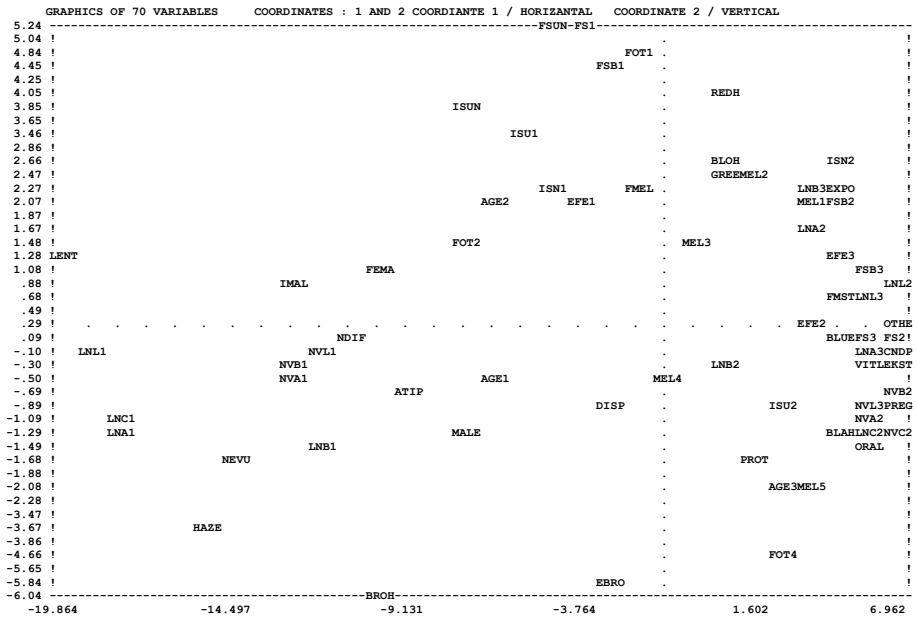
1-) AGE1 (0-24), AGE2 (25-49), AGE3 (50 $\geq$): Age
2-) MALE: Man, FEMA: Woman: Sex
3-) REDH: Red, BLOH: Blonde, BROH: Brown, BLAH: Black ; Hair color
4-) BLUE: Blue, GREE: Green, HAZE: Hazel, BLAE: Black: Eye color
5-) EFE1 - EFE4: Freckle(From weak to strong)
6-) FOT1 - FOT4: Skin color and reaction to the sun (From weak to strong)
7-) MEL1 - MEL6: Skin color (From light to dark)
8-) PROT: Protection from the sun completely
9-) EBRO: Easy bronze
10-) FSUN: First degree sunburnt,
FS1 - FS3 : $Age \leq 10$ and FSB1 - FSB3: $Age > 10$
11-) ISUN: Intensive sunburnt
ISU1 - ISU3: $Age > 10$ and ISN1 - ISN3: $Age \leq 10$,
12-) EXPO: Exposing to the sun
13-) NEVU: Birthmark
NVA1 - NVA3: Birthmark on arms, NVL1 - NVL3: Birthmark on legs,
NVC1 - NVC3: Birthmark on breast, NVB1 - NVB3: Birthmark on back
(From less to many)
14-) LENT: Sunburnt
LNA1 - LNA3: Sunburnt on arms, LNL1 - LNL3: Sunburnt on legs,
LNC1 - LNC3: Sunburnt on breast, LNB1 - LNB3: Sunburnt on back
(From less to many)

15-) ATIP: Malicious birthmark (malignant)
16-) DISP: Displease birthmark
17-) NDIF: Tendency of (a lot of) Nevus development in the family
18-) IMAL: Malignant in the internal organ
19-) FMEL: Malicious in the family
20-) EKST: Extirpation
21-) PREG: Pregnancy
22-) ORAL: Oral contraception
23-) CNDP: Changing of the birthmarks during the pregnancy
24-) VITL: Ilness connected with losing color on the skin
25-) HALO: Losing color around birthmark
26-) OILN: Other ilnesses
27-) FMST: Melanom story in the family

B: Visual and Hidden Variables

Visual	*Unvisual*	*	*Visual*	*Unvisual*	*	*Visual*	*Unvisual*
$LNL2$	$EFE4$	*	$EKST$	$NVA3$	*	$NVC2$	$ISN3$
$FS3$	$HALO$	*	$NVA1$	$NVC1$	*	$AGE3$	$FOT3$
$FS2$	$BLAE$	*	$PREG$	$NVC3$	*	$FOT4$	$MEL6$
$FS2$	$LNC3$	*	$NVA2$	$NVL2$	*		
$EKST$	$NVB3$	*	$NVC2$	$ISU3$	*		

APPENDIX C



References

[1] GREENACRE, M.J., Multiple and Joint Correspondence Analysis. In Correspondence Analysis in the Social Science-Recent Developments and Appl. Academic Press (Paris), (1994.)

[2] DEMOSTHENES, B.P., and CHRISTOS, P., Interpretation of Epidemiological Data Using Multiple Correspondence Analysis and Log-Linear Models Journal of Data Science 2 75-86 (2004).

[3] ESCOFIER-CORDIER, B., L'Analyse des Correspondences,These Universite de Rennes (Paris), (1965)

[4] BENZECRI, J.P., Sur L'Analyse des Tableaux Binaires Associes a une Correspondance Multiple Universite Pierre et Marie Curie (Paris),(1972).

[5] BENZECRI, J.P., Analyse des Donees, vol. 1: Analyse des Correspondances, Dunod,(Paris),(1973).

[6] ESCOFIER, B., PAGES, J., Analyses Factorielles Simples et Multiples; Objectifs, Methodes et Interpretation, Dunod (Paris),(1988)

[7] LEBART, L., MORINEAU, A., and PIRON, M., Statistique Exploratoire Multidimensionnelle, Dunod (Paris),(1998).

[8] CAZES, P., L'Analyse de Certain Tableaux Rectangulaires Decompose en Blocks:Generalisation des Proprietes Recontrees dans L'Etude des Correspondence Multiples, Les Cahiers de L'Analyse des Donnees 5: 145-161, (1980)

[9] BURT, C., The Factorial Analysis of Qualitative Data British Journal of Statistical Psychology Volume III 3, 166-185 (1955)

[10] CAZES, P., Trait. des Prob.s Geologiques Universite Paris IV (Paris), (1970)

[11] BENER, A., Etude par L'Analyse des Correspondence des Interactions dans un Tableaux Ternaire Application a des Donness Linguistiques, These Universite Pierre et Marie Curie (Paris), (1981)

[12] STEPHANE, D., DANIEL, C., ana JEAN, T., Co-Inertia Analysis and the Linking of Ecological Data Tables Ecology 84 (11) 3078-3089, (2003)

Ilyas Akhisar
Marmara University, Banking and Insurance School, Goztepe Campus, Istanbul, Turkey
E-mail: akhisar@marmara.edu.tr

Received 18 March 2013
Accepted 29 August 2013