Research Article

Exploring the Political Situation in Nigeria (2011) elections using Correspondence Analysis for Categorical Data

Anyiam Kizito E[1], Nworuh G E[2], Uchegbulem-Asiegbu Nwaoma P[3]

[1,2]Department of Information Management Technology, Federal University of Technology, Owerri (FUTO), NIGERIA.
[3]Department of Statistics, Imo State Polytechnic, Umuagwo, Owerri, NIGERIA.
Email: kizitondon@yahoo.com

Abstract Correspondence analysis is an exploratory technique related to to principal components analysis which finds a multidimensional representation of the association between the row and column categories of a two-way contingency table. This technique finds scores for the row and column categories on a small number of dimensions which account for the greatest proportion of the chi² for association between the row and column categories, just as principal components account for maximum variance. Empirical investigation is carried out using discrete data collected categorically. Graphical display of two or three dimensions are typically used to give a reduced rank approximation to the data.
Keywords: Discete Categorical data, Chi-square distances, correspondence matrix, profiles, masses and centroids

***Address for Correspondence:**
Dr. Anyiam Kizito E, Department of Information Management Technology, Federal University of Technology, Owerri (FUTO), NIGERIA.
Email: kizitondon@yahoo.com
Received Date: 01/10/2014 Accepted Date: 10/10/2014

Access this article online	
Quick Response Code:	Website: www.statperson.com
	DOI: 27 October 2014

INTRODUCTION

Correspondence analysis (CA) has proved to be very popular in research areas where large sets of discrete categorical data are collected, in particular linguistics (Haassal and ganesh,1996, Romney et al., 1997 and Greenacre,2007), ecology (Hoffman and Franke, 1986), archeology, marketing research, the social sciences (Terr Break, 1985, Fellenberg et al., 2001), healthcare and nursing studies (Javalgi et al., 1992, Watts, 1997), environmental management (Kishino et al.,1998) and genomics. Technically, correspondence analysis falls into the class of classical multivariate statistical methods of dimensions reduction based on the singular value decomposition. One of the benefits of CA is that it can simplify complex data from a potentially large table into a simpler display of categorical variables while preserving all of the valuable information in the data set. This is especially valuable when it would be inappropriate to use a table to display the data because the associations between variables would not be apparent due to the size of the table. The aim of this paper has been to discuss new developments of correspondence analysis for the application to discrete categorical two-way contingency tables. However, due to the nature of this procedure, only a visualization of the association between the categories of the nonordered variable can be made.

METHODOLOGY

Suppose a categorical data are collected in a two way $(r\,x\,s)$ contingency table N with r rows $(labelled\,A_1, A_2, \dots, A_r)$ and s columns $(labelled\,B_1, B_2, \dots, B_s)$ resulting in rs cells. The $ijth$ cell has entry n_{ij}, representing the observed frequency in row category A_i and column category $B_j, i = 1, 2, \dots, r,\ j = 1, 2, \dots, s.$ Then the marginal row total is $n_{i+} = \sum_{j=i}^{s} n_{ij}, i = 1, 2, \dots, r$ and the jth marginal column total is $n_{+j} = \sum_{i=1}^{r} n_{ij}, j = 1, 2, \dots, s$. If the individual row and column categories are classified $n = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}$, then the resulting table showing the cell frequencies, marginal total and total sample size is called a correspondence table.

Table 1: Two-way contingency table, showing observed cell frequencies, row and column marginal totals, and total sample size

	column Variable						
Row Variable	B_1	B_2	$\dots$	B_j	$\dots$	B_s	*Row Total*
A_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1s}	n_{1+}
A_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2s}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
A_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{is}	n_{i+}
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\dots$	n_{rj}	$\dots$	n_{rs}	n_{r+}
Column total	n_{+1}	n_{+2}	$\dots$	n_{+j}	$\dots$	n_{+s}	n

Suppose also that we denote by Π_{ij} the probability that an individual has the properties A_i and $B_j, i = 1, 2, \dots, r, j = 1, 2, \dots, s$. In the event of row variable A is independent of the column variable B, we have that $\Pi_{ij} = \Pi_{i1}\Pi_{1j}$ where $n_{i+} = \sum_j \Pi_{ij}$ and $\Pi_{+j} = \sum_i \Pi_{ij}$ for all $i = 1, 2, \dots, r$ and $j = 1, 2, \dots, s$.

This paper is generally interested assessing whether A and B are actually independent variables or to investigate the homogeneity of the row and column probability distributions; that is whether all rows have the same probability distributions across columns or conversely, whether all the columns have the same probability distributions across rows.

Row and Column dummy variables.

For the two way contingency table above, we are interested in the relationship between the row categories and column categories. We defines two sets of dummy variates, an r-vector $X_i = (X_{ij})$ to indicate which of the n observations fall into the ith row, and the s- vector $Y_j = (Y_{ij})$ to indicate which of the n observations fall into the jth columns; that is, indicator vectors.

$$X_{ij} = \begin{cases} 1, if\ the\ jth\ individaul\ belongs\ to\ A_i \\ 0, otherwise \end{cases}$$

$$Y_{ij} = \begin{cases} 1, if\ the\ jth\ individual\ belongs\ to\ B_j \\ 0, otherwise \end{cases}$$

$$i = 1, 2, \dots, r,\ j = 1, 2, \dots, s$$

These two indicator vectors are represented into two matrices, X_{rxn} and Y_{sxn}, and are given by

$$\underset{rxn}{X} = \begin{pmatrix} 1 & \dots & 1 & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & 1 & \dots & 1 \end{pmatrix} \qquad 2.1$$

$$\underset{sxn}{Y} = \begin{pmatrix} 1 & \dots & 1 & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & 1 & \dots & 1 \end{pmatrix} \qquad 2.2$$

respectively. The two derived matrices X and Y reproduce the observed cell frequencies and their marginal totals. The (rxs)- matrix, XY reproduces the observed cell frequencies of the contingency table

$$XY = \begin{pmatrix} n_{11} & n_{12} & \dots & n_{1s} \\ n_{21} & n_{22} & \dots & n_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ n_{r1} & n_{r2} & \dots & n_{rs} \end{pmatrix} = N \qquad 2.3$$

The matrix XX^T and YY^T are both diagonal having as diagonal entries the r marginal row totals and the s marginal column totals respectively.

$$XX^T = diag\left(n_{1+}, n_{2+}, \dots, n_{r+}\right) \qquad 2.4$$

$$XY^T = diag\left(n_{+1}, n_{+2}, \dots, n_{+s}\right) \qquad 2.5$$

Collecting (2.3), (2.4) and (2.5) together, we can form the $(r+s)x(r+s)$ block matrix called the **Brut Matrix** for the contingency table

$$\begin{pmatrix} x \\ y \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}^T = \begin{pmatrix} nD_r & N \\ N & nD_c \end{pmatrix} \qquad 2.6$$

Where,

$$D_r = n^{-1}xx^T \qquad 2.7$$

$$D_c = n^{-1}yy^T \qquad 2.8$$

The Brut matrix (2.6) above is non-negative symmetric.

Profiles, Masses and Centroids

Using the (rxs)-matrix

$$P = n^{-1}XY^T = n^{-1}N \qquad 2.9$$

The correspondence table N is converted to correspondence matrix

Table 2: Correspondence matrix, showing observed cell relative frequencies $P\left(p_{ij} = n_{ij}/n\right)$ row marginal totals $r\left(p_{i+} = n_{i+}/n\right)$ and column marginal totals $C^T\left(p_{+j} = n_{+j}/n\right)$

	column Variable						
Row Variable	B_1	B_2	$\dots$	B_j	$\dots$	B_s	Row Total
A_1	p_{11}	p_{12}	$\dots$	p_{1j}	$\dots$	p_{1s}	p_{1+}
A_2	p_{21}	p_{22}	$\dots$	p_{2j}	$\dots$	p_{2s}	p_{2+}
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
A_i	p_{i1}	p_{i2}	$\dots$	p_{ij}	$\dots$	p_{is}	p_{i+}
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
A_r	p_{r1}	p_{r2}	$\dots$	p_{rj}	$\dots$	p_{rs}	p_{r+}
Column total	p_{+1}	p_{+2}	$\dots$	p_{+j}	$\dots$	p_{+s}	1

Where $P = \left(p_{ij} = n_{ij}/n\right)$ are the cell relative frequencies, $r = \left(p_{i+} = n_{i+}/n\right)$ are row marginal totals and column marginal totals $c^T = \left(p_{+j} = n_{+j}/n\right)$

The matrix P can be characterized as either the Uniformly Minimum Variance Unbiased (UMVU) estimator or the Minimum likelihood (ML) estimator of Π_{ij}. The (rxs) matrix P_r of row profiles of N (or P) consists of the rows of N divided by their appropriate row totals $(example, n_{ij}/n_{i+})$ which under random sampling, can be characterized as either the UMVU or ML estimator of Π_{ij}/Π_{i+}, the conditional probability that an individual has property B_j given that he or she has property A_i, and can be completed as the regression coefficient matrix of y on x; i. e.

$$P_r = \left(XX^T\right)^{-1}XY^T = D_r^{-1}P = \begin{bmatrix} a_1^T \\ \vdots \\ a_r^T \end{bmatrix} \qquad 2.10$$

Where $a_i^T = \left(\frac{n_{i1}}{n_{i+}}, \quad \dots, \quad \frac{n_{is}}{n_{i+}}\right)$

Is the ith row profiles, under random sampling, can be characterized as UMVU or ML estimator of Π_{ij}/Π_{+j}, the conditional probability that an individual has probability A_i given that he or she has probability B_j, and computed as the regression coefficient matrix of x and y, that is

$$P_c = (Y'Y)^{-1} YX^{-1} = D_c^{-1} P^T = \begin{bmatrix} b_1^T \\ \vdots \\ b_s^T \end{bmatrix} \qquad 2.11$$

Where $b_j^T = \left(\frac{n_{ij}}{n_{+j}} \quad \ldots, \quad \frac{n_{rj}}{n_{+j}} \right)$

Is the jth column profile, $j = 1, 2, \ldots, s$

Then the row and column means of N are the row and column sums of P given by

$$PI_s = \begin{bmatrix} P_{1+} \\ \vdots \\ P_{r+} \end{bmatrix} = r \qquad 2.13$$

$$PI_r = \begin{bmatrix} P_{+1} \\ \vdots \\ P_{+s} \end{bmatrix} = c \qquad 2.14$$

Hence it can be shown that

$$r = P^T D_c^{-1} c \quad and \quad c = P^T D_r^{-1} r$$

Here the ith element $P_{i+} = \frac{n_{i+}}{n}$, of the r-vector r is called the **Row Mass** and is the estimate of the unconditional probability, Π_{i+}, of belonging to A_i. Similarly, the jth element, $P_{+j} = \frac{n_{+j}}{n}$ of the s-vector c is called the **Column Mass** and is an estimate of the unconditional probabilities Π_{+j}, of belonging to B_j. Moreso, r and c are also regarded as the average row profile and average column profile respectively of the contingency table.

Distances

Within Variable Distances

In correspondence analysis, it is pertinent to determine the distances between different row profiles or column profiles. To achieve this we use the Chi-square metric as a measure of distance (see. Greenacre and Hastie, 1987). Consider the ith row profiles. The within variable square Euclidean distance of the profile from a_i and a_i' is

$d^2(a_i, a_i') \equiv \sum_j^s \frac{n}{n_{+j}} \left(\frac{n_{ij}}{n_{i+}} - \frac{n_{i'j}}{n_{i'+}} \right)^2$ and the Chi-squared distance between a and c row centriod and summing over all row profiles yields

$$n \sum_{i=1}^r p_{i+} d^2(a_i, c) = \sum\sum \left(n_{ij} - \frac{n_{i+} n_{+j}}{n} \right)^2 \Big/ \left(\frac{n_{i+} n_{+j}}{n} \right) \qquad 2.15$$

Which is the Pearson Chi-squared statistic

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad 2.16$$

With $O_{ij} = n_{ij}$ the observed frequencies and $E_{ij} = \frac{n_{i+} n_{+j}}{n}$ the expected cell frequencies.

Consequently, the row profile coordinate close to the centroid supports the hypothesis of independence, while those situated far from the origin support its rejection. It can be shown that in a similar manner that squared Euclidean distance of the jth column profile from the centroid is

$$n \sum_{i=1}^r p_{+j} d^2(b_j, r) = \sum_{i=1}^r \sum_{j=1}^s \left(n_{ij} - \frac{n_{+j} n_{i+}}{n} \right)^2 \Big/ \left(\frac{n_{+j} n_{i+}}{n} \right) = \chi^2 \qquad 2.17$$

Where χ^2 is given by 2.16.

3. EMPIRICAL EXAMPLE AND RESULTS

A two –way contingency table N with r = 7 and s = 4 was adopted from the work of Haruna et al (2011). It relates to data on political zones and votes by major political parties of a sample of 35906694 valid voters scored by four major political parties in the six geo-political zones in Nigeria and FCT in the 2011 presidential election. It is given as a (7x4)-matrix by

$$N = XY^T = \begin{pmatrix} 4985226 & 20335 & 25507 & 20537 \\ 2836417 & 321609 & 1369943 & 30456 \\ 6014283 & 49691 & 143060 & 10819 \\ 2428350 & 1503319 & 83677 & 31583 \\ 1381297 & 3402933 & 55529 & 151956 \\ 3466924 & 6453437 & 120043 & 609246 \\ 253444 & 131576 & 2327 & 3170 \end{pmatrix}$$

The matrices XX^T and YY^T are given by:

$$XX^T = \begin{pmatrix} 5051605 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4558425 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 6217853 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4046929 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4991715 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1064965 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 390517 \end{pmatrix}$$

$$YY^T = \begin{pmatrix} 21365941 & 0 & 0 & 0 \\ 0 & 11882900 & 0 & 0 \\ 0 & 0 & 1800086 & 0 \\ 0 & 0 & 0 & 857767 \end{pmatrix}$$

Respectively, the matrices

D_r and D_c are obtained bydividing both XX^T and YY^T

by n=35906694.

$$D_r = \begin{pmatrix} 0.14068 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.12695 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.17316 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.11271 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.13902 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.02966 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.01086 \end{pmatrix}$$

$$D_c = \begin{pmatrix} 0.59504 & 0 & 0 & 0 \\ 0 & 0.33093 & 0 & 0 \\ 0 & 0 & 0.05013 & 0 \\ 0 & 0 & 0 & 0.02389 \end{pmatrix}$$

To convert the contingency table N into the correspondence matrix, then the (rxs)-matrix

$$P = n^{-1}N$$

$$= \begin{pmatrix} 0.13883 & 0.00061 & 0.00071 & 0.00057 \\ 0.07899 & 0.00895 & 0.03815 & 0.00084 \\ 0.16749 & 0.00138 & 0.00398 & 0.00031 \\ 0.06763 & 0.04187 & 0.00233 & 0.00088 \\ 0.03846 & 0.09477 & 0.00155 & 0.00423 \\ 0.09655 & 0.17972 & 0.00334 & 0.01697 \\ 0.00706 & 0.00366 & 0.00006 & 0.00009 \end{pmatrix}$$

The row and column profiles with their individual masses $D_r = diag\{r\}$ and $D_c = diag\{c\}$ are

Table 3: Row Profiles

Political Zones	Votes By Political Parties				
	PDP	CPC	ACN	ANPP	Active Margin
SOUTH EAST	.987	.004	.005	.004	1.000
SOUTH WEST	.622	.071	.301	.007	1.000
SOUTH SOUTH	.967	.008	.023	.002	1.000
NORTH CENTRAL	.600	.371	.021	.008	1.000

NORTH EAST	.277	.682	.011	.030	1.000
NORTH WEST	.326	.606	.011	.057	1.000
FCT	.649	.337	.006	.008	1.000
Mass	.595	.331	.050	.024	

Table 4: Column Profiles

Political Zones	Votes By Political Parties				
	PDP	CPC	ACN	ANPP	Mass
SOUTH EAST	.233	.002	.014	.024	.141
SOUTH WEST	.133	.027	.761	.036	.127
SOUTH SOUTH	.281	.004	.079	.013	.173
NORTH CENTRAL	.114	.127	.046	.037	.113
NORTH EAST	.065	.286	.031	.177	.139
NORTH WEST	.162	.543	.067	.710	.297
FCT	.012	.011	.001	.004	.011
Active Margin	1.000	1.000	1.000	1.000	

Which is the unconditional estimate of the probability Π_{i+}, of belonging to A_i and Π_{+j}, of belonging to B_j respectively.

Table 4 : Summary

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation
								2
1	.642	.412			.701	.701	.000	.158
2	.413	.170			.289	.990	.000	
3	.076	.006			.010	1.000		
Total		.588	21117658.794	.000[a]	1.000	1.000		

a. 18 degrees of freedom

From Table 4 above, the Pearson Chi-squared statistic of 21117658.794, and with a zero p-value, it is highly statistically significant. Therefore, with a total inertia of 0.588 which is the measure of variation in the N, there is a significant association between the votes scored by the major political parties and their political zones.

When correspondence analysis is applied as in this study, the squared singular values are $\lambda_1^2 = 0.4122, \quad \lambda_2^2 = 0.1706 \quad and \quad \lambda_3^2 = 0.0058$ and the two dimensional correspondence plot is given in Figure 1. There the first principal axis accounts for $0.4122/0.588*100 = 70.102\%$ of the two variables, and the second axis accounts for 29.014%. Therefore, the two dimensional plot of Figure 1 graphically depicts 99.116% of the association that exist between the votes by political parties and its political zones.

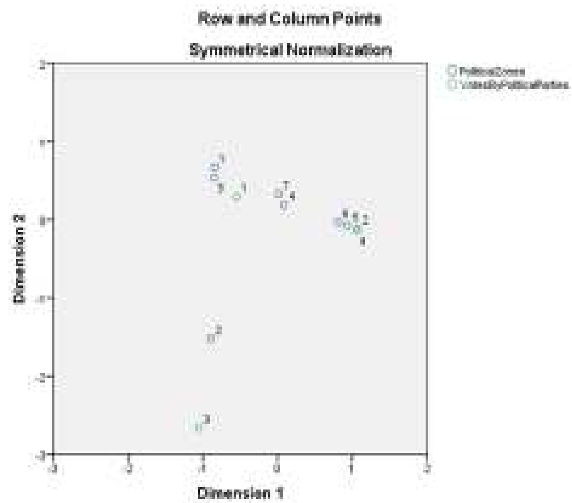

Figure 1: Symmetric Two dimensional Correspondence Map for the votes by political parties/political zones

CONCLUSION

The forgoing discussion we have developed a method for the analysis of two-contingency table using the correspondence analysis technique. The empirical result obtained shows that the pattern of voting in the 2011 presidential election were the same in the south east and south south, north central and FCT, north east and north west respectively, only south west demonstrated a different pattern. On the number of votes scored by the political parties, CPC and ANPP scored had the similar pattern while PDP and ACN seem to standout on their pattern.

REFERENCES

1. Alan J. I. (2008): Modern Multivariate Statistical Techniques; Regression, Classification and Manifold Learning. New York, Springer.
2. Greenacre M., and Hastic T. (1987). The geometric Interpretation of Correspondence Analysis. Journal of American Statistical
3. Association, 82, 437-447
4. Haruna I., Ozodinobi N. B., Sulaiman S. A. and Oyewole S. A (2011). Statistics and Electoral process, Nigerian Experience. Proceedings of the 35th conference on "Statistics: A varitable Tool for Achieving a stable Democracy. 349-360. Nigerian Statistical Association.
5. Hassel P. J. and Ganesh S. (1996). Correspondence analysis of English as a foreign language. The New Zealand Statistician, vol. 31, pp. 24-33.
6. Hoffman D. L. and Franke G. R. (1986). Correspondence analysis: graphical representation of categorical data in marketing research. The American Statistican, vol.23, pp 213-227.
7. Javalgi T., Whipple, T., McManamon, M. and Edick V. (1992). Hospital Image : A correspondence approach. Journal of Healthcare Marketing. Vol. 12, pp 34-41.
8. Kishino H., Hanyu K., Yamashita H. and Hayashi c. (1998). Correspondence analysis of paper recycling society: consumers and paper makers in Japan, resources, conservation and recycling. Vol. 23, no. 4, pp 193-208.
9. Romney, A. K., Moore, C. C. and Rusch C. D. (1997). Cultural Universals: Measuring the Semantic structure of emotion terms in English and Japanese. " Proceedings of the National Academy of Sciences of United States of America, vol. 94, No. 10, pp 5489-5494.
10. Watt, D. D. (1997). Correspondence Analysis: a graphical technique for examining categorical Data. Nursing Research, vol.46, no. 4, pp 235-239.

Source of Support: None Declared
Conflict of Interest: None Declared