Analysis of JAMB Examination Results: Using Cluster and Canonical Correlation Techniques.

Shehu Abdulazeez [1*] S.U. Gulumbe[1,2]

1. *Department of Mathematical Sciences & Information Technology, Federal University, Dutsin-ma, PMB* 5001, Katsina, Nigeria
2. Department of Mathematics, Usmanu Danfodiyo University Sokoto, PMB 2346, Sokoto, Nigeria

* shehu.abdulazeez@ymail.com

Abstract

This study investigates JAMB examination results for 2006. The data consist of scores of 225 students. Out of twenty-three subject areas which the students sat for, eleven subjects were used for the study. The data was subjected to cluster and canonical correlation analysis .The essence was to divide the subjects into smaller number of classes such that objects in the same class are similar to one another. The complete linkage method of cluster analysis was employed to this effect, while canonical correlation analysis was employed to investigate the relationship between these groups. The results show that two groups were formed. Group A consists of Government, religion, mathematics, commerce, and literature-in-English, while Group B consists of English, biology, chemistry, agriculture, economics, and physics. Five canonical roots were obtained and only one is statistically significant. However, group A is inversely related to group B indicating results from mathematics, commerce and religion show the strongest relationship and weighted heaviest among the subjects. While Chemistry, Economics and Agriculture appeared to be low with weak relationship and weighted lighter among the subjects.

Keywords canonical correlation, cluster analysis

1.0 Introduction

Multivariate statistics is a useful set of methods for analyzing a large amount of information in an integrated frame, focusing on the simplicity (Simon,1969) and latent order (Wheatley,1994) in seemingly complex array of variables. Benefits to using multivariate statistics include: Expanding sense of knowing, allowing rich and realistic research designs and having flexibility built on similar univariate methods. There are several disbelieves on multivariate method these includes: needing large samples, a belief that multivariate methods are challenging to learn, and results that may seem more complex to interpret than those from univariate methods. Most multivariate methods can embrace multiple theories and hypotheses; most can analyze several independent variables, and some allow several dependent variables for example canonical correlation, factor analysis and multivariate analysis of variance (MANOVA).Some allow the examination of several groups or samples (e.g logistic regression, analysis of covariance, and discriminant function analysis (Anastasi and Urbina,1997; Harlow,2005).

Canonical correlation analysis is the most generalized member of the family of multivariate statistical techniques. It sis the directly related to several dependence methods, similar to regression, canonical correlation's goal is to quantify the strength of the relationship, in this case between the two sets of variables. It corresponds factor analysis in the creation of composites of variables. It also resembles discriminant analysis in its ability to determine independent dimensions for each variable set, in this situation with the objective of producing the maximum correlation between the dimensions. Thus, canonical correlation identifies the optimum structure or the dimensionality of each variable set that maximizes the relationship between independent and dependent variable sets.

Cluster analysis classifies a set of observations into two or more mutually exclusive unknown groups based on combinations of interval variables. The purpose of cluster analysis is to discover a system of organizing observations, usually people, into groups where members of the groups share properties in common. It is cognitively easier for people to predict behavior or properties of people or objects based on group membership, all of whom share similar properties.

2.0 Applications of Canonical Correlation and Cluster Analysis

According to Hair et al, (1995,the first canonical correlation explains the maximum relationship between the canonical variates and each successive canonical correlation is estimated so as to be orthogonal yet explain the maximum relationship not accounted for by the previous canonical correlation. The overall strength of the relationship between X's and Y's is assessed by examining canonical correlation coefficients. A more detail examination of the canonical correlation structure may be accomplished by examining canonical loading and canonical crossloading. Canonical loading estimates the influence of each independent variable on the newly created variate and canonical crossloadings estimates the strength of the correlation between each dependent variable and independent variate set.

Jordan et al (2006), conducted a study to test empirically the hypotheses found in the literature regarding impart of moral orientation on the counseling process. The correlation between moral orientation parables (match, mismatch. care and justice) and counselor effectiveness rating on the CRF-S SEQ and RSRS were examined. The results suggested that therapists' effectiveness was not perceived differentially by clients. According to, whether or not the therapist matched the client's voice. Clients also found the therapists effectiveness whether the therapists spark predominately in the care or justice voice.

Aldenderfer et,al (1984) describe and asses tests proposed for identifying appropriate value of n from a hierarchy. Although this is the most common way in which tree diagrams have arisen in the data analysis, there has recently been an increasing interest in the investigation of tree that have been specified directly.

3.0 Materials and Methodology

3.1 Data Used For The Study

The data used for the study was collected from Joint Admission and Matriculation Board, Headquarter, Bwari-Abuja. The data consist of scores of 225 students for 2006 JAMB Exam in eleven subject areas namely, English language, Mathematics, Biology, Chemistry, Agriculture, Commerce, Economics, Government, Religion, Literature-in-English, and Physics.

3.2 Cluster Analysis Used For The Study

Cluster analysis quantifies card-sorting data by calculating the strength of the perceived relationships between pairs of subjects based on how often the members of each possible pair appear in a common group. The measure of the relationship between any two subjects is that pair's similarity score. Cluster analysis programs display output in the form of tree diagrams, in which the relationship between each pair and is represented graphically by the distance between the origin and the branching of the line leading to the two subjects.

The complete linkage, also called farthest neighbor, clustering method is the opposite of single linkage. Distance between groups is defined as the distance between the most distant pair of objects, one from each group. In the complete linkage method, ***D(r,s)*** is computed as ***D(r,s)*** = Max { *d(i,j)* : Where object *i* is in cluster ***r*** and object *j* is cluster ***s***}. Here the distance between every possible object pair *(i,j)* is computed, where object *i* is in cluster ***r*** and object *j* is in cluster ***s*** and the maximum value of these distances is said to be the distance between clusters ***r*** and ***s.*** In other words, the distance between two clusters is given by the value of the longest link between the clusters. At each stage of hierarchical clustering, the clusters ***r*** and ***s*** for which ***D(r,s)*** is minimum, are merged.

3.3 Canonical Correlation Analysis

Canonical correlation analysis is a multivariate statistical method that facilitates the study of interrelationship among sets of multiple dependent variables and multiple independent variables. It finds two bases one for each variable that are optimal with respect to correlations and at the same time, it finds the corresponding correlations. It also determines the correlation coefficients. More than one canonical will be found each corresponding to a different set of vector (or canonical variants)

Correlations between successively extracted canonical variates decrease. Correlation coefficients are proportional of correlation between the canonical variates accounted by a particular variable. The first canonical is always the one, which explain most of the relationships. The canonical correlations are then interpreted the same as Pearson's r. Their squares are the percent of variance in the canonical variates for the other set along the dimension represented by a given canonical correlation (usually the first). Another way to put it is to say that R_c square is the percent of variance shared by the canonical variates along this dimension. Some, researchers when reporting canonical correlation, report just the first canonical correlation but it is recommended that all meaningful and interpretable canonical correlations be reported.

Pooled R^2 represent all the orthogonal dimensions in the solution by which the two set of variables are related. Pooled R^2 is used to assess the extent to which one set of variables can be predicted or explained by the other set.

3.4 Method of Computation of Canonical Correlation

Consider two sets of variables $Y = Y_{n\times p}$ and $X = X_{n\times q}$ where $p \leq q$ construct the linear combinations i.e.

$$g = a'X \quad f = b'X \tag{1}$$

such that Γ_{gf} is a maximum.

The XY matrix is

$$XY = \begin{bmatrix} X_{11} & \cdots & X_{1q} & Y_{11} & \cdots & Y_{1p} \\ X_{21} & \cdots & X_{2q} & Y_{21} & \cdots & Y_{2P} \\ \vdots & & \vdots & \vdots & & \vdots \\ X_{n1} & \cdots & X_{nq} & Y_{n1} & \cdots & Y_{np} \end{bmatrix} \tag{2}$$

The correlation matrix of the above data and its partition is as shown in equation (2). Let R denote the correlation matrix. Then the partitioned correlation matrix R is given by

$$R = \begin{pmatrix} R_{xx} & \vdots & R_{xy} \\ \cdots & \vdots & \cdots \\ R_{yx} & \vdots & R_{yy} \end{pmatrix} \tag{3}$$

Where R_{XX} and R_{YY} are the correlation matrices within X and Y respectively and $R_{YX} = R_{XY}$ which is the correlation matrix between X and Y. To obtain the eigen values and their corresponding eigenvectors, it is required to form a symmetric metric $R_{xx}^{-1/2} R_{xy}^{-1} R_{yx} R_{xx}^{-1/2}$ .The eigenvalues and the corresponding we use the equation.

$$a_i = R_{11}^{-1/2} \gamma_1 \tag{4}$$

Where γ is the corresponding eigenvector for the independent variable, to find the corresponding eigenvectors for, we use the equation.

$$h_i = \frac{1}{\lambda} R_{yy}^{-\frac{1}{2}} \Sigma_{yx}\, a \tag{5}$$

We can obtain the canonical coefficients for the dependent variable Y by applying the equation

$$b_i = \frac{1}{\lambda} \Sigma_{yy}^{-1} \Sigma_{yx}\, a \tag{6}$$

Marda et al (1979).

3.5 Significant Test

The significant test of the canonical correlation is straightforward in principle. Simply stated, the different canonical correlations are tested, one by one, beginning with the largest one. Only those roots that are statistically significant are then retained for subsequent interpretation. First, evaluate the significance of all roots combined, then of the roots remaining after moving the first root, the second root and so on. Bartlett (1937) has outlined procedure for testing the significant of canonical correlations. He defined Lambda (Λ)

$$\Lambda = \prod_{i=1}^{q} (1 - \lambda_i) \qquad q < p \tag{7}$$

$$\chi 2 = -[N - .5(p + q + 1)] ln\Lambda \tag{8}$$

3.6 Interpretation

Having obtained the significant canonical correlations and their corresponding canonical coefficients for the two sets .We need to go further and interpret the correlation coefficients. The following relations are used for interpretation. The values of canonical coefficient indicate individual correlation between each of the variables in each set with respect to its canonical variants. For example:

$$U_1 = a_{11}Y_1 + a_{12}Y_2 + \cdots + a_{1p}Y_p \quad \text{(i=1)(1)p).} \tag{9}$$

These constant coefficients $a_{11} \; a_{12} \; \cdots \; a_{1p}$ are partial correlations between U_1 and the original variables $Y_1 \; Y_2 \; \cdots \; Y_p$. Each of a_{11} is the correlation between U_1 and the corresponding $Y_{1.}$

Canonical structure coefficients look at the relationship between each individual variable in a given set and variants (i.e., canonical variable). For example, suppose $U = a'Y_{,}$ is the canonical variable of the dependent set, Y then

$$Cor(Y, U) = Cor(Y, aY) = E(Y, \bar{Y})(a'Y \;\; a\bar{Y})' = E(Y - \bar{Y})(Y -)'a = R_{yy}a \tag{10}$$

Similarly

$$Cor(X, b'X) = R_{xx}b \tag{11}$$

The third relationship is the canonical cross loading. This looks at the relationship between the canonical variant and variables of its opposite set. For example,

$$Cor(X\; a'Y) = E(X - \bar{X})(a'Y - a\bar{Y})' = R_{xy}a = R_{xx}b\rho. \tag{12}$$

since $\Sigma_{21} a = \rho \Sigma_{22} b$

Similarly

$$Cor(Y, b'X) = R_{yy}a\rho \tag{13}$$

3.7 Statistical Packages Used For the Study.

The statistical Package for the Social Sciences (SPSS version 15) was used for the study of canonical correlation. In addition R version 2.10.1 was used for hierarchical clustering.

4.0 Results

The first stage in canonical correlation analysis is the clustering of the subjects,the complete linkage method was used to form the two groups.

From Figure 1 we have the following groupings

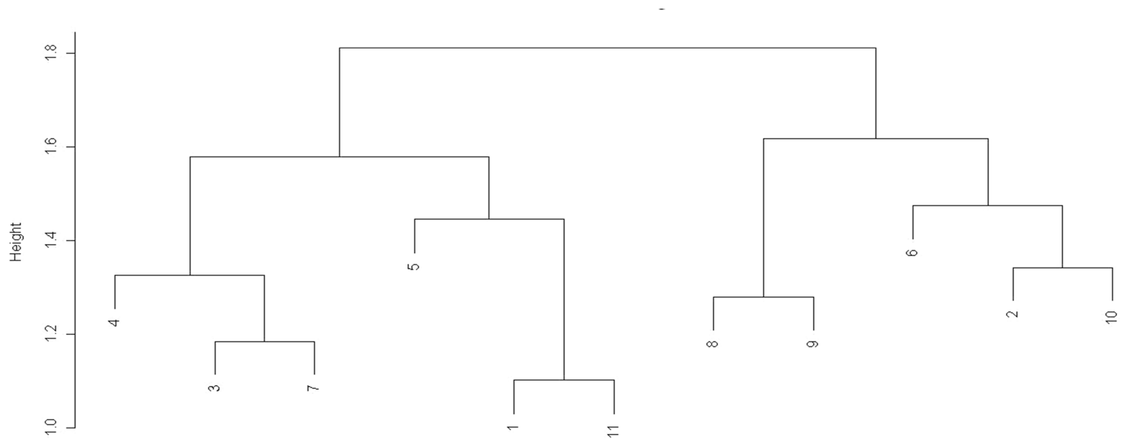

Figure 1: complete linkage Cluster dendrogram

Numbers 1, 2, 3…………..11 represents English language, mathematics, biology, Chemistry, agriculture, commerce, economics, government, religion, literature-in-english, physics on the dendrogram.

Figure 1 show complete linkage cluster dendrogram of the subjects, which ideally illustrates two clear groups with government, religion, mathematics, commerce and literature-in-english are seen forming a group, while English language, chemistry, biology, physics, agriculture and economics forming another group.

Analysis begins with a simple examination of the correlation significance.

Table 1. Canonical correlation coefficients of set1 and set2

Set1 = (Government, religion, mathematics, commerce, Literature-in-English).

No	Canonical Correlations	Eigen value	% of variance explained
1	.302	.116	56.73
2	.258	.047	22.69
3	.134	.032	15.71
4	.075	.009	4.47
5	.020	.001	0.39

Set 2 = (English, biology, chemistry physics, agriculture, economics).

The Table .1 shows the canonical correlation coefficients of the five Canonical variables and their corresponding eigenvalues. The eigen values of the canonical variables can be tested by employing Bartlett's criterion to test for the significant using Bartlett (1937).

H_0 :$\sum_{xy} = 0$ Against H_1 :$\sum_{xy} \neq 0$ at $\alpha = 0.05$, we have the following tables

Table 2. Test that the remaining correlation are zero.

s/no	N	P	Q	Df	χ^2_β	χ^2_α
1	225	6	5	30	41.325	40.77
2	225	5	4	20	20.482	31.41
3	225	4	3	12	5.346	21.03
4	225	3	2	6	1.358	12.29
5	225	2	1	2	0.0894	5.99

Since one of the canonical coefficients tested significant $\chi^2_\beta > \chi^2_\alpha$.It implies that the null hypothesis is rejected which indicates that one out of five canonical correlations is significantly different from zero.

Now consider the first canonical variate pair U_1 and V_1 with coefficient r_1=.302. So that the proportion of variance common to the first canonical variate pair is $r_1^2 = .0924$ showing about 9.24% of the proportion of variance captured by the first canonical variate .Similarly r_2 =.258 is the canonical correlation between the second canonical variate pair and so $r_2^2 = .0665$ which indicates about 6.7% of the proportion of variance captured, r_3 =.134 shows the correlation between the third canonical variate pair and so indicating 2% of the proportion of variance captured, r_4= .0175 captured about 1% of the variance, while r_5 = .020 only contributed less than 0.5% of the variance.

Although canonical correlation analysis has many tables for interpretations.

Table 3. Canonical loading for set1 and set2.

Set	Subject	r_1	r_2	r_3	r_4	r_5
Set1	Government	.001	-.311	-.044	-.938	.148
	Religion	.201	.729	-.128	-.296	-.571
	Commerce	.637	.161	.664	-.091	.344
	Mathematics	.747	-.232	-.616	.091	-.023
	Literature-in-english	.076	.375	-.438	-.186	.792
Set2	Chemistry	-.541	.304	-.151	-.723	-.233
	English language	-.144	-.492	-.460	-.154	-.043
	Physics	-.223	-.557	-.045	-.691	.353
	Agricultural sci	-.470	-.669	.204	.248	-.415
	Biology	-.098	-.214	.830	-.048	.009
	Economics	-.572	-.006	.176	.185	.748

The first pair of canonical variates can be written as follows

U_1 = .001government +.201religion +.637commerce+.747 mathematics +.076 literature-in-english

V_1 = -.541chemistry -.144english - .223physics -.098biology -.470agriculture - .572economics.

$Ø_1$=.302

The correlation $Ø_1$, between U_1 and V_1 is called first canonical correlation coefficient.

The canonical loadings in Table 4.3 provided information about the contribution of variables to each independent canonical relationship. Of the individual variable Mathematics loading heaviest with the value (.747) followed by Commerce (.637), Religion (.201) and English language (.144) and biology (.098) leading for the ordering for the criterion variables.

The values attached to each subjects are their partial correlation to their corresponding canonical variables and indicating the individual's contribution to the canonical pairs

Table 4. Canonical cross loading for set1 and set2

1	Subject	r_1	r_2	r_3	r_4	r_5
Set1	Government	.000	-.080	.006	-.070	-.003
	Religion	.061	.188	-.016	-.022	-.012
	Commerce	.192	.042	.089	-.007	.007
	Mathematics	.225	-.060	-.083	.007	.000
	Literature-in-english	.023	.097	-.059	-.014	.016
Set2	Chemistry	-.163	.078	.020	-.054	-.005
	English language	-.043	-.127	-.062	-.011	-.001
	Physics	-.067	-.144	-.006	-.051	.007
	Agriculture	-.142	-.173	.027	.018	-.008
	Biology	-.030	-.055	.111	-.004	.000
	Economics	-.173	-.001	.024	.014	.015

The table .4 shows the canonical cross loading of the two canonical functions .In the first canonical Function we see that both mathematics and commerce slightly high in correlations with independent Canonical variate .225 and .192 respectively.While the weakest correlation came from set2, economics with -0.173 followed by chemistry with -0.163.

4.1 Discussion of Result.

Figure 1 shows the dendogram of the cluster analysis of the two sets, the complete linkage method was used, the strength of the relationship between the two sets, trace a path from one of the sets to the other, following the branches of the dendogram and taking the shortest possible path. The distance from the origin to the outermost vertical line required by this path represents perceived degree of relatedness between the two sets.

The canonical correlation which examined the linear relationship between set1 and set2 variables by creating the combinations .The first canonical correlation explains the maximum relationship between the canonical variates and each successive canonical correlation is estimated so as to be orthogonal yet still explain the maximum relationship not accounted for by the previous canonical correlation.

This reflects the high shared variance between these variables .By squaring these terms we find percentage of the variance for each of the variables explained by function 1. The result shows that 31 percent in mathematics and 28 percent of the variance in commerce by function 1. Looking at the dependent variables we see that physics and biology have positive correlation of .067 and .03 with the dependent canonical variate from this information 25 percent of the variable is explained by the dependent variate (25 percent is obtained by squaring the correlation coefficient .038) .The remaining correlations of independent variables that is, chemistry, English language, agriculture appear to be low but after squaring these correlations , they only contributed 5.6 and 3.1 percent respective in the canonical variate.

The final issue of interpretation is examining the sign of the cross-loading. All dependent variables except literature-in-english have a positive signs, inverse relationship for the first function. Thus independent variables biology, economics, English language, physics, plus two dependent variables and government and mathematics are negative in the second function. The three highest cross-loading of the first independent variable correspond to the variables with the highest canonical loadings.

Thus all the relationships are inverse except for one direct relationship in the first canonical function.

4.2 Summary

The study is aimed at analyzing the JAMB examination results, the initial analysis considered the scores of the students in selected subject areas which are partitioned into two groups through the use of cluster analysis as shown in Figure 4.1. The canonical correlation analysis generated five correlation coefficients, which are test and found one out of the five correlation coefficients to be statistically significant from zero.

From Table 4.1 above it can be seen that first canonical pair captured the variability of about 57%, second canonical pair captured about 23% of variability, third canonical pair contributed only 16%, the fourth captured about 4% and fifth canonical pair captured less than 1%. Hence, the total variability captured by the five canonical pairs is 100%. The 57% of variability captured is due to the individual contribution s of the composite of religion, government, mathematics, commerce, & literature-in-English of set1 and English language, chemistry, economics, biology, agriculture, and, physics of set2, showing that there is relationship among the subjects selected for the analysis.

5.0 Conclusion

It can be seen that set1 and set2 are correlated as sought for. Canonical correlation analysis measured the level of the relationship of the first canonical pair and the subjects that contributed. The first pair with a measure of correlation of 0.302 with a proportion variability of about 39%, the second canonical pair with a measure correlation 0.258 having a proportion of variability of 33%, the third canonical pair shared about 17%, fourth shared about 9% and fifth canonical pair shared a proportion of variability of less than 3%.

Although the correlation is from low to moderate, with ,mathematics, commerce and religion showing relative more strongly associated with government and literature-in-English compared to the biology, English language, economics, chemistry agriculture and physics. Also from the analysis, the three highest cross-loading of the first independent variate (religion, commerce, mathematics) corresponding to the variables with the highest canonical loading as well as canonical weight which is an indication that the variables in the independent variates are related.

Mathematics, Commerce and Literature-in-english are directly related. Whereas the other subjects are also related to each other .However, it can be seen that mathematics literature-in-english and commerce are inversely related to other subjects. That is, an increase in the performance in mathematics, literature-in-english and commerce will lower the performance of the other subjects and vice- visa and this might be due to multiple-choice nature of JAMB questions where candidates are not allowed to express themselves. The non significant nature of the results might be due to different background of the students, some students came from disadvantage homes, some students attended public schools, where large number of students were taught under unconducieve environment.

References

Anastasi, A. and Urbina, S. (1997), *Psychological Testing upper saddle River* N.J PrenticeHall.

Aldenderfer,M.S.,and Blashfield,R.K (1984), Cluster analysis Beverty Hills sage.

Bartlett, M.S (1937), The statistical significant of canonical correlations, *Biometrical*32 - 29

Hair,J.F.,Anderson,R.E.,Tatham,R.L.,& Black,W.C.(1995).*Multivariate data Analysis, forth edition*.Upper Saddle River:Prentice Hall.

Harlow, L. L (2005), The Essence of Multivariate Thinking, Theme and Methods, Mahwah,N.J;Lawrence Erlbaum

Associates, pages 205,229.

Homer-,Dixion T.F (1995),"Strategies for studing causation in complex Ecological Political system", University of Toronto: Occasional paper for the project On Environment,Population and Security. http://www.padrigu.gu.se/EDCNews/Reviews/HomerDixion1995b.html (September 06, 2003).

Hotelling, H. (1936), Relation Between two of sets Variables,*Biometrika* 28 ,312 -377

Jordan J.; Martin D.and Sermivan S. (2006),The effect of moral orientation variables in counseling college students, *Journal of College character*, volume vii .

Marda K.V,Kent J.T,and Bibby J.M (1979),*Multivariate Analysis,* fifth edition, academic press Inc, London Ltd.

Simon, H.A (1969), The Science of the Artificial, Cambridge M.A., M.I.T press.

Wheatley, M.J (1994), Leadership and the new science about organization from an orderly universe, san-francisco CA.Bertt Koehley publisher,Inc.