Comparison Of Canonical Correlation Analysis And The Generalized Canonical Correlation Analysis Using The Lognormal And Cauchy Distributed Data

S. I. ONYEAGU, *G. A. OSUJI AND O.M. ONYIA.

DEPARTMENT OF STATISTICS, NNAMDI AZIKIWE UNIVERSITY, AWKA, NIGERIA.

FOR CORRESPONDENCE: *G.A. OSUJI (george.osuji99@yahoo.com)

ABSTRACT

This work compares the Canonical Correlation Analysis and Generalized Canonical Correlation Analysis for two data sets using Lognormal and Cauchy distributions. Data of different sample sizes (10, 20, 30, 40 and 50) were simulated and used to perform the analysis. Their relative efficiencies were calculated and it was discovered to be the same for two data sets except for scaling.

Keywords: Relative efficiency, Scaling, Canonical Correlation Analysis, Generalized Canonical Correlation Analysis, Lognormal, Cauchy

1.0 INTRODUCTION

Hotellings (1936) developed the process of maximizing the correlation between two linear functions called Canonical Correlation analysis and the Generalized Canonical Correlation analysis which are principally designed and/or used for the maximization of two or more linear functions are compared in the work to see this advantage(s) of one over the other. In Canonical Correlation Analysis, many works have been done as on how to maximize the correlation between two linear functions of two sets of random variables. The common goal of all these works is to find two sets of weights say a and b or α and β such that each canonical variate is maximally correlated subject to the restriction that each variate be orthogonal with the previous linear combinations. Apart from using the Canonical Correlation Analysis in testing for independence of two sets of variables, more importantly, it is also used as a data reduction method which reduces the task of the researcher and yet clarifies its interpretation in a coordinate system. However, it is important to note that as the Canonical Correlation decreases in size so the relationship between the corresponding canonical variates become weaker and the consequent predictions become less accurate. More so, although, the technique may be of some interest in the study of relationships between two sets of variables, and may even provide some useful predictive models, it can be seen that their scope is very limited. This is because they only predict linear combinations of the Y_i, and furthermore the linear combinations that they predict are determined by the data and are not under the control of the investigator. (See Krzanowski 1993). Based on the above demerits of the Canonical Correlation, Kettenring (1971) developed and compared extensions of Canonical Correlation to three or more sets of variates, and has given iterative schemes for the computation of the correlations and coefficient that is user friendly. However, it was Van de Velden and Biljmolt (2006) that explained that Caroll (1970) introduced the Generalized Canonical Correlation Analysis which allows for several sets of variables to be analyzed simultaneously. This makes the method suited for the analysis of various types of data especially in situation where subjects may be asked to rate a set of objects on a set of attributes. In this case, for each individual, a data matrix can be constructed where objects are represented row-wise and attributes column-wise. Then using Generalized Canonical Correlation Analysis, a graphical representation, sometimes referred to as a perceptual map can be made on the basis of the individuals' observation matrices. The advantage of the Caroll's approach to Generalized Canonical Correlation Analysis is that it has some attractive properties that makes the method well fit for the analysis of multiple sets of data. This is because computationally, the method is straight forward and its solution is based on an eigen equation and the method is closely related to several well known multivariate techniques such as principal component analysis and multiple correspondence analysis. Based on the above, this work compares the

Canonical Correlation analysis and the Generalized Canonical Correlation Analysis based on their relative efficiency using the Lognormal and Cauchy distributions whose probability densities are given as:

$$f(x) = \frac{1}{x\sqrt{2\Pi}}\ _{\delta}\ \exp\left\{\frac{(\log x - \varepsilon)^2}{2\delta^2}\right\}, \quad x > 0$$

and

$$f(x) = \frac{1}{\pi\lambda} \cdot \frac{1}{1+(\frac{x-\theta}{\lambda})^2} \qquad -\infty < x < \infty$$

Respectively (See Balakrishnan & Chin-Diew Lai 2010) for samples sizes of 10, 20, 30, 40 and 50.

2.0 REVIEW OF RELATED LITERATURE

There are many literatures on the Canonical Correlation Analysis but few works that are relevant to our work upon which we build our comparison are the works mentioned by Krznowski (1993). He pointed out in his text that the technique of Canonical Correlation Analysis transforms the p-variates in x_1, and the q-variates in x_2 to s pairs of variates $(U_1, V_1), \ldots, (U_s, V_s)$. The n x (p + q) data matrix is thereby transformed to a new n x (2s) data matrix. Writing $U' = (U_1,\ldots,U_s)$, $V' = (V_1,\ldots,V_s)$ and $y = (U', V')$, this transformation reduces the sample covariance matrix

$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$ of the x_i to one of the form

$$\begin{pmatrix}
1 & 0. \ . \ . \ 0 & R_1 & 0. \ . \ . \ 0 \\
0 & 1. \ . \ . \ 0 & 0 & R_2. \ . \ . \ 0 \\
. & & & \\
. & & & \\
. & & & \\
0 & 0. \ . \ . \ 1 & 0 & 0. \ . \ . \ R_s \\
R_1 & 0. \ . \ . \ 0 & 1 & 0. \ . \ . \ 0 \\
0 & R_2. \ . \ . \ 0 & 0 & 1. \ . \ . \ 0 \\
. & & & \\
. & & & \\
. & & & \\
0 & 0. \ . \ . \ R_s & 0 & 0. \ . \ . \ 1
\end{pmatrix}$$

For the y_i

Pedhazur (1997) noted the following properties of canonical functions "the first canonical function identifies linear combinations of the study's variables that yield the largest squared correlation R^2_c possible. The second canonical function identifies linear combinations of the study's variables that are not correlated with the first pair of canonical variates and yield the second largest R^2_c possible, given the residual variance left over from the first function, the same is true for subsequent canonical functions, such that the mth canonical function identifies linear combinations of the study's variables that are not correlated with prior pairs of canonical variates and yields the mth largest possible. Takaine and Hwang (2002) have proposed a method for generalized constrained Canonical Correlation Analysis (GCCANO) which incorporates external information on both rows and columns of data matrices. They demonstrated the method with two illustrations and observed that in Canonical Correlation Analysis,the interpretation of the canonical variates obtained from GCCANO can be difficult and a simple rotation of either the canonical pattern or the structure matrix (by varimax) may be used to make it easier to interpret. Mekeaon (1966) explained that it is now possible to define the Generalized Canonical Correlation for m set of variables as the maximum value of generalized product moment correlation for m linear composites.

$a_i^{'} x_i$; i = 1, 2,…,m with respect to variation of a_i. He added that $\ell(x_1, x_2,\dots, x_m) = \max r_i\,(a'_1x_1, a'_2x_2, \dots, a'_mx_m)$; which is the same with

$$\max \frac{1}{m-1}\left(\frac{\sum_{ij} a_i^{'} s_{ij} a_i}{\sum_i a_i^{'} s_{ii} a_i} - 1\right)$$

which is equivalent to maximizing the quantity

$$r = \left(\frac{\sum_{ij} a_i^{'} s_{ij} a_i}{\sum_i a_i^{'} s_{ii} a_i}\right) = (m-1)r_1 \;\; +1$$

letting S be the sum products matrix for the m set combined and Sd be a diagonal supermatrix with element S_{ii}. The Horst's term "super matrix" refers to a matrix whose elements is matrices and a' = [a_1', a_2', … , a_m',] is the vector of combined weights. Hence

$$r = \frac{a^{'} sa}{a^{'} sda}.$$

Also an alternative generalization of the Canonical correlation to more than two sets of variates is based upon the generalized association measure

$$r_1 = \frac{2\sum_{i<j} \sigma_{ij}}{(m-1)\sum_i \sigma_i^{\;2}} = \frac{\sigma_i^2 - \sum_i \sigma_i^{\;2}}{(m-1)\sum_i \sigma_i^2}$$

Most importantly, Van de Veldon (2011) speaking on the equivalent of the Canonical Correlation Analysis (CCA) and the Generalized Canonical Correlation Analysis argued that the Caroll's Generalize Canonical Correlation Analysis takes ordinary correlation analysis as a special case. Hence, he proved that when there are only two sets of variables, orthogonality of the group configuration implies the orthogonality (with respect to the appropriate variance matrices) of the canonical variates. Furthermore, to show the relationship between CCA and GCCA for two sets of variables, he first showed that the CCA solution is also a GCCA solution. It should be noted that the group configuration used in GCCA is not present in CCA. On the other hand, the canonical weights are restricted in CCA but not in GCCA. That is in CCA, the correlation is maximized under the restrictions

$A_1'R_{11}A_1= 1$ and $A'_2R_{22}A_2 =1$

In GCCA, only the group configuration is constrained to be orthonormal. The eigen equation is given as $(\sum_{i=1}^{n} p_i)$ x Y = Y x Γ (2.1)

Where Y is the group configuration matrix. Using the CCA solution where

$W = X_1A_1$ and $Z = X_2A_2$, we have $(P_1 + P_2)(W + Z) = (X_1R_{11}^{-1}X_1' + X_2R_{22}^{-1}X_2')(W + Z)$

$= X_1R_{11}^{\;1/2}U + X_2R_{22}^{\;-1} R_{21} R_{11}^{\;1/2}U + X_1R_{11}^{-1}R_{12} R_{22}^{1/2}V + X_2R_{22}^{\;1/2}V$ (2.2)

Since $U = X_1'W$ and $V = X_2'Z$

Equation (2.2) reduces to

$(P_1 + P_2)(W+Z) = (W+Z)(I + \Lambda)$

Furthermore $W'W = Z'Z = 1$ and $W'Z = Z'W = \Lambda$ it follows that

$(W + Z)'(W + Z) = 2(1 + \Lambda)$

Hence $(\sum_{i=1}^{2} P_i)\, x\, Y = Y\, x\, \Gamma$

Where $\Gamma = 1 + \Lambda$ and $Y = \frac{1}{\sqrt{2}}(W + Z)(1 + \Lambda)^{1/2}$

So that $Y'Y = 1$

Then $A_{GCCA,i} = (X_i'X_i)^{-1}X_i'Y$ so that

$\sqrt{2}\, A_{GCCA,1} = (X_i'X_i)^{-1}X_i'(W + Z)(1 + \Lambda)^{1/2}$

$= (R_{11}^{\ 1/2}U + R_{22}^{\ -1} R_{12} R_{22}^{\ 1/2}V)(1 + \Lambda)^{1/2}$

$=(R_{11}^{\ 1/2}U + R_{11}^{\ 1/2}U\Lambda)(1 + \Lambda)^{1/2}$

$= R_{11}^{\ 1/2}U(1 + \Lambda)^{1/2}$

$=A_1(1 + \Lambda)^{1/2}$ Similarly

$\sqrt{2}\, A_{GCCA,2} = A_2(1 + \Lambda)^{1/2}$ Hence, he concluded that the only difference between methods concerns the scaling, since in GCCA

$A'_1R_{11}A_1 = \frac{1}{2}(1 + \Lambda) = \frac{1}{2}\Gamma$

For other literatures, see Roy [11, p.26], Anderson [1], Horst [4,5], Meredith [9], Mckeon [1965] etcetera.

3.0 METHODOLOGY

3.1 CANONICAL CORRELATION ANALYSIS

Let $Y' = [Y_1, Y_2, \dots, Y_p]$ and $X' = [X_1, X_2, \dots, X_q]$ where the elements $p \leq q$. We define two linear functions $U = a'y$ and $V = b'x$ with unit variance such that the correlation between U and V is maximum. We then maximize

$$F_{uv} = \max_{a,b}\, a' \sum_{12} b$$

Subject to the constraints that $a'\Sigma_{11}a = b'\Sigma_{22}b = 1$ using the Lagrange multipliers

$$F = a'\Sigma_{12}b - \frac{\ell_1}{2}(a'\Sigma_{11}a - 1) - \frac{\ell_2}{2}(b'\Sigma_{22}b - 1) \qquad 3.11$$

Hence,

$$\frac{df}{da} = \Sigma_{12}b - \ell_1\Sigma_{11}a = 0 \qquad 3.12$$

$$\frac{df}{db} = \sum_{12}a - \ell_2\sum_{22}b = 0 \qquad 3.13$$

$$\frac{df}{d\ell 1} = a'\sum_{11}a - 1 = 0 \qquad 3.14$$

$$\frac{df}{d\ell 2} = b'\sum_{22}b - 1 = 0 \qquad 3.15$$

Multiplying (3.12) by a', (3.13) by b' and applying the constraint, the system of equations to solve becomes

$$\left.\begin{array}{l} -\ell\sum_{11}a + \sum_{12}b = 0 \\ \sum_{21}a - \ell\sum_{22}b = 0 \end{array}\right\} \qquad 3.16$$

Where

$\ell = \ell_1 = \ell_2 = a'\sum_{12}b$

Multiplying the first equation in (3.16) by ℓ and the second by $\sum_{12}\sum_{22}^{-1}$, equation (3.16) becomes

$$\left.\begin{array}{l} -\rho^2\sum_{11}a + \rho\sum_{12}b = 0 \\ \sum_{12}\sum_{22}^{-1}\sum_{21}a - \ell\sum_{12}\sum_{22}^{-1}b = 0 \end{array}\right\} \qquad 3.17$$

adding equation (3.17) together yields the equation

$$(\sum_{12}\sum_{22}^{-1}\sum_{21} - \ell^2\sum_{11})a = 0$$

Therefore

$\ell_1^2, \ell_2^2, \dots, \ell_p^2$, and $a_1, a_2, \dots, a_p$ are the roots and vectors, respectively of the characteristic equation $\left|\sum_{12}\sum_{22}^{-1}\sum_{21} - \ell^2\sum_{11}\right| = 0$ 3.18

let A = $[a_1, a_2, \dots, a_p]$ then

$A'\sum_{11}A = I_p$ and $A'\sum_{12}\sum_{22}^{-1}\sum_{21}A = \Lambda_1$

Where Λ_1 is a diagonal matrix with roots $\rho_1^2, \rho_2^2, \dots, \rho_p^2$

In like manner, by multiplying the second equation in (3.16) by ℓ and the first by $\sum_{21}\sum_{11}^{-1}$, (3.16) becomes

$$\left.\begin{array}{l} -\rho\sum_{21}a + \sum_{21}\sum_{11}^{-1}\sum_{12}b = 0 \\ \rho\sum_{21}a - \ell^2\sum_{22}b = 0 \end{array}\right\} \qquad 3.19$$

adding equation (3.19) leads to the characteristic equation

$$\left|\sum_{21}\sum_{11}^{-1}\sum_{12} - \ell^2\sum_{22}\right| = 0$$

The roots of the equation are $\rho_1^2, \rho_2^2, \dots, \rho_q^2$, with corresponding vectors $b_1, b_2, \dots, b_q$.

If we let B = $[b_1, b_2, \dots, b_q]$, then

$B'\sum_{22}B = I_q$ and $B'\sum_{21}\sum_{11}^{-1}\sum_{12}B = \Lambda_2$

Where Λ_2 is a diagonal matrix with roots $\rho_1^2, \rho_2^2, \ldots, \rho_q^2$

The non zero positive square roots ρ_i of the roots ρ_i^2 are called the canonical correlation between the canonical variates $U_i = a_i y$ and $V_i = b_i x$ for I = 1, …, $p \leq q$.

From (3.17), the relationship between a_i and b_i is given by

$$b_i = \frac{\sum_{22}^{-1} \sum_{21} a_i}{\rho_i} \qquad 3.20$$

The set of canonical variates U_i and V_i are clearly uncorrelated and have

$$\mathrm{Cov}(U_i, U_j) = \mathrm{Cov}(V_i, V_j) = \begin{cases} 1 \text{ if } I \in j \\ 0 \text{ if } i \neq j \end{cases}$$

Furthermore, the covariance between V_i and U_i is ℓ_i, for I = 1, … p and 0 otherwise

$$\begin{cases} \mathrm{Cov}(U_i, V_i) = \rho_i \quad I = 1, \ldots, p \\ \mathrm{Cov}(U_j, V_i) = 0 \quad i \neq j \end{cases}$$

See (Onyeagu 2003)

3.2 GENERALIZED CANONICAL CORRELATION ANALYSIS

Roy [11, p. 26] has proposed formal generalization of the notion of canonical correlation to three or more sets of variates. Considering n column centered observations X_i of order m x p_i, then the generalized Canonical Correlation is given as

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

S. t. $Y'Y = I_k$ 3.22

Where Y is the group configuration matrix (Van de Veldon 2011).

However, carol (1968) has shown that the matrices A_i can be calculated as

$$A_i = (X_i' X_i)^{-1} X_i' Y \qquad 3.23$$

The group configuration matrix y can be obtained using eigen equation as

$$\sum_{i=1}^{n} X_i (X'_i X_i)^{-1} X'_i) \, Y = Y\Lambda \qquad 3.24$$

Where Λ is a diagonal matrix with diagonal elements $\boldsymbol{\lambda}_j$, being the k largest eigenvalues of

$\sum_{i=1}^{n} (X_i (X'_i X_i)^{-1} X'_i)$, X'_is assumed full column rank and the column of Y are corresponding eigenvectors (see Van de Veldon and Tammo 2009).

3.3 COMPUTATIONAL ALGORITHMS

Calling the (CCA) function in R – 2 . 13 . 0 programming language and using the generalized canonical correlation between two data matrices as given by Lahti and Huovilainer (2013), we can advance our argument as follows: Datasets: A list containing the data matrices to be analyzed. Each matrix needs to have the same number of rows (samples), but the number of columns (features) can differ. Each row needs to correspond to the same sample in every matrix. Reg: Regularization parameter for the whitening step used to remove data set specific variation. The value of parameter must be between 0 and 1. The default value is set to 0 which means no regularization will be used. If a non-zero value is given, it means that some of the dimensions with the lowest variance are ignored when whitening. In more terms, the dimensions whose total contribution to sum of eigenvalues of the covariance matrix of each data set below reg will not be used for whitening. The eigenvalue in case of two data sets (eigval-1) would give the correlation.

4.0 SAMPLING EXPERIMENT AND RESULTS

In order to compare the performance of the two methods namely: the Canonical Correlation Analysis and the Generalized Canonical Correlation Analysis, we imputed data on R-command window, calling for the CCA and GCCA function for lognormal distribution and the Cauchy distribution for samples of sizes 10,20,30,40 and 50 that we generated with which we obtained the results given below,

SUMMARY OF RESULT FROM THE ANALYSIS

Distribution		Correlation	eigenvalue	X-mean Vector		Y mean vector	
lognormal	Sample	CCA	GCCA	CCA	GCCA	CCA	GCCA
	10	1	2	148.27	148.27	944.32	944.32
		0.95	1.95	9040.78	9040.78	1894.14	1894.14
		0.80	1.80	1155.57	1155.57	139.26	139.26
		0.32	1.32	2916.61	2916.61	632.92	632.92
		SD=0.31	SD=0.31			148.22	148.22
						2291.79	2291.79
	20	0.85	1.85	191.73	191.73	454.23	454.23
		0.65	1.65	551.09	55.09	880.31	880.31
		0.45	1.45	423.19	423.19	231.07	231.07
		0.30	1.30	316.44	316.44	464.11	464.11
		SD=0.24	SD=0.24			530.05	530.05
						335.79	335.79
	30	0.60	1.60	2428.51	2428.51	1228.25	1228.25
		0.26	1.26	1224.25	1224.25	724.23	724.23
		0.20	1.20	2295.55	2295.25	1166.23	1166.23
		0.09	1.09	927.38	927.38	849.82	849.82
		SD=0.22	SD=0.22			2758.71	2758.71
						1169.09	1169.09
	40	0.51	1.51	1230.56	1230.56	297.54	297.54
		0.48	1.48	705.58	705.58	1163.90	1163.90
		0.19	1.19	603.08	603.08	816.78	816.78
		0.12	1.12	970.78	970.78	492.01	492.01
		SD=0.20	SD=0.20			957.19	957.19
						376.20	376.20
	50	0.54	1.54	766.77	766.77	1432.35	1432.35
		0.31	1.31	674.07	674.07	1080.15	1080.15
		0.18	1.18	3594.64	3594.64	1378.61	1378.61
		0.03	1.03	416.17	416.17	861.38	861.38
		SD=0.22	SD=22			374.07	374.07
						1685.36	1685.36
	Cauchy 10	1	2	2.48	2.48	-1.57	-1.57
		0.97	1.97	2.94	2.94	5.38	5.38
		0.76	1.76	4.11	4.11	3.60	3.60
		0.34	1.34	4.06	4.06	-85.65	-85.65
		SD=0.30	SD=0.30			5.22	5.22
						-16.60	-16.60
	20	0.85	1.85	9.62	9.62	12.87	12.87

		0.67	1.67	3.12	3.12	5.53	5.53
		0.56	1.56	6.32	6.32	8.57	8.57
		0.19	1.19	5.05	5.05	6.23	6.23
		SD=0.28	SD=0.28			6.94	6.94
						9.70	9.70
	30	0.69	1.69	10.07	10.07	14.38	14.38
		0.55	1.55	4..81	4.81	2.82	2.82
		0.30	1.30	3.84	3.83	0.82	0.82
		0.15	1.15	7.60	7.60	6.26	6.26
		SD=0.24	SD=0.24			7.25	7.25
						3.66	3.66
	40	0.48	1.48	-2.15	-2.15	6.56	6.56
		0.46	1.46	3.54	3.54	5.38	5.38
		0.14	1.14	5.20	5.20	-2.53	-2.53
		0.11	1.11	7.93	7.93	4.54	4.54
		SD=0.20	SD=0.20			52.18	52.18
						-1.79	-1.79
	50	0.69	1.69	2.65	2.65	4.68	4.68
		0.45	1.45	3.60	3.60	16.13	16.13
		0.38	1.38	2.88	2.88	5.75	5.75
		0.18	1.18	1.86	1.86	-7.30	-7.30
		SD=0.21	SD=0.21			10.92	10.92
						8.89	8.89

CONCLUSION

We conclude here that since the standard deviation of the correlations and eigenvalues are the same for the methods, it implies that the relative efficiency of the CCA and GCCA is the same for the lognormal and Cauchy distributed data. We also observed that the X-variates of the CCA and GCCA do not differ. This results correlated with the findings of Van de Velden (2011) who showed that when there are only two sets of variables, the orthogonality of the group configuration of the generalized canonical correlation implies the orthogonality of the canonical correlation with scaling as the only difference between the methods.

REFERECES

1. Anderson, T.W. “An Introduction to Multivariate Statistical Analysis, “John Wiley & Sons, Inc., New York, 1958.
2. Caroll, J.D. and Chang, J.J. (1970) Analysis of individual differences in multidimensional Scaling via an N-way generalization of the “Eckart-Young” decomposition, Psychometrika, 35,283-319.
3. Horst, P: “Generalized Canonical Correlations and their applications to experimental data, Journal of clinical Psychology (Monograph Supplement), N0 14(October, 1961), pp331-347.
4. Hotelling, H. (1936a): Simplified Calculation of Principal Components, Psychometrika, Vol. 1, pp27-35
5. Kettenring, J.R.(1971) ”Canonical Correlation Analysis of several variables, Biometrika, vol 56, pp. 433-451.
6. Krzanowski, W.J ;(1993)”Principles of Multivariate Analysis” Oxford University Press inc. New York.
7. Lahti, L. and Houvilainen (2013): Depending Modeling Tool Kit. Statistical Machine learning and Bioinformatics; 14:965-1003.
8. Mackeon, J.J.(1995): Canonical Analysis: Some Relationship between Canonical Correlation, Factor Analysis, Discriminant Function Analysis and Scaling Theory” Psychometric Monographs, N0.13
9. Onyeagu, S.1. (2003): A First course in Multivariate Statistical Analysis. Mega concept, Nigeria.

10. Takane, Y; and Hwan, H,(2002).”Generalized Constrained Canonical Correlation Analysis” Multivariate Research,37(2), 163-195 .

11. Vande Velden, M. (2011).”Generalized Canonical Correlation Analysis” International Statistical Institute. Proceedings of the 58th world Statistical congress, Dublin (session 1PS042); pages 758-765

12. Vande Velden, M., and Bijmolt, T.H.(2006) Generalized canonical correlation analysis of matrices with missing rows: A simulation study. Psychometrica 71, 323-33