ISSN:2321-1156
International Journal of Innovative Research in Technology & Science (IJIRTS)

Comparing two Approaches of Data Reduction Techniques in Principal Component Analysis

Shehu Usman Gulumbe[1], Hussaini Garba Dikko[2], Haruna Suleiman[2]
Professor, Department of Mathematics, Usmanu Danfodiya,University,Sokoto,Nigeria[1]
Dr. Department of Mathematics, Ahmadu Bello university,Zaria,Nigeria[2]
Student, Department of Mathematics, Ahmadu Bello university,Zaria,Nigeria[2]

Abstract: Principal components Analysis is a mathematical procedure of data reduction technique that uses an orthogonal transformation to convert a set of observation of possibly correlated variables into a set of values of uncorrelated variables. This transformation is defined in such a way that the first principal component has the largest possible variance. Different approaches for selecting the principal component to be retained exist in many literatures; in this paper, a Comparative study between two approaches was carried out. The two approaches considered are proportion of variance accounted for and Eigen value one-criterion. Data of cholesterol level of human body was used to select the Principal Components and model adequacy checking was also used to test the fitted models. It was found that three Principal Components where retained using proportion of variance accounted for and R-square (coefficient of determination) of the fitted model is 22.7% and R-squared adjusted was found to be 16.9%. Likewise for the Eigen value one criterion two Principal Components were retained, R-square and R-square adjusted were 23.3%, and18.5% respectively. By considering the result obtained it is clear that Eigen value one criterion is more preferred and desirable in reducing the Dimensionality of every data set in Principal component Analysis.

Keywords: *Principal component, Eigen value, Eigen vector, Proportion of variance, Correlation Matrix, Variance Co-variance Matrix*

1. Introduction

PCA, known as Principal Component Analysis – is a statistical analytical tool that is used to explore, sort and group data. What PCA does is, it takes a large number of correlated variables and transforms this data into a smaller number of uncorrelated variables, known as (Principal Components) or Artificial Variables while retaining maximal amount of variation, thus making it easier to operate the data and make predictions accounted to.

1.1. Development of PCA (Historical Perspective of PCA

According to Jolliffe [5], it is generally accepted that PCA was first described by pearson in (1901), and also discusses the graphical representation of data and lines that best represent the data. At the same year, also concludes that "the best-fitting straight line to a system of points coincides in direction with the maximum axis of the correlation ellipsoid". And also states that the analysis used in his paper can be applied to multiple variables.

However, PCA was not widely used until the development of computers. It is not really feasible to do PCA by hand when a number of variables is greater than four, as Such PCA is only useful for larger amount of Variables.

According to Jolliffe [5], significant contributions to the development of PCA were made by Hotelling et al [4] before the expansion in the interest towards PCA. In 1960s as the interest in PC's rose, important contributors were Anderson [2] with a theoretical discussion and Rao [8] with numerous new ideas concerning uses, interpretations and extensions of PCA. Gower [3] discusses about link between PCA and other statistical techniques and Jeffers [6] with a practical application in two case studies.

2. Procedure

The PCA procedure involves finding the Eigen Value of the sample covariance matrix most especially when the variables are standardized. The variances of the principal components are Eigen values of the covariance matrix; there is *P* of them, some of which may be zero (0).Assuming that the Eigen values are ordered $X\_1 \geq X\_2 \geq X\_3, \geq, \ldots, \geq, X\_P \geq 0$ then X_i corresponds to the i^{th} Principal component, the constants a_{11}, a_{12}, $a_{13}\ldots$ a_{ip} are the elements corresponding to Eigen Vectors scale so that,

$$a_{11}^2 + a_{12}^2 + \ldots + a_{1p}^2 = 1.$$

2.1 The Summary of the procedure are;

1. Start by coding the variables $X_1 \ldots X_P$ to have a zero mean and a unit variance.
2. Calculate the covariance of matrix of the coded variances i.e. the correlation matrix.
3. Find the Eigen Value $X_1 \ldots X_P$ and the corresponding Eigen Vectors a_{11}, $a_{12}\ldots a_{ip.}$ The coefficient of the i^{th} PC are then given as $X_{i,j}$ while the X_i is its variance.
4. Discard any component that only account for a small proportion of the variation in the data.

Let X_1, X_2, X_3… X_P be the variables under study then first principal component may be defined as;

$$Z_1 = a_{11}x_1 + a_{11}x_2 + a_{ip}\, x_p$$

such that variance of Z_1 is as large as possible subject to the condition that;

$$a_{11}^2 + a_{12}^2 + \ldots + a_{ip}^2 = 1$$

This constant is introduced because if this is not done, then $V(Z_1)$ can be increased simply by multiplying any $a_{i,j}$'s by a constant factor. The second principal component is defined as;

$$Z_2 = a_{21}x_1 + a_{22}\, x_2 + \ldots + a_{2p}x_p$$

Such that $V(Z_2)$ is as large as possible next to $V(Z_1)$ subject to the constraint that

$$a_{21}^2 + a_{22}^2 + \ldots + a_{2p}^2 = 1$$

and Cov $(Z_1, Z_2) = 0$ and so on

2.2 The Eigen value-one criterion.

In principal components analysis, one of the most commonly used criteria for solving the number-of-components problem is the Eigen value-one criterion, also known as the Kaiser criterion (1960). The rationale for this criterion is straightforward. Each observed variable contributes one unit of variance to the total variance in the data set. Any component that displays an Eigen value greater than 1.00 is accounting for a greater amount of variance than had been contributed by one variable. Such a component is therefore accounting for a meaningful amount of variance, and is worthy of being retained.

On the other hand, a component with an Eigen value less than 1.00 is accounting for less variance than had been contributed by one variable. The purpose of principal components analysis is to reduce a number of observed variables into a relatively smaller number of components. This cannot be effectively achieved if components that account for less variance than had been contributed by individual variables are retained. For this reason, components with eigenvalues less than 1.00 are viewed as trivial, and are not retained.

2.3 Proportion of Variance Accounted for.

Another Approach is Proportion of Variance Accounted for which involves retaining components that accounts for specified proportion (or percentage) of variance in the data set. For instance, components that accounts for at least 70% to 90% of the total variance are decided to be retained. Note that the computation starts with the highest eigenvalue and continues until the specified and desired proportion is reached. This proportion can be calculated as;

$$proportion = \frac{eigenvalue\ for\ the\ component\ of\ interest}{total\ eigenvalues\ of\ the\ correlation\ matrix}$$

3. Analyses and Discussion of Result:

Data Sets of twelve (12) different variables X_1, X_2,…, X_{12} with cholesterol level in mg as dependent variable. X_1 represents Age, H_2= Height in m, X_3= Weight in kg, X_4= diastolic pressure, X_5= Body Mass Index, X_6= length of the leg, X_{10}= length of the elbow, X_{11}= Wrist in cm and X_{12}= Arm in cm.

3.1 Methodology of the Analysis

The estimated covariance matrix of the data above is given by;

TABLE 1: Variance-covariance matrix.

$$S_{ij} = \begin{pmatrix} 0.1781 & 0.0551 & 0.3757 & 0.1922 & 0.1086 & 0.1768 & 0.1444 & 0.0638 & 0.0205 & 0.0088 & 0.0053 & 0.0618 \\ 0.0551 & 0.1047 & 0.2620 & 0.1389 & 0.1243 & 0.1761 & 0.1055 & 0.0401 & 0.0661 & 0.0104 & 0.0082 & 0.0470 \\ 0.3757 & 0.2620 & 1.9014 & 0.8236 & 0.3377 & 0.7831 & 0.4292 & 0.3035 & 0.1423 & 0.0422 & 0.0263 & 0.2783 \\ 0.1922 & 0.1389 & 0.8236 & 0.4076 & 0.1832 & 0.4034 & 0.2353 & 0.1384 & 0.0705 & 0.0203 & 0.0133 & 0.1322 \\ 0.1086 & 0.1243 & 0.3377 & 0.1832 & 0.2943 & 0.2433 & 0.1674 & 0.0553 & 0.0602 & 0.0121 & 0.0099 & 0.0578 \\ 0.1768 & 0.1761 & 0.7831 & 0.4034 & 0.2433 & 0.5850 & 0.3346 & 0.1314 & 0.0888 & 0.0221 & 0.0152 & 0.1262 \\ 0.1444 & 0.1055 & 0.4292 & 0.2353 & 0.1674 & 0.3346 & 0.2427 & 0.0738 & 0.0450 & 0.0124 & 0.0084 & 0.0714 \\ 0.0638 & 0.0401 & 0.3035 & 0.1384 & 0.0553 & 0.1314 & 0.0738 & 0.0532 & 0.0197 & 0.0070 & 0.0043 & 0.0487 \\ 0.0205 & 0.0661 & 0.1423 & 0.0705 & 0.0602 & 0.0888 & 0.0450 & 0.0197 & 0.0508 & 0.0064 & 0.0052 & 0.0259 \\ 0.0088 & 0.0104 & 0.0422 & 0.0203 & 0.0121 & 0.0221 & 0.0124 & 0.0070 & 0.0064 & 0.0014 & 0.0009 & 0.0073 \\ 0.0053 & 0.0082 & 0.0263 & 0.0133 & 0.0099 & 0.0152 & 0.0084 & 0.0043 & 0.0052 & 0.0009 & 0.0007 & 0.0048 \\ 0.0618 & 0.0470 & 0.2783 & 0.1322 & 0.0578 & 0.1262 & 0.0714 & 0.0487 & 0.0259 & 0.0073 & 0.0048 & 0.0489 \end{pmatrix}$$

where S_{ij} is the sample variance or covariance between variables x_i and x_j .With The aid of Mat lab package, each and every element in the matrix above is multiplied by 1000 .

From the above Variance-Covariance Matrix, table 2 consisting of the correlation matrix is obtained.

$$R = \begin{pmatrix} 1.0000 & 0.4036 & 0.6455 & 0.7134 & 0.4741 & 0.5475 & 0.6946 & 0.6556 & 0.2152 & 0.5666 & 0.4619 & 0.6623 \\ 0.4036 & 1.0000 & 0.5872 & 0.6722 & 0.7083 & 0.7113 & 0.6617 & 0.5368 & 0.9065 & 0.8715 & 0.9339 & 0.6570 \\ 0.6455 & 0.5872 & 1.0000 & 0.9355 & 0.4514 & 0.7425 & 0.6319 & 0.9540 & 0.4580 & 0.8281 & 0.7045 & 0.9124 \\ 0.7134 & 0.6722 & 0.9355 & 1.0000 & 0.5288 & 0.8260 & 0.7480 & 0.9395 & 0.4900 & 0.8613 & 0.7674 & 0.9362 \\ 0.4741 & 0.7083 & 0.4514 & 0.5288 & 1.0000 & 0.5863 & 0.6266 & 0.4421 & 0.4920 & 0.6018 & 0.6734 & 0.4815 \\ 0.5475 & 0.7113 & 0.7425 & 0.8260 & 0.5863 & 1.0000 & 0.8880 & 0.7446 & 0.5151 & 0.7803 & 0.7360 & 0.7457 \\ 0.6946 & 0.6617 & 0.6319 & 0.7480 & 0.6266 & 0.8880 & 1.0000 & 0.6496 & 0.4050 & 0.6794 & 0.6281 & 0.6553 \\ 0.6556 & 0.5368 & 0.9540 & 0.9395 & 0.4421 & 0.7446 & 0.6496 & 1.0000 & 0.3788 & 0.8175 & 0.6962 & 0.9549 \\ 0.2152 & 0.9065 & 0.4580 & 0.4900 & 0.4920 & 0.5151 & 0.4050 & 0.3788 & 1.0000 & 0.7627 & 0.8492 & 0.5199 \\ 0.5666 & 0.8715 & 0.8281 & 0.8613 & 0.6018 & 0.7803 & 0.6794 & 0.8175 & 0.7627 & 1.0000 & 0.9441 & 0.8967 \\ 0.4619 & 0.9339 & 0.7045 & 0.7674 & 0.6734 & 0.7360 & 0.6281 & 0.6962 & 0.8492 & 0.9441 & 1.0000 & 0.7961 \\ 0.6623 & 0.6570 & 0.9124 & 0.9362 & 0.4815 & 0.7457 & 0.6553 & 0.9549 & 0.5199 & 0.8967 & 0.7961 & 1.0000 \end{pmatrix}$$

From Table 2 above we obtain the eigen values of the above correlation matrix as; vv $\lambda_1 = 8.594$, $\lambda_2 = 1.443$, $\lambda_3 = 0.836$, $\lambda_4 = 0.439$, $\lambda_5 = 0.349$, $\lambda_6 = 0.114$, $\lambda_7 = 0.128$,
$\lambda_8 = 0.168$, $\lambda_9 = 0.160$, $\lambda_{10} = 0.637$, $\lambda_{11} = 0.052$, $\lambda_{12} = 0.046$

The computation is by considering the λ's with highest value

i. $$\frac{\lambda_1}{\lambda_1+,...,+\lambda_{12}} = .717 \approx 70\%$$

ii. $$\frac{\lambda_1+\lambda_2}{\lambda_1+,...,+\lambda_{12}} = .8374 \approx 80\%$$

iii. $$\frac{\lambda_1+\lambda_2+\lambda_3}{\lambda_1+,...,+\lambda_{12}} = .9071 \approx 90\%$$

From the above result, it means that the first three Z's could replace 12 variables by sacrificing negligible information about the total variation in the system. The scores of the principal components can be obtained by substituting the values of X's in equation of Z_i's denoted by Z_1, Z_2, and Z_3.

Z_1 = $(-0.235)X_1 + (-0.2900)X_2 + (-0.3019)X_3 + (-0.3203)X_4 + (-0.2338)X_5 + (-0.2987)X_6 + (-0.2777)X_7 + (-0.2997)X_8 + (-0.2342)X_9 + (-0.3256)X_{10} + (-0.3101)X_{11} + (-0.3142)X_{12}$

Z_2 = $(0.3567)X_1 + (-0.4167)X_2 + (0.2558)X_3 + (0.2179)X_4 + (0.2288)X_5 + (0.0479)X_6 + (0.1033)X_7 + (0.3157)X_8 + (-0.5388)X_9 + (-0.1138)X_{10} + (-0.2888)X_{11} + (0.1860)\ X_{12}$

Z_3 = $(-0.3381)X_1 + (-0.0410)X_2 + (0.2599)X_3 + (0.1054)X_4 + (-0.5408)X_5 + (-0.2068)X_6 + (-0.4956)X_7 + (0.2421)X_8 + (0.2153)X_9 + (0.1890)X_{10} + (0.1246)\ X_{11} + (0.2612)X_{12}$

However, using Eigen value one criterion, only two PC's with corresponding Eigen values greater than one can be retained. Equation of Z_i's of the corresponding PC's are:

Z_1 = $(-0.235)X_1 + (-0.2900)X_2 + (-0.3019)X_3 + (-0.3203)X_4 + (-0.2338)X_5 + (-0.2987)X_6 + (-0.2777)X_7 + (-0.2997)X_8 + (-0.2342)X_9 + (-0.3256)X_{10} + (-0.3101)X_{11} + (-0.3142)X_{12}$

Z_2 = $(0.3567)X_1 + (-0.4167)X_2 + (0.2558)X_3 + (0.2179)X_4 + (0.2288)X_5 + (0.0479)X_6 + (0.1033)X_7 + (0.3157)X_8 + (-0.5388)X_9 + (-0.1138)X_{10} + (-0.2888)X_{11} + (0.1860)\ X_{12}$

4. Conclusion

The computation using the proportion of variance accounted for shows that, 3 PC's are to be retained for further analysis which yields or accounts for most of the variance of the original data set. A model was fitted using the three retained PC's and coefficient of determination was used to measure the adequacy of the fitted model; it was found that R^2 (coefficient of determination) is 22.7% and R^2 (adjusted) is 16.9%. Similarly, it also shows that 2PC's are to be retained for Eigen value one criterion and R^2= 23.3% and R^2(adjusted) = 18.5% respectively. The adequacy of the model conclude comparatively that Eigen value one criterion approach of data reduction techniques of principal component analysis is more suitable and preferable in carrying out an analysis of data with a high dimension.

REFERENCES

[1] Alvin C, Rencher. (2002): Method of Multivariate Analysis, Second Edition, *A John Wiley and sons, Inc. publications), page 112.*

[2] Anderson T.W.(1963). Asymptotic theory for principal component analysis, *Ann. Math. Statist.,34:122-148.*

[3] Gower J.C. (1966) Some distance properties of latent root and vector methods used in multivariate data analysis. *Biometrikal 53: 315-328.*

[4] Hoteling , H, (1933) Analysis of complex statistical variables into principal components *journal of educational psychology, volume 4: 419-441 and 498-520 (10.97/year).*

[5] I.T.Jolliffe (2002): Principal Component Analysis, Second Edition, *Springer.*

[6] JEFFER, J.N.R., (1967). Two case studies in the application of principal component analysis. *Journal of Applied statistics.*

[7] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6): 559-572.

[8] Rao .C.R., (1964). The use and Interpretation of Principal Component Analysis in Applied Research. *Sankhya A 26:329-58,*

[9] T. W Anderson (1950): An introduction to multivariate statistical analysis. *A wiley publication in mathematical statistics; New-york. London and Sydney.*