Principal Components Analysis a method Useful in Identification of Clusters of Variables

Dr. S.V.Kakade[1], Dr. Mrs. J.A.Salunkhe[2], Dr V.B.Jagadale[3], T.S.Bhosale[4]
[1]Associate Professor of Statistics, Krishna Institute of Medical Sciences, Karad, Maharashtra
[2]Professor, Krishna Institute of Nursing Sciences, Karad, Maharashtra
[3]Professor & Head, Dept. of Statistics, Yashwantrao Chavan College of Science, Karad, Maharashtra
[4]Statistician, Office of Director of Research, Krishna Institute of Medical Sciences Deemed University, Karad, Maharashtra.

Abstract:- There are many different methods that can be used to conduct a factor analysis which is a data reduction or structure detection method. The commonly used method for factor analysis is 'Principal Components Analysis (PCA)'. The principal components account for most of the variance in the original variables.

The data on some baseline variables and a 15 questions about 'Attitude towards female feticide' measured on Likert scale was collected from women admitted for delivery in KH&MRC, Karad; a teaching hospital. Principal components were extracted by using varimax rotation method. Components with eigenvalue ≥ 1.00 were identified as new (latent) variables.

The PCA derived six components. It revealed that original variables in each component were inter-related with each other.

The application of PCA is useful in detection of latent variables whose original variables are technically inter-related. This helps in identification of factors whose participant variables are measuring the entity which cannot be directly measured and also provides major contribution in the study.

Key words:- *Factor analysis, varimax rotation, Likert scale*.

I. INTRODUCTION

Study of linear combinations of variables is useful for predicting the category of the qualitative dependent variable or predicting the amount of the quantitative dependent variable in a multivariate space. Linear combinations of variables are also useful for characterizing or accounting the variation i.e. spread of each dimension in a multivariate space. Principal components analysis (PCA) does this by identifying the linear combinations sequentially such that the first linear combination of variables accounts for the largest amount of variation in the sample; the second for the next largest amount of variance in a dimension independent of the first, and so on.[1] Thus successive components explain smaller and smaller quantity of the total variance and are independent of each other.

PCA is one of the methods of factor analysis. These methods are often used in exploratory data analysis to:

- Study the correlations among the variables by grouping them into a few factors (components). The variables within each component are highly correlated with each other.
- Interpret each component according to the meaning of the variables highly contributing that component.

Thus PCA helps to simplify and understand the structure of correlation or covariance matrix.

Basically, the principal components are extracted by rotating the original variable space using varimax (variance maximizing) rotation method.[2] This type of rotation is called variance maximizing because the criterion for the rotation is to maximize the variance (variability) of the "new" variable (factor or component), while minimizing the variance around the new variable.

PCA, a popular data processing and dimension reduction technique, has numerous applications in engineering, biology and social science.[3] PCA has been used in gene expression data analysis also.[4]

For getting valid result from PCA, the data must fulfill following assumptions[5]: 1. Have multiple variables that should be measured at the continuous level; although ordinal variables are very frequently used._____Ordinal variables commonly used in PCA include a wide range of Likert scales. 2. Data needs to have a linear relationship between all variables. In practice, this assumption is somewhat relaxed when variables contain ordinal data. 3. Should have sampling adequacy; i.e. large enough sample sizes are required to produce reliable result. 4. Need to have adequate correlations between the variables in order to reduce variables into a smaller number of components. And 5. There should be no significant outliers.

The present study was undertaken to demonstrate computational procedure of PCA in as simple as possible manner to the researchers as well as to specify them utilization of PCA in determining clusters i.e. components that can be used in the further analysis for better analysis as well as interpretation of the study.

II. MATERIAL AND METHODS

The data on some baseline variables (Age, Education of mother, Marriage years, Gravida and Family income) and a 15 questions assessing 'Attitude towards female feticide' (Table 1) measured on Likert scale was collected from 500 multi-gravid women admitted for delivery in KH&MRC, Karad; a teaching hospital during six months period.

Attitude No.	Description of Attitude
At1	I am having right to have male or female child
At2	I feel that it is our fundamental right to have a male child in our family
At3	I feel that T.V. and radio are good means to provide information about female feticide.
At4	I should undergo USG for sex detection
At5	I feel that there is difference between girl child and boy child.
At6	If I come to know that I have female child in my womb I should undergo abortion
At7	I should do sex determination in second pregnancy if I am having first female child.
At8	I feel that I should have more deliveries till I get male child.
At9	I feel that female feticide is violence against women
At10	I feel that first birth of daughter should not be avoided.
At11	I feel that selection of sex practice is not only common among the uneducated.
At12	I feel that woman alone cannot play an active role in reducing gender discrimination.
At13	I believe that religious and socio economic factors are responsible for sex determination.
At14	I feel that aborting the fetus is crime
At15	I feel that modern technology is responsible for killing unwanted baby girls

Table 1. Attitudes towards female feticide

SPSS Version 20 was used to carry out principal component analysis.[5] In the beginning univariate descriptive analysis of all study variables was carried out with purpose of confirming how many cases were actually going to be included in the principal components analysis. Further the correlations (Pearson's correlation coefficient) between the variables were computed as very high or very low correlation creates burden on entire analysis. Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy and Bartlett's Test of Sphericity were carried out to assess applicability of principal components analysis. Principal components were detected on basis of eigenvalue i.e. those who were having eigenvalue ≥ 1.00. Using the Varimax rotation method the total variance accounted by these components was maximized component wise and it was redistributed over each of these components. These principal components were the new (latent) variables. Further Rotated Component Matrix was developed to know the original variables those generated the principal components. Finally they were properly named by considering the group of original variables.

III. RESULTS

In all total 500 multi-gravid women participated in the study. The univariate analysis revealed that the descriptive statistics for each study variable was determined from data of these all 500 women. This indicated that there was not a single missing value found on any of the variable. Thus in the analysis there was no loss of data of a single case (woman).

The correlations between the variables were checked before conducting a principal components analysis with the concept that: 1. the high correlations (above .9) indicate two variables seem to be measuring the same thing, under circumstances it was better to remove one of the variables from the analysis. Another alternative would be to combine the variables in some way (perhaps by taking the average). 2. If the correlations were too low, say below 0.1, then one or more of the variables might load only onto one principal component; in other words, make its own principal component. This was not helpful, as the whole point of the analysis is to reduce the number of variables. In the present analysis it was observed that At1 (I am having right to have male or female child) and At3 (I feel that T.V. and radio are good means to provide information about female feticide) were showing low correlations with most of the variables.

The eligibility of data, measured on various study variables, for determination of principal components was the essential step before conducting PCA. This was achieved by employing Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy and Bartlett's Test of Sphericity. Principally these both should be carried out after deleting variables from data set those are non useful as per characteristics of correlation matrix. Theoretically the KMO Measure of Sampling Adequacy varies between 0 and 1. The values closer to 1 are better. A value of 0.6 is a suggested minimum. The KMO value in present study, in presence of all study variables, was 0.670. The Bartlett's Test of Sphericity tested the null hypothesis that 'the correlation matrix is an identity matrix' which was expected to be rejected. This test revealed its chi-square value of 2054.604 with $p<0.001$. Hence null hypothesis was rejected. These two tests, together, passed the minimum standard; hence the data was suitable to be used to conduct principal components analysis.

The principal component analysis, on the basis of all study variables, revealed 7 components with eigenvalue >1. The percentage of variance accounted by these 7 components when all variables were introduced in the analysis was

61.938%. While after deletion of At1 and At3, alone, as they resulted into their own components, the variances accounted were 58.391% and 58.601%, respectively. However, when these both variables were removed from the analysis the variance accounted by resultant 6 components was 60.126% which sacrificed only 1.812% of variance (Table 2). During this trial every time KMO Measure of Sampling Adequacy and Bartlett's Test of Sphericity was carried out. When both At1 and At3 were removed from data set KMO value was 0.695 and chi-square value was 1811.049 with p<0.001. This indicated data was suitable to carry the principal components analysis. Thus the final analysis was carried out by removing attitudes At1 and At3.

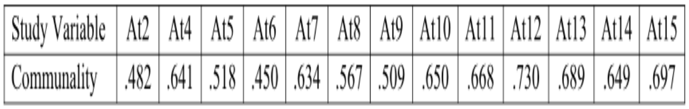
Component	Initial Eigenvalues		
	Total*	% of Variance	Cumulative %
1	3.481	19.337	19.337
2	1.812	10.066	29.403
3	1.616	8.977	38.380
4	1.474	8.189	46.569
5	1.344	7.468	54.037
6	1.096	Study Variable Communality 6.089	60.126
7	0.970	5.391	65.516
8	0.818	4.545	70.061
9	0.753	4.184	74.246
10	0.703	3.907	78.153
11	0.675	3.749	81.902
12	0.587	3.259	85.160
13	0.559	3.105	88.265
14	0.518	2.879	91.144
15	0.504	2.801	93.945
16	0.420	2.332	96.277
17	0.358	1.986	98.263
18	0.313	1.737	100.00

Table 2. Percentage variance accounted by the principal components.

- *Component wise total variance accounted*

The proportion of almost all variable's variance that can be explained by these 6 principal components which is the 'sum of squared factor loadings' also known as 'communalities' was more than 50% (Table 3).

Study Variable	Age	Education Mother	Marriage years	Gravida	Family Income
Communality	.641	.587	.562	.502	.648



Study Variable	At2	At4	At5	At6	At7	At8	At9	At10	At11	At12	At13	At14	At15
Communality	.482	.641	.518	.450	.634	.567	.509	.650	.668	.730	.689	.649	.697

Table 3. Communalities of study variables based on 6 principal components.

The number of components extracted during principal components analysis was equal to the number of variables that were put into analysis. In present study, 18 variables were put into the analysis, so there were 18 components. The sum of variance accounted by these components was also 18 (sum of column 'Total' in Table 2). The first component accounted for the most variance, eigenvalue, 3.481 out of 18, which contributed 19.337% of the variance explained by all 18 components (Table 2). Next component accounted as much of the left variance, 18 - 3.481 = 14.519, which was 1.812, and so on. Each successive component accounted for less and less variance. Varimax rotation maximized the variance of each of the components. The total amount of variance accounted for was redistributed over the six extracted components. The panel of 'Rotation Sums of Squared Loadings' represents the component wise distribution of the variance after the varimax rotation. (Table 4)

In fact the basis of selection criterion of principal components could be the eigenvalue or scree plot.

The scree plot graphed the eigenvalue against the component number (Fig 1). From the second factor onwards, the line was almost flat i.e the slope showed more or less similar declining trend, meaning the each successive factor was accounting for smaller and smaller amounts of the total variance. This revealed that only two components were identified by the scree plot method. However, it could account only 29.403% of the total variance. There was much disparity between eigenvalue criterion and scree plot criterion in deciding the number of components. But as eigenvalue criterion of >1 had good accountability of the variance (60.126%) and number of components identified was acceptable, components were decided to be finalized accordingly. The scree plot, independently, showed its inability to detect/identify enough number of components.

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.481	19.337	19.337	3.481	19.337	19.337	2.896	16.088	16.088
2	1.812	10.066	29.403	1.812	10.066	29.403	1.970	10.947	27.035
3	1.616	8.977	38.380	1.616	8.977	38.380	1.564	8.688	35.722
4	1.474	8.189	46.569	1.474	8.189	46.569	1.522	8.457	44.179
5	1.344	7.468	54.037	1.344	7.468	54.037	1.500	8.334	52.513
6	1.096	6.089	60.126	1.096	6.089	60.126	1.370	7.612	60.126
7	.970	5.391	65.516						
8	.818	4.545	70.061						
9	.753	4.184	74.246						
10	.703	3.907	78.153						
11	.675	3.749	81.902						
12	.587	3.259	85.160						
13	.559	3.105	88.265						
14	.518	2.879	91.144						
15	.504	2.801	93.945						
16	.420	2.332	96.277						
17	.358	1.986	98.263						
18	.313	1.737	100.00						

Table 4. Eigenvalues of principal components - initial and after varimax rotation

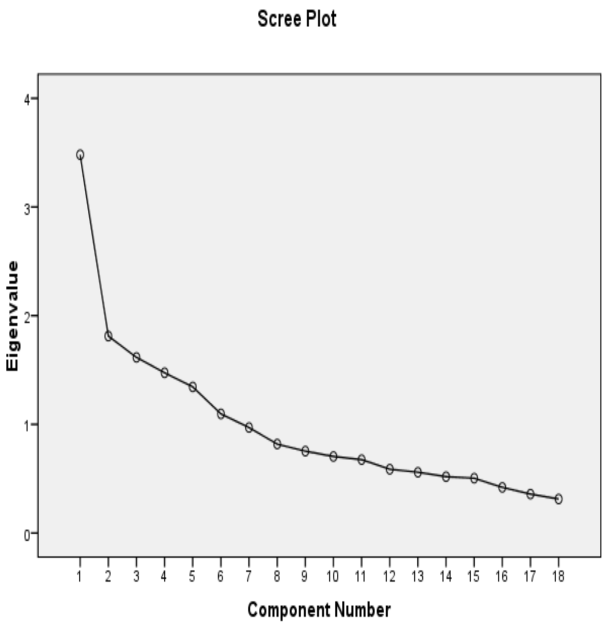


Fig 1:- Scree Plot shows component wise eigenvalues.

Rotated Component Matrix (Table 5) contains the rotated component loadings, which represented both how the variables are weighted for each component and also the correlation between the variables and the component. Because these are correlations, possible values range from -1 to +1. The correlations that are above +0.5 were presented in dark as they were considered to be strongly correlated. This makes the output easier to read by removing the clutter of low correlations that are probably not meaningful anyway. However, these were the components i.e. latent variables that were required to be determined to understand un-majorable entity for further utility; instead of original variables; which improves prediction of the outcome. These original variables (showed in dark) should be considered as major contributors for creating the respective component. These variables revealed the logically as well as statistically inter-relationship amongst them. Finally these components were named meaningfully in view of the contributing variables (Table 6).

Study Variable	Component					
	1	2	3	4	5	6
At4	.790	.051	.049	-.027	-.088	-.059
At7	.738	.105	.148	.003	-.031	.236
At5	.672	.192	.076	.126	-.043	.072
At2	.655	.003	-.177	-.075	-.041	-.118
At8	.621	.234	.102	.108	.222	.235
At10	.119	.776	-.133	.066	.095	-.044
At14	.222	.709	.256	-.041	-.125	-.118
At6	.422	.490	.053	-.065	.048	.150
At13	.090	.008	.808	.164	-.009	.030
At15	.018	-.284	-.644	.253	.057	.367
Age	.076	-.057	.079	.772	.158	-.071
Mar_years	.049	.010	-.030	.731	-.148	-.050
Gravida	-.239	.282	-.007	.473	-.310	.213
Edu_mother	.124	-.045	-.094	-.064	.741	-.084
Family_incm	-.168	.329	.219	.078	.667	.110
At 9	-.269	-.382	-.112	-.127	.510	.045
At12	.130	.035	-.148	-.068	-.052	.826
At11	.111	-.313	.491	-.049	.090	.554

Table 5. Rotated Component Matrix

IV. DISCUSSION

When the number of variables is very large, it may be advantageous to find sets of linear combination of variables (latent variables) having some properties in terms of correlation, covariance or variance.[6]

The essential purpose of PCA is to describe the covariance relationship among many variables in terms of unobservable variables called principal components or factors.[7] The principal component model motivates to argue that: Variables can be grouped by their correlations. Hence all variables in the particular group are highly correlated among themselves and have relatively small correlations with variables in other group(s). Thus it signifies that each group of variables represents a single underlying construct or component or factor, which is responsible for the observed correlations.

The major advantage of PCA is that; the components formed from the variables, represents the parameters that are difficult or hard to measure or cannot be measured directly. Hence, PCA cannot be carried out only to reduce the data dimensions but also to quantify inestimable variable(s) which is/are beyond the capacity of existing standard measure(s).

The important step in PCA is to choose principal components. This should be achieved by comparative assessment of eigenvalue criterion and scree plot criterion. However, the total variance accounted by selected components should not be less than 60% for better consequences. The components with eigenvalue ≥ 1 should be taken up along with the graph displayed by scree plot. The components selected on the basis of scree plot criterion are those which are at and before the flatness starts in the graph. If these all requirements are satisfying, finalize the analysis. Otherwise, select components fulfilling any one of the eigenvalue or scree plot criterion along with criterion of total variance. The original variables may be replaced by finalized components and analysis may be repeated. On the basis of comparative assessment of results viz. percentage of correct predictability of the outcome, the researcher could decide whether the study should be concluded using original variables or using the components.

V. CONCLUSION

It is difficult to infer relationship or impact of large number of independent (study) variables on the outcome (dependent) variable. Also sometimes it is difficult or hard or impossible to measure the variable(s) required to be studied. Such variable(s) can be measured indirectly i.e. by measuring some other associated variables. In the first situation principal component analysis helps in reducing the number of study variables by detecting new variables known as components. However, in second situation the component generated by principal component analysis suggests which associated variables can be used to estimate the target variable.

Attitude	Component Name (New variable)
At 4: I should undergo USG for sex detection	Male favouring attitude
At 7: I should do sex determination in second pregnancy if I am having first female child.	
At 5: I feel that there is difference between girl child and boy child.	
At 2: I feel that it is our fundamental right to have a male child in our family	
At 8: I feel that I should have more deliveries till I get male child.	
At 10: I feel that first birth of daughter should not be avoided.	Pregnancy continuation attitude
At 14: I feel that aborting the fetus is crime	
At 13: I believe that religious and socio economic factors are responsible for sex determination.	Interruption in female live birth
At 15: I feel that modern technology is responsible for killing unwanted baby girls	
Age	Marital life
Marriage Years	
Education of Mother	Advanced life style & thoughts
Family Income	
At 9: I feel that female feticide is violence against women	
At 12: I feel that woman alone cannot play an active role in reducing gender discrimination.	Traditional thoughts about new birth
At 11: I feel that selection of sex practice is not only common among the uneducated.	

Table 6. New (Latent) Variables [Principal Components]

REFERENCES

[1]. Diana D. Suhr, Ph.D. University of Northern Colorado. Paper 203-30: Principal Component Analysis vs. Exploratory Factor Analysis. (http://www2.sas.com/proceedings/sugi30/203-30.pdf).

[2]. Principal Components and Factor Analysis. www.uta.edu/faculty/sawasthi/Statistics/stfacan.htm

[3]. Hui Zou, Trevor Hastie and Robert Tibshirani. Sparse Principal Component Analysis. Journal of Computational and Graphical statistics, 2006; 15(2), 265-286.

[4]. Alter O, Brown P and Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. In proceedings of the National Academy of Sciences (2000), 97, 10101-10106.

[5]. Principal Components Analysis (PCA) using SPSS Statistics. https://statistics.laerd.com/spss-tutorials/principal-components-analysis-pca-using-spss-statistics.php.

[6]. Luigi D'Ambra, Pietro Amenta, and Michele Gallo. Dimensionality Reduction Methods. Metodoloski zvezki, Vol. 2, No. 1, 2005, 115-123.

[7]. Principal component and Factor Analysis. https://www.linkedin.com/pulse/principal-component-factor-analysis-fayaz-ahmad.