African Journal of Mathematics and Computer Science Research Vol. 3(3), pp. 049-053, March 2010
Available online at http://www.academicjournals.org/AJMCSR ISSN 2006-9731

Full Length Research Paper

Principal component procedure in factor analysis and robustness

Joy Chioma Nwabueze

Department of Statistics, Abia State University, Uturu, Abia State, Nigeria. E-mail: teeubueze@yahoo.co.uk.
Tel: +2348068164190.

Accepted 6 November, 2009

Principal component procedure has been widely used in factor analysis as a data reduction procedure. The estimation of the covariance and correlation matrix in factor analysis using principal component procedure is strongly influenced by outliers. This study investigates the robustness of principal component procedure in factor analysis by generating random variables from five different distributions which are used to determine the common and specific factors in factors analysis using principal component procedure. The results revealed that the variance of the first factor was widely distributed from distribution to distribution ranging from 0.6730 to 5.9352. The contribution of the first factor to the total variance varied widely from 15 to 98%. We conclude that the principal component procedure is not robust in factor analysis.

Key words: Principal component, factor analysis, robustness, random variables, distributions.

INTRODUCTION

Factor analysis performs a decomposition of the data matrix into a matrix of loadings which describes the connections between the variables and the new co-ordinate system and a matrix of factors scores which consists of the variable values in the new co-ordinate system (Filzmoser, 1999). Factor analysis has been used in empirical researches as a statistical tool that can be used to analyze interrelationships among a large number of variables and to explain these variables in terms of their common understanding dimensions (factors) with a minimum loss of information. Factor analysis can be used to refer to a class of models that include ordinary principal components, weighed principal components, maximum likelihood factors analysis, certain multi dimensional scaling models and others.

Factors analysis has been used in recent works by several authors such as Budweiser et al. (2005), who used factor and discriminant analysis to find the long term reduction of hyperinflation in stable chronic obstructive pulmonary disease (COPD) by non invasion nocturnal and factor analysis which are sensitive to outliers.

Dutter (1987) has shown a lot of possibilities to robust classical multivariate methods and has also applied these techniques to the analysis of geostatistical variables. He went further to show that the estimation of the covariance and correlation matrix are strongly influenced by outliers. As a consequence, the estimation of Eigen vectors and Eigen values is also strongly dependent on outliers in the data. Further errors might appear with the estimation of the mean vector which is necessary for both the centering of the data and the calculation of the classical covariance matrix.

This study investigates the robustness of principal component method in factor analysis by using artificial data generated independently from five distributions namely: normal, uniform exponential, Laplace and Gamma distributions. Robustness is the quality of being able to withstand stresses, pressures or changes in procedure or circumstance. A system, organization or design may be said to be robust if it is capable of coping well with variations (Sometimes unpredictable variations) in its operating environment with minimal damage.

A robust statistical techniques is one that performs well even if its assumptions are somewhat isolated by the true model from which the data is generated (Wikipedia, 2009). There are different degrees of robustness. A measure for the determination of the robustness of an estimator is given by the breakdown value (Donoho and Huber, 1983). It is defined as the minimum proportion of contaminated data which causes the estimator to give arbitrary values. Nwabueze et al. (2009) investigated the robustness of the maximum likelihood method of estimation in factor analysis. Their findings showed that maximum likelihood method of estimation is robust in factor analysis.

METHODOLOGY/ EXPERIMENTAL DESIGN

Correlation and factor loadings

Random variates were generated independently from the five distributions used in the study. Five response variables were generated using a sample size of 200. The experiment was replicated fifty times for each of the five distributions. These random variates were used to calculate the correlation matrix for the five distributions. These correlation matrices were used to calculate matrix of constants (factor loadings) and the ith specific factor which is only associated with the ith response for each of the five distributions. The contribution of total sample variance due to the first and second factors was also obtained.

Factor analytical model

The factor analytic model is given by

$$\begin{aligned} X_1 - \mu_1 &= L_{11}F_1 + L_{12}F_2 + \ldots L_{im}F_m + \delta_1 \\ X_2 - \mu_2 &= L_{21}F_1 + L_{22}F_2 + \ldots + L_{2m}F_m + \delta_2 \\ \vdots \qquad & \quad \vdots \qquad \vdots \qquad\qquad \vdots \qquad \vdots \\ X_p - \mu_p &= L_{pi}F_1 + L_{p2}F_2 + \ldots + L_{pm} + \delta_p \end{aligned} \quad (1)$$

Which in matrix form is equivalent to:

$$X - \mu_{(PX1)} = L_{(PXM)}F_{(mX1)} + \sum(px1) \quad (2)$$

Where; *X* is the observable random vector with p components, μ is the mean of *X*, *L* is the loading of the ith variable on the jth factor. *F* and *m* are unobservable random variable called the common factors while δ are p additional sources of variation called error or sometimes specific factors.

Principal component (principal factor) method

In factor analysis using principal component procedure, the correlation matrix can be factored out using the spectral decomposition theorem.

$$\sum = \lambda_1 \ell_1 \ell_1' + \lambda_2 \ell_2 \ell_2' + .. + \lambda_p \ell_p \ell_p' \quad (3)$$

$$\sum = \left(\sqrt{\lambda_1 \ell_1}\sqrt{\lambda_2 \ell_2} \ldots \sqrt{\lambda_p \ell_p}\right) \begin{pmatrix} \sqrt{\lambda_1 \ell_1'} \\ \sqrt{\lambda_2 \ell_2'} \\ \sqrt{\lambda_p \ell_p'} \end{pmatrix} \quad (4)$$

Where $\sum$ is the covariance matrix of the compositions, $(\lambda_1 \ell_1)$ is the eigenvalue-eigenvector pair of $\sum$ and $\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_p \geq 0$.

Eigen values are the new variances gotten when the un-rotated or old variances have been rotated and are termed component loadings which are used to calculate the component scores. When the specific factors are included in the model and their variables are taken to be the diagonal elements, the approximation becomes:

$$\sum = LL' + \Psi \quad (5)$$

$$\sum = \left(\sqrt{\lambda_1 \ell_1}\sqrt{\lambda_2 \ell_2} \ldots \sqrt{\lambda_p \ell_p}\right) \begin{pmatrix} \sqrt{\lambda_1 \ell_1'} \\ \sqrt{\lambda_2 \ell_2'} \\ \vdots \\ \sqrt{\lambda_p \ell_p'} \end{pmatrix} + \begin{pmatrix} \Psi_1 & 0\ldots\ldots\ldots 0 \\ 0 & \Psi_2\ldots\ldots\ldots 0 \\ \vdots & \vdots \quad \vdots \\ 0 & \Psi p \quad 0 \end{pmatrix} \quad (6)$$

Where Ψ_i is the variances or error from the compositions which is the specific variance and L is the matrix of the factor loadings. The proportion of total sample variance $S_{11} + S_{22} + .. + Spp = trS$ from the first and common factor is then:

$$L_{11}^2 + L_{21}^2 + \ldots + L_{p1}^2 = \left(\sqrt{\hat{\lambda}_1 \hat{\ell}_1}\right)\left(\sqrt{\hat{\lambda} \hat{\ell}_1}\right) = \hat{\lambda}_1 \quad (7)$$

Since the eigenvector ℓi has unit length, in general, the proportion of total sample variance due to the jth factor is $\dfrac{\hat{\lambda}_j}{S_{11} + S_{12} \ldots + S_{pp}}$ for a factor analysis of S.

DISTRIBUTIONS USED IN THE STUDY

Normal distribution

A random variable is said to follow a normal distribution with mean μ and variance σ^2 if the probability density function is given by:

$$f\left(X / \mu, \sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}} \ell' \frac{(X - \mu)^2}{2\sigma^2} - \infty < X < \infty$$

Exponential distribution

An exponential distribution with parameter λ has the probability function given as:

$f(X) = \lambda \ell^{-\lambda X}$, X > 0.

Uniform distribution

A continuous random variable has a uniform distribution if and if its probability density is given by:

$$f(X, \alpha, \beta) = \frac{1}{\beta - \alpha} \qquad \alpha < X < \beta$$

Table 1. Correlation matrix R_1 from the normal distribution.

	X1	X2	X3	X4	X5
X1	1				
X2	0.1980	1			
X3	-0.1808	-0.0880	1		
X4	-0.3360	-0.5839	-0.0262	1	
X5	-0.2272	0.1745	-0.0106	0.2070	1

Table2. Correlation matrix R_2 from exponential distribution.

	X1	X2	X3	X4	X5
X1					
X2	-0.1840	1			
X3	-0.3127	0.0585	1		
X4	0.2070	0.0992	0.0846	1	
X5	0.1125	-0.1032	-0.0880	0.5839	1

Table 3. Correlation matrix R_3 from uniform distribution.

	X1	X2	X3	X4	X5
X1	1				
X2	0.0585	1			
X3	-0.1840	0.1980	1		
X4	0.1125	-0.5839	-0.2503	1	
X5	-0.0688	0.1745	-0.0391	0.0393	1

Gamma distribution

A random variable X has a Gamma distribution if its probability density is given by:

$$f(X) = \frac{1}{\beta^x \Gamma_{(\alpha)}} \chi^{\alpha-1} \ell^{-X/\beta} X > 0$$

Where $\alpha > 0$ and $\beta > 0$.

Laplace distribution

A random variable has a Laplace (μ , *b*) distribution if its probability function is given by:

ANALYSIS OF DATA

The analysis of the data generated from the distribution using Monto Carlo was preformed using MINITAB statistical software. The correlation matrix for the five distributions were obtained and displayed on Tables 1 - 5.

Table 4. Correlation matrix R_4 from gamma distribution.

	X1	X2	X3	X4	X5
X1	1				
X2	0.1543	1			
X3	0.6370	0.2637	1		
X4	0.1170	0.4370	0.2781	1	
X5	0.1830	0.5432	0.2190	0.5043	1

Table 5. Correlation matrix R5 from Laplace distribution.

	X1	X2	X3	X4	X5
X1	1				
X2	0.0872	1			
X3	0.2774	0.0576	1		
X4	0.3987	0.0593	0.0992	1	
X5	-0.0861	-0.1629	0.2070	-0.01189	1

RESULTS AND DISCUSSION

Table 1 shows the correlation matrix from the random variates generated from the normal distribution. It is observed that the variable X_1 is negatively correlated with other variables except X_2. The highest negative correlation of 58% was recorded between X_2 and X_4 while a highest positive correlation of 21% was recorded between X_4 and X_5. The correlation matrix between the variables drawn from the exponential distribution is displayed on Table 2.

$$f\left(X/\mu,b\right) = \frac{1}{2b}\begin{cases} \exp\left[\dfrac{-\mu - x}{b}\right] & \text{if } X < \mu \\ \exp\left[\dfrac{-X-\mu}{b}\right] & \text{if } X \geq \mu \end{cases}$$

The variable X_1 is negatively correlated with X_2 and X_3 while it positively correlated with X_4 and X_5. The variable X_2 positively correlated with X_3 and X_4 but negatively correlated with X_5. A significant negative correlation coefficient of 58% was recorded between the variables X_4 and X_5. Table 3 showed that correlation matrix calculated from the random variates generated from the uniform distribution. A least positive correlation coefficient of 6% was recorded between X_1 and X_2 while X_1 recorded a highest negative correlation of 18% between X_1 and X_3. The variable X_3 is negatively correlated with both X_4 and X_5 while X_4 is positively correlated with X_5.

The correlation matrix of the random variates drawn from the Gamma distribution is displayed on Table 4. The variables generated from the Gamma distribution showed positive correlation between the variables. Variable X_1 recorded the highest positive correlation coefficient of 64%

Table 6. Rotated factor loading for the five distributions.

Variable	Normal		Exponential		Uniform		Gamma		Laplace	
	1	2	1	2	1	2	1	2	1	2
X_1	0.9324	0.4321	0.5434	0.2462	0.9763	0.3654	0.5321	0.0346	0.6511	0.1834
X_2	0.7265	0.8734	-0.4326	0.0183	0.8996	0.4392	0.6361	0.0381	-0.5372	0.0674
X_3	-0.7632	0.4326	0.2342	-0.3589	-0.7693	0.1245	0.7231	0.0219	-0.3515	0.0321
X_4	0.6500	-0.3214	-0.1056	0.7234	0.9672	0.0568	0.0632	0.0451	0.3688	0.0986
X_5	0.4631	-0.2143	0.2645	-0.4683	0.8163	0.0334	0.0792	0.0321	0.7614	0.1326
Variance	1.8321	0.8542	0.6730	0.3657	0.1832	0.2615	5.9352	2.3452	2.3333	0.8264
% Contribution of variance	0.9752	0.4983	0.5957	0.8981	0.6920	0.0605	0.1474	0.0008	0.4911	0.0210

with X_3 and the least positive correlation coefficient of 12% with X_4. Variable X_2 had the highest positive correlation of 74% with X_4 and a least positive correlation coefficient of 26% with X_3. X_3 had a positive correlation coefficient of 28 and 22% with X_4 and X_5, respectively.

The correlation matrix calculated from the random variates generated from Laplace distribution is shown in Table 5. The variable X_1 positively correlated with variables X_2, X_3 and X_4 while it negatively correlated with X_5.

Rotated Factor loading for the five distributions pre-sented in Table 6 contained the entire rotated component matrix. The idea of rotation is to reduce the number of factors to only these factors on which the variables under investigation have high loading. For the purpose of comparison of the five distributions, the first two factors are retained for each of the distribution. In the normal distri-bution (Table 6), factor 1 is made up of variables X_1, X_2 and X_3 because of their high loadings while factor 2 is made up mainly of variables X_2 because of its high factor loading. In the exponential distribution the first factor is made up of mainly variable X_1 while factor 2 is made up mainly of variable X_4.

The five variables in the uniform distribution contributed significantly for factor 1. Variables X_1, X_2 and X_3 contri-buted significantly to factor 1 in the Gamma distribution. Factor 1 of the Laplace distribution was made up mainly of variables X_1, X_2 and X_5. Table 6 also showed the percentage contribution of the variance f each factor in the five distributions.

Using the variance for each factor and the factor loadings the contribution of the total sample variance due to each of the factors was obtained for the distributions. For the normal distribution, contribution of total sample variance due to the first factor is:

$$\frac{0.9324^2 + 0.7265^2 + 0.7632^2 + 0.6500^2 + 0.4631^2}{1.8321 + 0.8542} \times \frac{100}{1} = 98\ \%$$

The contribution of total sample variance due to second factor for the normal distribution is:

$$\frac{0.4321^2 + 0.8734^2 + 0.4326^2 + 0.3214^2 + 0.3143^2}{1.8321 + 0.8542} \times \frac{100}{1} = 50\ \%$$

The contribution of the total sample variance due to first and second factors in the other distribution were obtained and displayed in Table 6.

For factor 1, the contributions of the total variance are 98, 90, 69, 15 and 49% for normal exponential, uniform, Gamma and Laplace distributions, respectively. For factor 2, the contributions of the total variance are 50, 60, 6, 1 and 2% for normal, exponential, uniform, Gamma and Laplace distributions, respectively. The result of the analysis showed that the contribution of the first and second factors to the total variance differ much from distribution to distribution. These contributions from the five distributions were widely distributed and do not fall within the range.

Conclusion

The contributions of the first and second factors to the total sample variance using principal component procedure differ from distribution to distribution very widely. The results are sensitive to the distributions used in the study. We conclude that the principal component procedure on factor analysis is not robust to all the distribution considered.

REFERENCES

Budweiser S, Heinermann F, Fischer W, Laub M (2005). Long term Reduction of Hyper inflation in Stable Chronic Obstructive Pulmonary Disease by Non-invasive noctural Home Ventilation. Resprimed pp 976-984

Donoho DL, Huber PJ (1983). The notion of breakdown point. In a Fests Chrift for Erich Lehmann, eds P. Bickel, K. Doksum and J.L. Hodges Jr. Belmont: Wadsworth.

Dutter R (1987). Robust Statistical methods applied in the analysis of geochemical Variables; in Contributions to Stochastics, eds. W. Sendler, Heidelberg. Physica-Verlag pp.89-100.

Filzmoser P (1999). Robust Principal Component and factor analysis in the GeoStatistical treatment of environmental Data Environmetrics 10: 363-375.

Nwabueze JC, Onyeagu SI, Ikpegbu O (2009). Robustness of the maximum Likelihood estimation procedure in factor analysis. Afr. J. Math. Comput. Sci. Res. 2(5): 081-087