American Journal of Theoretical and Applied Statistics
2017; 6(2): 90-94
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20170602.13
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

Scale Independent Principal Component Analysis and Factor Analysis with Preserved Inherent Variability of the Indicators

Priyadarshana Dharmawardena[1], Raphel Ouseph Thattil[2], Sembakutti Samita[2]

[1]Department of Census and Statistics, Battaramulla, Sri Lanka

[2]Postgraduate Institute of Agriculture, University of Peradeniya, Peradeniya, Sri Lanka

Email address:

priya_npd@yahoo.com (P. Dharmawardena), thattil@pgia.ac.lk (R. O. Thattil), ssamita@pdn.ac.lk (S. Samita)

To cite this article:

Priyadarshana Dharmawardena, Raphel Ouseph Thattil, Sembakutti Samita. Scale Independent Principal Component Analysis and Factor Analysis with Preserved Inherent Variability of the Indicators. *American Journal of Theoretical and Applied Statistics.* Vol. 6, No. 2, 2017, pp. 90-94. doi: 10.11648/j.ajtas.20170602.13

Received: February 2, 2017; **Accepted**: February 17, 2017; **Published**: March 2, 2017

Abstract: Principal Component Analysis (PCA) and Factor Analysis (FA) are common multivariate techniques used for dimensionality reduction. With these techniques it is expected to identify actual number of dimensions while accounting almost all observed variability. Standard PCA is based either on correlation matrix (CORM) or covariance matrix (COVM). When it is based on CORM, scale dependency can be removed but inherent variability cannot be preserved. On the other hand, when PCA is based on COVM, inherent variability can be preserved but scale dependency cannot be removed. As a solution to this issue, this paper suggests scaling each indicator by its mean, resulting in new mean equal to 1 and standard deviation equal to the coefficient of variance (CV). This leads to PCs, which are scale independent while retaining the observed variability. The computation of PCs and factors under the suggested method is derived in the study. The procedure is illustrated using the lowest level administrative division census data of Western province of Sri Lanka.

Keywords: Scaling Indicators, Coefficient of Variation, Multivariate Techniques, Dimensional Reduction, Computation of PCAs and Factors

1. Introduction

Principal component analysis (PCA) and factor analysis (FA) are common multivariate techniques used for dimensionality reduction for many purposes. Multivariate techniques have been used in many divergent fields in the construction of composite indices [1]. In constructing CIs, it is important to identify a small number of transformed indicators out of the considered set of indicators. One of the most important applications of PCA and FA is construction of composite indices.

1.1. Principal Component Analysis (PCA)

PCA involves a mathematical procedure that transforms a set of correlated variables into a smaller set of uncorrelated variables. Its goal is to extract the important information from the data table and to express this information as a set of new orthogonal variables called principal components (PCs) [10]. These principal components are linear combinations of the original variables. Hence the results of PCA depend on the scales that the variables are measured on.

1.2. Factor Analysis (FA)

Factor analysis is also a variable reduction technique and is similar to PCA. It is a useful tool for investigating variable relationships for complex concepts such as socio-economic status, dietary patterns, or psychological scales [11]. It allows researchers to investigate concepts that are not easily measured directly by collapsing a large number of variables into a few interpretable, uncorrelated underlying factors. In factor analysis, a factor is a latent (unmeasured) variable that expresses itself through its relationship with other measured variables. Contrary to the PCA, the FA model assumes that the data is based on the underlying factors of the model, and that the data variance can be decomposed into that accounted

for by common and unique factors [12]. Performing FA based on PCA is one of a commonly used methods.

1.3. Composite Index (CI)

CI measures multi-dimensional aspects which cannot be captured properly by a single variable. CI should be based on a theoretical framework or definition, which allows individual indicators or variables to be selected, combined and weighted in a manner which reflects the dimensions or structure of the phenomena being measured [2]. With CIs decision makers should be able to have a better understanding of complex, multi-dimensional realities as it is easier to interpret than a set of separate indicators. The most important fact of CI is, it's ability of reducing the visible size of a set of indicators without dropping the base of underlying information. Farrugia [7] pointed out that in the context of policy analysis, CIs are useful in identifying trends and drawing attention to particular issues and they can also be helpful in setting policy priorities and in benchmarking or monitoring performance. However, if the CI is constructed in a manner which does not reflect the real situation and the construction process lacks proper statistical or conceptual principles, those CIs may indicate misleading information for policy decisions. Therefore more attention should be paid on constructing CIs.

1.4. Issues

PCA is performed on a relationship (or association) matrix, which captures the interrelationships between variables. Mainly correlation matrix (CORM) or covariance matrix (COVM) is used as the relationship matrix. But depending on the considered matrix, results of the PCA differ. Jolliffe [3] says that when performing a PCA, a major argument for using CORM rather than COVM is that the results of analyses for different sets of random variables are more directly comparable. Because PCA based on COVM is sensitive to the units of measurement used for each variable. Therefore in CORM approach, PCA operates on standardized data, scaled by their standard deviation. Then all the variables become scale less with zero means and unit variances. On the other hand Jolliffe [3] argues that if there are large differences between the variances among the variables, then those variables whose variances are largest will tend to dominate the first few PCs. In that situation, those inherent variability cannot be captured performing PCA with standardized data. Then drawing conclusions about the dominance of variation for the actual, unstandardized data tends to be misleading. Hence, COVM approach may be entirely appropriate for the set of variables with different variances but measured in the same scale. Another disadvantage of PC's derived using the CORM is that they give coefficients for standardized variables and are therefore less easy to interpret directly [3]. Therefore this problem has to be addressed in constructing scale independent composite indices, while preserving the inherent variability of the variables

1.5. Objective

The objective of this study is to find out a solution to the problem of scale dependency of performing PCA without standardizing the variables while preserving the information with respect to inherent variability of the variables.

2. Proposed Method

As a solution to the issues mentioned in section 1, data of each variable were scaled by its mean. Then the new mean will be equal to 1 and standard deviation equal to the CV. Consequently scale independent new set of variables can be obtained preserving inherent variability.

Suppose the original variables are $X_1, X_2, \ldots, X_m$ with means and variances equal to μ_i and σ_i^2. where $i=1, 2, \ldots, m$.

Let's divide the each variable by their means and symbolized the transformed new set of variables as Y_i. Then,

$$Y_i = \frac{X_i}{\mu_i} \tag{1}$$

Here, Y_i s are independent of the scale.

$$E(Y_i) = E\left(\frac{X_i}{\mu_i}\right) = \frac{E(X_i)}{\mu_i} = \frac{\mu_i}{\mu_i} = 1 \tag{2}$$

$$V(Y_i) = V\left(\frac{X_i}{\mu_i}\right) = \frac{V(X_i)}{\mu_i^2} = \frac{\sigma_i^2}{\mu_i^2} = \left(\frac{\sigma_i}{\mu_i}\right)^2 = CV_i^2 \tag{3}$$

Then, the standard deviation of $Y_i = CV_i$

Unlike the standardized variables, there are different values for the variances of Y_i.

The matrix, X

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}_{n\times m}$$

and

$$\bar{X} = \begin{bmatrix} \bar{x}_1 & 0 & \cdots & 0 \\ 0 & \bar{x}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \bar{x}_m \end{bmatrix}_{m\times m}$$

Where, n = number of observations

m = number of variables

Then the matrix after the transformation,

$$Y = X\bar{X}^{-1} \tag{4}$$

$$Y = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}_{n\times m} \begin{bmatrix} \frac{1}{\bar{x}_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\bar{x}_2} & \cdots & 0 \\ & & \ddots & \vdots \\ \vdots & \vdots & & 1 \\ 0 & 0 & \cdots & \frac{1}{\bar{x}_m} \end{bmatrix}_{m\times m}$$

$$Y = \begin{bmatrix} \frac{x_{11}}{\bar{x}_1} & \frac{x_{12}}{\bar{x}_2} & \cdots & \frac{x_{1m}}{\bar{x}_m} \\ \frac{x_{21}}{\bar{x}_1} & \frac{x_{22}}{\bar{x}_2} & \cdots & \frac{x_{2m}}{\bar{x}_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_{n1}}{\bar{x}_1} & \frac{x_{n2}}{\bar{x}_2} & \cdots & \frac{x_{nm}}{\bar{x}_m} \end{bmatrix}_{n\times m}$$

The covariance between two transformed variables, Y_i and Y_j is given by,

$$Cov(Y_i, Y_j) = E(Y_iY_j) - E(Y_i)E(Y_j)$$

$$= E\left(\frac{X_i}{\mu_i} \times \frac{X_j}{\mu_j}\right) - 1 \times 1$$

$$= \frac{1}{\mu_i\mu_j} E(X_iX_j) - 1$$

$$= \frac{1}{\mu_i\mu_j}[E(X_iX_j) - \mu_i\mu_j]$$

$$Cov(Y_i, Y_j) = \frac{Cov(X_i,X_j)}{\mu_i\mu_j} \quad (5)$$

If Pearson Correlation Coefficient of X_i and X_j is ρ_{ij}; which is equal to

$$\rho_{ij} = \frac{Cov(X_i, X_j)}{\sqrt{\sigma_i^2\sigma_i^2}}$$

$$Cov(X_i, X_j) = \rho_{ij}\sqrt{\sigma_i^2\sigma_i^2} = \rho_{ij}\sigma_i\sigma_j \quad (6)$$

from (5),

$$Cov(Y_i, Y_j) = \frac{Cov(X_i, X_j)}{\mu_i\mu_j}$$

Substituting (6) to this equation, we obtain

$$Cov(Y_i, Y_j) = \frac{\rho_{ij}\sigma_i\sigma_j}{\mu_i\mu_j}$$

$$Cov(Y_i, Y_j) = \rho_i\left(\frac{\sigma_i}{\mu_i}\right)\left(\frac{\sigma_j}{\mu_j}\right)$$

$$Cov(Y_i, Y_j) = \rho_{ij}CV_iCV_j \quad (7)$$

Then the Variance-Covariance matrix of Y), ituting

$$\sum = \begin{bmatrix} CV_1^2 & \rho_{12}CV_1CV_2 & \cdots & \rho_{1m}CV_1CV_m \\ \rho_{21}CV_1CV_2 & CV_2^2 & \cdots & \rho_{2m}CV_2CV_m \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{m1}CV_1CV_m & \rho_{m2}CV_2CV_m & \ldots & CV_m^2 \end{bmatrix}_{m\times m} \quad (8)$$

FA was performed followed by the PCA using the variance-covariance matrix (8).

3. Validation

In order to validate the proposed method, analysis was performed using a dataset relevant to the problem. To achieve this task, set of variables had to be identified with different scales and different variances.

3.1. Data

In Sri Lanka, urban / rural classification is not based on a proper statistical methodology. Urban areas are defined on the basis of administrative boundaries of local authorities (LAs). There are three types of Local Authorities in the country at present, namely Municipal Council (MC), Urban Council (UC) and *Pradeshiya Sabha* (PS). MCs and UCs are considered urban LAs while PSs are considered rural LAs. It could be seen that some areas with urban characteristics were in PS divisions while some rural categories were in MC and UCs. Because, variability of those attributes are significant within a LA. Therefore, we need to go to the lowest administrative level in a LA for the classification. Then, the variability of the considered variables within the LAs could be taken into account. In the Sri Lankan context, being the smallest administrative unit, *Grama Niladhari* (GN) division is the most appropriate level to be considered.

3.2. Variables

Considering the following variables, data were collected by GN divisions in the Western province of Sri Lanka. All the variables were adjusted in a manner which explaining the high degree of urban nature. Number of data points was 2495. Data were obtained from the Department of Census and Statistics in Sri Lanka. The description of the variables are as follows.

Pop_Dencty: Population Density

stories3_HU_Pcnt: Percentage of three and above storied housing unites out of total housing units in the GN division

HU_Bld_pcnt_INVS: 100 - Percentage of housing units out of total buildings in the GN division

In_Migration: Percentage of in-migrated population

WS_Rtail_HH: Number of wholesale and retail outlets per housing unit in a GN division

Edu_HH: Number of education centers (Private) per housing unit in a GN division

Health_HH: Number of medical centers (Private) per housing unit in a GN division

Recreation_HH: Number of recreation centers per housing unit in a GN division

IND_Above_5: Number of industries with 5 and above number of employees in a GN division

4. Results

All the considered variables were in different scales. The appropriateness of them for this study was identified using descriptive statistics.

4.1. Descriptive Statistics

Descriptive statistics of the considered variables are given in the table 1 to identify the nature of variability.

Table 1. *Descriptive statistics of the considered variables.*

Variable	Scale	Mean	Standard Deviation	CV
Pop_Dencty	No. of people per Km^2	3128.432	4580.556	1.464
stories3_HU_Pcnt	%	0.543	1.219	2.245
HU_Bld_pcnt_INVS	%	21.038	9.247	0.440
In_Migration	%	7.320	7.731	1.056
WS_Rtail_HH	Per Housing unit	0.076	0.135	1.768
Edu_HH	Per Housing unit	0.010	0.012	1.205
Health_HH	Per Housing unit	0.003	0.007	2.141
Recreation_HH	Per Housing unit	0.003	0.009	3.640
IND_Above_5	Number	9.691	24.178	2.495

Table 1, clearly indicates that the considered set of variables were in different scales. Also they were with highly dispersed variability. That was not only due to magnitude of the numbers but also due to inherent property of the variable. Therefore, that nature could be captured using the CV included in the fifth coloumn in table 1. As an example the highest standard deviation was recorded from the variable "Pop-Dencty" (Population density), which is 4580.556 number of people per Km^2. But it's CV was not the highest. The variable, "Recreation_HH" (Number of recreation centers per housing unit in a GN division) recorded the highest CV of 3.640. But the standard deviation of it was very low. (0.009 per housing unit). Hence, the set of variables given in table 1 was suitable to validate the proposed method.

4.2. Application of PCA

PCA was performed to identify the minimum linear combination of considered variables with higher explanation of the original variation of the data. Proposed method was applied followed by the conventional approaches, those are with the relationship matrices of CORM and COVM. In CORM approach, data were standardized whereas in COVM approach, they were not. Variables under the proposed method were transformed by dividing by their means. Then scale dependency problem was solved and the inherent variability of the variables was also taken into account. Then the variances of the new set of variables are the square term of CV of the original variables. Using COVM, PCA was performed to the transformed data set under the proposed method and the results were included in table 2.

Table 2. *Eigen values under conventional and proposed methods.*

PC No.	Conventional method				Proposed method	
	CORM approach		COVM approach			
	Eigen value	Cumulative % of variance	Eigen value	Cumulative % of variance	Eigen value	Cumulative % of variance
1	3.980	44.260	20981550.512	99.997	14.650	49.930
2	1.960	66.020	567.619	100.000	7.570	75.710
3	0.800	74.900	67.319	100.000	2.600	84.570
4	0.560	81.110	42.162	100.000	1.580	89.960
5	0.540	87.120	0.628	100.000	1.090	93.680
6	0.390	91.480	0.012	100.000	0.790	96.370
7	0.320	95.060	0.000	100.000	0.510	98.120
8	0.290	98.300	0.000	100.000	0.470	99.700
9	0.150	100.000	0.000	100.000	0.090	100.000

4.2.1. Conventional Method

Considering the results of PCA, first two PCs those eigen values are above 1, explained only 66 percent of total variation. This is not supposed to be a good approach due to two reasons. One of these was requirement of selecting higher number of PCs to get reasonably higher degree of explanation out of total variation though the objective is to reduce variable at minimum level while explaining the greater degree of variability. The other reason was neglecting the inherent variability due to standardizing variables. Therefore, the possible alternative was to perform PCA using covariance matrix approach with unstandardized data. But here, whole variability was dominated by one PC due to the variable having the highest variance (Table 1). However this approach cannot be applied since the set of variables was in different scales.

4.2.2. Proposed Method

Under the proposed method, 75.71 percent of total variance was explained by the first two PCs while in the conventional method with CORM approach, it was 66.02 percent. This is more than 9 percent of improvement which can be considered as sufficient.

4.3. Application of FA

FA was performed using the method of principal component. Therefore two factors were considered under CORM approach in conventional method and proposed method. Due to dominance of one PC under the COVM approach, only one factor could be considered.

Table 3. *Under the factor analysis contribution of variables for first two factors for each method.*

Variable	Conventional method			Proposed method	
	CORM approach		COVM approach		
	Factor 1	Factor 2	Factor 1	Factor 1	Factor 2
Pop_Dencty	-0.115	0.706*	1.000*	-0.051	0.583*
stories3_HU_Pcnt	0.155	0.868*	0.564	0.141	0.882*
HU_Bld_pcnt_INVS	0.711*	0.378	0.101	0.588*	0.363
In_Migration	0.139	0.685*	0.251	0.148	0.546*
WS_Rtail_HH	0.882*	0.099	0.052	0.927*	0.087
Edu_HH	0.816*	0.038	0.003	0.723*	0.066
Health_HH	0.834*	0.058	-0.015	0.771*	0.109
Recreation_HH	0.852*	0.098	0.021	0.963*	0.101
IND_Above_5	0.264	0.744*	0.288	0.242	0.858*

* Significant contribution

In factor analysis, for both methods (except COVM approach) few variables indicated significant contribution on two factors which should not to be. Therefore Varimax rotation was used to overcome that issue. In the proposed method and the CORM approach under conventional method, all the variables could be adequately explained by two factors. Since the large variance of the variable Pop_dencty, in COVM approach, only that variable indicated highly significant contribution to the identified single factor.

In PCA, with the application of the proposed method, there was a significant improvement over the conventional method. Considering the FA, in the proposed method, contribution of the variables on factors was not dominated by few variables as COVM approach under conventional method.

5. Conclusions

Conducting PCA and FA as a variable reduction techniques, with CORM approach is not always acceptable due to ignoring the inherent variability of variables. COVM approach is a good solution to the above problem, but it also has the drawback of scale dependency. To get scale independent set of indicators, all the indicators were converted in to new set dividing the data of original indicators by their means. The means of the new set of variables were unit, while the standard deviations were CVs. Hence the inherent variability of the original indicators were preserved under the proposed method. Therefore, in the application of PCA and FA, converting new set of indicators scaling by their means can generate meaningful information.

References

[1] Silva, G. (2000) Construction of Human Development Indices using Multivariate Techniques, PGIA, University of Peradeniya.

[2] The Organization for Economic Co-operation and Development, (2004) The OECD-JRC Handbook on Practices for Developing Composite Indicators. https://stats.oecd.org/glossary/detail.asp?ID=6278.

[3] Jolliffe, T. (2002) Principal Component Analysis, Springer Verlag, New York.

[4] Josseph, F. et al. (2010) Multivariate Data Analysis. A Global Perspective, Pearson Education Inc, New Jersey.

[5] Fernando, S., Samita, S and Abenayake R (2011). Modified factor analysis to construct composite indices: Illustration on Urbanization index. Tropical Agricultural Research, 24, 271–281.

[6] Tuan, A. and Magi, S. (2009) Principal Component Analysis: Final Paper in Financial Pricing. National Cheng Kung University, 3-26.

[7] Farrugia. N, (2007) conceptual issues in constructing composite indices. Occasional papers on islands and small states, 2/2007.

[8] Yutaka, K., Yusuke, M. and Shohei, S. (2003) factor rotation and ica. 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), Japan.

[9] Delchambre, L. (2014) Weighted principal component analysis: a weighted covariance eigen decomposition approach, University of Liege, Belgium.

[10] Abdi, H., and Williams, J., Principal component analysis John Wiley & Sons; WIREs Comp Stat 2010 2 433–459 2010.

[11] http://www.theanalysisfactor.com/factor-analysis-1-introduction/.

[12] The Organization for Economic Co-operation and Development, (2008) Handbook on constructing composite indicators: methodology and user guide.