ISSN: 2348 9510
International Journal Of Core Engineering & Management (IJCEM)
Volume 3, Issue 4, July 2016

THE EFFECTIVENESS OF FACTOR ANALYSIS AS A STATISTICAL TOOL OF VARIABLE REDUCTION TECHNIQUE

Weeraratne N.C.
Department of Economics & Statistics, FSSL
Sabaragamuwa University of Sri Lanka
BelihulOya, Sri Lanka

ABSTRACT

The idea of Factor Analysis (FA) is to describe a set of p variables in terms of a fewer no: of factors. So Factor Analysis is also a variable reduction tool in statistical analysis. Most researchers try to get one factor from several correlated variables by using Factor Analysis. But this variable reduction method is not suitable for any cases. So this paper explains how to select effective number of factors and when Factor Analysis (FA) is appropriate as a variable reduction tool. When p=2, if correlation between these two variables is greater than or equal 0.6 ($r \geq 0.6$), two original variables can be described by one common factor and when p=2, if correlation between each of these two variables is greater than or equal 0.7 ($r \geq 0.7$), three original variables can be described by one common factor. According this situation, if no: of original variables to be reduced are high required pairwise correlations also high. If it is not, p original variables can be described by a few common factors (more than one). That means Factor Analysis (FA) is effective for highly correlated variables as a variable reduction technique.

Key Words: Factor Analysis (FA), Variable Reduction Technique, Correlation

I. INTRODUCTION

Factor Analysis (FA) has similar aims as Principal Component Analysis (PCA). The idea is to derive new variables called factors which can give a better understanding of the data. Here too it is expected that p original variables can be described by a few common factors. One of the key differences between FA and PCA is that FA is based on a proper statistical model whereas PCA produces an orthogonal transformation of the variables which depends on no underlying model. FA is more concerned with explaining the covariance structure of the variables than explaining the variances. Any variance which is unexplained by the common factors is described by the error term. The factors are sometimes called latent factors or unobservable factors. Specially, when the variables are related it is possible that all the variables have one underlying factor. In fact the purpose of FA is to identify that underlying factor or factors.

The FA model assumes that there are m underlying factors (where m < p) which may be denoted as $f_1, f_2, \ldots\ldots, f_m$ and that each observable variable is a linear function of these factors together with a residual

ISSN: 2348 9510
International Journal Of Core Engineering & Management (IJCEM)
Volume 3, Issue 4, July 2016

variates, so that,

$$X_j = a_{j1}f_1 + a_{j2}f_2 + \cdots + a_{jm}f_m + \epsilon_j$$

The coefficients $a_{j1}, a_{j2}, \ldots\ldots, a_{jm}$ are called factor loadings, so that a_{jk} is the loading of the j^{th} variable on the k^{th} factor. The variates ϵ_j describes the residual variation specific to the j^{th} variable. Usually $f_1, f_2, \ldots\ldots, f_m$ are called the common factors and ϵ_j are called the specific factors. Note that equation for the factor model is the inverse transformation of the equation for the PCs. It is assumed that specific factors are independent from one another and the common factors. It is also assumed that common factors are independent of one another, though this assumption is sometimes relaxed when factors are rotated. As Xs have zero mean, it is also convenient to assume that the factors also have zero mean. According to the factor model, there is an arbitrary scale factor related to each common factor and so it is usual to choose constrain each common factor to also have unit variance. However, variance of specific factors may vary. From the factor model, for actual observations, we can write,

$$x_{rj} = \sum_{k=1}^{m} a_{jk}f_{rk} + \epsilon_{rj}$$

Where x_{rj} is the value of r^{th} observation on the j^{th} variable, f_{rk} is the score of the k^{th} common factor for the r^{th} observation and ϵ_{rj} is the value for the j^{th} specific factor for the r^{th} observation. From the factor model, using the independence of different models and since the common factors have unit variance, we have,

$$Var(X_j) = a_{j1}^2 + a_{j2}^2 + \ \ldots + a_{jm}^2 + Var(\epsilon_j)$$
$$= \sum_{k=1}^{m} a_{jk}^2 + \Psi_j$$

Where Ψ_j is $Var(\epsilon_j)$. The part of the variance explained by the common factors, namely $\sum_{k=1}^{m} a_{jk}^2$, is called the communality of the j^{th} variable. From the factor model it can also be found,

$$Cov\,(X_i, X_j) = \sum_{k=1}^{m} a_{ik}a_{jk}$$

Thus the covariance of the matrix of **X,** which is **Σ**, is given by;

$$\boldsymbol{\Sigma} = \boldsymbol{AA}^T + \boldsymbol{\Psi}$$

Accordingly, common factors "explain" the off-diagonal elements, namely the (covariance) of **Σ** exactly, since **Ψ** is diagonal. When m=1, i.e., one factor solution exists, then the solutions are usually unique. If $m > 1$ and a solution exists, then the solution is not unique. It can be shown that any orthogonal rotation of

ISSN: 2348 9510
International Journal Of Core Engineering & Management (IJCEM)
Volume 3, Issue 4, July 2016

the factors in the relevant m-space will give us a new set of factors which will also satisfy $\boldsymbol{\Sigma = AA^T + \Psi}$.

The unobservable random vectors **F** and $\boldsymbol{\Sigma}$ satisfy the following conditions.

[1]. **F** and $\boldsymbol{\Sigma}$ are independent
[2]. $E(\boldsymbol{F}) = \boldsymbol{0}$
[3]. $Cov(\boldsymbol{F}) = \boldsymbol{I}$
[4]. $E(\boldsymbol{\Sigma}) = \boldsymbol{0}$
[5]. $Cov(\boldsymbol{\Sigma}) = \boldsymbol{\Psi}$; $\boldsymbol{\Psi}$ is a diagonal matrix

Methods of selecting the number of factors (m):

[1]. By theory or the work of other researchers.
[2]. m equal to the number of Eigen values of R greater than one if the sample correlation matrix is factored or equal to the number of positive Eigen values of S if the sample covariance matrix is factored. (These rules of thumb should not be applied indiscriminately).
[3]. The choice of m can be based on the estimated Eigen values. (Cumulative proportion of total sample variance up to m^{th} factor > 0.80).

$$\begin{pmatrix} Proportion\ of\ total\ sample \\ variance\ due\ to\ j^{th}\ factor \end{pmatrix} = \begin{cases} \dfrac{\hat{\lambda}_j}{S_{11}+S_{22}+\ldots.+S_{pp}} ; for\ a\ FA\ of\ S \\ \dfrac{\lambda_j}{p} ; for\ a\ FA\ of\ R \end{cases}$$

[4]. Residual Matrix [R] = $S - \left(\hat{L}\hat{L}^T + \psi\right)$. We may subjectively take the m factors model to be appropriate, analytically we have,

$$\text{Sum of squared entries of } \left(S - \left(\hat{L}\hat{L}^T + \psi\right)\right) \leq \hat{\lambda}_{m+1}^2 + \cdots + \hat{\lambda}_p^2.$$

Unfortunately for the Factor Analysis (FA), most covariance matrices cannot be factored as $\hat{L}\hat{L}^T + \psi$, where the no: of factors (m) is much less than no: of original variables (p).

II. OBJECTIVE OF THE STUDY

To identify, how to select effective/ appropriate number of factors and when Factor Analysis (FA) is appropriate as a variable reduction tool.

ISSN: 2348 9510
International Journal Of Core Engineering & Management (IJCEM)
Volume 3, Issue 4, July 2016

III. MATERIALS AND METHODS

This study was mainly based on generated data. The observations of variables were generated from a multivariate normal distribution to accommodate the correlation of the variables.

IV. RESULTS AND FINDINGS

Two Variable Case (P = 2)

Illustration – 01[$r = 0$]

Principal Component Factor Analysis of the Correlation Matrix

Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Communality
Price_5	0.000	1.000	1.000
Income_5	1.000	0.000	1.000
Variance	1.0000	1.0000	2.0000
% Var	0.500	0.500	1.000

Factor Score Coefficients

Variable	Factor1	Factor2
Price_5	0.000	1.000
Income_5	1.000	0.000

Illustration – 01[$r = -0.186$]

Principal Component Factor Analysis of the Correlation Matrix

Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Communality
Price_4	0.711	0.703	1.000
Income_4	-0.711	0.703	1.000
Variance	1.0115	0.9885	2.0000
% Var	0.506	0.494	1.000

Rotated Factor Loadings and Communalities
Varimax Rotation

Variable	Factor1	Factor2	Communality
Price_4	1.000	-0.006	1.000
Income_4	-0.006	1.000	1.000
Variance	1.0000	1.0000	2.0000
% Var	0.500	0.500	1.000

ISSN: 2348 9510
International Journal Of Core Engineering & Management (IJCEM)
Volume 3, Issue 4, July 2016

Factor Score Coefficients

Variable	Factor1	Factor2
Price_4	1.000	0.006
Income_4	0.006	1.000

Illustration – 01[$r = -0.5$]

Principal Component Factor Analysis of the Correlation Matrix

Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Communality
Price_3	0.866	0.500	1.000
Income_3	-0.866	0.500	1.000
Variance	1.4997	0.5003	2.0000
% Var	0.750	0.250	1.000

Rotated Factor Loadings and Communalities
Varimax Rotation

Variable	Factor1	Factor2	Communality
Price_3	-0.259	0.966	1.000
Income_3	0.966	-0.259	1.000
Variance	1.0000	1.0000	2.0000
% Var	0.500	0.500	1.000

Factor Score Coefficients

Variable	Factor1	Factor2
Price_3	0.299	1.115
Income_3	1.115	0.299

Illustration – 01[$r = -0.857$]

Principal Component Factor Analysis of the Correlation Matrix

Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Communality
Price_2	0.964	0.267	1.000
Income_2	-0.964	0.267	1.000
Variance	1.8570	0.1430	2.0000
% Var	0.928	0.072	1.000

Rotated Factor Loadings and Communalities
Varimax Rotation

Variable	Factor1	Factor2	Communality
Price_2	0.870	-0.492	1.000
Income_2	-0.492	0.870	1.000
Variance	1.0000	1.0000	2.0000
% Var	0.500	0.500	1.000

ISSN: 2348 9510
International Journal Of Core Engineering & Management (IJCEM)
Volume 3, Issue 4, July 2016

Factor Score Coefficients

Variable	Factor1	Factor2
Price_2	1.689	0.955
Income_2	0.955	1.689

Illustration – 01[$r = 1$]

Principal Component Factor Analysis of the Correlation Matrix

Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Communality
Price_1	1.000	-0.000	1.000
Income_1	1.000	0.000	1.000
Variance	2.0000	0.0000	2.0000
% Var	1.000	0.000	1.000

Factor Score Coefficients

Variable	Factor1	Factor2
Price_1	1.000	-0.022
Income_1	-0.000	0.022

Table 1: Summary Factor Analysis

Illustration No:	Correlation (r)	Cumulative proportion of total sample variance Due to 1st Factor	No: of factors (m)
1	0.000	0.500	Exactly 2
2	-0.186	0.506	2
3	-0.500	0.750	2
4	-0.857	0.928	1
5	1.000	1.000	Exactly 1

* Generally FA expects cumulative proportion of total sample variance explained is greater than 0.80.

When r = 0.6

Principal Component Factor Analysis of the Correlation Matrix

Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Communality
Price_4_1	-0.894	0.447	1.000
Income_4_1	0.894	0.447	1.000
Variance	1.5998	0.4002	2.0000
% Var	0.800	0.200	1.000

ISSN: 2348 9510
International Journal Of Core Engineering & Management (IJCEM)
Volume 3, Issue 4, July 2016

Rotated Factor Loadings and Communalities
Varimax Rotation

Variable	Factor1	Factor2	Communality
Price_4_1	0.949	-0.316	1.000
Income_4_1	-0.316	0.949	1.000
Variance	1.0000	1.0000	2.0000
% Var	0.500	0.500	1.000

Factor Score Coefficients

Variable	Factor1	Factor2
Price_4_1	1.186	0.395
Income_4_1	0.395	1.186

Because cumulative proportion of total sample variance explained due to first factor is greater than or equal to 0.8, It is clear that When r ≥ 0.6, two original variables can be described by one common factor.

Three Variable Case (P = 3)

Principal Component Factor Analysis of the Correlation Matrix

Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Factor3	Communality
x1_1	-0.879	-0.464	0.114	1.000
x2_1	-0.910	0.116	-0.398	1.000
x3_1	-0.895	0.337	0.293	1.000
Variance	2.4006	0.3422	0.2573	3.0000
% Var	0.800	0.114	0.086	1.000

Rotated Factor Loadings and Communalities
Varimax Rotation

Variable	Factor1	Factor2	Factor3	Communality
x1_1	0.897	0.304	0.321	1.000
x2_1	0.348	0.368	0.862	1.000
x3_1	0.314	0.882	0.351	1.000
Variance	1.0237	1.0060	0.9703	3.0000
% Var	0.341	0.335	0.323	1.000

Factor Score Coefficients

Variable	Factor1	Factor2	Factor3
x1_1	1.374	-0.324	-0.415
x2_1	-0.384	-0.466	1.513
x3_1	-0.314	1.440	-0.488

ISSN: 2348 9510
International Journal Of Core Engineering & Management (IJCEM)
Volume 3, Issue 4, July 2016

```
Matrix CORR

1.00000  0.70000  0.70000
0.70061  1.00000  0.70000
0.70000  0.70000  1.00000
```

According to the above result, Corr(X_1, X_2), Corr(X_1, X_3) and Corr(X_2, X_3) are equal to the 0.7. Because cumulative proportion of total sample variance explained due to first factor is greater than or equal to 0.8, It is clear that When r $\geq$ 0.7, three original variables can be described by one common factor.

V. CONCLUSIONS

High correlation among variables indicates homogeneous sets of variables and each set of variables measure the same underlying dimension. Low correlation among variables indicates not much is common between term or a group of heterogeneous variable and thus data are not appropriate for Factor Analysis (FA). Partial correlation matrix is another such measure and it is expected that low partial correlations among variables is necessary for Factor Analysis (FA) to be appropriate. It is suggested a value above 0.8 is a good fit. A value between 0.6 and 0.8 is supposed to be tolerable and a value below 0.5 is unacceptable.

REFERENCES

- Arrindell, W. A., & van der Ende, J. (1985). An empirical test of the utility of the observations-to-variables ratio in factor and components analysis. Applied Psychological Measurement, 9, 165-178.
- Borgatta, E. F., Kercher, K., & Stull, D. E. (1986). A cautionary note on the use of principal components analysis. Sociological Methods and Research, 15, 160-168.
- Ford, J K., MacCallum, R. C., & Tail, M. (1986). The applications of exploratory factor analysis in applied psychology: A critical review and analysis. Personnel Psychology, 39, 291-314.
- Harris, C. W., & Kaiser, H. F. (1964). Oblique factor analytic solutions by orthogonal transformations. Psychometrika, 29, 347-362.
- Longman, R. S., Cota, A. A., Holden, R. R., & Fekken, G. C. (1989). A regression equation for the parallel analysis criterion in principal components analysis: Mean and 95th percentile eigenvalues. Multivariate Behavioral Research, 24, 59-69.
- Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. Psychological Bulletin, 103, 391-410.
- Skinner, H. A. (1980). Factor analysis and studies of alcohol. Journal of Studies on Alcohol, 41, 1091-1101.
- Velicer, W. F. (1977). An empirical comparison of the similarity of principal component, image, and factor patterns. Multivariate Behavioral Research, 12, 3—22.
- Velicer, W. F., & Jackson, D. N. (1990a). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. Multivariate Behavioral Research, 25, 1-28.

ISSN: 2348 9510
International Journal Of Core Engineering & Management (IJCEM)
Volume 3, Issue 4, July 2016

- Widaman, K. F. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? Multivariate Behavioral Research, 28, 263-311.