Evaluating Optimal Number of Factors in an Orthogonal Factor Analysis: An Information Criteria Approach

Nwosu Dozie F.,
Department of Mathematics and Statistics Federal Polytechnic Nekede P.M.B 1234 Owerri Imo State.
E-mail : fedocon2003@gmail.com

Onyeagu Sidney I.,
Department of Statistics Nnamdi Azikiwe University P.M.B 5025 Awka, Anambra State.
E-mail : onyeagusidney@gmail.com

Osuji George A.
Department of Statistics Nnamdi Azikiwe University P.M.B 5025 Awka, Anambra State.
E-mail : george.osuji99@yahoo.com

Corresponding Author: Nwosu Dozie F.; fedocon2003@gmail.com

ABSTRACT
An exact quantitative basis for deciding the number of factors to extract in a factor analysis has been an unresolved issue. Hair, et al (2010) in their work proposed four stopping criteria for the number of factors to extract namely: latent root criterion, a priori criterion, percentage of variance criterion and the scree test criterion. They inferred that an exact quantitative basis for deciding the number of factors to extract has not been developed. Hence, we tried solving this problem using three information criteria approaches namely Akaike information criterion, Schwarz information criterion and the Hanna-Quinn information criterion for various factor loadings estimation methods at sample sizes of 30, 50, 70 and two variable (p) cases of 10 and 15. From the results of the analysis, the 10 variable cases shows that the majority of the estimation methods at different sample sizes retained the 5 factor solution as the best whereas for the 15 variable cases the 9 factor solution was selected as the number of factors to retain. The values of the Maximum Likelihood method are consistent with all the sample sizes for the two variable cases considered. Also, from the values of the AIC, SIC, and HQIC for the two variable cases considered, it is observed that the performance of the HQIC was the best since it has the least value among the three information criteria considered in this work. Also, comparing the information criteria approach with the Hair’s rules, the information criteria approach performed best among the five methods compared for most of the four estimation methods of the two variables cases (p = 10 and 15) and the sample sizes considered. Also, the results show that the method of maximum likelihood estimation adjudged to be the best among the four methods of orthogonal factor model extraction considered in this work using the information criteria approach.
Keywords: Factor Analysis, Factor Rotation; Principal Component Factors Method; Principal Factor Method; Iterated Principal Factor Method; Maximum Likelihood Method; Akaike; Schwarz; and Hannan Quinne Information Criteria.

1.0. INTRODUCTION

Factor analysis attempts to simplify complex and diverse relationships that exist among a set of observed variables by uncovering common dimensions or factors that link together the seemingly unrelated variables and consequently provides insight into the underlying structure of the data. The goal of factor analysis is to reduce a large number of variables to a smaller number of factors, to concisely describe the relationship among observed variables or to test theory about underlying processes.

Factor analysis is a collection of methods used to examine how underlying constructs influences the responses on a number of measured variables. Factor analysis is performed by examining the pattern of correlations (or covariance) between the observed measures. Measures that are highly correlated (either positively or negatively) are likely influenced by the same factors, while those that are not highly correlated are likely influenced by different factors.

In the factor analysis, we represent the variables \( y_1, y_2, \ldots, y_p \) as a linear combination of a few random variables \( f_1, f_2, \ldots, f_m \) (\( m < p \)) called factors. The factors are underlying constructs or latent variables that ‘generate’ the \( y \)’s. Like the original variables, the factors vary from individual to individual; but unlike the variables, the factors cannot be measured or observed. The existence of these hypothetical variables is therefore open to question. If the original variables \( y_1, y_2, \ldots, y_p \) are at least moderately correlated, the basic dimensionality of the system is less than \( p \). The goal of factor analysis is to reduce the redundancy among variables by using a smaller number of factors. Factor analysis can be viewed as an extension of principal components analysis. It related to the principal components analysis in that both seek a simpler structure in a set of variables but they differ in many respects. Both of them have the goal of reducing dimensionality. Because the objectives are similar, many authors discuss principal components analysis as another type of factor analysis.

1.1. THE MATHEMATICAL MODEL FOR FACTOR STRUCTURE:

Suppose that the multivariate system consist of a random sample \( y_1, y_2, \ldots, y_n \) from a homogeneous population with mean \( \mu \) and covariance matrix \( \Sigma \). The factor analysis model expresses each variable as a linear combination of underlying common factors \( f_1, f_2, \ldots, f_m \), with an accompanying error term to account for that part of the variable that is unique (not in common with the other variables). For \( y_1, y_2, \ldots, y_p \) in any observation vector \( y \), the model is as follows
ideally, $m$ should be substantially smaller than $p$. $f_j$ = $j$-th common-factor variates. $\varepsilon_i$ = $i$-th specific factor variates. The coefficients $\lambda_{ij}$ are called loadings and serve as weights showing how each $y_i$ individually depends on the $f$'s. $\lambda_{ij}$ indicates the important of $j$-th factor $f_j$ to the $i$-th variable $y_i$ and can be used in interpretation of $f_j$.

It is assumed that for $j = 1, 2, \ldots, m$

$$E(f_j) = 0, \quad \text{Var}(f_j) = 1 \quad \text{and} \quad \text{Cov}(f_j, f_k) = 0, j \neq k.$$  

The assumptions for $\varepsilon_i$, $i = 1, 2, \ldots, p$, are the same, except that we must allow each $\varepsilon_i$ to have a different variance, since it shows the residual part of $y_i$ that is not in common with the other variables. Hence, we assume that

$$E(\varepsilon_i) = 0, \quad \text{Var}(\varepsilon_i) = \psi_i \quad \text{and} \quad \text{Cov}(\varepsilon_i, \varepsilon_k) = 0, i \neq k.$$  

Also,

$$\text{Cov}(\varepsilon_i, f_j) = 0 \text{ for all } i \text{ and } j$$

We refer to $\psi_i$ as the specific variance. Since $E(y_i - \mu_i) = 0$, we need $E(f_j) = 0$, $j = 1, 2, \ldots, m$.

The assumption $\text{Cov}(f_j, f_k) = 0$ is made for parsimony in expressing the $y$'s as functions of few factors as possible.

The assumptions $\text{Var}(f_j) = 1$, $\text{Var}(\varepsilon_i) = \psi_i$, $\text{Cov}(f_j, f_k) = 0$, and $\text{Cov}(\varepsilon_i, f_j) = 0$ yield a simple expression for variance of $y_i$.

$$\text{Var}(y_i) = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2 + \psi_i$$  

(1.1.2)

Note that the assumption $\text{Cov}(\varepsilon_i, \varepsilon_k) = 0$ implies that the factors account for all the correlations among the $y$'s, that is, all that the $y$'s have in common. Thus the emphasis in factor analysis is on modelling the covariance or correlation among the $y$'s.

For matrix version of the model in equation (1.1.1) above;

Let,

$$f^1 = (f_1, f_2, \ldots, f_m), \quad y^1 = (y_1, y_2, \ldots, y_p)$$  

$$\varepsilon^1 = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p), \quad \mu^1 = (\mu_1, \mu_2, \ldots, \mu_p)$$
and \[ \Lambda = \begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1m} \\ \vdots & \ddots & \vdots \\ \lambda_{p1} & \cdots & \lambda_{pm} \end{bmatrix}. \]

Then, the factor model can be written as
\[ (y - \mu) = \Lambda f + \varepsilon. \] (1.1.3)

Hence,
\[ \text{E}(f_i) = 0, \quad \text{Var}(f_i) = 1, i = 1,2, \ldots, m \quad \text{and} \quad \text{Cov}(f_i, f_k) = 0, j \neq k \]
\[ \text{Var}(F) = I. \]
\[ \text{E}(\varepsilon_i) = 0, i = 1,2, \ldots, p \quad \text{becomes} \quad \text{E}(\varepsilon) = 0; \quad \text{Var}(\varepsilon_i) = \psi_i, i = 1,2, \ldots, p, \quad \text{and} \]
\[ \text{Cov}(\varepsilon_i, \varepsilon_k) = 0, i \neq k \quad \text{become} \]
\[ \text{Cov}(\varepsilon) = \Psi = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix} \]
and
\[ \text{Cov}(\varepsilon_i, f_j) = 0, \quad \text{for} \ i \text{ and} \ j \text{ becomes} \quad \text{Cov}(\varepsilon, f) = 0. \]

The notation \( \text{Cov}(\varepsilon, f) \) indicates a rectangular matrix containing the covariance of the \( f \)'s with the \( \varepsilon \)'s.

\[ \text{Cov}(\varepsilon, f) = \begin{bmatrix} \sigma_{f_1\varepsilon_1} & \sigma_{f_1\varepsilon_2} & \cdots & \sigma_{f_1\varepsilon_p} \\ \sigma_{f_2\varepsilon_1} & \sigma_{f_2\varepsilon_2} & \cdots & \sigma_{f_2\varepsilon_p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{f_m\varepsilon_1} & \sigma_{f_m\varepsilon_2} & \cdots & \sigma_{f_m\varepsilon_p} \end{bmatrix}. \]

Since the emphasis in factor analysis is on modelling the covariance among the \( y \)'s, we wish to express the \( \frac{1}{2} p(p-1) \) covariance (and the \( p \) variances) of the variances \( y_1, y_2, \ldots, y_p \) in term of a simplified structure involving the \( pm \) loadings \( \lambda_{ij} \) and the \( p \)-specific variances \( \psi_i \); that is, we wish to express \( \Sigma \) in terms of \( \Lambda \) and \( \Psi \). Since \( \mu \) does not affect variances and covariances of \( y \), we have
\[ \Sigma = \text{Cov}(y) = \text{Cov}(\Lambda f + \varepsilon) \]
\[ = \Lambda \text{Cov}(f) \Lambda^T + \Psi \]
\[ = \Lambda I \Lambda^T + \Psi \]
\[ = \Lambda \Lambda^T + \Psi \] (1.1.4)
If $\Lambda$ has only a few columns, say two or three, then $\Sigma = \Lambda \Lambda^1 + \Psi$ represents a simplified structure for $\Sigma$, in which the covariances are modelled by $\lambda_{ij}$'s alone since $\Psi$ is diagonal.

Also we can find the covariances of $y$'s with the $f$'s in terms of the $\lambda$'s. the loading themselves represent covariances of the variables with the factors, it then implies that $\text{Cov}(y_i,f_j) = \lambda_{ij}$, $i = 1,2,\ldots, p$ and $j = 1,2,\ldots, m$. Since $\lambda_{ij}$ is the $(ij)$th elements of $\Lambda$, we then have that

$$\text{Cov}(y_i,f_j) = \lambda_{ij} \quad (1.1.5)$$

If standardized variables are used, equation (1.1.5) is replaced by $p_F = \Lambda \Lambda^1 + \Psi$, and the loadings become correlations:

$$\text{Cov}(y_i,f_j) = \lambda_{ij} \quad (1.1.6)$$

Partitioning the variance of $y_i$ into a component due to the common factors called the communality, and a component unique to $y_i$, called the specific variance.

Introducing communality, $\text{Cov}(y_i) = \Lambda \Lambda^1 + \psi$ can be written as

$$\text{Var}(y_i) = \lambda_{i1}^2 + \ldots + \lambda_{im}^2 + \psi_i^2; \quad \text{Cov}(y_{ij},y_{ik}) = \lambda_{ij} \lambda_{ik1} + \ldots + \lambda_{ijm} \lambda_{ikm}$$

and $\text{Cov}(y_i,F_i) = \Lambda$ can be written as $\text{Cov}(y_i,f_i) = \lambda_{ij}$.

Communality is the portion of the variance of the variable contributed by the $m$ common factors.

Suppose the i-th communality is $h_i^2$, then

$$\sigma_i = \text{Var}(y_i) = (\lambda_{i1}^2 + \lambda_{i2}^2 + \ldots + \lambda_{im}^2) + \psi_i$$

$$= h_i^2 + \psi_i$$

where, communality $= h_i^2 = (\lambda_{i1}^2 + \lambda_{i2}^2 + \ldots + \lambda_{im}^2)$ and specific variance $= \psi_i$.

The i-th communality is the sum of square of the loading of the ith variable on the m common factor. When the number of factors $m > 1$, there are multiple factor loadings that generate the same covariance matrix.

The loading in the model (1.1.3) can be multiply by an orthogonal matrix without impairing their ability to reproduce the covariance matrix in $\Sigma = \Lambda \Lambda^1 + \Psi$.

Let $\beta$ be any mxm orthogonal matrix. If we let $\Lambda^* = \Lambda \beta$ and $f^* = \beta^1 f$, then $f^*$ has the same statistical properties as $f$ since

$$E(f^*) = E(\beta^1 f) = \beta^1 E(f) = 0.$$  
$$\text{Cov}(f^*) = \text{Cov}(\beta^1 f) = \beta^1 \text{Cov}(f) \beta = \beta^1 \beta = I_{mxm}$$

$\Lambda$ and $\Lambda^1$ yield the same covariance because $\Lambda \Lambda^1 = \Lambda^* \Lambda^{*1}$. The factor model

$$(y - \mu) = \Lambda f + \epsilon = \Lambda \beta \beta^1 f + \epsilon$$

$= \Lambda^* f^* + \epsilon,$ produces the same covariance matrix $\Sigma$, since $\Sigma = \Lambda \Lambda^1 + \psi = \Lambda^* \Lambda^{*1} + \psi$. 

© 2014 British Journals ISSN 2047-3745
2.1 THE PRINCIPAL FACTORS METHOD:

The principal factor method (also called the principal axis method) uses an initial estimate $\hat{\Psi}$ and factors $S - \hat{\Psi}$ or $R - \hat{\Psi}$ to obtain

$$S - \hat{\Psi} \cong \hat{\Lambda} \hat{\Lambda}^T,$$

$$R - \hat{\Psi} \cong \hat{\Lambda} \hat{\Lambda}^T.$$  \hspace{1cm} (2.1.1)

where $\hat{\Lambda}$ is pxm and is calculated using the eigenvalues and eigenvectors of $S - \hat{\Psi}$ or $R - \hat{\Psi}$.

The i-th diagonal element of $S - \hat{\Psi}$ is given by $S_{ii} - \hat{\Psi}_{ii}$, which is the i-th communality $\hat{h}_i^2 = S_{ii} - \hat{\Psi}_{ii}$. Likewise, the diagonal elements of $R - \hat{\Psi}$ are the communalities $\hat{h}_i^2 = 1 - \hat{\Psi}_{ii}$. (Clearly, $\hat{\Psi}_{ii}$ and $\hat{h}_i^2$ have different values for S than for R). With these diagonal values, $S - \hat{\Psi}$ and $R - \hat{\Psi}$ have the form

$$S - \hat{\Psi} = \begin{bmatrix} \hat{h}_1^2 & S_{12} & \ldots & S_{1p} \\ S_{21} & \hat{h}_2^2 & \ldots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \ldots & \hat{h}_p^2 \end{bmatrix}$$  \hspace{1cm} (2.1.3)

$$R - \hat{\Psi} = \begin{bmatrix} \hat{r}_1^2 & r_{12} & \ldots & r_{1p} \\ r_{21} & \hat{r}_2^2 & \ldots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \ldots & \hat{r}_p^2 \end{bmatrix}$$  \hspace{1cm} (2.1.4)

A popular initial estimate for a communality in $R - \hat{\Psi}$ is $\hat{h}_i^2 = \hat{r}_i^2$, the squared multiple correlation between $y_i$ and the other $p - 1$ variables. This can be found as

$$\hat{h}_i^2 = \hat{r}_i^2 = 1 - \frac{1}{\hat{S}_{ii}}$$  \hspace{1cm} (2.1.5)

where, $r_{ii}$ is the i-th diagonal element of $R^{-1}$.

For $S - \hat{\Psi}$, an initial estimate of the communality is

$$\hat{h}_i^2 = S_{ii} - \frac{1}{\hat{S}_{ii}}$$  \hspace{1cm} (2.1.6)

where $S_{ii}$ is the i-th diagonal element of $S$ and $S_{ii}^{-1}$ is the i-th diagonal element of $S^{-1}$.

It can be shown that equation (2.1.6) is equivalent to

$$\hat{h}_i^2 = S_{ii} - \frac{1}{\hat{S}_{ii}} = S_{ii} \hat{r}_i^2$$  \hspace{1cm} (2.1.7)

which is a reasonable estimate of the amount of variance that $y_i$ has in common with the other $y$'s.

© 2014 British Journals ISSN 2047-3745
to use equation (2.1.5) or (2.1.6), \( R \) or \( S \) must be non-singular. If \( R \) is singular, we can use the absolute value or the square of the largest correlation in the i-th row of \( R \) as an estimate of communality.

After obtaining communality estimates, we calculate eigenvalues and eigenvectors of \( S - \widehat{\Psi} \) or \( R - \widehat{\Psi} \) and use it to obtain estimates of factor loadings, \( \widehat{\Lambda} \). Then, the columns and rows of \( \widehat{\Lambda} \) can be used to obtain new eigen values (variance explained) and communalities, respectively. The sum of squares of the j-th column of \( \widehat{\Lambda} \) is the j-th eigen values of \( S - \widehat{\Psi} \) or \( R - \widehat{\Psi} \), and the sum of squares of the i-th row of \( \widehat{\Lambda} \) is the communality of \( \psi_i \). The proportion of variance explained by the j-th factor is

\[
\frac{\sum_{i=1}^{p} \lambda_i^2}{\text{Tr}(S - \widehat{\Psi})} \quad \text{or} \quad \frac{\sum_{i=1}^{p} \lambda_i^2}{\text{Tr}(R - \widehat{\Psi})} = \frac{\theta_j}{\sum_{i=1}^{p} \theta_i},
\]

(2.1.8)

where \( \theta_j \) is the j-th eigenvalues of \( S - \widehat{\Psi} \) or \( R - \widehat{\Psi} \). The matrices \( S - \widehat{\Psi} \) and \( R - \widehat{\Psi} \) are not necessarily positive semi-definite and will often have some small negative eigenvalues. In such a case, the cumulative proportion of variance will exceed 1 and then decline to 1 as the negative eigen values are added.

2.2 THE PRINCIPAL COMPONENT FACTORS METHOD:

The first technique we consider is commonly called the principal component method. The covariance matrix \( \Sigma \) is represented by the spectral decomposition. If \( \Sigma \) has eigen values \( \lambda_i \) with \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0 \) and corresponding eigen vector \( e_i \), then,

\[
\Sigma = \lambda_1 e_1^t e_1 + \lambda_2 e_2^t e_2 + \cdots + \lambda_p e_p^t e_p = \Lambda_{(p \times p)} \Lambda_{(p \times p)}^t,
\]

where \( \Lambda_{(p \times p)} = (\sqrt{\lambda_1} e_1, \sqrt{\lambda_2} e_2, \ldots, \sqrt{\lambda_p} e_p) \). Because this matrix product form represents the covariance matrix perfectly, there is no uniqueness. However, this for is not useful for factor analysis because the number of factors should be less than the number of variables. It is desirable to have a model that explain the covariance with small number of common factors and leave the differences between two as uniqueness. Suppose the number of variables is p and the number of factors is m, then we find m such that the last \( p - m \) eigen values are so small that we can ignore the contribution of \( \Sigma = \lambda_{m+1} e_{m+1} e_{m+1}^t + \lambda_{m+2} e_{m+2} e_{m+2}^t + \cdots + \lambda_p e_p e_p^t \) to \( \Sigma \).

We then have,

\[
\Sigma = \lambda_1 e_1^t e_1 + \lambda_2 e_2^t e_2 + \cdots + \lambda_m e_m e_m^t + \text{diag}(\Psi_1, \Psi_2, \ldots, \Psi_p)
\]

\[
= \Lambda_{(p \times m)} \Lambda_{(p \times m)}^t + \Psi;
\]
where \( \Lambda_{(p \times p)} = (\sqrt{\lambda_1} e_1, \sqrt{\lambda_2} e_2, ..., \sqrt{\lambda_m} e_m) \) and \( \Psi_1, \Psi_2, ..., \Psi_p \) are very small. This representation is principal components solution.

Because the units of the variables of the original data may be different, standardization is required for a factor model. That is, \( Z_{ij} = \frac{(y_{ij} - \bar{y}_j)}{\sqrt{s_{jj}}} \), \( i = 1, ..., m \) and \( j = 1, ..., p \), is required because some variables with large variances influences the determination of factor loading too much. In this case, the sample covariance matrix, \( S \), becomes the sample correlation matrix, \( R \).

For the principal components method, the estimate of each factor loading is fixed independent of the number of factors.

If \( \tilde{\Lambda} = \left( \sqrt{\hat{\lambda}_1} e_1, \sqrt{\hat{\lambda}_2} e_2, ..., \sqrt{\hat{\lambda}_q} e_q \right) \) for \( m = q \), and
\[
\tilde{\Lambda} = \left( \sqrt{\hat{\lambda}_1} e_1, \sqrt{\hat{\lambda}_2} e_2, ..., \sqrt{\hat{\lambda}_q} e_q \right)
\]
for \( m = n, q < n \), then \( \left( \sqrt{\hat{\lambda}_1} e_1, \sqrt{\hat{\lambda}_2} e_2, ..., \sqrt{\hat{\lambda}_m} e_m \right) \) are the same for both cases.

One way of determining the number of factors \( m \) is to consider the residual matrix \( S - (\tilde{\Lambda}\tilde{\Lambda}^T + \Psi) \). If \( m \)'s are chosen to ensure that the residual matrices are small enough, the least number of \( m \) among all \( m \)'s that satisfy the small residual matrix condition is appropriate. The sum of squared entries of
\[
S - \left( \tilde{\Lambda}\tilde{\Lambda}^T + \Psi \right) \leq \hat{\lambda}_{m+1}^2 + \hat{\lambda}_{m+2}^2 + ... + \hat{\lambda}_p^2
\]
This means that if the right hand side of the inequality is small, then the left side should also be small.

The contribution to the total sample variance from the k-th common factor is
\[
\hat{\lambda}_{1k}^2 + \hat{\lambda}_{2k}^2 + ... + \hat{\lambda}_{pk}^2 = \left( \sqrt{\hat{\lambda}_k e_k} \right)^T \left( \sqrt{\hat{\lambda}_k e_k} \right) = \hat{\lambda}_k.
\]
Therefore, the proportion of the total sample variance due to the k-th factors equal \( \frac{\hat{\lambda}_k}{S_{11} + S_{22} + ... + S_{pp}} \), for a sample covariance matrix, \( S \), and \( \frac{\hat{\lambda}_k}{p} \) for a sample correlation matrix \( R \). From this proportion, \( m \) is chosen to obtain the appropriate high proportion.

2.3 ITERATED PRINCIPAL FACTORS METHOD

The principal factor method can easily be iterated to improve the estimates of communality. After obtaining \( \tilde{\Lambda} \) from \( S - \tilde{\Psi} \) or \( R - \tilde{\Psi} \) in \( S - \tilde{\Psi} \cong \tilde{\Lambda}\tilde{\Lambda}^T \), or \( R - \tilde{\Psi} \cong \tilde{\Lambda}\tilde{\Lambda}^T \) using initial communality estimates, we can obtain new communality estimates from the loadings in \( \tilde{\Lambda} \) using \( \hat{\lambda}_{ik}^2 = \sum_{i=1}^{p} \hat{\lambda}_{ij}^2 \). The value of
are substituted into the diagonal of $S - \hat{\Psi}$ or $R - \hat{\Psi}$ from which we obtain a new value of $\hat{\Lambda}$ using $\hat{\Lambda} = C_1 D_1^{\frac{1}{2}} = (\sqrt{\theta_1 c_1}, \sqrt{\theta_2 c_2}, ..., \sqrt{\theta_m c_m})$, where $D_1 = diag(\theta_1, \theta_2, ..., \theta_m)$ with the $m$ largest eigenvalues $\theta_1 > \theta_2 > ... > \theta_m$ and $C_1 = (e_1, e_2, ..., e_m)$ containing the corresponding eigenvectors. $\hat{\Lambda}$ is $pxm$, $C_1$ is $pxm$ and $D_1^{\frac{1}{2}}$ is $mxm$.

This process is continued until the communality estimates converge. (For some data sets, the iterative procedure does not converge). Then the eigen values and eigenvectors of the final version of $S - \hat{\Psi}$ or $R - \hat{\Psi}$ are used as in $\hat{\Lambda} = C_1 D_1^{\frac{1}{2}} = (\sqrt{\theta_1 c_1}, \sqrt{\theta_2 c_2}, ..., \sqrt{\theta_m c_m})$ to obtain the loadings.

The principal factor method and iterated principal factor method will typically yield results very close to those from the principal component methods when either of the following is true.

a. The correlations are fairly large, with a resulting small value of $m$.
b. The number of variables, $p$, is large.

A shortcoming of the iterative approach is that sometimes it leads to a communality estimate $\hat{h}_{ii}^2$ exceeding 1 (when factoring $R$); such a result is known as a Heywood case. If $\hat{h}_{ii}^2 > 1$, then $\hat{\Psi} < 0$ by $\hat{\Psi} = S_{ii} - \sum_{i=1}^{p} \hat{\lambda}_{ij}^2$ and $\hat{h}_{ii}^2 = \sum_{i=1}^{p} \hat{\lambda}_{ij}^2$ which is clearly improper since we cannot have a negative specific variance. Thus, when communality exceeds 1, the iteration process should stop, with the program reporting that a solution cannot be reached. Some software programs have an option of continuing the iteration by setting the communality equal to 1 in all subsequent iterations.

2.4 THE MAXIMUM LIKELIHOOD METHOD:

If the distribution of $X$ and the specific factors, $\varepsilon$, are assumed to be normal, the estimates of the factor loadings and the uniqueness can be obtained using the maximum likelihood method (Johnson and Winchern, 2007). The distribution of $Y$ given $\Lambda, \psi \hat{\Psi}$ is normal with mean 0 and covariance matrix $\Lambda \Lambda^T + \psi \hat{\Psi}$, and the likelihood is

$$L(\Lambda, \psi \hat{\Psi}) \propto \det(\Sigma)^{-\frac{n}{2}} e^{-\frac{1}{2} \sum_{i=1}^{n}(y_i - \mu_y)^T \Sigma^{-1} (y_i - \mu_y)}.$$  

The resulting log likelihood (LL) is

$$LL(\Lambda, \psi \hat{\Psi}) \propto -\frac{n}{2} \ln \det(\Sigma) - \frac{1}{2} \sum_{i=1}^{n}(y_i - \mu_y)^T \Sigma^{-1} (y_i - \mu_y)^T.$$  

The maximum likelihood estimates of $\Lambda, \psi \hat{\Psi}$ (called $\hat{\Lambda}$ and $\hat{\psi} \hat{\Psi}$) can be obtained by maximizing the LL. It can be shown that the estimates $\hat{\Lambda}$ and $\hat{\psi} \hat{\Psi}$ satisfy the following;
Based on the invariance property of maximum likelihood estimates, the maximum likelihood estimate of the communality due to j-th factor is 

$$\lambda^2_j = \hat{\lambda}^2_j.$$  

As a result, the proportion of the total sample variance due the j-th factor is 

$$\frac{\lambda^2_j}{\sum \lambda^2_j},$$

where \(\lambda^2_j\) is the (i,i)-th entry of the sample covariance matrix which is an estimated of the unknown population covariance matrix, \(\Sigma\).

If \(y_i\) is standardized to be 

$$Z_i = V^{-\frac{1}{2}}(y_i - \mu_y),$$

the covariance matrix \(Q\) is given as 

$$Q = V^{-\frac{1}{2}} \Sigma V^{-\frac{1}{2}} = \left(V^{-\frac{1}{2}} \Lambda \right)^{1/2} + V^{-\frac{1}{2}} \Psi^2 V^{-\frac{1}{2}},$$

where \(V^{-\frac{1}{2}}\) is the diagonal matrix with the reciprocal of the sample standard deviation on the main diagonal of \(y_i\). Based on the invariance property of maximum likelihood estimators, the maximum likelihood estimator of \(Q\) is 

$$\tilde{Q} = \hat{\Lambda}_z \hat{\Lambda}_z + \hat{\Psi}^2_z,$$

where \(\hat{\Lambda}_z = V^{-\frac{1}{2}} \hat{\Lambda} V^{-\frac{1}{2}}\).

The proportion of total standardized sample variance due to j-th factor is 

$$\frac{\sum \lambda^2_j}{tr(S)} = \frac{\lambda^2_j + \lambda^2_q}{S_{11} + S_{22} + \cdots + S_{qq}} = \lambda^2_j + \lambda^2_q, \quad \text{where} \quad S_{ii} = \lambda^2_j + \lambda^2_q.$$ 

2.5 Information Criteria.

The necessity of introducing the concept of model evaluation has been recognized as one of the important technical areas and the problem is posed on the choice of the best approximating model among a class of competing models by suitable model evaluation criteria given a data set. Based on the usual factor analysis model, choosing a model with too few parameters can involve making unrealistically simple assumptions and lead to high bias, poor prediction, and missed opportunities for insight. Such models are not flexible enough to describe the sample or the population well. A model with too many parameters can fit the observed data very well, but be too closely tailored to it; such models may generalize poorly. Penalized-likelihood information criteria, such as Akaike’s information criterion (AIC), the Schwarz’s information criterion (SIC), the Hannan-Quinn information criterion (HQIC) and so on are widely used for model selection. The comparison of the models using information criterion can be viewed as equivalent to a likelihood ratio test and understanding the differences.
among the criteria may make it easier to compare the results and to use them to make informed decisions.

AKAIKE’S information criterion is probably the most relevant and famous as for the comparison and selection between different models and is constructed on log likelihood

\[
\text{AIC} = -2 \log \max L + 2k
\]

where \( L \) denotes the likelihood function of the factor model and \( k \) is the number of the model’s parameter/factors. \( \log \max L(k) = -\frac{1}{2} N \log |\Sigma_k| + \text{tr} \Sigma_k^{-1} S \), where \( S \) denotes the sample covariance matrix of \( Y \) and \( \Sigma_k = \Lambda_k \Lambda_k^{-1} + \psi_i^2 \); \( \Lambda_k \) is the matrix factor of factor loading. The first term can be interpreted as a goodness-of-fit measurement, while the second gives a growing penalty to increasing numbers of parameters according to the parsimony principle. In the choice of model, a minimization rule is used to select the model with the minimum Akaike information criterion value.

Still in the likelihood based procedures, Schwarz (1978) proposed the alternative information criterion given by

\[
\text{SIC} = -\log \max L + \frac{1}{2} k \log N.
\]

Unlike the AIC, SIC considers the number of \( N \) of observations and is therefore less favourable to factors inclusion.

Finally, the third criterion is the Hannan-Quinn information criterion (HQC); it is a criterion for model selection. It is an alternative to Akaike information criterion (AIC) and Bayesian information criterion (BIC). It is given as

\[
\text{HQC} = -\log \max L + 2k \log \log N,
\]

where \( k \) is the number of parameters, \( N \) is the number of observations.
3.1 THE RESULTS FROM THE FOUR METHODS OF ESTIMATION AT SAMPLE SIZES OF 30, 50 AND 70 FOR THE 10 VARIABLE CASE.

Table 3.1.1. Method of Principle Factors

<table>
<thead>
<tr>
<th>n= 30, p= 10</th>
<th>K</th>
<th>LR</th>
<th>AIC</th>
<th>SIC</th>
<th>HQIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-118.6680</td>
<td>241.3360</td>
<td>120.1451</td>
<td>119.3457</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-115.2820</td>
<td>236.5640</td>
<td>117.4977</td>
<td>116.2985</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-111.6615</td>
<td>233.3230</td>
<td>115.3543</td>
<td>113.3557</td>
<td></td>
</tr>
</tbody>
</table>

n= 50, p = 10

| 2            | -239.8450 | 483.6900 | 241.5440 | 240.7657 |
| 3            | -236.4975 | 478.9950 | 239.0460 | 237.8786 |
| 5            | -237.0900 | 484.1800 | 241.3374 | 239.3919 |

n= 70, p = 10

| 2            | -328.3934 | 660.7868 | 330.2385 | 329.4575 |
| 3            | -326.2455 | 656.4910 | 329.0131 | 327.8416 |
| 5            | -323.6940 | 651.3880 | 328.3067 | 326.3542 |

Table 3.1.2. Method of principle Component Factors

<table>
<thead>
<tr>
<th>n= 30, p= 10</th>
<th>K</th>
<th>LR</th>
<th>AIC</th>
<th>SIC</th>
<th>HQIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-121.3815</td>
<td>246.7630</td>
<td>122.8586</td>
<td>122.0592</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-121.1655</td>
<td>246.3310</td>
<td>123.3812</td>
<td>122.1820</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-126.6940</td>
<td>263.3050</td>
<td>130.3453</td>
<td>128.3467</td>
<td></td>
</tr>
</tbody>
</table>

n= 50, p = 10

| 2            | -249.2975 | 502.5950 | 250.9965 | 250.2182 |
| 3            | -250.6500 | 507.3000 | 253.1985 | 252.0311 |
| 5            | -268.3175 | 546.6350 | 272.5649 | 270.6194 |

n= 70, p = 10

| 2            | -341.9745 | 687.9490 | 343.8196 | 343.0386 |
| 3            | -349.9615 | 705.9230 | 352.7291 | 351.5576 |
| 5            | -367.6505 | 745.3010 | 372.2632 | 370.3107 |

Table 3.1.3. Method of Iterated Principle Factors

<table>
<thead>
<tr>
<th>n= 30, p= 10</th>
<th>K</th>
<th>LR</th>
<th>AIC</th>
<th>SIC</th>
<th>HQIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-117.3765</td>
<td>238.7530</td>
<td>118.8536</td>
<td>118.0542</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-112.9410</td>
<td>231.8820</td>
<td>115.1567</td>
<td>113.9575</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-116.6615</td>
<td>233.3230</td>
<td>115.3543</td>
<td>113.3557</td>
<td></td>
</tr>
</tbody>
</table>

n= 50, p = 10

| 2            | -237.8175 | 479.6350 | 239.5165 | 238.7382 |
| 3            | -234.1150 | 474.2300 | 236.6635 | 235.4961 |
| 5            | -230.0950 | 470.1900 | 234.3424 | 232.3969 |

n= 70, p = 10

| 2            | -326.9945 | 657.9890 | 328.8396 | 325.0586 |
| 3            | -323.7360 | 653.4720 | 326.5036 | 325.3321 |
| 5            | -319.8685 | 649.7370 | 324.4812 | 322.5287 |

Table 3.1.4. Method of Maximum Likelihood Factors

<table>
<thead>
<tr>
<th>n= 30, p= 10</th>
<th>K</th>
<th>LR</th>
<th>AIC</th>
<th>SIC</th>
<th>HQIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-116.6265</td>
<td>237.2430</td>
<td>118.1036</td>
<td>117.3042</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-112.6245</td>
<td>231.2490</td>
<td>114.8402</td>
<td>113.6420</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-108.2130</td>
<td>226.4260</td>
<td>111.9058</td>
<td>109.9072</td>
<td></td>
</tr>
</tbody>
</table>

n= 50, p = 10

| 2            | -237.8175 | 479.6350 | 239.5165 | 238.7382 |
| 3            | -234.1150 | 474.2300 | 236.6635 | 235.4961 |
| 5            | -230.0950 | 470.1900 | 234.3424 | 232.3969 |

n= 70, p = 10

| 2            | -326.9945 | 657.9890 | 328.8396 | 325.0586 |
| 3            | -323.7360 | 653.4720 | 326.5036 | 325.3321 |
| 5            | -319.8685 | 649.7370 | 324.4812 | 322.5287 |
### 3.2 THE RESULTS FROM THE FOUR METHODS OF ESTIMATION AT SAMPLE SIZES OF 30, 50 AND 70 FOR THE 15 VARIABLE CASE.

<table>
<thead>
<tr>
<th>Table 3.2.1. Method of Principle Factors</th>
<th>Table 3.2.2. Method of Principle Component Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>n = 30, p = 15</strong></td>
<td><strong>n = 30, p = 15</strong></td>
</tr>
<tr>
<td><strong>K</strong></td>
<td><strong>LR</strong></td>
</tr>
<tr>
<td>5</td>
<td>-193.4180</td>
</tr>
<tr>
<td>7</td>
<td>-183.7143</td>
</tr>
<tr>
<td>9</td>
<td>-176.6939</td>
</tr>
<tr>
<td>10</td>
<td>-176.0506</td>
</tr>
<tr>
<td><strong>n = 50, p = 15</strong></td>
<td><strong>n = 50, p = 15</strong></td>
</tr>
<tr>
<td>5</td>
<td>-335.1125</td>
</tr>
<tr>
<td>7</td>
<td>-328.5867</td>
</tr>
<tr>
<td>9</td>
<td>-321.2146</td>
</tr>
<tr>
<td>10</td>
<td>-622.1136</td>
</tr>
<tr>
<td><strong>n = 70, p = 15</strong></td>
<td><strong>n = 70, p = 15</strong></td>
</tr>
<tr>
<td>5</td>
<td>-486.5374</td>
</tr>
<tr>
<td>7</td>
<td>-480.6700</td>
</tr>
<tr>
<td>9</td>
<td>-479.9556</td>
</tr>
<tr>
<td>10</td>
<td>-502.6715</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3.2.3. Method of Iterated Principle Factor</th>
<th>Table 3.2.4. Method of Maximum Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>n = 30, p = 15</strong></td>
<td><strong>n = 30, p = 15</strong></td>
</tr>
<tr>
<td><strong>K</strong></td>
<td><strong>LR</strong></td>
</tr>
<tr>
<td>5</td>
<td>-193.4210</td>
</tr>
<tr>
<td>7</td>
<td>-183.7161</td>
</tr>
<tr>
<td>9</td>
<td>-176.6943</td>
</tr>
<tr>
<td>10</td>
<td>-176.0512</td>
</tr>
<tr>
<td><strong>n = 50, p = 15</strong></td>
<td><strong>n = 50, p = 15</strong></td>
</tr>
<tr>
<td>5</td>
<td>-326.9147</td>
</tr>
<tr>
<td>7</td>
<td>-328.5517</td>
</tr>
<tr>
<td>9</td>
<td>-325.9611</td>
</tr>
<tr>
<td><strong>n = 70, p = 15</strong></td>
<td><strong>n = 70, p = 15</strong></td>
</tr>
<tr>
<td>5</td>
<td>-486.5358</td>
</tr>
<tr>
<td>7</td>
<td>-480.6671</td>
</tr>
<tr>
<td>9</td>
<td>-479.9917</td>
</tr>
<tr>
<td>10</td>
<td>-480.0345</td>
</tr>
</tbody>
</table>

© 2014 British Journals ISSN 2047-3745
3.3. THE NUMBER OF FACTORS MODELS \((M)\) RETAINED BY THE FIVE METHODS AND FOR THE FOUR METHODS OF ESTIMATION Considered and for \(P = 10\).

<table>
<thead>
<tr>
<th>N</th>
<th>Method I</th>
<th>Method II</th>
<th>Method III</th>
<th>Method IV</th>
<th>Method V</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>50</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>70</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

3.4. THE NUMBER OF FACTORS MODELS \((M)\) RETAINED BY THE FIVE METHODS AND FOR THE FOUR METHODS OF ESTIMATION Considered and for \(P = 15\).

<table>
<thead>
<tr>
<th>N</th>
<th>Method I</th>
<th>Method II</th>
<th>Method III</th>
<th>Method IV</th>
<th>Method V</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>7</td>
<td>2</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>70</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N</th>
<th>Method I</th>
<th>Method II</th>
<th>Method III</th>
<th>Method IV</th>
<th>Method V</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>11</td>
<td>11</td>
<td>6</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>50</td>
<td>9</td>
<td>10</td>
<td>8</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>70</td>
<td>10</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>

(Note: Method I - percentage of variance criterion, Method II - eigenvalues greater than the average of the eigenvalues criterion, Method III is latent root criterion, Method IV - the scree test criterion, Method V – Information Criterion.)

© 2014 British Journals ISSN 2047-3745
4.0 SUMMARY

The result for the 10 variable cases as shown in Table 3.1.1 to Table 3.1.4 shows that the method of principal component factor estimation (n=30) accepts the 2 factor solution as the best for SIC and HQIC and also the three information criteria for n=50 and n=70. The 3 factor solution is accepted by the iterated principal factor (n=70) and principal factor (n=50) for the three information criteria; iterated principal factor (n=30) for the AIC and SIC and principal component factor (n=30) for the AIC. Finally, the 5 factor solution is selected as the best number of factors to retain by iterated principal factor (n=50), principal factor (n=30, 70), maximum likelihood (n=30, 50 and 70) for all the three information criteria of AIC, Sic and HQIC; iterated principal factor (n=70) only for the HQIC.

From Table 3.3.1 - Table 3.3.4, the results show that method I underestimated the number of factors for the methods of principal factors and iterated principal factors and overestimated the factors for the method of principal component factors; the extraction is good for the method of maximum likelihood. In method II, the factor extraction are okay for the method of iterated principal factors and are underestimated for the methods of principal factors, principal component factors and maximum likelihood. Method III underestimated the factors for all the methods except for the method of principal component factors. Method IV underestimated the factors for the method of maximum likelihood and overestimated the factors for the methods of principal component factors at (n=50) and iterated principal factors at (n=30). Finally, method V underestimated the factor extraction for only the method of principal component factors.

From the result for the 15 variable case as shown in Table 3.2.1 to Table 3.2.4, the 9 factor solution is selected as best number of factors to retain by the methods of principal factor (n=30 and 50), iterated principal factor (n=30) and maximum likelihood (n=30, 50 and 70) for all the three information criteria. The seven factor solution by principal factor (n= 70), iterated principal factor (n=70) for the three information criteria and principal component factor (n=70) for the SIC and HQIC. The 5 factor solution was chosen by principal component factor (n=30 and 50), iterated principal factor (n=50) for the three information criteria and principal component factor (n=70) for only the Akaike Information Criterion; whereas none of the estimation methods selected the 10 factor solution. Finally, looking at the values of the three information criteria that are AIC, SIC and HQIC, we observed that the values increases as the sample size increases for both the 10 and 15 variable cases.

From Table 3.4.1 to Table 3.4.4 on the comparison of the five methods for the 15 variables case, method I underestimated the number of factors for the method of principal factors, iterated principal factors, maximum likelihood at (n=70) and overestimated the number of factors for the method of principal component factor.
at (n=30 and 70). Method II overestimated the factors for the method of principal component factor at (n=30 and 50) and underestimate the factors for the maximum likelihood, and principal factors at (n=30 and 70) methods. In method III, we obtained an underestimated number of factors for the methods of principal factors, iterated principal factors, and maximum likelihood at (n=70) and overestimate for the maximum likelihood at (n=30). In method IV almost all the methods underestimated the factors. And finally, method V shows a good extraction of factors for almost all the methods and sample sizes except for the method of principal component factors at (n=30 and 50).

5.0. CONCLUSION.

The results of the analysis for the 10 variable cases shows that the majority of the estimation methods at different sample sizes chooses the five factor solution as the best whereas for the 15 variable cases the nine factor solution was selected as the number of factors to retain using the information criteria approach. The values of the maximum likelihood method are consistent with all the sample sizes for the two variable cases considered. Hence, we conclude that for the two variable cases considered, the best number of factors to retain is 5 factors and 9 factors for the 10 variable and 15 variable cases respectively. Also, comparing the information criteria approach with the Hair’s rules, the information criteria approach performed best among the five methods compared for most of the four estimation methods of the two variables cases (p = 10 and 15) and the sample sizes considered. Also, the results show that the method of maximum likelihood estimation is adjudged the best among the four methods of orthogonal factor model extraction considered in this work using the information criteria approach.

6.0. REFERENCES


© 2014 British Journals ISSN 2047-3745


ABOUT THE AUTHORS

1. Mr. Nwosu Dozie F. is a lecturer at Federal Polytechnic Nekede P.M.B. 1234 Owerri, Imo State. His contact number is 08038410702. He is at the final stage of his Ph.D at Nnamdi Azikiwe University Awka.

2. Professor Onyeagu Sidney I. is a professor of Statistics Nnamdi Azikiwe University P.M.B 5025 Awka Anambra State. His area of specialization is multivariate statistics. He can be contacted through the phone number 08036758066.

3. Doctor Osuji George A. is a senior lecturer at the department of statistics Nnamdi Azikiwe University P.M.B 5025 Awka Anambra State. He specializes in mathematical statistics and probability. His phone number is 08033895073.