# A Graphical Procedure for Determining Useful Principal Components

**Nahed. A. Mokhlis**
Department of Mathematics, Faculty of Science, Ain Shams University, Egypt
**Sahar A. N. Ibrahim**
Department of Mathematical Statistics, Institute of Statistical Studies and Research, Cairo University, Egypt
**Nadia B. Gregni**
Department of Statistics, Faculty of Science, El Fateh University, Libya

Abstract

One of the purposes of principal component analysis is to reduce the dimensionality of the set of variables. Several approaches have been suggested by different authors for determining the number of principal components that should be kept for further analysis. In this paper we present a graphical procedure depending on the computation of the coefficient of multiple determination of each variable when this variable is regressed on the other variables. A comparison of our criterion with the eigenvalue-one criterion, the Scree test criterion and the percentage criterion is given through examples. Our criterion can be considered as a lower bound for principal components retained. It is a precise one, in the sense that when different people analyze the same data they will obtain the same result. In addition it takes into account the components with variance smaller than one but important.

Key words: Eigenvalues, Scree Test, Communality, Coefficient of Multiple Determination.

## 1. Introduction

Principal component analysis (PCA) is a technique that is useful for the compression and classification of data. The dimensionality reductions of multivariate data attract the attention of many authors. The main idea depends on finding a new set of variables, smaller than the original ones, which retains most of the sample's information. Information means the variation presented in the sample and given by the correlations between the original variables. These new variables, called principal components, are uncorrelated.

Several criteria in the literature have been proposed for determining the number of principal components (pc's) that should be kept for further analysis. A large sample test for hypothesis that the last roots are equal was developed by Bartlett (1950). This test can be performed after each stage. If the remaining roots are not significantly different from each other, the procedure is terminated at that point. Most practitioners would agree that Bartlett's test ends up retaining too many pc's, that is, it may retain pc's that explain very little of the total variance. Kaiser (1960) used a correlation matrix and proposed dropping principal components whose eigenvalues are less than one. His idea is based on the fact that if the eigenvalue is less than one, then the corresponding principal component will provide less information than that provided by a single variable. Jolliffe (1972) claimed that Kaiser's criterion is too large. He suggested using a cutoff on the eigenvalues of 0.7. Other authors noted that if the largest eigenvalue is close to one, then holding to a cutoff of one may cause useful principal components to be dropped. However, if the largest eigenvalue is several times larger than one, then those near one may be reasonably dropped. Cattell (1966) proposed the Scree test, which is a graphical technique. With the Scree test one plots the eigenvalues against the components numbers and retains the components which the line in the Scree graph is steep to the

left but not steep to the right. Studying this chart is probably the most popular criterion for determining the number of principal components, but it is subjective, causing different people to analyze the same data with different results. Also it can be considered as a graphical substitute for the significance test. The percentage of the total variation criterion (Manly 1994), (called also "proportion of trace explained" criterion in Jackson (1991)), is based on the idea that a certain percentage of the variation that must be accounted for is preseted, and then enough principal components are retained so that this percentage of variation is achieved. Usually, however, this cutoff percentage is used as a lower limit. That is, if the designated number of principal components do not account for at least 50% of the variance, then the whole analysis is aborted. Another procedure (Jackson 1991) is based on the amount of the explained and unexplained variability. In this procedure, one determines in advance the amount of residual variability that one is willing to tolerate. Characteristic roots and vectors are obtained until the residual has been reduced to that quantity. This procedure should be carried out after the significance test. Parallel Analysis is a Monte Carlo simulation technique that aides researcher in determining the number of factors to retain in principal component and exploratory factor analysis. This technique provides a superior alternative to other techniques. Horn (1965) suggested generating a random data set having the same number of variables and observations as the set being analyzed. These variables should be normally distributed but uncorrelated. A Scree plot of these roots will generally approach a straight line over the entire range. The intersection of this line and the Scree plot for the original data should indicate point separating the retained and deleted pc's. The reasoning being that any roots for the real data that are above the line obtained for the random data represent roots that are larger than they would be by chance alone. Ledesma and Pedro (2007) described an easy to use computer program capable of carrying out parallel analysis- the ViSta-PARAN program. Its user interface is fully graphic and includes a dialog box to specify parameters.

In this article we introduce a graphical procedure for determining the number of principal components retained for further analysis. Our approach is based on the fact that the determination of optimal number of pc's retained needs a measure that takes into account the intercorrelation between the variables and the information redundancy. We see that the coefficient of multiple determinations is a good measure for detecting the presence of intercorrelation within a set of variates. It also discloses the important pc's for a given variable through the determination of the highly correlated variables. Guttman (1953) used the word "index" to refer to the square of a correlation coefficient. He investigated possibilities to capitalize the structural of the multiple-correlation approach. He also studied the common and alien parts of the observed variates as defined by multiple correlations. In our procedure the coefficients of multiple determination of each variable when this variable is regressed on the other variables are computed, arranged in decreasing order, then the cumulative proportions are computed and graphed against the components numbers. The point of intersection of this curve with the curve of eigenvalues against the components numbers is considered. All components before this point are retained for further analysis and the others are dropped. A comparison of our criterion with the eigenvalue-one criterion, the Scree test criterion and the percentage criterion is given through examples. Our criterion may be considered as a lower bound for pc's retained. It is a precise one. in the sense that when different people analyze the same data they will obtain the same result. The logic of the proposed criterion is given in the following section.

## 2. The Rationale of the Criterion

Let $R$ be the sample correlation matrix of the normally distributed random vector $X' = [X_1, X_2, ....., X_p]$ for $n$ individuals. Let $\lambda_1, \lambda_2, \ldots, \lambda_p$ be the eigenvalues of $R$ and without loss of generality consider $\lambda_1 > \lambda_2 > .... > \lambda_p > 0$. The values of the principal components are calculated from the standardized variables $X^*$ where $X_{ij}^* = \frac{X_{ij} - \overline{X}_j}{S_j}$, $S_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X_j})^2$ and $\overline{X_j} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$ , $j = 1,2,..., p$. The $j-th$ principal component is the linear combination $Y_j = u_j' X^*$ which is obtained by maximizing $\text{var}(Y_j)$ where $\text{var}(Y_j) = \text{var}(u_j' X^*) = u_j' R u_j$ subjected to $u_j' u_j = 1$ and $u_i' u_k = 0$, and $u_j' = [u_{1j}, u_{2j}, ..., u_{pj}]$ is the eigenvector corresponding to the eigenvalue $\lambda_j$.

Let us denote the $p \times p$ orthogonal matrix, whose columns are the eigenvectors corresponding to the eigenvalues $\lambda_j$'s of the sample correlation matrix $R$. by $U$. where $U = [u_1, u_2, ....., u_p]$ and the $(p \times 1)$ vector of pc's by $Y$. Then $Y = U' X^*$. The $(p \times p)$ covariance matrix of $Y$, denoted by $\Delta$, is given by $\Delta = Diag(\lambda_1, \lambda_2, ..., \lambda_p)$. Notice that the diagonal matrix $\Delta$ results from diagonalizing the matrix $R$ by the orthogonal matrix $U$. i.e., $\Delta = U'RU$. This can also be written as $R = U \Delta U'$. It can be shown that (Jackson 1991)

$$\sum_{j=1}^{p} \text{var}(Y_j) = \sum_{j=1}^{p} \text{var}(X_j^*)$$

i.e.,

$$\sum_{j=1}^{p} \lambda_j = p$$

and

$$|R| = \prod_{j=1}^{p} \text{var}(Y_j) = \prod_{j=1}^{p} \lambda_j ,$$

where $|R|$ is the determinant of $R$.

Also $\frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j} = \frac{\lambda_j}{p}$ is the amount of variation of the $j-th$ pc to the total variation and $\frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{p} \lambda_j} = \frac{\sum_{j=1}^{m} \lambda_j}{p}$ is the amount of variation accounted by the first $m$ pc's to the total variation.

Now the $p$ principal components $Y_1, Y_2, ..., Y_p$, which are linear combinations of the standardized $p$ variables $X_1^*, X_2^*, ..., X_p^*$, have the form

$$Y_i = u_{1i}X_1^* + u_{2i}X_2^* + ... + u_{pi}X_p^* \quad , i = 1,2,..., p \tag{1}$$

This transformation from $X^*$ values to $Y$ values is orthogonal, so that the inverse relationship is simply

$$X_j^* = u_{j1}Y_1 + u_{j2}Y_2 + ... + u_{jp}Y_p \quad , j = 1,2,..., p$$

Let $m$ be the hypothetical number of pc's that have most of the variability of the standardized variables $X^*s$. So the last equation becomes.

$$X_j^* = u_{j1}Y_1 + u_{j2}Y_2 + ... + u_{jm}Y_m + e_j \quad , j = 1,2,..., p \tag{2}$$

where $e_j$ is a linear combination of the pc's $Y_{m+1}$ to $Y_p$. Now we need to scale the principal components $Y_1, Y_2, ..., Y_m$ to have unit variances. So Equation (2) can be rewritten as

$$X_j^* = u_{j1}\sqrt{\lambda_1}\, Y_1^* + u_{j2}\sqrt{\lambda_2}\, Y_2^* + ... + u_{jm}\sqrt{\lambda_m}\, Y_m^* + e_j \quad j = 1,2,..., p.$$

where $Y_i^* = \dfrac{Y_i}{\sqrt{\lambda_i}}$ , $i = 1,2,..., m$ and if we put $v_{ji} = u_{ji}\sqrt{\lambda_i}$, which is the correlation coefficient between $X_j$ and $Y_i$ we get

$$X_j^* = v_{j1}\, Y_1^* + v_{j2}\, Y_2^* + ... + v_{jm}\, Y_m^* + e_j \quad , \ j = 1,2,..., p \tag{3}$$

Since $\text{var}(X_j^*) = 1$ and $\text{var}(Y_i^*) = 1$, for $j = 1,2,..., p$ and $i = 1,2,..., m$ then

$$1 = \sum_{i=1}^{m} v_{ji}^2 + \text{var}(e_j) \quad , \quad j = 1,2,..., p \tag{4}$$

Thus $\sum_{i=1}^{m} v_{ji}^2 \le 1$ is the part of the variance of $X_j^*$ that is related to $Y_1^*, Y_2^*, ..., Y_m^*$. (called the communality of $X_j^*$ in the sense of factor analysis (Manly 1994))while $\text{var}(e_j)$ is the part of the variance of $X_j^*$ that is unrelated to $Y_1^*, Y_2^*, ..., Y_m^*$. It can be shown that, (Jackson 1991), the coefficients of multiple determination of the standardized variables $X_j^*s$ when they are regressed on the retained scaling principal components $Y_1^*, Y_2^*, ..., Y_m^*$ (denote by $R_{jm}^{*2}$) is equal to the part of variance of $X_j^*s$ that is related to $Y_i^*$ , $i = 1,2,..., m$ i.e.,

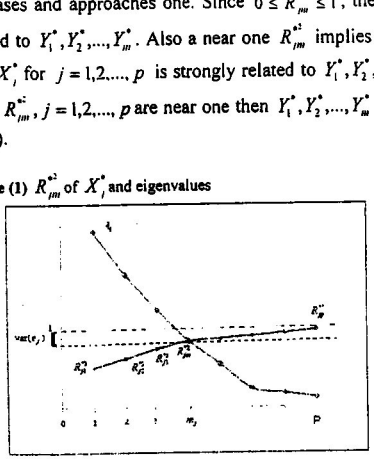
$$R^{*2}_{jm} \equiv R^{*2}_{X_j^*/Y_1^*,Y_2^*,\ldots,Y_m^*} = \sum_{i=1}^{m} v_{ji}^2 = \sum_{i=1}^{m} \lambda_i u_{ji}^2 \qquad (5)$$

That is to say, $R^{*2}_{jm}$ measures the proportionate reduction of total variation in $X_j^*$ associated with the use of the $m$ scaling principal components $Y_1^*, Y_2^*, \ldots, Y_m^*$ (Neter et al. 1983). So Equation (4) becomes

$$1 = R^{*2}_{jm} + \text{var}(e_j) \qquad , \; j = 1,2,\ldots,p \qquad (6)$$

Increasing $m$ will increase the part of the variance of $X_j^*s$ that is related to $Y_1^*, Y_2^*, \ldots, Y_m^*$ and consequently will increase $R^{*2}_{jm}$ and at the same time will decrease $\text{var}(e_j)$. Figure 1 presents the graph of $R^{*2}_{jm}$ against components numbers $m = 1,2,\ldots,p$. It is clear that as $m$ increases $R^{*2}_{jm}$ increases and approaches one. Since $0 \le R^{*2}_{jm} \le 1$, then a near zero $R^{*2}_{jm}$ implies that $X_j^*$ is weakly related to $Y_1^*, Y_2^*, \ldots, Y_m^*$. Also a near one $R^{*2}_{jm}$ implies that $X_j^*$ is strongly related to $Y_1^*, Y_2^*, \ldots, Y_m^*$. If each $X_j^*$ for $j = 1,2,\ldots,p$ is strongly related to $Y_1^*, Y_2^*, \ldots, Y_m^*$, but not related to $Y_{m+1}^*, Y_{m+2}^*, \ldots, Y_p^*$, i.e., if all $R^{*2}_{jm}$, $j = 1,2,\ldots,p$ are near one then $Y_1^*, Y_2^*, \ldots, Y_m^*$ is the best choice for the pc's retained (Manly 1994).

**Figure (1)** $R^{*2}_{jm}$ of $X_j^*$ and eigenvalues




For a given variable $X_j^*$, if we plot the curve $R^{*2}_{ji}$, $i = 1,2,\ldots,m_1,\ldots,p$, which is increasing, against components numbers $i$ together with the curve of eigenvalues $\lambda_i, i = 1,2,\ldots,p$, which is decreasing one against its components numbers $i$, we find that they intersect at a point for which it indicates the optimal numbers of pc's to retain (see Figure 1). At the point of intersection $R^{*2}_{jm_1} = \lambda_{m_1}$ then Equation (6) becomes

$$1 = \lambda_{m_1} + \text{var}(e_j)$$

So

$$\text{var}(e_j) = 1 - \lambda_{m_1}$$

Since

$$\lambda_{m_1} = R^{*2}_{jm_1} = \sum_{i=1}^{m_1} \lambda_i u_{ji}^2 \quad , \quad 0 \le \lambda_{m_1} \le 1 \text{ and } 0 \le u_{ji}^2 \le 1.$$

Then $\sum_{i=1}^{m_1} \lambda_i u_{ji}^2$ convergence to one, that is, the point of intersection should be close to one which means that $X_j^*$ is strongly related to $Y_1^*, Y_2^*, ..., Y_{m_1}^*$. From all above we conclude that the $m$-component model of Equation (3) is a good fit of $X_j^*$.

Now, if we do the same plot for all $X_j^*$'s , $j = 1,2,...,p$, we obtain different points of intersections. Here we face two problems. First, which one of these points of intersections we consider. Second as long as the number of variables increases, the computations of $R_{ji}^{*2}$, $i, j = 1,2,...,p$, will be tedious. So we try to find another measure replacing all these quantities and having the same range as $R_{ji}^{*2}$.

On other hand if we let $V$ be the $p \times p$ matrix of correlation coefficients of $X_i^*$'s and $Y_i^*$'s , $i, j = 1,2,...,p$. It is clear that $v_{ij} \neq v_{ji}$ for $i \neq j$ and $VV' = R$ is the source of the explained correlations among the variables and $V'V = \Delta$. It can be shown that

$$\sum_{j=1}^{p} v_{ji}^2 = \sum_{j=1}^{p} u_{ji}^2 \lambda_i = \lambda_i \sum_{j=1}^{p} u_{ji}^2 = \lambda_i = \text{var}(Y_i)$$

and

$$\sum_{i=1}^{p} v_{ji}^2 = \sum_{i=1}^{p} u_{ji}^2 \lambda_i = 1$$

then

$$\sum_{j=1}^{p} R_{ji}^{*2} = \sum_{j=1}^{p} \sum_{i=1}^{m} v_{ji}^2 = \sum_{i=1}^{m} \sum_{j=1}^{p} v_{ji}^2 = \sum_{i=1}^{m} \lambda_i \qquad (7)$$

Equation (7) means that the variation accounted by the first $m$ pc's is equal to the sum over all coefficients of multiple determination of $X_j^*$'s, $j = 1,2,...,p$, $R_{ji}^{*2}$. Also if we consider $Y_1^*, Y_2^*, ..., Y_m^*$ are the best choice then the correlation between $X_j^*$ and $X_k^*$ is given approximately by

$$v_{j1}v_{k1} + v_{j2}v_{k2} + ... + v_{jm}v_{km} = \sum_{i=1}^{m} v_{ji}v_{ki}$$

This means that $X_j^*$ and $X_k^*$ can only be highly correlated if and only if they have high loading on the same $Y_1^*, Y_2^*, ..., Y_m^*$ i.e., $R_{jm}^{*2}$ and $R_{km}^{*2}$ are large. So to obtain the optimal number $m$ of the retained pc's we need to determine all highly correlated variables $X_j^*$'s that have high loading on the same $Y_1^*, Y_2^*, ..... Y_m^*$.

So to overcome the above two problems we replace $R_{jj}^{2}$ by the coefficients of multiple determination of $X_j^*$'s when they are regressed on the other variables, denoted by $R_j^2 \equiv R^2_{X_j^*/X_1^*X_2^*...X_{j-1}^*X_{j+1}^*...X_p^*}$ , $j = 1,2,...,p$ which is a good measure for detecting the presence of intercorrelation between variables. As $R_{jj}^{2}$, $0 \le R_j^2 \le 1$, so a near zero $R_j^2$ means that $X_j^*$ is weakly correlated with the other $X^*$'s and a near one $R_j^2$ means that $X_j^*$ is highly correlated with the other $X^*$'s.

Notice that $R_j^2$ can be computed using different formulas.

$$R_j^2 = 1 - \frac{1}{c_{jj}} \qquad \text{(Neter 1983);} \tag{8}$$

where $c_{jj}$ is the $jth$ diagonal element of the inverse of the matrix $R$ .

$$R_j^2 = 1 - \frac{1}{\sum_{i=1}^{p} \frac{u_{ji}^2}{\lambda_i}} \qquad \text{(Jackson 1991)} \tag{9}$$

Jackson (1991) noticed that two or more variables are highly correlated i.e., two or more $R_j^2$ will be near one, resulting in one or more eigenvalues being positive but quite small. Therefore, as we arrange $\lambda_j$'s in decreasing order to capture the maximum amount of variation accounted by the first $m$ pc's, we arrange $R_j^2$'s in decreasing order say $R_{(1)}^2 > R_{(2)}^2 > ... > R_{(p)}^2$ and take the sum over the first $m$ $R_{(j)}^2$'s. Since this sum will exceed one and the measure we need should have the range [0,1], so it is evident to relate this sum to the total sum of $R_{(j)}^2$'s, $j = 1,2,...,p$. That is, we compute $CR_m^2$ given in the following equation

$$CR_m^2 = \frac{\sum_{j=1}^{m} R_{(j)}^2}{\sum_{j=1}^{p} R_{(j)}^2} \qquad , m = 1,2,...,p \tag{10}$$

Plotting the curve of $CR_m^2$ for $m = 1,2,...,p$, which is an increasing one, and the curve of eigenvalues $\lambda_j$'s, which is decreasing one, we find that as long as $CR_m^2$ is increasing to approach the maximum value one, the variability of the components (eigenvalues) are decreasing to approach zero. This means that the optimal number of pc's retained will be before the point of intersection.

The following lemma shows that $R_j^2$ is a function of all $R_{im}^{*2}$ $,i=1,2,...,p$. Consequently $CR_{ni}^2$ given in Equation (10) is a function of all $R_{im}^{*2}$ $,i=1,2,...,p$. The proof of lemma is given in the appendix.

Lemma

Considering the standard correlation model

$$X_{kj}^* = \sum_{\substack{i=1 \\ i\neq j}}^{p} \beta_i^* X_{ki}^* + e_{kj}^* \quad , j=1,2,...,p, \quad k=1,2,...,n$$

where $\beta_i^*$ is the parameter of the model, $e_{kj}^*$ is the error and $X_{kj}^* = \dfrac{X_{kj} - \overline{X}_j}{S_j\sqrt{n-1}}$. For a given $j\,; j=1,2,...,p$ the relation between the coefficient of multiple determination of $X_j^*$ given $X_1^*,...,X_{j-1}^*,X_{j+1}^*,...,X_p^*$ and the coefficient of multiple determination of $X_j^*$ given $Y_1^*,Y_2^*,....,Y_m^*$ is given as

$$R_j^2 = \sum_{h=1}^{p} a_h R_{hm}^{*2} - t_j$$

where $\sum_{h=1}^{p} a_h R_{hm}^{*2} = a_j R_{jm}^{*2} + \sum_{\substack{i=1 \\ i\neq j}}^{p} a_i R_{im}^{*2}$ , $a_j = \dfrac{1}{2}\sum_{\substack{i=1 \\ i\neq j}}^{p} \hat{\beta}_i^*$ , $a_i = \dfrac{1}{2}\hat{\beta}_i^*$,

$t_j = \dfrac{1}{2}\sum_{\substack{i=1 \\ i\neq j}}^{p}\sum_{k=1}^{m} (v_{jk} - v_{ik})^2 \hat{\beta}_i^*$ and $\hat{\beta}_i^*$ is the estimate of $\beta_i^*$ □

***The Algorithm***

Now we can state the steps of our method

1. Standardize the variables $X_1, X_2,..., X_p$ to have zero means and unit variances. Denote them by $X_1^*, X_2^*,..., X_p^*$

2. Calculate the correlation matrix $R$ of $X_1, X_2,..., X_p$ which is a covariance matrix of $X_1^*, X_2^*,..., X_p^*$.

3. Find the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ and the corresponding eigenvectors $u_1, u_2, \ldots, u_p$ of $R$.

4. Compute the inverse of $R$.

5. Calculate the coefficient of multiple determination $R_j^2$ for all $X_j^*$ , $j = 1,2,\ldots, p$, using Equation (8) or (9).

6. Arrange the $R_j^2$ in decreasing order, say $R_{(1)}^2 > R_{(2)}^2 > \ldots > R_{(p)}^2$.

7. Calculate $CR_m^2$ using Equation (10).

8. Graph the curves of $CR_m^2$ and $\lambda_m$ , $m = 1,2,\ldots, p$ against the components numbers.

9. The number of components before the point of intersection of the two curves is the number of principal components retained.

## 3. Illustrating Examples

To illustrate our criterion and to compare it with the other criteria we consider the following examples:

***Example (1)***

Consider the data (Manly 1994) of the percentages of the labors force in nine different types of industry for 26 European countries. The correlation matrix for the nine variables is shown in Table 1. Since the overall values in this matrix are not particularly high, one could say that several principal components will be required to account for the variation but this is not true since the examination of simple correlation coefficients will not necessarily disclose the existence of relations among group of independent variables. The eigenvalues of the correlation matrix are shown in Table 2 second column. It can be shown that, using Kaiser criterion, only the first three components are important because they are the only ones with eigenvalues greater than 1.00. but 'rule of thumb' suggests that four principal components should be considered, since the fourth eigenvalue is almost equal to the third. So either three or four principal components can reasonably be allowed. For the Scree graph of Cattell, Figure 2 shows a plot of the eigenvalues (variation) against the components numbers presented in Table 2 column 1 and 2. It is clear that there are two obvious breaks in the plot that separates the meaningful components from the trivial components. There is a gentle change of steepness at the third component and another sharp change of steepness at the fourth component. This example illustrates that the Scree graph approach for deciding the number of principal components is very subjective. The result of the percentage of the total variation criterion is given in Table 2 third column. It is clear that 75% of the total variation is captured by three components and 86% of the total variation is captured by four components. Table 2 fourth column presents the results of $CR_m^2$ and Figure 3 presents the graph of $CR_m^2$ and the $\lambda's$ against the components numbers. We see that the two curves intersect after the fourth component. So we decide precisely that four principal components must be retained.

Table (1) The correlation matrix for percentages employed in nine industry groups in 26 countries in Europe, in lower diagonal form.(Manly 1994)

| | AGR | MIN | MAN | PS | CON | SER | FIN | SPS | TC |
|---|---|---|---|---|---|---|---|---|---|
| Agriculture | 1.000 | | | | | | | | |
| Mining | 0.036 | 1.000 | | | | | | | |
| Manufacturing | -0.671 | 0.445 | 1.000 | | | | | | |
| Power supplies | -0.400 | 0.406 | 0.385 | 1.000 | | | | | |
| Construction | -0.538 | -0.026 | 0.495 | 0.060 | 1.000 | | | | |
| Service industries | -0.737 | -0.397 | 0.204 | 0.202 | 0.356 | 1.000 | | | |
| Finance | -0.220 | -0.443 | -0.156 | 0.110 | 0.016 | 0.366 | 1.000 | | |
| Social &personal services | -0.747 | -0.281 | 0.154 | 0.132 | 0.158 | 0.572 | 0.108 | 1.000 | |
| Transport & communications | -0.565 | 0.157 | 0.351 | 0.375 | 0.388 | 0.188 | -0.246 | 0.568 | 1.00 |

Table (2) Eigenvalues, cumulative proportion, and $CR^2_m$ of the correlation matrix of Table 1.

| Components numbers | $\lambda_m$ | $C\lambda_m$ | $CR^2_m$ |
|---|---|---|---|
| 1 | 3.487 | 0.387 | 0.11278 |
| 2 | 2.130 | 0.624 | 0.22553 |
| 3 | 1.099 | 0.746 | 0.33828 |
| 4 | 0.995 | 0.857 | 0.45097 |
| 5 | 0.543 | 0.917 | 0.56352 |
| 6 | 0.383 | 0.960 | 0.67556 |
| 7 | 0.226 | 0.985 | 0.78739 |
| 8 | 0.137 | 1.000 | 0.89810 |
| 9 | 0.000 | 1.000 | 1.00000 |

Figure (2) Scree graph for the industry data

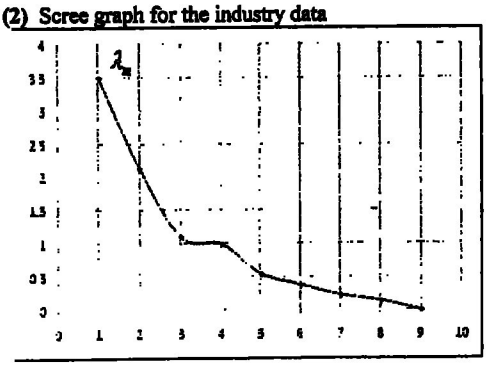
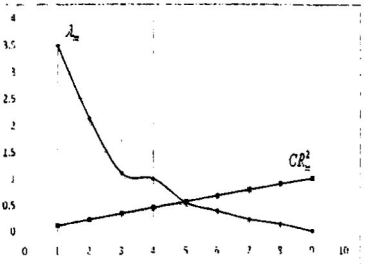


**Figure (3)** $CR_m^2$ and eigenvalues graph for the industry data




***Example (2)***

Consider the audiometric example analyzed in Jackson (1991; chap. 5). The data consists of measurements of lower hearing threshold on 100 men. Observations were obtained, on each ear, at frequencies 500, 1000, 2000 and 4000Hz, so that eight variables were recorded for each individual. The correlation matrix of the eight variables is given in Table 3. Table 4 presents the eigenvalues, $\lambda_m$ the cumulative proportion of variation, $C\lambda_m$, and the $CR_m^2$. Figure 4 presents the graph of $CR_m^2$ and $\lambda_m$ against the components numbers. The two curves intersect after three components, so according to our criterion we retain three principal components only. According to Kaiser criterion, there are two eigenvalues greater than one and since the third eigenvalue is almost equal to one so either two or three pc's can be retained. It is clear, also, that two components capture only 69% of the total variation while three components capture 82% of the total variation. The Scree graph of $\lambda_m$ shows that there is a sharp change of steepness at the second and at third component and a gentle change of steepness at fourth component so either two or three or four components can be retained.

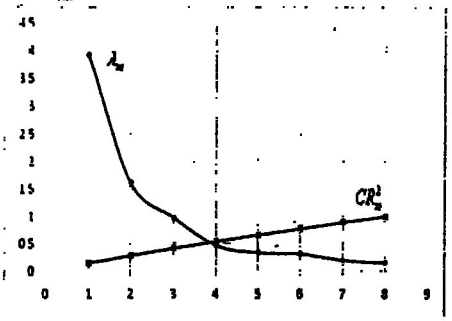
***Table (3)*** The correlation matrix of audiometric data (Jackson 1991)

| Left Ear | | | | Right Ear | | | |
|---|---|---|---|---|---|---|---|
| 500 | 1000 | 2000 | 4000 | 500 | 1000 | 2000 | 4000 |
| 1 | 0.78 | 0.40 | 0.26 | 0.70 | 0.64 | 0.24 | 0.20 |
| | 1 | 0.54 | 0.27 | 0.55 | 0.71 | 0.36 | 0.22 |
| | | 1 | 0.42 | 0.24 | 0.45 | 0.70 | 0.33 |
| | | | 1 | 0.18 | 0.26 | 0.32 | 0.71 |
| | | | | 1 | 0.66 | 0.16 | 0.13 |
| | | | | | 1 | 0.41 | 0.22 |
| | | | | | | 1 | 0.37 |
| | | | | | | | 1 |

*Table (4)* Eigenvalues. cumulative proportion and $CR_m^2$ of Table 3.

| Components numbers | $\lambda_m$ | $C\lambda_m$ | $CR_m^2$ |
|---|---|---|---|
| 1 | 3.92901 | 0.491 | 0.14490 |
| 2 | 1.61832 | 0.693 | 0.28871 |
| 3 | 0.97532 | 0.815 | 0.41893 |
| 4 | 0.46678 | 0.874 | 0.54488 |
| 5 | 0.34009 | 0.916 | 0.66402 |
| 6 | 0.31589 | 0.956 | 0.77747 |
| 7 | 0.20011 | 0.981 | 0.89064 |
| 8 | 0.15447 | 1.000 | 1.00000 |

Figure (4) $CR_m^2$ and eigenvalues graph for audiometric data




## A Simulation Study

We have made a simulation study of normally distributed data with different values of means, variances and different values of correlation coefficients. We have taken different sample size (30,50,100) and different values of p variables (4,6,7,8,9,10,12,20), (the simulation study is available upon request). From this simulation we observe that

1. Our criterion gives precise results in the sense that any one can obtain the same results.
2. Our criterion takes into account the useful eigenvalues that are less than one.
3. If we consider the value 0.9 close to one. Kaiser's criterion gives the same results as that of our criterion.
4. The numbers of pc's retained by Kaiser's criterion or by our criterion capture at least 70% of the total variation using the percentage criterion.
5. Using Scree test criterion, the steepness is not clear in some cases.

*Conclusion*

In this paper, we introduce a criterion depending on computing the cumulative proportion of the coefficient of multiple determination of each standardized variable. A number of studies have been carried out to compare our criterion with some other criteria (The eigenvalue-one criterion, the Scree test and the percentage criterion). One of these studies was a simulation study. From these studies we conclude that, our criterion gives a precise number of pc's retained, in the sense that when different people analyze the same data they will obtain the same result. It also captures amount of variability greater than the other methods. In addition it takes into account the components with variance smaller than one but important.

## Appendix

Proof of the lemma: To prove that $R_j^2$ is a function of all $R_{im}^{*2}$ $, i = 1,2,...,p$, where $R_j^2 \equiv R^2_{X_j^* / X_1^*, X_2^*, ..., X_{j-1}^*, X_{j+1}^*, ..., X_p^*}$ and $R_{im}^{*2} \equiv R^2_{X_i^* / Y_1^*, Y_2^*, ..., Y_m^*}$, consider the standard correlation model

$$X_{kj}^* = \sum_{\substack{i=1 \\ i \neq j}}^{p} \beta_i^* X_{ki}^* + e_{kj}^* \quad , j = 1,2,...,p \quad , k = 1,2,...,n$$

where $\beta_i^*$ is the parameter of the model, $e_{kj}^*$ is the error and $X_{kj}^* = \dfrac{X_{kj} - \overline{X}_j}{S_j \sqrt{n-1}}$.

The coefficient of multiple determination of $X_j^*$ given $X_1^*, ..., X_{j-1}^*, X_{j+1}^*, ..., X_p^*$ is given by

$$R_j^2 = \sum_{\substack{i=1 \\ i \neq j}}^{p} r_{ij} \hat{\beta}_i^* , \tag{11}$$

Where $\hat{\beta}_i^*$ is the estimate of $\beta_i^*$ and $r_{ij}$ is correlation between $X_i^*$ and $X_j^*$. Given that $Y_1^*, Y_2^*, ..... Y_m^*$ are the best choice of pc's then $r_{ij}$ can be given approximately as

$$r_{ij} = \sum_{k=1}^{m} v_{jk} v_{ik} \qquad \text{(Manly 1994)} \tag{12}$$

Now let

$$2 v_{jk} v_{ik} = v_{jk}^2 + v_{ik}^2 - (v_{jk} - v_{ik})^2$$

$$\sum_{k=1}^{m} v_{jk} v_{ik} = \frac{1}{2} \sum_{k=1}^{m} \left[ v_{jk}^2 + v_{ik}^2 - (v_{jk} - v_{ik})^2 \right] \tag{13}$$

Substituting Eq. (13) in Eq. (12) then $r_{ij}$ can be written as

$$r_{ij} = \frac{1}{2}\left[\sum_{k=1}^{m} v_{jk}^2 + \sum_{k=1}^{m} v_{ik}^2 - \sum_{k=1}^{m} (v_{jk} - v_{ik})^2\right]$$

From Eq. (5), we get

$$r_{ij} = \frac{1}{2}\left[R_{jm}^{*2} + R_{im}^{*2} - \sum_{k=1}^{m}(v_{jk} - v_{ik})^2\right] \qquad (14)$$

Substituting Eq. (14) in Eq. (11) we get

$$R_j^2 = \sum_{\substack{i=1 \\ i\neq j}}^{p} \frac{1}{2}\left[R_{jm}^{*2} + R_{im}^{*2} - \sum_{k=1}^{m}(v_{jk} - v_{ik})^2\right]\hat{\beta}_i^*$$

$$R_j^2 = \frac{1}{2}\sum_{\substack{i=1 \\ i\neq j}}^{p} \hat{\beta}_i^* R_{jm}^{*2} + \frac{1}{2}\sum_{\substack{i=1 \\ i\neq j}}^{p} \hat{\beta}_i^* R_{im}^{*2} - \frac{1}{2}\sum_{\substack{i=1 \\ i\neq j}}^{p} \sum_{k=1}^{m}(v_{jk} - v_{ik})^2 \hat{\beta}_i^*$$

Then

$$R_j^2 = \frac{1}{2} R_{jm}^{*2} \sum_{\substack{i=1 \\ i\neq j}}^{p} \hat{\beta}_i^* + \frac{1}{2}\sum_{\substack{i=1 \\ i\neq j}}^{p} \hat{\beta}_i^* R_{im}^{*2} - \frac{1}{2}\sum_{\substack{i=1 \\ i\neq j}}^{p} \sum_{k=1}^{m}(v_{jk} - v_{ik})^2 \hat{\beta}_i^*$$

$$R_j^2 = a_j R_{jm}^{*2} + \sum_{\substack{i=1 \\ i\neq j}}^{p} a_i R_{im}^{*2} - t_j$$

where $a_j = \frac{1}{2}\sum_{\substack{i=1 \\ i\neq j}}^{p} \hat{\beta}_i^*$ , $a_i = \frac{1}{2}\hat{\beta}_i^*$, $t_j = \frac{1}{2}\sum_{\substack{i=1 \\ i\neq j}}^{p}\sum_{k=1}^{m}(v_{jk} - v_{ik})^2 \hat{\beta}_i^*$

So

$$R_j^2 = \sum_{h=1}^{p} a_h R_{hm}^{*2} - t_j \quad ; \sum_{h=1}^{p} a_h R_{hm}^{*2} = a_j R_{jm}^{*2} + \sum_{\substack{i=1 \\ i\neq j}}^{p} a_i R_{im}^{*2}$$

Which means that $R_j^2$ is a function of all $R_{im}^{*2}$ $, i = 1,2,..., p$

## References


1. Bartlett, M. S. (1950). Test of significance in factor analysis, Br. *J. Psych. Stat.* Sec. 3, 77-85

2. Cattell, R. B. (1966), The Scree test for the number of factors. *Multivariate Behavioral Research*. 1. 245-276.

3. Guttman, Louis (1953), Image theory for the structure of quantitive variates. *Psychometrika*. 18 (4). 277-296.

4. Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179-185.

5. Jackson, J. E. (1991), *A User's Guide to Principal Components* (1st edition). Wiley. New York.

6. Jolliffe, I. T. (1972), Discarding variables in principal component analysis. I: Artificial data. *Appl. Stat.*, 21, 160-173.

7. Kaiser, H. F. (1960), The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20, 141-151

8. Ledesma, Ruben D. and Pedro, Valero-Mora (2007). Determining the number of factors to retain in EFA: an easy-to-use computer program for carrying out Parallel Analysis. Practical Assessment, Research & Evaluation,12(2). Available on line: http://pareonline. net/pdf/v12n2.pdf

9. Manly, Bryan F. J. (1994), *Multivariate Statistical Methods* (2nd ed). Chapman and Hall, New York.

10. Neter, J., Wasserman, W. and Kutner, M. H. (1983). *Applied Linear Regression Models*. U.S.A: Richard D. Irwin.