On the Kaiser-Meier-Olkin's Measure of Sampling Adequacy

Bismark Kwao Nkansah
Department of Statistics, University of Cape Coast, Ghana
E-mail of author: bnkansah@ucc.edu.gh

Abstract

The paper examines the suitability of the Kaiser-Meier Olkin's Measure of Sampling Adequacy (KMO) as a measure of suitability for factor analysis for a number of selected multivariate datasets. It first explores a systematic approach that determines the initial dimensionality of the dataset. It then identifies two sets of indicators that could create distortions in assessing factor-suitability: variables that do not influence any dimension; and those that influence multiple dimensions. Dimensionality is also affected by negatively correlated indicators leading to a small suitability measure, which portrays such datasets as unsuitable for factor analysis. It is found that for KMO to be high, the zero- and first-order partial correlations must be almost the same for indicators that influence the same dimension. It follows that generally, a KMO value within the range 0.6 – 0.7 is a typically good measure of factor-suitability. The results show that the overall KMO generally reflects factor-suitability. The study does not find the expected intuitive relation that should exist between the individual KMO value and the communality for a suitably selected factor solution. A high variable KMO appears to be associated with moderate value of coefficient of multiple determination of its model in terms of the others. A reasonable assessment of the KMO should therefore be made only by a good understanding of the correlation structure of the indicator variables.

Keywords: KMO, Factor-suitability, Factor analysis, Dimensionality

1. Introduction

The key concept of factor analysis is that multiple indicator variables have similar patterns of responses as they are all associated with a latent (i.e., not directly measured) variable. Factor analysis is based on the correlation matrix of the indicator variables. The dimensionality of this matrix can be reduced by "looking for variables that correlate highly with a group of other variables, but correlate very badly with variables outside of that group" (Field, 2000). These variables with high inter-correlations could well measure one underlying variable, which is called a 'factor'. The factors f_j, $j = 1, 2, \ldots, m$, are constructed from a set of variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)$, such that

$$x_i = \sum_{j=1}^{m} l_{ij} f_j + \varepsilon_i; \qquad i = 1, 2, \ldots, p, \tag{1}$$

where $m \le p$, and ε_i are the factors specific to the individual indicator x_i. The factors are labelled by the size of their loadings (l_{ij}) on the indicators. Usually, a cut-off value of 0.5 is used to associate a variable with a factor. Indicators with loadings higher than 0.5 are considered to be influential in the formation of the factor. However, the choice of the cut-off depends on the size of the correlation coefficients. Using the *m* factors, the correlation matrix **R** could be approximated by the fundamental factor analysis equation

$$\mathbf{R} = \mathbf{\Lambda\Lambda}' + \boldsymbol{\psi}, \tag{2}$$

The matrices $\boldsymbol{\psi}$ and $\mathbf{\Lambda}$ are defined as the diagonal matrix of specific variances and loading matrix, respectively. In Equations (1) and (2), we could have $m = p$, or $m < p$.

The suitability of factor analysis for a dataset is influenced by the sample size (Tabachnick & Fidell, 2007; Comrey & Lee, 1992), the nature of the data (van der Eijk & Rose, 2015) with particular reference to ordered categorical survey data, and the type of correlation coefficients involved which could be polychoric, or Pearson correlations.

A number of guidelines are used to determine the suitability of a dataset for factor extraction. One of the commonly used guidelines is the Kaiser-Meier-Olkin's Measure of Sampling adequacy, usually referred to as the KMO. It is a diagnostic measure for assessing the extent to which the indicators of a dimension belong together. A small value of the KMO indicates that the correlation between pairs of variables cannot be explained by a well-defined latent factor and that factor analysis may not be appropriate. Table **1** gives a guideline for the KMO measure (Kaiser, 1974).

Table 1: A Guide for Interpreting KMO Measure

KMO Measure	Recommendation
≥ 0.90	Marvelous
0.80+	Meritorious
0.70+	Middling
0.60+	Mediocre
0.50+	Miserable
≤ 0.50	Unacceptable

By the guideline in Table 1, it is generally expected that to have satisfactory results, the overall KMO measure should be 0.8 or higher. This rule of thumb appears to have been accepted widely, although a measure of above 0.6 is acceptable (Rencher, 2002). The index compares the magnitude of the observed correlation coefficients to the magnitude of the partial correlation coefficients. An equation for calculating the KMO is given by

$$KMO = \frac{\sum_{i<j} r_{ij}^2}{\sum_{i<j} r_{ij}^2 + \sum_{i<j} pr_{ij}^2}, \qquad (3)$$

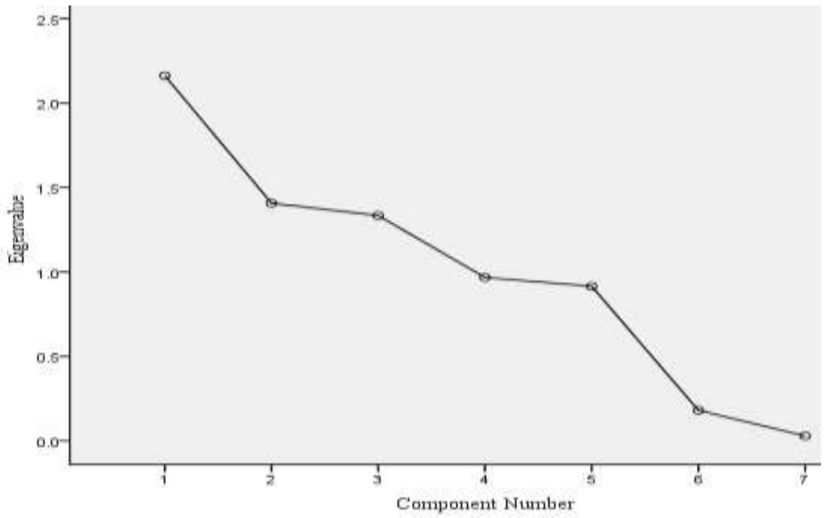
where r_{ij}^2 is the square of the correlation coefficient between any pair of variables (X_i, X_j), and is an element of the correlation matrix, **R**. The corresponding value pr_{ij}^2 is the square of the partial correlation coefficient and is an element of the matrix $\mathbf{Q} = \mathbf{D}\mathbf{R}^{-1}\mathbf{D}$, where $\mathbf{D} = \left[\left(\text{diag}\, \mathbf{R}^{-1} \right)^{1/2} \right]^{-1}$. The partial correlations may be obtained, for example, from the anti-image matrix in SPSS or the MATLAB codes given as follows:

$$\begin{aligned}
\mathbf{R} &= corr(data);\\
\mathbf{RI} &= inv(\mathbf{R});\\
Diag\mathbf{RI} &= diag(diag(\mathbf{RI}));\\
SqrtDiag\mathbf{RI} &= (Diag\mathbf{RI})\wedge 0.5;\\
\mathbf{D} &= inv(SqrtDiag\mathbf{RI});\\
\mathbf{Q} &= \mathbf{D}*\mathbf{RI}*\mathbf{D}
\end{aligned}$$

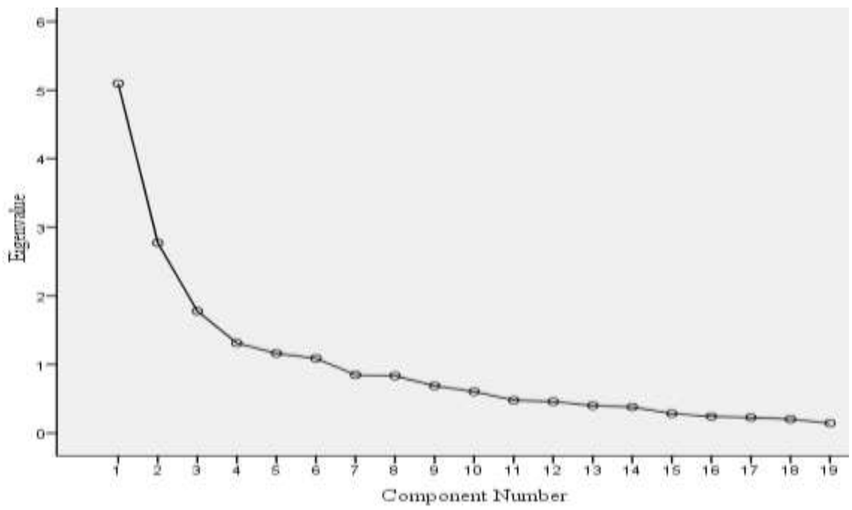
The relation shows that as $\mathbf{R}^{-1}$ approaches a diagonal matrix, KMO approaches one. Thus, drastic reduction in partial correlations is required for a high KMO. This has implications for individual variable KMO, as will be pointed. In the Anti-image matrix, the diagonal elements are the KMO of the individual variable. The study intends to explain the practical representation of the overall KMO and variable KMO for determining the factor-suitability of the dataset and the variable, respectively. It is already the opinion of some (e.g., Sharma 1996) who suggest that one could subjectively examine the correlation matrix to determine its factor suitability, suggesting a restraint on the use of the KMO.

It is possible to obtain a graphical view of the factor suitability of a dataset. This is usually obtained by a scree plot of the eigenvalues of the factors against their respective factor numbers. If the scree plot does show a pronounced bend or the eigenvalues show a large gap around one, then the correlation matrix is likely to be factor-suitable. The point where the 'elbow' is located gives an indication of the number of factors that could be extracted from the data.

Figures 1 and 2 are scree plots of some datasets used in the study that show the extent of their factor-suitability. In Figures 1(a) and 1(b), there is no clear bend in the plots. Figure 1(a), in particular, suggests that the corresponding dataset is highly unsuitable for factor extraction, as there is no systematic decrease in eigenvalue for higher numbers of the factors. It will be realised, however, that the lack of suitability is not as a result of low correlations among the variables. Thus, the source of factor suitability may be attributable to causes other than the correlation coefficient on which KMO is based.



(a)



(b)

Figure 1: Graphs indicating lack of factor suitability of data.

On the other hand, Figure 2 shows a scree plot that has quite a clear bend, suggesting that the respective dataset is factor-suitable. Thus, in Figures 1(a) and 1(b), the KMO values are expected to be small, whiles in Figures 2, the KMO value is 'expected' to be large.

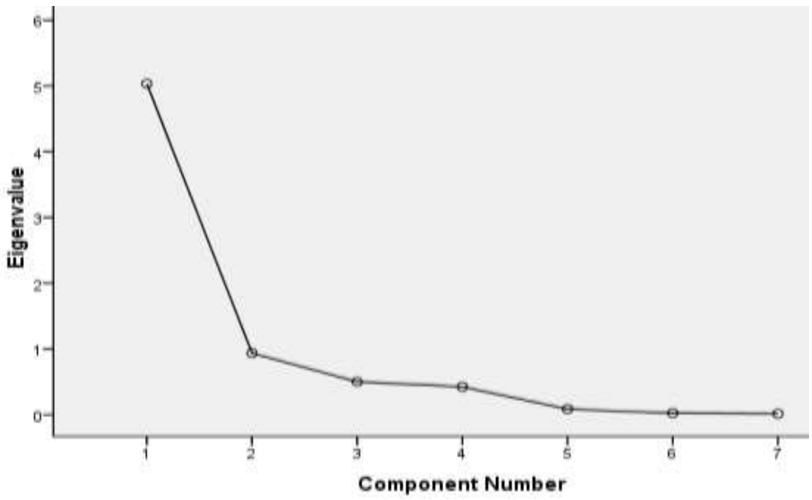


Figure 2: Graphs indicating suitability of data for factor analysis.

Unlike Figure 2, the determination of the elbow point could be quite subjective in many scree plots, which calls for other methods such as the parallel analysis. A well-defined elbow point, unfortunately, does not suggest a high KMO value. In Figure 2, for example, the KMO associated with the data (see description of Dataset 1) is just 0.616. Thus, one may suspect that the KMO value (and as interpreted in Table 1) may not provide a fair representation of the factor-suitability of some datasets. As a result, assessment of homogeneous groupings of indicators has been suggested (e.g., Field, 2000). This paper demonstrates a way of carrying out such an assessment by outlining a procedure for determining the dimensionality of the dataset. Subsequently, the KMO value of the data is computed and compared with the value given in the software output. This is intended to verify the consistency of the preliminary dimensionality assessment and the reliability of the KMO value given in the output. To proceed, descriptions of the datasets used in the study are given next.

1.1 Description of Datasets

A number of datasets have been used in Section 4 to carry out the study. The following provides explanation to the background of these datasets and the rationale for their selection. The datasets have been numbered in the section for convenience of reference in Section 4.

Dataset 1 (Performance of Sales Personnel): The data covers assessment of performance of sales personnel employees of a marketing company (Johnson & Wichern, 2007). The firm attempts to evaluate the quality of its sales staff and tries to find an examination, or series of tests, that may reveal the potential for good performance in sales. It has selected a random sample of 50 salespeople and has evaluated each on three measures of performance: growth of sales, profitability of sales, and new account sales. These measures have been converted to a scale, on which 100 indicates “average” performance. Each of the 50 individuals would take each of four tests, which purportedly measures creativity, mechanical reasoning, abstract reasoning, and mathematical ability, respectively. The $n = 50$ observations on $p = 7$ variables are listed. The data is interesting in that it has a well-defined single dimension (see Figure 2) which, however, is not significant under a confirmatory test.

Dataset 2 (Performance of High School Students in Nine Subjects): This is unpublished data which covers marks scored out of 100% obtained by 72 students in a senior high school on nine subjects. These subjects include Information Communication Technology (ICT), Economics, Elective Mathematics, English Language, Geography, Integrated Science, Core Mathematics, Physical Education (PE), and Social Science. By design, this data is typically suited for principal components, and hence, factor analysis.

Dataset 3 (Benefits of Students Industrial Attachment): The data is obtained from 525 students of a Technical University in Ghana. Structured questionnaires are used which contained 48 indicators of benefits and challenges of students industrial attachment. Twenty of the indicators cover issues of benefits whilst twenty-eight cover issues of challenges. Data on these indicators are obtained on a five-point Likert scale. The data is used (Frempong, Nkansah, & Nkansah, 2017) to determine the salient latent dimensions of benefits of the programme.

Dataset 4 (Prices of Food Items in Ghana in 2012): The main variables of study are the prices of selected commodities collected from 91 leading market centres across the country. The food items include those that form the basis for the computation of the monthly Food Price Index (FPI) by the Ghana Statistical Service (GSS). Nineteen food items are studied which include: Maize, White Yam, Cassava, Tomato, Garden Egg, Dried Pepper, Red Groundnut, White Cowpea, Palm Oil, Orange, Banana, Smoked Herring, Salted fish, Onion, Eggs, Plantain, Gari, Local Rice, and Imported Rice, with appropriate respective unit of sale for each item. The information is obtained from the Statistical, Research and Information Directorate (SRID) of the Ministry of Food and Agriculture (MoFA). The selection of the markets is based on results of previous related studies (Seglah, 2013). Data for Year 2012 is particularly selected as a result of negative correlations observed among prices of the items in that year, which suits this and related studies.

Dataset 5 (Concrete Compressive Strength): The concrete compressive strength (CCS) is a highly nonlinear function of age and ingredients and was studied as a regression problem. These ingredients include Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticiser, Coarse Aggregate, and fine aggregate. The actual concrete compressive strength (MPa) for a given mixture under a specific age (days) was determined from laboratory. Each ingredient is measured in kg in m^3 mixture. Thus, there are eight input quantitative variables and one output variable, and covers 1030 observations. The data was studied by Yeh (1998a) and subsequently in Yeh (1998b, 1999, 2003a, 2003b, 2006). In the paper, only the seven ingredient variables will be studied.

Dataset 6 (Challenges of Students Industrial Attachment): The data were obtained in the same study (Frempong, Nkansah & Nkansah, 2017) that made use of Dataset 3. There are twenty-eight indicator variables involved in this part of the study.

Section 2 presents some useful mathematical background on the subject. In section 3, an outline of a procedure for examining the homogeneity of groupings is explored. Based on the observations in Section 3, Section 4 will examine the consistency in the KMO value using the datasets described. Conclusions are drawn in Section 5.

2 Some Mathematical Background

Suppose that data is obtained on the variable $\mathbf{X} = (X_1, X_2, \ldots, X_p)$. The Kaiser-Meyer-Olkin's measure of sampling adequacy, usually simply referred to as the KMO, is given by Equation (3). The simple correlation coefficient (r_{ij}) between X_i and X_j is given as

$$r_{X_i X_j} = \frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}} \tag{4}$$

The value s_{ij} is the (i, j) element of the matrix of sum of squares and cross-product matrix of the data which may be given as $\mathbf{S_{XX}} = (\mathbf{X} - \bar{\mathbf{x}}\mathbf{1}')(\mathbf{X} - \bar{\mathbf{x}}\mathbf{1}')'$, where $\mathbf{1} = \text{ones}(p, 1)$ is a vector of ones. Next, we examine the partial correlation component $\sum pr_{ij}^2$ in Equation (3).

The sample partial correlation coefficient (PCC) between X_i and X_j controlling for the other variables, $\mathbf{Y} = \mathbf{X} \setminus \{X_i, X_j\}$, is given by

$$r_{X_i X_j \cdot \mathbf{Y}} = \frac{s_{X_i X_j \cdot \mathbf{Y}}}{\sqrt{s_{X_i X_i \cdot \mathbf{Y}}}\sqrt{s_{X_j X_j \cdot \mathbf{Y}}}} \tag{5}$$

The element $S_{X_i X_i \cdot \mathbf{Y}}$ in Equation (5) is the (i, j) entry of the variance-covariance matrix

$$\mathbf{S_{XX}} - \mathbf{S_{XY}}\mathbf{S_{YY}^{-1}}\mathbf{S_{YX}}. \tag{6}$$

Proof

Consider the pair (X_i, X_j), $i \neq j$. Then define the vectors $\mathbf{X^{(1)}} = (X_i, X_j)$ and $\mathbf{X^{(2)}} = \mathbf{X} \setminus \mathbf{X^{(1)}}$ with corresponding mean vectors $\boldsymbol{\mu}_1 = E(\mathbf{X}^{(1)})$ and $\boldsymbol{\mu}_2 = E(\mathbf{X}^{(2)})$, and variance-covariance matrices, $\boldsymbol{\Sigma}_{11} = \text{cov}(\mathbf{X}^{(1)})$ and $\boldsymbol{\Sigma}_{22} = \text{cov}(\mathbf{X}^{(2)})$. We partition the vector of variables as $\mathbf{X} = \left(\mathbf{X}^{(1)} \mid \mathbf{X}^{(2)}\right)$. Subsequently, define the vectors (e.g., Johnson & Wichern, 2007)

$$\mathbf{Y_1} = \mathbf{X}^{(1)} - \mathbf{B}\mathbf{X}^{(2)} \quad \text{and} \quad \mathbf{Y}_2 = \mathbf{X}^{(2)}, \tag{7}$$

so that information about $\mathbf{X}^{(1)}$ could be extracted through $\mathbf{Y}_1$. The covariance between $\mathbf{Y}_1$ and $\mathbf{Y}_2$ is given by

$$
\begin{aligned}
\operatorname{cov}(\mathbf{Y}_1, \mathbf{Y}_2) &= E\left(\mathbf{X}^{(1)} - \mathbf{B}\mathbf{X}^{(2)}\right)\mathbf{X}^{(2)'} - E\left(\mathbf{X}^{(1)} - \mathbf{B}\mathbf{X}^{(2)}\right)E\left(\mathbf{X}^{(2)'}\right) \\
&= E\left(\mathbf{X}^{(1)}\mathbf{X}^{(2)'}\right) - E\left(\mathbf{X}^{(1)}\right)E\left(\mathbf{X}^{(2)'}\right) - \mathbf{B}E\left(\mathbf{X}^{(2)}\mathbf{X}^{(2)'}\right) + \mathbf{B}E\left(\mathbf{X}^{(2)}\right)E\left(\mathbf{X}^{(2)'}\right) \\
&= \mathbf{\Sigma}_{12} - \mathbf{B}\mathbf{\Sigma}_{22}
\end{aligned}
$$

Since $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ must be independent, $\operatorname{cov}(\mathbf{Y}_1, \mathbf{Y}_2) = 0$. Hence,

$$\mathbf{B} = \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}. \tag{8}$$

Finding the variance-covariance matrix of $\mathbf{Y}_1$, we obtain

$$\operatorname{cov}(\mathbf{Y}_1) = E\left(\mathbf{X}^{(1)} - \mathbf{B}\mathbf{X}^{(2)}\right)\left(\mathbf{X}^{(1)} - \mathbf{B}\mathbf{X}^{(2)}\right)' - E\left(\mathbf{X}^{(1)} - \mathbf{B}\mathbf{X}^{(2)}\right)E\left(\mathbf{X}^{(1)} - \mathbf{B}\mathbf{X}^{(2)}\right)'$$

Expanding and simplifying gives

$$\operatorname{cov}(\mathbf{Y}_1) = \mathbf{\Sigma}_{11} + \mathbf{B}\mathbf{\Sigma}_{22}\mathbf{B}' - \mathbf{B}\mathbf{\Sigma}_{21} - \mathbf{\Sigma}_{12}\mathbf{B}'$$

Substituting for **B** from Equation (8) and simplifying gives

$$\mathbf{\Sigma}_{\mathbf{Y}_1} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$$

Taking expectation of $\mathbf{Y_1}$, and making substitution for **B**, gives

$$E(\mathbf{Y}_1) = E\left(\mathbf{X}^{(1)} - \mathbf{B}\mathbf{X}^{(2)}\right) = \boldsymbol{\mu}_1 - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2$$

Therefore, $\mathbf{Y_1}$ is normally distributed as

$$\mathbf{Y}_1 \middle| \mathbf{X}^{(2)} = c \sim N\left(\boldsymbol{\mu}_1 - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2,\ \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}\right)$$

Thus, if $\mathbf{X}^{(2)} = \mathbf{c}$, for a given vector **c**, then $\mathbf{X}^{(1)} = \mathbf{Y_1} + \mathbf{Bc}$ is a translation of $\mathbf{Y_1}$ through **Bc**. Now, taking the conditional expectation of $\mathbf{X}^{(1)}$ gives

$$
\begin{aligned}
E(\mathbf{X}^{(1)} | \mathbf{c}) &= E(\mathbf{Y_1} + \mathbf{Bc}) \\
&= \boldsymbol{\mu}_1 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathbf{c} - \boldsymbol{\mu}_2)
\end{aligned}
$$

Thus,

$$\boldsymbol{\mu}_{1|\mathbf{c}} = \boldsymbol{\mu}_1 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathbf{c} - \boldsymbol{\mu}_2) \tag{9}$$

Since, $\mathbf{X}^{(1)}$ is a translation of $\mathbf{Y_1}$, the conditional variance-covariance of $\mathbf{X}^{(1)}|\mathbf{c}$ is the same as that of $\mathbf{Y_1}$. Therefore,

$$\mathbf{\Sigma}_{11|\mathbf{c}} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} \tag{10}$$

This ends the proof.

Equation (10) is the same as that of the sample conditional variance-covariance matrix in Equation (6).

Remark 2.1

If we define the sub-vectors $\mathbf{X}^{(1)} = (X_i, X_j)$ and $\mathbf{X}^{(2)} = \mathbf{X} \setminus \mathbf{X}^{(1)}$, the vector of variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ and its variance-covariance matrix may be partitioned, respectively, as

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}^{(1)} \\ \hdashline \mathbf{X}^{(2)} \end{pmatrix} \text{ and } \mathbf{S} = \left(\begin{array}{c:c} \underset{2\times 2}{\mathbf{S_{xx}}} & \underset{2\times(p-2)}{\mathbf{S_{xy}}} \\ \hdashline \underset{(p-2)\times 2}{\mathbf{S_{yx}}} & \underset{(p-2)\times(p-2)}{\mathbf{S_{yy}}} \end{array}\right),$$

with indicated dimensions of variance-covariance sub-matrices.

Remark 2.2

Given the conditional variance-covariance in Equation (10), the partial correlation coefficient (PCC) between the pair $(X_i, X_j),\ i, j = 1, 2, \ldots, p,$ after controlling for $\mathbf{X^{(2)}} = \mathbf{X} \setminus \mathbf{X^{(1)}}$, may be restated as

$$r_{x_i x_j \cdot \mathbf{c}} = \frac{(\mathbf{\Sigma}_{11|\mathbf{c}})_{ij}}{\sqrt{(\mathbf{\Sigma}_{11|\mathbf{c}})_{ii}(\mathbf{\Sigma}_{11|\mathbf{c}})_{jj}}}, \tag{11}$$

where $(\boldsymbol{\Sigma}_{11|\mathbf{c}})_{ij}$ is the (i, j) element of the conditional matrix $\boldsymbol{\Sigma}_{11|\mathbf{c}}$. This is the $(p-2)th$-order partial correlation since $\mathbf{X}^{(2)}$ contains $p-2$ variables.

Remark 2.3

In this remark, the nature of the expression in Equation (11) is examined if it involves the same component variable X_i. Does it become equal to 1, as in the case of the zero-order correlation coefficient? Without loss of generality, consider the variable component X_i, $i=1$. Then re-define the vectors $\mathbf{X}^{(1)}=(X_1)$ and $\mathbf{X}^{(2)}=(X_2, X_3, \ldots, X_p)$, with corresponding mean vectors $\mu_1 = E(\mathbf{X}^{(1)})$ and $\boldsymbol{\mu}_2 = E(\mathbf{X}^{(2)})$, and variance-covariances $\boldsymbol{\Sigma}_{11}=\sigma_{11}=\operatorname{cov}(\mathbf{X}^{(1)})$ and $\boldsymbol{\Sigma}_{22}=\operatorname{cov}(\mathbf{X}^{(2)})$. In this case, the mean and variance of $\mathbf{X}^{(1)}$ are constants. Similarly, the condition for independence $\mathbf{B}=\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$ becomes a vector of dimension $1\times(p-1)$, which is write as $\mathbf{b}=\mathbf{v}_{12}\boldsymbol{\Sigma}_{22}^{-1}$, where $\mathbf{v}_{12}=\boldsymbol{\Sigma}_{12}$, and

$$\mu_{1|\mathbf{c}} = \mu_1 + \mathbf{v}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{c}-\boldsymbol{\mu}_2) \tag{12}$$

is a constant (i.e., just one model).

Now, the numerator $(\boldsymbol{\Sigma}_{11|\mathbf{c}})_{ij}$ in Equation (11) may be seen as the covariance between $\mathbf{X}^{(1)}=(X_1)$ and the estimated model given in Equation (12) in terms of the set $\mathbf{X}^{(2)}$. Denote the model by

$$\hat{\mathbf{X}}^{(1)}\big|\mathbf{X}^{(2)} = \mu_1 + \mathbf{v}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{X}^{(2)}-\boldsymbol{\mu}_2).$$

Now, the variance of the estimated random variable is given as

$$\begin{aligned}\operatorname{var}\left(\hat{\mathbf{X}}^{(1)}\big|\mathbf{X}^{(2)}\right) &= \mathbf{v}_{12}\boldsymbol{\Sigma}_{22}^{-1}D\left(\mathbf{X}^{(2)}\right)\boldsymbol{\Sigma}_{22}^{-1}\mathbf{v}_{12}' \\ &= \mathbf{v}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{v}_{12}' \\ &= \mathbf{v}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{v}_{12}'\end{aligned}$$

$$\begin{aligned}\operatorname{cov}(\mathbf{X}^{(1)}, \hat{\mathbf{X}}^{(1)}\big|\mathbf{X}^{(2)}) &= E(\mathbf{X}^{(1)}(\mu_1 + \mathbf{v}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{X}^{(2)}-\boldsymbol{\mu}_2)') \\ &\quad - E(\mathbf{X}^{(1)})E(\mu_1 + \mathbf{v}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{X}^{(2)}-\boldsymbol{\mu}_2)' \\ &= \mu_1 E(\mathbf{X}^{(1)}) + \mathbf{v}_{12}\boldsymbol{\Sigma}_{22}^{-1}((E(\mathbf{X}^{(1)}\mathbf{X}'^{(2)}) - \boldsymbol{\mu}_2' E(\mathbf{X}^{(1)})) - \mu_1 E(\mathbf{X}^{(1)}) \\ &= \mu_1^2 + \mathbf{v}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(E(\mathbf{X}^{(1)}\mathbf{X}'^{(2)}) - \boldsymbol{\mu}_1\boldsymbol{\mu}_2'\right) - \mu_1^2 \\ &= \mathbf{v}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{v}_{12}'\end{aligned}$$

Therefore,

$$\operatorname{cov}(\mathbf{X}^{(1)}, \hat{\mathbf{X}}^{(1)}\big|\mathbf{X}^{(2)}) = \mathbf{v}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{v}_{12}' \tag{13}$$

The partial correlation of $\mathbf{X}^{(1)}=(X_1)$ is then the conditional correlation between $\mathbf{X}^{(1)}\big|\mathbf{c}$ and $\hat{\mathbf{X}}^{(1)}\big|\mathbf{X}^{(2)}$ and given by

$$r_{\mathbf{x}^{(1)}|\mathbf{c}} = \frac{\operatorname{cov}\left(\mathbf{X}^{(1)}, \hat{\mathbf{X}}^{(1)}\big|\mathbf{X}^{(2)}\right)}{\sqrt{\operatorname{var}\left(\mathbf{X}^{(1)}\right)}\sqrt{\operatorname{var}\left(\hat{\mathbf{X}}^{(1)}\big|\mathbf{X}^{(2)}\right)}} = \frac{\mathbf{v}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{v}_{12}'}{\sqrt{\sigma_{11}}\sqrt{\mathbf{v}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{v}_{12}'}}$$

Simplifying and squaring the result becomes

$$r^2_{\mathbf{x}^{(1)}|\mathbf{c}} = \frac{\mathbf{v}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{v}_{12}'}{\sigma_{11}}. \tag{14}$$

The result in Equation (14) is the coefficient of multiple determination (CMD) between $\mathbf{X}^{(1)}=(X_1)$ and the estimated regression function in terms of $\mathbf{X}^{(2)}=\mathbf{X}\setminus\mathbf{X}^{(1)}$.

The notion of the CMD could be linked to the individual KMO of variable X_i (KMO_{x_i}) which is similarly obtained by Equation (3) but includes only correlations that involve variable X_i. Denoting $I = \{1, 2, 3, \ldots, p\}$ the index set of all variables, KMO_{x_i} is given by

$$KMO_{x_i} = \frac{\sum_{j \in I \setminus i} r_{ij}^2}{\sum_{j \in I \setminus i} r_{ij}^2 + \sum_{j \in I \setminus i} pr_{ij}^2}, \tag{15}$$

where $I \setminus i$ denotes the index set that excludes *i*. It should be noted from the discussion that if X_i has a high correlation coefficient with each of the other variables, its $r_{\mathbf{x}^{(1)}|\mathbf{c}}$ would be high, and the (sum of) PCC with other variables will be high, leading to a low individual KMO value, suggesting that the variable is not factor-suitable. Thus, we expect the CMD of a variable to be related with its individual KMO. The question is: How can a variable that has very high correlations with other variables be represented as not too good for factoring? A high $r_{\mathbf{x}^{(1)}|\mathbf{c}}$ means that the variable could 'overlap' in almost all sub-groupings and hence may not be a clear indicator of a particular underlying dimension. Thus, if CMD is high and the data is factor-suitable, we expect the individual KMO to be of moderate values.

Effect of Partial Correlation Coefficient on the KMO

The effect of the partial correlation coefficient is clear from Equation (3). The effect may also be explained from the point of view of the first-order PCC given in Equation (16), expressed in terms of correlation coefficients between X_i and X_j after controlling for X_k as

$$r_{X_i X_j \cdot X_k} = \frac{r_{X_i X_j} - r_{X_i X_k} \times r_{X_j X_k}}{\sqrt{\left(1 - r_{X_i X_k}^2\right)}\sqrt{\left(1 - r_{X_j X_k}^2\right)}}, \tag{16}$$

where $r_{X_i X_j}$ is the zero-order (or simple) correlation coefficient between X_i and X_j and X_k represents the controlled component. If $r_{X_i X_k}$ and $r_{X_j X_k}$ are both large, then the denominator is very small and could lead to a large value of $r_{X_i X_j \cdot X_k}$, and hence a small value of the KMO. This suggests that (X_i, X_j) does not form a strong group. On the other hand, if $r_{X_i X_k}$ and $r_{X_j X_k}$ are both small, the denominator is large and could lead to a small $r_{X_i X_j \cdot X_k}$, and hence a large KMO. This means that (X_i, X_j) could constitute a strong group. The implication is that PCC between two variables X_i and X_j after controlling for other variables, depends to a large extent on the correlation coefficient between each of X_i and X_j and other variables. A high value of the KMO also indicates that the PCCs are generally low. Consequently, we will conclude that the variables (X_i, X_j) belong together. Therefore, a high KMO value is an indication that there are distinct groupings among the variables, and hence, a justification for using a dimensionality reduction technique. There are two scenarios to the case of a low KMO value: it is an indication of generally low or high simple correlations coefficients, and hence there are no distinct groupings among the variables. Thus, the suitability of factor analysis technique requires generally moderate simple correlation coefficients and low elements of the matrix of PCC.

In this section, the overall KMO and the individual variable KMO has been studied. It has been shown that the CMD between a variable $\mathbf{X}^{(1)} = (X_1)$ and the estimated regression function in terms of $\mathbf{X}^{(2)} = \mathbf{X} \setminus \mathbf{X}^{(1)}$ is closely linked to the factor-suitability of a variable. The partial correlation coefficient between two variables is another important element of factor-suitability. It is observed that for any pair of variables, if their association is not much influenced by the other variables (i.e., high partial correlations), then the overall KMO is likely to be low and reflect a lack of factor-suitability. It is also shown that the CMD of a variable could be high when in fact, the associated individual KMO is low. It shows that moderate values of KMO may actually reflect factor-suitability.

In the next section, the datasets described in Section 1 will be used to verify some of the observations made in the theoretical Section 2.

3. Exploration of Initial Dimensions in Datasets

The computation of the KMO requires knowledge of homogeneity of sub-groupings among the indicators. This section therefore explores a procedure for identifying groupings among indicators that could suggest factor-suitability.

Illustration1

Dataset 1 (Performance of Sales Personnel)

Table 2 is the correlation matrix of seven indicators $(x_1, x_2, \ldots, x_7)$ in Dataset 1.

Table 2: Correlation Matrix of Dataset 1

	x_1	x_2	x_3	x_4	x_5	x_6
x_2	0.926					
x_3	0.884	0.843				
x_4	0.572	0.542	0.700			
x_5	0.708	0.746	0.637	0.591		
x_6	0.674	0.465	0.641	0.147	0.386	
x_7	0.927	0.944	0.853	0.413	0.575	0.566

Generally, the correlation coefficients in the table are high, enough to justify the use of factor. Using a cut-off value of 0.5, we construct sets of indicators that are pair-wisely correlated with correlation coefficient of at least 0.5. First, take the pair $(X_i, X_j),\ i, j \in I = (1, 2, \ldots, 7)$ with the highest correlation coefficient. This pair is (x_2, x_7). Thus, we obtain the first set $S_1 = \{x_2, x_7\}$. Let $I_1 = (2, 7)$. If any other variable x_i is such that the correlation coefficients $r_{x_k, x_i} > 0.5,\ \forall k \in I_1,\ i \in I \setminus I_1$, then $x_i \in S_1$, otherwise, $x_i \notin S_1$. Using this rule, it is noticed that $x_1 \in S_1$. The set S_1 is then updated as $S_1 = \{x_2, x_7, x_1\}$ and the index set is also updated as $I_1 = (2, 7, 1)$. Following the process again for the updated set, we notice that $x_3 \in S_1$. The updated sets are

$S_1 = \{x_2, x_7, x_1, x_3\}$ and $I_1 = (2, 7, 1, 3)$. Clearly, $x_4 \notin S_1$ since $r_{x_7, x_4} = 0.413 < 0.5$. By the rule, $x_5 \in S_1$ and we obtain the updated sets $S_1 = \{x_2, x_7, x_1, x_3, x_5\}$ and $I_1 = (2, 7, 1, 3, 5)$. Lastly, $x_6 \notin S_1$ since $r_{x_2, x_6} = 0.465 < 0.5$. Thus, the first homogeneous set is $S_1 = \{x_2, x_7, x_1, x_3, x_5\}$. We should form a new set S_2 from the elements $x_i \notin S_1$, $i \in I \setminus I_1 = \{4, 6\}$, and naturally, $I_1 \cap I \setminus I_1 = \phi$. Denote $I_2 = I \setminus I_1$. However, for these two elements, $r_{x_i, x_j} < 0.5$, $\forall\, i, j \in I_2$. Therefore, we conclude that in this dataset, only one dimension underlies the correlations among the variables.

It should be noted that the two variables which are not part of this single dimension have a low correlation coefficient between them, but quite high with other variables that constitute the main dimension.

There may yet be other features of this rule that are yet to emerge. The use of another dataset with more variables will, hopefully, highlight all the desired features of the procedure.

Illustration 2 (*Dataset 3*)

In this dataset, there are twenty indicator variables. Denote the variables as $(x_1, x_2, \dots, x_{20})$. The correlation coefficients are generally low (see Appendix), with the highest coefficient being 0.492. However, they are all statistically significant. Even the smallest coefficient of 0.107 is significant (with *p*-value of 0.015). On the basis of this, one may attempt to conduct factor analysis on the data. To determine the expected dimensionality in the dataset, we will use a cut-off value of 0.34, on the basis of the low correlation coefficients. We follow the same rule prescribed for Illustration 1. We construct sets of indicators that are pair-wisely correlated with correlation coefficient of at least 0.34. First, take the pair (X_i, X_j), $i, j \in I = (1, 2, \dots, 20)$ with the highest correlation coefficient. This pair is (x_5, x_6). Thus, the first set is $S_1 = \{x_5, x_6\}$. Let $I_1 = (5, 6)$. If any other variable x_i is such that the correlation coefficients $r_{x_k, x_i} > 0.34$, $\forall\, k \in I_1$, $i \in I \setminus I_1$, then $x_i \in S_1$, otherwise, $x_i \notin S_1$. By this rule, $x_9 \in S_1$. The set S_1 is then updated as $S_1 = \{x_5, x_6, x_9\}$ and $I_1 = (5, 6, 9)$. Following the process again for the updated sets, $x_{13} \in S_1$, since $r_{x_k, x_{13}} > 0.34$, $\forall\, k \in I_1$, $i \in I \setminus I_1$. Updating gives $S_1 = \{x_5, x_6, x_9, x_{13}\}$ and $I_1 = (5, 6, 9, 13)$. Now, $r_{x_k, x_i} < 0.34$, for some $k \in I_1$ and some $i \in I \setminus I_1$. Thus, the first homogeneous set is $S_1 = \{x_5, x_6, x_9, x_{13}\}$.

Next, form a new set S_2 from the elements $x_i \notin S_1$, $i \in I \setminus I_1$. Denote $T_1 = I \setminus I_1$. Take the pair (X_i, X_j), $i, j \in I \setminus I_1$ with the highest correlation coefficient. This pair is (x_1, x_2). Thus, we obtain the second set $S_2 = \{x_1, x_2\}$. Let $I_2 = (1, 2)$. Now, $r_{x_k, x_i} < 0.34$, for some $k \in I_2$ and some $i \in I \setminus I_2$. Thus, the second homogeneous set is $S_2 = \{x_1, x_2\}$.

Proceeding similarly, we obtain the sets $S_3 = \{x_{16}, x_{18}, x_{11}, x_{15}, x_{14}\}$, $S_4 = \{x_{19}, x_{20}\}$ and $S_5 = \{x_3, x_7, x_9\}$.

Next, to form a sixth set S_6 from the elements $x_i \notin \left(\bigcup_{k=1}^{5} S_k\right)$, $i \in I \setminus \left(\bigcup_{k=1}^{5} I_k\right)$, denote $T_5 = I \setminus \left(\bigcup_{k=1}^{5} I_k\right)$.

Then take the pair $(X_i, X_j),\ i, j \in T_5$ with the highest correlation coefficient that meets the cut-off value.

Now, for $r_{x_i, x_j} < 0.34,\ \forall i, j \in T_5$. The procedure therefore terminates. It is therefore expected that there would be five main dimensions in this dataset.

The following remarks are about the detection of dimensions in this dataset.

Remarks 3.1

The five remaining variables in $T_5 = \{4, 8, 10, 12, 17\}$ are not independent of each other. By the procedure, there is an incidence of overlapping element. This occurs between the sets S_1 and S_5 with the overlapping element being $V = \{x_9\}$. The cut-off value is chosen particularly to minimise the incidence of overlapping sets.

Illustration 3 (Dataset 5)

In this dataset, there are seven indicator variables. In the order described in Section 1, we will denote the variables as $(x_1, x_2, \ldots, x_7)$, where x_1 denotes 'Cement' and x_7 denotes 'Fine Aggregate' components of the concrete strength. From Table 3, almost all the coefficients are negative, indicating that for any two components, one is very low on the ingredient in the other component. Another observation is that the correlation coefficients are generally low, with the highest coefficient being -0.658. However, they are all

Table 3: Correlation Matrix of Dataset 5

	x_1	x_2	x_3	x_4	x_5	x_6
x_2	-0.275					
x_3	**-0.397**	-0.324				
x_4	-0.082	0.107	-0.257			
x_5	0.092	0.043	0.378	**-0.658**		
x_6	-0.109	-0.284	-0.010	-0.182	-0.266	
x_7	-0.223	-0.282	0.079	-0.451	0.223	-0.178

statistically significant, with exception of the coefficient of -0.010 (with *p*-value of 0.75) between x_3 and x_6. To determine the expected dimensionality in the dataset, we will use a cut-off value of 0.2, on the basis of the low correlation coefficients. In this case, let us suppose that an absolute value of the coefficient is considered.

Following the rule prescribed for previous illustrations, construct sets of indicators that are pair-wisely correlated with absolute correlation coefficient of at least 0.2. Denote $I = (1, 2, \ldots, 7)$. Thus, the first set is $S_1 = \{x_4, x_5\}$. Let $I_1 = (4, 5)$. Now, $\left|r_{x_k, x_7}\right| > 0.2,\ \forall k \in I_1,\ i \in I \setminus I_1$, so $x_7 \in S_1$. Updating gives $S_1 = \{x_4, x_5, x_7\}$ and $I_1 = (4, 5, 7)$. Now, $\left|r_{x_k, x_i}\right| < 0.2$, for some $k \in I_1$ and some $i \in I \setminus I_1$. Thus, the first homogeneous set is $S_1 = \{x_4, x_5, x_7\}$.

We form a new set S_2 from the elements $x_i \notin S_1$, $i \in I \setminus I_1$. Denote $T_1 = I \setminus I_1$. The pair $(X_i, X_j) = (x_1, x_3),\ i, j \in I \setminus I_1$ has the highest absolute correlation coefficient. Thus, the initial second set is

$S_2 = \{x_1, x_3\}$. Let $I_2 = (1, 3)$. Now, $r_{x_k, x_2} > 0.2,\ \forall k \in I_2$, and hence $x_2 \in S_2$. Thus, $S_2 = \{x_1, x_3, x_2\}$ and $I_2 = (1, 3, 2)$. Since $\left|r_{x_k, x_i}\right| < 0.2$, for some $k \in I_2$ and some $i \in I \setminus I_2$, the second homogeneous set is $S_2 = \{x_1, x_3, x_2\}$. We therefore expect that there would be two main dimensions in this dataset. It is also observed that the dimensions are distinct.

The remaining variable in $T_2 = I \setminus (I_1 \cup I_2) = \{6\}$ is not independent of all the other variables since its correlation coefficient with all others are not insignificant. Again, the first subgroup $S_1 = \{x_4, x_5, x_7\}$ in this dataset is made up of variables some of which are negatively correlated and others positively correlated.

Remark 3.2

In order that the procedure would work in this dataset which has widespread negative correlations among indicators, there was the need to fix a cut-off value using the absolute correlation coefficient. Homogeneous sets thus formed are made up of variables that are negatively correlated. The question is, can variables that are negatively correlated constitute homogeneous set? Again, can variables that are both negatively and positively correlated among themselves constitute a homogeneous set? To attempt an answer, we follow the procedure with positive cut-off value as used in the previous illustrations. First, take the pair $(X_i, X_j),\ i, j \in I = (1, 2, \ldots, 7)$ with the highest correlation coefficient. This pair is (x_3, x_5). Thus, the first set is $S_1 = \{x_3, x_5\}$. Let $I_1 = (3, 5)$. Now $r_{x_k, x_i} < 0.2,$, for some $k \in I_1$ $(k = 3)$ and all $i \in I \setminus I_1$. Thus, the first homogeneous set is $S_1 = \{x_3, x_5\}$.

Attempt to form a new set S_2 from the elements $x_i \notin S_1,\ i \in I \setminus I_1$. Denote $T_1 = I \setminus I_1$. The only pair $(X_i, X_j),\ i, j \in I \setminus I_1$ with the highest correlation coefficientis (x_5, x_7). However, since $5 \in I_1$, we cannot have this starting pair. Thus, there is only a single homogeneous set given as $S_1 = \{x_3, x_5\}$. Therefore, only one main dimension is expected in this dataset.

Generalisation of the Rule for Determining Expected Dimensions of Datasets

Suppose the dataset is generated on a set of p variables $(X_1, X_2, \ldots, X_p)$ with correlation coefficients that are generally significant. On the basis of the level of correlation coefficients, we fix a cut-off value τ for which variables may be considered to belong together if their pair-wise correlation coefficients exceed τ. First, take the pair $(X_i, X_j),\ i, j \in I = (1, 2, \ldots, p)$ with the highest correlation coefficient. Let this pair be (x_{i_1}, x_{i_2}), and label the set as $S_1 = \{x_{i_1}, x_{i_2}\}$ and the index set $I_1 = \{i_1, i_2\}$. If the correlation coefficients $r_{x_k, x_i} > \tau,\ \forall k \in I_1,\ i \in I \setminus I_1$, then $x_i \in S_1$, otherwise, $x_i \notin S_1$. The sets S_1 and I_1 are updated each time. Now, if $r_{x_k, x_i} < \tau$, for some $k \in I_1$ and all $i \in I \setminus I_1$, then the final first homogeneous set is $S_1 = \{x_{i_1}, x_{i_2}, \ldots, x_{i_{g1}}\}$ with index set $I_1 = \{i_1, i_2, \ldots, i_{g1}\} \subset I$.

We will form a new set S_2 from the elements $x_i \notin S_1$, $i \in I \setminus I_1$. Denote $T_1 = I \setminus I_1$. Consider the pair (x_i, x_j), $i, j \in I \setminus I_1$ with the highest correlation coefficient that meets the cut-off value τ. Suppose this pair is (x_{f_1}, x_{f_2}). Thus, we obtain the second set $S_2 = \{x_{f_1}, x_{f_2}\}$, and an index set $I_2 = \{f_1, f_2\}$. Now, if the correlation coefficients $r_{x_k, x_i} > \tau$, $\forall k \in I_2$, $i \in I \setminus I_2$, then $x_i \in S_1$, otherwise, $x_i \notin S_1$. The sets S_2 and I_2 are updated each time. Now, if $r_{x_k, x_i} < \tau$, for some $k \in I_2$ and all $i \in I \setminus I_2$ then we obtain a final second homogeneous set $S_2 = \{x_{f_1}, x_{f_2}, \ldots, x_{f_{g2}}\}$ with index set $I_2 = \{f_1, f_2, \ldots, f_{g2}\} \subset I$.

Consider all elements $x_i \notin (S_1 \cup S_2)$, $i \in I \setminus (I_1 \cup I_2)$. Denote $T_2 = I \setminus (I_1 \cup I_2)$. To form the new set, take the pair (x_i, x_j), $i, j \in T_2$ with the highest correlation coefficient that meets the cut-off value τ. Let the pair be (x_{t_1}, x_{t_2}). Then $S_3 = \{x_{t_1}, x_{t_2}\}$, and $I_3 = \{t_1, t_2\}$. If the correlation coefficients $r_{x_k, x_i} > \tau$, $\forall k \in I_3$, $i \in I \setminus I_3$, then $x_i \in S_3$, otherwise, $x_i \notin S_3$. The set S_3 and I_3 are updated each time. Now, if $r_{x_k, x_i} < \tau$, for some $k \in I_3$ and all $i \in I \setminus I_3$ then we obtain the final third homogeneous set $S_3 = \{x_{t_1}, x_{t_2}, \ldots, x_{t_{g2}}\}$ with index set $I_3 = \{t_1, t_2, \ldots, t_{g3}\} \subset I$.

We attempt to form the qth set S_q from the elements $x_i \notin \left(\bigcup_{k=1}^{q-1} S_k \right)$, $i \in I \setminus \bigcup_{k=1}^{q-1} I_k$. Denote $T_{q-1} = I \setminus \bigcup_{k=1}^{q-1} I_k$. Take the pair (x_i, x_j), $i, j \in T_{q-1}$ with the highest correlation coefficient that meets the cut-off value τ. Thus, $S_q = \{x_{d_1}, x_{d_2}\}$, and $I_q = \{d_1, d_2\}$. Now, if $r_{x_k, x_i} < \tau$, for some $k \in I_q$ and all $i \in I \setminus I_q$, then we obtain the final qth homogeneous set $S_q = \{x_{d_1}, x_{d_2}, \ldots, x_{d_{gq}}\}$ with index set $I_q = \{d_1, d_2, \ldots, d_{gq}\} \subset I$.

Now, if for some set S_{l+1} and index set I_{l+1}, and for $x_i \notin \left(\bigcup_{k=1}^{l} S_k \right)$, $r_{x_i, x_i} < \tau$, for all $i, j \in I \setminus \bigcup_{k=1}^{l} I_k = T_l$, then S_l is the last set of variables in the original set of p variables and there are a total of l dimensions underlining the correlation matrix.

Remarks 3.3

By the outlined procedure, there is an incidence of overlapping elements in two or more of the sets. The remaining variables in T_l that do not influence any dimension may not be independent of the others. If these 'non-classified' variables are independent of the others, they potentially constitute a one-variable dimension. It is expected that the overall level of homogeneity of the dataset that determines its factor suitability would be affected by this non-classified set of elements and the overlapping elements.

4 Computation of the KMO Measure of Sampling Adequacy

First, Table 4 presents a summary of some statistics obtained from the datasets used in the study. The number of homogeneous sets is obtained by the procedure of dimensionality determination discussed in Section 2.

Table **4**: Summary Statistics of Datasets used in the Study

Data Number	Data Description	No. of Indicators	KMO	Cut-off	Number of Sub-groups
1	Sales personnel Performance	7	0.616	0.50	1
2	SubjectScores	9	0.822	0.50	2
3	StudBenInd Attach	20	0.924	0.34	5
4	ComPrice	19	0.734	0.30	5
5	ConcStrength	7	0.140	0.20	1
6	StdtChall in Ind. Attach	28	0.797	0.20	8

In this section, we examine various values of the KMO and determine how practical the interpretation is as given in Table 1.

The discussion in the methods show scenarios of correlations under which we could expect low or high value of the KMO. It is pertinent therefore, to expect the value to be influenced by the number of sub-groups among the original set of variables. Again, the number of variables in each group may also influence the value of the KMO. This is the motivation for the discussion in this section.

We can deduce from our previous discussion on Equation (3) that a KMO value which is not too high (see Table 1) may be an indication that there is generally one (or few) major dimension underlining the correlation matrix. This case is demonstrated using Dataset 1.

Illustration 4

Dataset 1 (Performance of Sales Personnel)

Generally, the direct correlations given by the PCC matrix in Table 5 are high reflecting generally high zero-order correlations as pointed out earlier.

Table 5: Partial Correlation Matrix for Dataset 1

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	0.663						
x_2	0.248	0.783					
x_3	-0.570	0.103	0.630				
x_4	0.723	-0.077	0.876	0.410			
x_5	0.123	0.681	0.014	0.018	0.750		
x_6	0.794	-0.398	0.763	**-0.799**	0.140	0.417	
x_7	0.609	0.528	0.623	-0.669	-0.515	-0.413	0.632

It should be noted further that the diagonal elements are the KMO of the individual variables. These are not the coefficients of multiple determination of the variables (See Table 9) which are pointed out in the mathematical background.

From Table 2, we obtain $\sum r_{ij}^2 = 0.926^2 + 0.884^2 \ldots 0.575^2 + 0.566^2 = 9.82497$

and from Table 5, we obtain $\sum pr_{ij}^2 = 0.248^2 + 0.570^2 \ldots 0.515^2 + 0.413^2 = 6.122616$,

Using Equation (3), we have $KMO = 0.6161$. This value is the same as that in Table 4 generated in SPSS.

Now, consider the two sets $S_1 = \{x_2, x_7, x_1, x_3, x_5\}$ and $S_2 = \{x_4, x_6\}$ of variables identified earlier in the dataset with indexed sets $I_1 = (2, 7, 1, 3, 5)$ and $I_2 = (4, 6)$. Consequently, KMO is computed in terms of each of these components. That is, we consider the results for $KMO_{I_k} = \dfrac{\sum_{i,j\in I_k} r_{ij}^2}{\sum_{i,j\in I_k} r_{ij}^2 + \sum_{i,j\in I_k} pr_{ij}^2}$; $k = 1, 2$. The two components of $\sum pr_{ij}^2$ gives

$\sum_{i,j\in I_1} pr_{ij}^2 = 0.248^2 + 0.570^2 + \ldots + 0.515^2 + 0.413^2 = 5.484215$, excluding values involving $I_2 = (4, 6)$, and $\sum_{i,j\in I_2} pr_{ij}^2 = 0.723^2 + 0.077^2 + \cdots + 0.14^2 + 0.413^2 = 3.943498$, including all values that involves variables in I_2. The values involved in the sum for the set I_2 are highlighted in the table with the intersection (0.799) counted once. Each of the values is used once in the summations. The three KMO values are summarised in Table 6 shown.

Table 6: KMO for Dataset 1 Based on Sub-groupings of Variables

SN	Grouping	$\sum_{i,j\in I_k} r_{ij}^2$	$\sum_{i,j\in I_k} pr_{ij}^2$	KMO Value
1	All	9.8250	6.1226	0.6161
2	S_1 only	9.8034	5.4842	0.6413
3	$(S_1)'$	3.2031	3.9435	0.4482

Since only one main dimension is detected in the dataset, it can be deduced that on the basis of the single dimension in the data, the KMO of this dataset could be approximately 0.64. This value has the same description as the original value as being 'mediocre' (Table 1).

Of interest is to assess the KMO of the individual variables of this dataset. In this dataset, variable x_1 has the highest correlations with all other variables. Its partial correlation coefficients are therefore expected to be high in general. We see this in Table 5. It can be verified that the variable KMO, $KMO_{x_i} = 0.663$ as seen in Table 9. We can obtain the CMD for the variable as in Table 9.

To highlight the point further we obtain the CMD for the models for each of the variables. The coefficient estimates of the regression model for x_7 in terms of the other variables and corresponding significance as well as the partial correlation coefficients are given in Table 7.

Table 7: Coefficients and Description of Model for x_7 in Dataset 1

Model	Unstandardized Coefficients B	Std. Error	t	Sig.	Correlations Zero-order	Partial
(Constant)	-173.303	14.408	-12.029	0.000		
x_1	0.790	0.157	5.038	0.000	0.927	0.609
x_2	0.471	0.115	4.080	0.000	0.944	0.528
x_3	0.989	0.189	5.228	0.000	0.853	0.623
x_4	-0.882	0.150	-5.897	0.000	0.413	-0.669
x_5	-0.522	0.132	-3.941	0.000	0.575	-0.515
x_6	-0.914	0.307	-2.977	0.005	0.566	-0.413

It is noticed that all of the variables are significant in the model (all sig. are less than 0.05), and that the variables have little reduction in size (in absolute terms) in PCC compared to the zero-order correlation. We present one other model for variable x_5 with high KMO value in Tables 9. It is noted that as much as four variables (out of six) are insignificant in the model, and these are those with drastic reduction in partial correlations.

It must be pointed out that there could be a complete change in sign (from positive to negative) for which the partial correlation in absolute terms shows rather a small decrease or an increase and does not cause the variable to be insignificant.

Table 8: Coefficients and Description of Model for x_5 in Dataset 1

Model	Unstandardized Coefficients B	Std. Error	t	Sig.	Correlations Zero-order	Partial
(Constant)	-55.956	28.468	-1.966	0.056		
x_1	0.158	0.194	.814	0.420	0.708	0.123
x_2	0.600	0.098	6.103	0.000	0.746	0.681
x_3	0.022	0.239	.091	0.928	0.637	0.014
x_4	0.024	0.199	.118	0.906	0.591	0.018
x_6	0.307	0.330	.930	0.357	0.386	0.140
x_7	-0.509	0.129	-3.941	0.000	0.575	-0.515

Remark 4.1

The models show that variables in the partial models for a variable (x_i) would be significant for those with little reduction in size in partial correlation coefficients compared to the zero-order correlation. On the other hand, variables will not be significant in a model for those with drastic reduction in size of the partial correlations. A small reduction in the partial correlation, for example, is an indication of 'consistency' of relationship between the variable and the others. In this case, all groupings of variables can explain the variation in the variable, hence a high CMD. Thus, the presence of other variables does not influence its relationship with another. This will translate into a low KMO (about 0.5). A low KMO value therefore reflects a uniform relationship with all other indicators and may reflect a 'lack' of meaningfulness of factor-suitability of the variable. A high CMD could therefore translate into a low to moderate individual KMO value.

We examine these observations in Table 9 which gives the CMD (R^2) of all (partial) models for each of the variables in terms of the other variables and their corresponding KMO value, as well as the communalities from a specified factor solution. Table 9 shows that there is no apparent connection between variable KMO and R^2 in general. As expected, however, a low KMO value is associated with a high R^2. It can be observed that a high KMO is not necessarily associated with a high communality, even though for a suitable factor solution, some association is quite discernible. The result shows that a one-factor model which was initially identified for this dataset would be more consistent with the KMO values. We can however conclude from this dataset that there is no definite representation of the individual KMO value. A much lower KMO (less than 0.5) in this case definitely suggests that the variable does not influence any dimension.

Table 9: Summary of Factor and Model Statistics for each variable in Dataset 1

			Communality	
Variable	KMO	R^2	One Factor Solution	Two-Factor Solution
1	0.663	0.972	0.947	0.959
2	0.783	0.968	0.889	0.890
3	0.630	0.954	0.893	0.893
4	0.410	0.905	0.436	0.853
5	0.750	0.781	0.614	0.695
6	0.417	0.885	0.421	0.806
7	0.632	0.977	0.836	0.873

Illustration 5

Dataset 2 (Performance of High School Students in Nine Subjects)

Table 10 is the zero-order correlation matrix of Dataset 2. Generally, the correlation coefficients are high. However, the coefficients are not too high among sub-groups of variables. Thus, by the argument presented earlier, the moderate high coefficients would cause a significant reduction in the PCC. This is precisely what we observe in Table 11. To verify the value of the KMO for this dataset, we compute from Table 10, the following sums:

$\sum r_{ij}^2 = 0.135^2 + 0.160^2 + \ldots + 0.422^2 + 0.067^2 = 5.577275,$ and from Table 11,

$\sum pr_{ij}^2 = 0.248^2 + 0.570^2 \ldots 0.515^2 + 0.413^2 = 1.206068.$. Using Equation (3), the KMO value is 0.8222. This value is the same as that in Table 4.

Table 10: Zero-Order Correlation Coefficient Matrix of Dataset 2

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_2	0.135							
x_3	0.160	0.637						
x_4	-0.085	0.549	0.402					
x_5	0.180	0.431	0.318	0.407				
x_6	0.126	0.693	0.616	0.381	0.289			
x_7	0.020	0.627	**0.746**	0.447	0.317	0.604		
x_8	-0.113	0.010	-0.018	-0.029	-0.028	-0.011	-0.019	
x_9	0.045	0.692	0.464	0.504	0.386	0.395	0.422	0.067

Table 11: Partial Correlation Coefficient Matrix of Dataset 2

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
x_1	0.410								
x_2	0.106	0.808							
x_3	0.170	0.089	0.831						
x_4	-0.213	0.178	0.008	0.876					
x_5	0.191	0.100	-0.005	0.227	0.881				
x_6	0.032	0.451	0.184	-0.001	-0.023	0.849			
x_7	-0.162	0.148	0.542	0.103	0.048	0.153	0.824		
x_8	-0.118	0.008	-0.002	-0.072	-0.011	0.004	-0.022	0.415	
x_9	-0.047	0.503	0.105	0.173	0.098	-0.168	-0.064	0.089	0.806

In this datasets, only two groups of variables are identified that could constitute the main dimensions. These are variables with index $I_1 = (3,\ 7,\ 6)$, and $I_2 = (2,\ 9,\ 4)$. Thus, the set of variables that are not classified has the index set $T_2 = \{1,\ 5,\ 8\}$.

Hence, the following results given by $KMO_{I_k} = \dfrac{\sum\limits_{i,j \in I_k} r_{ij}^2}{\sum\limits_{i,j \in I_k} r_{ij}^2 + \sum\limits_{i,j \in I_k} pr_{ij}^2}$; $k = 1,\ 2,\ 3$ are considered. The values of the three components of $\sum pr_{ij}^2$ are given in Table 12. It also gives the overall KMO value that involves the two sets, the combined set $S_1 \cup S_2$ and the non-classified set $(S_1 \cup S_2)'$ given respectively by

$$KMO_{S_1\cup S_2}=\frac{\sum_{i,j\in(I_1\cup I_2)} r_{ij}^2}{\sum_{i,j\in(I_1\cup I_2)} r_{ij}^2+\sum_{i,j\in(I_1\cup I_2)} pr_{ij}^2} \text{ and } KMO_{(S_1\cup S_2)'}=\frac{\sum_{i,j\in(I_1\cup I_2)'} r_{ij}^2}{\sum_{i,j\in(I_1\cup I_2)'} r_{ij}^2+\sum_{i,j\in(I_1\cup I_2)'} pr_{ij}^2}.$$

Table 12: KMO for Dataset 2 Based on Sub-groupings of Variables

SN	Grouping	$\sum_{i,j\in I_k} r_{ij}^2$	$\sum_{i,j\in I_k} pr_{ij}^2$	KMO Value
1	All	5.5773	1.2061	0.8222
2	S_1 only	3.9637	0.6978	0.8503
3	S_2 only	3.9027	0.7626	0.8365
4	$\bigcup_{i=1}^{2} S_i$ only	5.5313	1.1555	0.8272
5	$\left(\bigcup_{i=1}^{2} S_i\right)'$ only	0.9071	0.2532	0.7818

Since only two main dimensions are detected in the dataset, it can be deduced that the KMO of this dataset could be approximately 0.8503.

As in Dataset 1, the individual variable KMO could be assessed for this dataset. Presented in Table 13 are the CMD (R^2) of all (partial) models for each of the variables in terms of the others and their corresponding KMO value, as well as the communalities from two- and three-factor solutions. In the table, there appears to be some significant correlation (obtained as 0.749 with *p*-value 0.02) between R^2 and KMO.

Table 13: Summary of Factor and Model Statistics for each variable in Dataset 2

Variable	KMO	R^2	Communality 2-Factor Solution	Communality 3-Factor Solution
1	0.410	0.137	0.669	0.900
2	0.808	0.732	0.795	0.797
3	0.831	0.632	0.667	0.675
4	0.876	0.405	0.519	0.624
5	0.881	0.264	0.342	0.342
6	0.849	0.558	0.599	0.608
7	0.824	0.624	0.644	0.646
8	0.415	0.026	0.388	0.974
9	0.806	0.529	0.560	0.561

The low R^2 values are for those variables in the set $T_2=\{1, 5, 8\}$ which are non-classified in along any major dimensions. From the table, there is no clear association between communality and the R^2. A high KMO is not necessarily associated with a high communality, even though the two-factor model, which was initially identified for this dataset, would be more consistent with the KMO values. We can however conclude from this dataset that generally, there does not appear to be a clear representation of the KMO regarding the factor-suitability of the individual variables. However, a very low KMO definitely suggests the variable does not influence any dimension. The overall KMO, however, appears to reflect the general factor-suitability of the dataset.

Table 14 gives the computation of the KMO for various groupings in Dataset 3 (Industrial Attachment Benefits) as identified in Section 3.

Table 14: KMO for Dataset 3 Based on Sub-groupings of Variables

SN	Grouping	$\sum_{i,j \in I_k} r_{ij}^2$	$\sum_{i,j \in I_k} pr_{ij}^2$	KMO Value
1	All	17.7609	1.3987	0.9270
2	S_1 only	7.1020	0.5357	0.9299
3	S_2 only	2.7479	0.2419	0.9191
4	S_3 only	8.5097	0.5509	0.9392
5	S_4 only	2.4937	0.2393	0.9125
6	S_5 only	5.1622	0.4830	0.9144
7	$\bigcup_{i=1}^{5} S_i$ only	17.0100	1.3631	0.9258
8	$\left(\bigcup_{i=1}^{5} S_i\right)'$ only	7.2650	0.5326	0.9317

An interesting observation is the high value of the KMO of the non-classified variables. Table 15 shows very high individual KMO values for all indicators. The impression is that the variables would be adequately explained by a suitably identified factor solution. However, from the table, there is rather a negative relationship between communality and KMO (correlation obtained as -0.8, with *p*-value 0.000 for both factor solutions). Thus, a high KMO is rather associated with a low communality, and this association is not connected to a particular factor solution. We can therefore conclude from this dataset that there does not appear to be a clear representation of the KMO regarding the factor suitability of the individual variables, as it is expected that a high KMO would be associated positively with a high communality. The table also shows rather low values of the CMD. This is a further indication that most of the variables are not too highly pair-wisely correlated, which does not support the formation of parsimonious homogeneous groups. It is therefore not surprising that in this dataset, several groupings were identified, with several unclassified indicators. Thus, these low to moderate R^2 are to be expected. It should be noticed that the lowest R^2 values are associated with those variables in the unclassified set $I_6 = \{4, 8, 10, 12, 17\}$. The overall KMO, however, appears to reflect the general factor-suitability of the dataset.

Mathematical Theory and Modeling www.iiste.org
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.8, No.7, 2018

Table 15: Summary of Factor and Model Statistics for Variables in Dataset 3

Variable	KMO	R^2	Communality 4-Factor Solution	Communality 5-Factor Solution
1	0.888	0.325	0.594	0.607
2	0.902	0.336	0.556	0.653
3	0.927	0.338	0.498	0.621
4	0.926	0.313	0.533	0.538
5	0.908	0.431	0.531	0.638
⋮	⋮	⋮	⋮	⋮
16	0.924	0.375	0.530	0.539
17	0.938	0.319	0.351	0.430
18	0.932	0.367	0.449	0.460
19	0.894	0.345	0.578	0.579
20	0.868	0.243	0.646	0.682

Illustration 6

Dataset 5 (The Concrete Compressive Strength)

It is recalled that in this dataset, only one dimension is formed involving only two variables (x_3, x_5). Thus, there are five unclassified variables with index $I_2 = \{1, 2, 4, 6, 7\}$. In Table 16, we have the overall KMO value of Dataset 5 as well as the KMO of the single dimension.

Table 16: KMO for Dataset 5 Based on Sub-groupings of Variables

SN	Grouping	$\sum_{i,j \in I_k} r_{ij}^2$	$\sum_{i,j \in I_k} pr_{ij}^2$	KMO Value
1	All	1.6254	9.9743	0.1401
2	S_1 only	1.0416	3.4600	0.2314
3	$(S_1)'$	1.4266	9.3193	0.1328

That PCC matrix shows very large correlations between variables after controlling for others. As pointed out, the high PCC values suggest highly unstable correlations between pairs of variables and are influenced by the presence of others. This has resulted in very low KMO values. The result shows that the identification of the single dimension in the dataset is appropriate and consistent with the factor structure of the data.

Table 17 shows very low individual KMO values for all indicators, with exception of variable 5 (Plasticiser). The impression is that the variables could not be adequately explained by a suitably identified factor solution. However, from the table, though the relationship between communality and KMO appears high particularly for the one-factor solution (correlations obtained as 0.646 with p-values 0.117) the relationship is not significant. One- and Two-Factor solutions are examined as the initial dimension does not exceed two. Thus, the individual KMO is not clearly linked with the communality.

Table 17: Summary of Factor and Model Statistics for each variable in Dataset 5

			Communality	
Variable	KMO	R^2	1-Factor Solution	2-Factor Solution
1	0.083	0.866	0.016	0.019
2	0.095	0.861	0.105	0.692
3	0.136	0.836	0.372	0.402
4	0.201	0.857	0.681	0.682
5	0.613	0.662	0.600	0.761
6	0.066	0.802	0.003	0.628
7	0.108	0.856	0.386	0.386

The table also shows rather high values of CMD, suggesting that each variable can be reliably predicted by the others. It further indicates that most of the variables are significantly pair-wisely correlated. However, this does not translate into formation of well-defined homogeneous groups. There is negative correlation (obtained as -0.906 with *p*-value 0.005) between R^2 and KMO. The element with the highest KMO (variable 5) rather has the least (moderate) R^2 value, which is expected. The KMO is consistent with the factor structure of the data as it has very few elements constituting the single dimension.

5 Conclusions and Recommendations

The paper attempts to identify problems encountered with the use of KMO as a suitability measure for Factor Analysis technique. It has made use of a number of datasets that are selected to highlight various problems.

In order to understand the factor structure of any data, the paper has systematically described an approach that explores the dimensionality of the dataset that could justify the use of Factor Analysis. The procedure shows that there are some indicators that may not influence any of the dimensions. It is observed that this set of 'unclassified' indicators could be excluded in the determination of the factor-suitability of the data. In addition, there could be a number of indicators that influence multiple dimensions, and could adversely affect the factor-suitability, especially when they are many. This may be avoided if variables are constructed so that they do not correlate too highly. Exploring dimensions in future studies will provide an algorithm that will make easy applications in datasets with several variables.

Another important observation is that the dimensionality of datasets could be affected by prevalence of negative correlations among indicators. Negative correlations distort the notion of homogeneity. As a result, in datasets with negative correlations, the determination of dimensionality is eventually based on a few significant positive correlations, leading to a small adequacy measure. This has the tendency to portray such datasets as unsuitable for factor analysis.

It is found that for KMO to be high, the zero-order and partial correlations must be almost the same for indicators that influence the same dimension. Following this pattern, it is found that generally, a KMO value within the range 0.6 – 0.7 is typically a good measure of factor-suitability. Although the overall KMO typically reflects factor-suitability, the KMO of the individual variables does not appear to have a general representation. A high variable KMO is found to be associated with moderate coefficient of multiple determination, but its relation to the communality is not discernible for a suitably selected factor solution.

The study further shows that the KMO may not be a golden rule for determining factor-suitability. The nature of the relationship among indicators and the design of the study could inform the factor-suitability of the data.

References

Comrey, A. L. & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Frempong, M., Nkansah, H. & Nkansah, B.K. (2017). Multivariate Assessment of Industrial Attachment by Students of Tertiary Institutions in Ghana: *Mathematical Theory and Modelling*, **7**(6), ISSN 2224-5804.

Field, A. (2000). *Discovering Statistics using SPSS for Windows*. London – Thousand Oaks – New Delhi: Sage publications.

Johnson, R.A., & Wichern, D.W. (2007). *Applied Multivariate Statistical Analysis* (6th ed.). NJ: Prentice-Hall Int., Inc.

Kaiser, H. 1974. An index of factor simplicity. Psychometrika 39: 31–36.

Kaiser, H. F. & Rice, J. (1974): "Little Jiffy Mark IV", *Psychometrica* 35 (December), 401 – 415.

MacCallum, R.C., Widaman, K.F., Zhang, S.B., & Hong, S.H. (1999). Sample size in factor analysis. *Psychological Methods*, **4**(1), 84–99.

Rencher, A.C. (2002). *Methods of Multivariate Analysis*: (2nd Ed.). New Jersey: John Wiley & Sons.

Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics* (5^{th} ed.). Boston, Pearson.

Seglah, F. (2014).Multivariate Analysis of Food Prices in Ghana. MPhil Thesis, University of Cape Coast.

Sharma, S. (1996). *Applied Multivariate Techniques*. Wiley, New York.

van der Eijk, C., & Rose, J. (2015). Risky Business: Factor Analysis of Survey Data – Assessing the Probability of Incorrect Dimensionalisation. PLoS ONE 10(3): e0118900. doi:10.1371/journal.pone.0118900

Yeh, I-Cheng (1998).Modelling of strength of high performance concrete using artificial neural networks. *Cement and Concrete Research*, **28**(12) 1797-1808.

Yeh, I-Cheng (1998b). Modelling Concrete Strength with Augment-Neuron Networks, *Journal of Materials in Civil Engineering*, ASCE, **10**(4), 263-268.

Yeh, I-Cheng (1999). Design of High Performance Concrete Mixture using Neural Networks, *Journal of Computing in Civil Engineering*, ASCE, **13**(1), 36-42.

Yeh, I-Cheng (2003a). Prediction of Strength of Fly Ash and Slag Concrete by the use of Artificial Neural Networks, *Journal of the Chinese Institute of Civil and Hydraulic Engineering*, **15**(4), 659-663.

Yeh, I-Cheng (2003b). A mix Proportioning Methodology for Fly Ash and Slag Concrete using Artificial Neural Networks, *Chung Hua Journal of Science and Engineering*, **1**(1), 77-84.

Yeh, I-Cheng (2006). "Analysis of strength of concrete using design of experiments and neural networks, *Journal of Materials in Civil Engineering*, ASCE, **18**(4), 597-604.

Appendix

Table A1: Correlations among Indicators of Dataset 3

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
2	0.476																		
3	0.276	0.347																	
4	0.269	0.293	0.396																
5	0.361	0.327	0.360	0.416															
6	0.345	0.332	0.243	0.293	0.492														
7	0.219	0.274	0.400	0.289	0.254	0.296													
8	0.250	0.285	0.203	0.179	0.326	0.334	0.352												
9	0.255	0.268	0.338	0.237	0.402	0.388	0.368	0.326											
10	0.232	0.287	0.224	0.203	0.310	0.266	0.290	0.324	0.316										
11	0.162	0.239	0.292	0.318	0.278	0.278	0.352	0.175	0.314	0.288									
12	0.279	0.274	0.331	0.259	0.290	0.263	0.345	0.278	0.292	0.298	0.372								
13	0.277	0.255	0.340	0.325	0.401	0.369	0.302	0.297	0.412	0.269	0.415	0.455							
14	0.200	0.245	0.373	0.311	0.379	0.206	0.313	0.222	0.381	0.245	0.338	0.372	0.394						
15	0.277	0.249	0.307	0.352	0.391	0.324	0.306	0.282	0.339	0.334	0.370	0.313	0.445	0.392					
16	0.204	0.206	0.255	0.318	0.237	0.247	0.308	0.201	0.271	0.274	0.427	0.284	0.373	0.346	0.427				
17	0.266	0.301	0.286	0.257	0.296	0.246	0.307	0.234	0.336	0.264	0.305	0.288	0.257	0.394	0.391	0.339			
18	0.255	0.246	0.299	0.245	0.321	0.283	0.351	0.198	0.295	0.268	0.366	0.306	0.306	0.380	0.387	0.453	0.424		
19	0.160	0.269	0.277	0.313	0.242	0.239	0.324	0.282	0.257	0.252	0.278	0.279	0.257	0.314	0.370	0.375	0.266	0.319	
20	0.193	0.193	0.162	0.228	0.232	0.180	0.170	0.107	0.182	0.209	0.219	0.197	0.225	0.272	0.238	0.298	0.187	0.286	0.438

Table A2: Partial Correlation Matrix of Dataset 3

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0.888																			
2	0.326	0.902																		
3	0.042	0.128	0.927																	
4	0.044	0.058	0.176	0.926																
5	0.096	0.007	0.088	0.202	0.908															
6	0.106	0.090	-0.062	0.029	0.275	0.911														
7	-0.004	0.009	0.194	0.058	-0.088	0.060	0.929													
8	0.039	0.076	-0.062	-0.053	0.096	0.098	0.180	0.906												
9	0.011	-0.005	0.082	-0.065	0.098	0.143	0.107	0.077	0.942											
10	0.012	0.098	-0.010	-0.041	0.078	-0.008	0.048	0.137	0.085	0.946										
11	-0.099	0.041	0.012	0.087	-0.008	0.049	0.105	-0.072	0.050	0.084	0.938									
12	0.095	0.017	0.077	-0.007	-0.017	-0.005	0.078	0.063	-0.033	0.078	0.119	0.937								
13	0.038	-0.019	0.059	0.036	0.071	0.093	-0.030	0.064	0.144	-0.029	0.141	0.228	0.924							
14	-0.051	0.016	0.105	0.031	0.135	-0.127	0.023	-0.004	0.129	-0.037	0.035	0.114	0.088	0.937						
15	0.059	-0.044	0.011	0.072	0.085	0.035	-0.005	0.031	0.007	0.100	0.051	-0.024	0.162	0.074	0.948					
16	0.015	-0.034	-0.026	0.091	-0.090	0.013	0.029	0.002	-0.013	0.044	0.179	-0.020	0.109	0.044	0.136	0.924				
17	0.052	0.088	0.008	0.017	0.003	-0.016	0.037	0.027	0.101	0.013	0.039	0.033	-0.081	0.133	0.141	0.056	0.938			
18	0.046	-0.011	0.042	-0.066	0.071	0.058	0.105	-0.044	-0.005	0.028	0.063	0.033	-0.026	0.094	0.054	0.214	0.198	0.932		
19	-0.094	0.082	0.039	0.095	-0.050	0.030	0.088	0.122	0.018	0.019	-0.010	0.056	-0.060	0.045	0.124	0.120	0.014	0.019	0.894	
20	0.075	0.011	-0.058	0.026	0.051	0.002	-0.030	-0.093	-0.012	0.066	0.021	-0.000	0.033	0.087	-0.031	0.068	-0.017	0.085	0.314	0.868