Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7778458

AUTOMATED THRESHOLD METHOD FOR DIMENSIONALITY DETECTION IN MULTIVARIATE DATA

Alfred Asiwome Adu[1], David Kwamena Mensah[2], Henrietta Nkansah[3]
Bismark Kwao Nkansah[2*]

1 Department of Mathematics and Science Education, Valley View University, Accra, Ghana

2 Department of Statistics, University of Cape Coast, Ghana

3 Department of Mathematics, University of Cape Coast, Ghana

* Corresponding Author (bnkansah@ucc.edu.gh)

Abstract

A measure of well-defined homogenous subsets among indicator variables on which multivariate data is generated is given by Kaiser-Meier-Olkin's measure of sampling adequacy (KMO). This measure relies on a subtle use of a cut-off value. This cut-off value as well as the expected number of dimensions in the data constitute important background information for dimensionality detection that is not reported in the application of dimensionality reduction techniques. The implication is that these techniques do not establish a priori the existence of dimensionality in the data, and hence could be misapplied. In this regard, the study proposes an automated threshold-setting approach with an algorithm that generates a data-specific optimal threshold from the data structure for detecting the dimensionality of multivariate data for more accurate results. Three different threshold settings are implemented for various correlation profiles of the data. The known techniques may now be useful for purposes of interpretation of the extracted reduced dimensions. Results are further explained using confirmatory factor analysis. The proposed method completely removes the challenge of subjectivity associated with dimensionality detection.

KEY WORDS: DIMENSIONALITY DETECTION, KMO, SIMILARITY DETECTOR, THRESHOLD SETTING

1 Introduction

The dimensionality of a dataset has been described as the minimum number of unobserved traits that is needed to describe all statistical dependencies in the data (Lord & Novick, 1968; Zhang & Stout, 1999). From a practical point of view, the determination of dimensionality helps to understand the structure of the phenomenon (Pett, Lackey, Sullivan, 2003). Performance of dimensionality assessment methods have been the focus of a number of notable studies (e.g., Zhang, 2016; Zupluoglu, 2013; Tabachnick & Fidell, 2001).

Dimensionality detection in multivariate data has commonly made use of statistical techniques such as principal component analysis (PCA), factor analysis (FA) and item response theory (IRT) modelling. However, these statistical techniques do not establish the exact number of dimensions that exist in the data prior to their application. In FA, for example, a value greater than or equal to 0.6 of the Kaiser-Meier-Olkin's measure of sampling adequacy (KMO) is seen as an indication of existence of dimensionality in the data. It has been demonstrated (Nkansah, 2018) that for some datasets, it may be difficult to determine the dimensionality

even though the KMO measure may suggest that such datasets have underlying dimensions. In addition, the suitability criteria provided by the KMO does not also provide indication of the expected number of dimensions that possibly underlie the data. It is the view of this paper that a prior knowledge of the number of dimensions must be available so that application of the existing multivariate techniques may just be useful for extracting the actual dimensions. One does not have to extract dimensions from a dataset only to find out that those dimensions do not exist, or are not interpretable. Perhaps, it is this gap that has inspired further works to determine the dimensionality of multivariate models (Rutledge, Roger, & Lesnoff, 2021) with minimal error.

Additionally, for the same data, different techniques may yield different dimensionalities, even where basic conditions of data size (Comrey & Lee, 1992; Nkansah, Zakaria & Howard, 2019) and type of data and correlation profile (van der Eijk & Rose, 2015) are duly considered. Even though the relative importance of the dimensions may differ from technique to technique, the basic number of dimensions should be the same, and this information is what appears unavailable.

One of few works that specifically focus on dimensionality detection by the use of KMO is that by Nkansah (2018), which observed some drawbacks on the subject. In particular, the study uses a subjective exprimenter-specific threshold to demonstrate that the usual KMO is one that is obtained from the entire data and that it may not be a fair measure of a well-defined dimensionality for the data. The demonstration in that paper obviously involves some computational intensity. The goal of this research, therefore, is to make more explicit the use of a non-subjective cut-off value for the determination of a more representative KMO value for a dataset by proposing an automated data-specific threshold which is generated from the data structure. This study also investigates the sensitivity and robustness of the method based on the correlation profile adopted for the data.

The remaining sections are arranged as follows: Section 2 presents the methodology used for this research. The development and implementation of the algorithms are carried out in Section 3 with relevant codes written in R. Finally, Section 4 presents the conclusion and recommendation.

2 Methodology

As noted in the introduction, this work is motivated by the work of Nkansah (2018) on the computation of the KMO. The summary of the generalized rule prescribed in that study for determining the expected dimensions in multivariate data is presented in order to highlight (in Remark 1) the main drawbacks on the general use of the KMO. The underlying concepts of the KMO are orders zero and one correlation coefficients. This section examines related concepts that border on the construction of homogenous sets that yields well-defined dimensions from a given set of variables.

A Generalized Rule for Determining Expected Dimensions

Suppose a multivariate dataset is generated on a set of p variables $\mathbf{X}=(X_1, X_2, \ldots, X_p)$ with a given correlation profile. On the basis of the level of correlation coefficients, a cut-off value of τ is fixed for which variables may be considered to belong together if their pair-wise

correlation coefficient exceeds 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 labelled as the set $S_1 = (x_u, x_v)$ and the index set $I_1 = (u, v)$. 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 some $i \in I \setminus I_1$, then we obtain a final first homogeneous set $S_1 = \{x_{i_1}, x_{i_2}, \ldots, x_{i_{g1}}\}$ with index set $I_1 = \{i_1, i_2, \ldots, i_{g1}\} \subset I$.

We 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, we obtain $S_q = \{x_{d_1}, x_{d_2}\}$, and the index set $I_q = \{d_1, d_2\}$. Now, if $r_{x_k, x_i} < \tau$, for some $k \in I_q$ and some $i \in I \setminus I_q$ then we obtain the final qth homogeneous set $S_q = \{x_{d_1}, x_{d_2}, \ldots, x_{d_{gq}}\}$ with index set $I_q = \{d_1, d_2, \ldots, d_{gq}\} \subset I$.

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

Remark 1

The procedure outlined is influenced by a subjective experimenter-specific fixed cut-off value, τ by which variables may be considered to belong together if their pair-wise correlation coefficient exceeds it. The problem of subjectivity would be resolved by generating a data-specific threshold δ_0 that serves as a cut-off value.

There are some variables that are likely to contaminate the measure of homogeneity in the data. These are elements identified in S_{l+1} with indexed set T_l and are not found in any of the homogeneous sets. The new procedure would screen the variables to include only those that are identified with a particular homogeneous group.

CONFIRMATORY FACTOR ANALYSIS

The test of adequacy of m-factor model for data on p indicator variables is equivalent to the test of the hypothesis

$$H_o: \underset{p\times p}{\boldsymbol{\rho}} = \underset{p\times m}{\boldsymbol{\Lambda}}\ \underset{m\times p}{\boldsymbol{\Lambda}'} + \underset{p\times p}{\boldsymbol{\Psi}} \quad \text{against} \quad H_a: \underset{p\times p}{\boldsymbol{\rho}} \neq \underset{p\times m}{\boldsymbol{\Lambda}}\ \underset{m\times p}{\boldsymbol{\Lambda}'} + \underset{p\times p}{\boldsymbol{\Psi}} \tag{2.1}$$

The matrices $\boldsymbol{\Lambda\Lambda}'$ and $\boldsymbol{\Psi}$ are respectively, the reproduced correlation matrix based on the model and diagonal matrix of specific variances. The null hypothesis means that the m factors are adequate in approximating the original correlation matrix. If H_o is rejected, it means that the factor model does not significantly represent the underlying dimensions of the correlation

matrix. Under H_o, the maximum of the likelihood function is obtained with $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ and $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Psi}}$, where the estimates $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\Psi}}$ are given in standards texts (e.g., Johnson & Wichern, 2014; Anderson, 2003). The hypothesis H_o is thus rejected at α level of significance if

$$\left[n-1-\tfrac{1}{6}(2p+4m+5)\right]\ln\frac{\left|\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'+\hat{\boldsymbol{\Psi}}\right|}{|\mathbf{S}_n|} > \chi^2_{\frac{1}{2}[(p-m)^2-p-m]}(\alpha) \tag{2.2}$$

provided n and $(n-p)$ are large, and $\mathbf{S}_n$ is the sample variance-covariance matrix.

Remark 2

In this study, it will be observed that, for the same data, several factor models may be significant (i.e., H_o will not be rejected) for a number of values of *m*. The proposed algorithm helps in this case, to decide on the optimal value of *m* that gives the most plausible factor solution.

THE KAISER-MEIER-OLKIN'S MEASURE OF SAMPLING ADEQUACY

The Kaiser-Meier-Olkin's Measure of Sampling Adequacy (KMO) is a detection metric for determining the degree to which the indicators of a dimension are homogeneous in a multivariate data. KMO value within the interval [0.6, 1.0] is a good measure (Rencher, 2002; Kaiser, 1974). The guide for interpreting KMO measure (Kaiser, 1974) is well-known. In this study, the KMO index is stated as

$$KMO = \frac{1}{1+\sum_{i<j} pr_{ij}^2 \Big/ \sum_{i<j} r_{ij}^2}, \tag{2.3}$$

where r_{ij}^2 is the square of the observed correlation coefficient (OCC) between any pair of variables (x_i, x_j) and pr_{ij}^2 is the corresponding partial correlation coefficient (PCC). In Equation (2.3), the proposed methodology for computing the KMO is based on a comparison of the size of the OCC and PCC.

ORDER STATISTICS CORRELATION COEFFICIENT

Let (x_{ij}, x_{kj}), $j = 1, 2, \ldots, n$; $i, k = 1, 2, \ldots, p$ be *n* observations on any two variables from the set $\mathbf{X}$. By rearranging pair-wisely the observations on the two variables with respect to the magnitudes of x_i, we obtain two new sets of data $(x_{i(j)}, x_{k[j]})$ where $x_{i(1)} \le x_{i(2)} \le \cdots x_{i(n)}$ are the order statistics of x_i and $x_{k[1]}, x_{k[2]}, \ldots, x_{k[n]}$ are the associated concomitants of x_k. The order statistics correlation coefficient may be defined for (x, y) as

$$r_x(x,y) = \frac{\sum_{i=1}^{N}[x_{(i)} - x_{(N-i+1)}y_{[i]}]}{\sum_{i=1}^{N}[x_{(i)} - x_{(N-i+1)}y_{(i)}]}, \tag{2.4}$$

and has the usual basic properties of a correlation coefficient.

3 Development and Implementation of Dimensionality Detection Algorithms

The proposed dimensionality detection methods are similarity measures which hinge on correlation profiles. The study employs the Pearson's correlation and Order statistic profiles. In the implementation, attention is focused on identifying 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

SIMILARITY BASED DIMENSIONALITY DETECTOR

Consider *n* observations made on a *p*-variate random variable, $Y = Y_1, Y_2, \ldots, Y_P$. Let C_Y denote a $p \times p$ matrix of pairwise similarity measures based on Y. The number of dimensions underlying the data is at most the number of variables defining the data. It is also possible that the variables are inter-related in some simple or complex sense. This relationship may be informative about the intrinsic dimension underlying the data and hence, an appealing basis for building a dimensionality detection scheme for detecting dimensions in such data. The similarity-based algorithm for detecting dimension is outlined in the given algorithm, with the following conversion for notation: $N(x)$ denotes the number of variables in X. and $Y \setminus Y_i$ denotes the remaining variables without Y_i, $i = 1, 2, \ldots, p$.

ALGORITHM

Initialization: Data: $Y = Y_1, Y_2, \ldots, Y_P$. Set threshold, $\delta = \delta_0$

Compute similarity matrix, $C_Y = \nu(Y) = \nu(Y_1, Y_2, \ldots, Y_p)$.

Compute lower triangular matrix of C_Y, D_Y

Compute fundamental spanning set, $S_f = \{(Y_i, Y_J) : D = \max(D_Y),\ i \neq j\}$.

Set $m_f = N(S_f)$, $\kappa = 0$.

Compute reduced dataset,

$Y^* = Y \setminus S_f = Y \setminus (Y_i, Y_j),\ i \neq j$

Set $n^* = p - m_f$, $H_s = S_f$ and $H_{ns} = NULL$

Do while $n^* > 2$

1. $D_{Yk,S_f} = D_{Yk,Yi},\ D_{Yk,Yj},\ Y_k \in Y^*$
2. **if** $D_{Yk,S_f} \geq \delta_0$
   - $S_f = \{(Y_i, Y_j, Y_k)\}$
   - $m_f = N(S_f)$
   - $H_s = S_f$
   - $Y^* = Y^* \setminus Y_k$

- $n^* = p - m_f$

3. else

- $H_{ns} = Y_k$
- $Y^* = Y^* \setminus Y_k$
- $n^* = N(Y^*)$

4. Go to step 1 . Otherwise return H_s, H_{ns}, C_Y

AUTOMATED THRESHOLD SETTINGS

The use of threshold is primal in dimensionality detection since the generation as well as the detection of homogeneous sets from a given multivariate dataset is threshold driven. It is important to note that not all thresholds will yield homogeneous set. Also, it is likely that a single threshold may generate multiple homogeneous sets. Three automated threshold setting algorithm procedures are specified as follows:

$\delta_1 = [a_1, a_1 + \alpha_1, a_1 + 2\alpha_1, \ldots, a_n]$
$\delta_2 = [a_1, a_1 + \alpha_2, a_1 + 2\alpha_2, \ldots, a_n]$
$\delta_3 = [\delta_1 \geq \widetilde{\delta_1}]$
$\alpha_2 = \frac{a_n - a_1}{k\delta}$; where $\alpha_1 = 0.01$, $\alpha_2 = \frac{a_n - a_1}{k\delta}$ $a_1 = \min(D_Y)$
$a_n = \max(D_Y)$
$\widetilde{\delta_1} = median\ of\ \delta_1$
We set $k\delta = 12$

AUTOMATED THRESHOLD SETTING 1 (δ_1)

The algorithm picks the lowest pairwise correlation, generates series of thresholds using a step value of 0.01 until all values in the correlation matrix are accommodated. The algorithm is then used to generate homogeneous sets for each of the thresholds. Since multidimensionality is expected, some thresholds could yield more than one homogeneous set. The KMO values are then calculated for each homogeneous set for each threshold.

AUTOMATED THRESHOLD SETTING TWO (δ_2)

The correlation profile used is the Pearson's correlation which is normally distributed. Statistically, majority (about 99.7%) of the data points lie 3 standard deviations about the mean. This gives 6 standard deviations; we add an allowance of 2 standard deviations to cater for the rest of the data points. The algorithm then uses a step value of the ratio of the range for the correlation matrix to the resultant standard deviation to generate series of thresholds. The dimensionality detection algorithm is then used to generate homogeneous sets for each of the thresholds. Since multidimensionality is expected some thresholds could yield more than one

homogeneous set. The KMO values are then calculated for each homogeneous set that corresponds to each threshold.

AUTOMATED THRESHOLD SETTING THREE (δ_3)

This procedure is based on Threshold Setting 1. Statistically the correlation matrix used which hinges on Pearson's correlation is symmetric. The algorithm determines the median for thresholds generated using automated Threshold Setting I and selects those thresholds that are at least equal to the median. This is similar to the usage of the lower triangular matrix of the correlation matrix. For each of these thresholds, homogeneous sets are then generated along with the respective KMO values.

Implementation I

Figure 3.1 shows the plots of KMOs against the various thresholds generated by the algorithms for Dataset 1 (Johnson & Wichern, 2014; Anderson, 2003; Nkansah, 2018) that concerns performance of sales personnel employees of a marketing company. For Threshold Setting I, for example, the threshold values are 0.15(0.01)0.94.

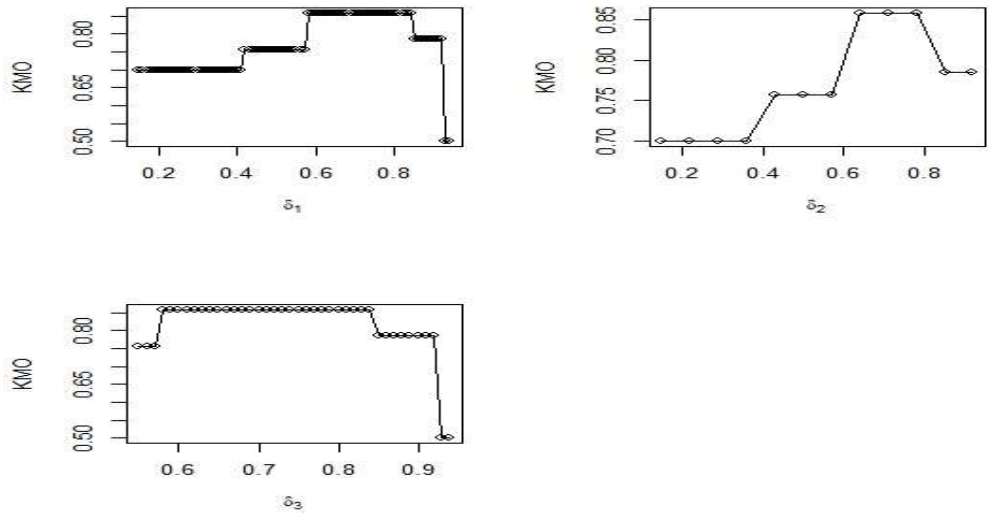

FIGURE 3.1: Plot of implementation FOR DATASET 1

Although there are a couple of saturation points, the interest is in the one that corresponds to the highest KMO. It could be observed that the highest saturation points for all three graphs corresponding to the highest KMO lie within [0.6, 0.85]. This means that any threshold within this range could be the optimal threshold for dimensionality detection for this dataset.

For each threshold, we obtain the number of homogeneous sets along with corresponding KMO values and the number of variables in each set. Table 3.1 shows some specific thresholds with corresponding KMO values and the number of homogeneous sets, in addition to the number of variables in each set, generated for the respective threshold value based on Threshold Setting I.

Table 3.1: Dimensionality detection for Dataset 1

SN/ Threshold	No. of hom. sets	No. of var in hom. set	KMO
[1] 0.15	1	6	0.6995
[2] 0.16	1	6	0.6995
[27] 0.41	1	6	0.6995
[28] 0.42	1	5	0.7571
[36] 0.50	1	5	0.7571
[43] 0.57	1	5	0.7571
[44] 0.58	2	4, 2	0.8587 0.50000
[70] 0.84	2	4, 2	0.8587 0.50000
[71] 0.85	2	3, 2	0.7848 0.50000
[78] 0.92	2	3, 2	0.7848 0.50000
[79] 0.93	3	2, 2, 2	0.50 0.50 0.50

The table shows that the highest number of homogenous set given by any threshold is 3, an indication that the dimensionality cannot exceed 3, if it exists. However, dimensionality does not exist in this data since no threshold yields a unique highest KMO value. Consistent with this result, the highest number of variables (6) for a homogenous set does not generate the highest KMO. The claim of a lack of dimensionality in the data is further buttressed with CFA.

CONFIRMATORY TEST OF MODEL ADEQUACY FOR DATASET 1

Since the highest expected number of dimensions for the data does not exceed three, the CFA is therefore carried out for a maximum of three factor solutions in Dataset 1. Thus, we test the adequacy of one, two and three factor models equivalent to one, two and three dimensions. The result of the test is given in Table **3.2.** As indicated earlier, each threshold within the highest saturation point could yield different factor solutions.

Table 3.2: Significance test of factor models for Dataset 1

Model	Chi-Square	Df	Sig.
1	162.715	14	0.000
2	117.114	8	0.000
3	61.651	3	0.000

In Table 3.2, no specific factor solution is seen to fit the data due to the small *p*-values on the basis of the hypotheses stated in Equation (2.1). This confirms that there is no unique

homogeneous set that has the highest number of variables and hence no unique highest KMO. It implies that this data may not be practically suitable for factor extraction.

IMPLEMENTATION II

Figure 3.2 shows the plots of KMOs against the various thresholds generated by the algorithm for Dataset 2 (Nkansah, 2018) that concerns performance of high school students in nine subjects. For this data, only Threshold Setting III converges as a result of challenges of negative correlation coefficients. The threshold values obtained are 0.38(0.1)0.75.

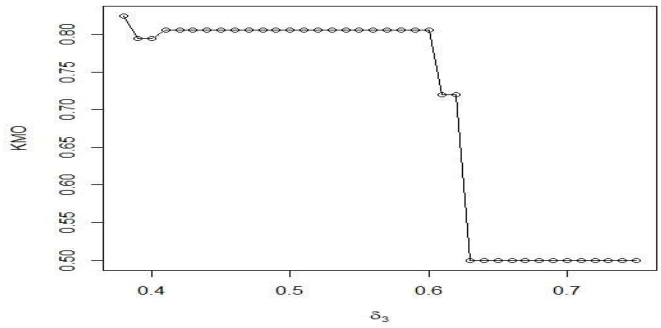

FIGURE 3.2: PLOT OF IMPLEMENTATION FOR DATASET 2

Figure 3.2 shows that there is a saturation point between [0.42, 0.63]. However, the range does not contain the highest KMO (0.8242). It could therefore be observed that a threshold of 0.38 is the unique optimal threshold that corresponds to the highest KMO. There is therefore a clear dimensionality in this dataset. Table 3.3 displays some specific thresholds with corresponding KMO values and the number of homogeneous sets, in addition to the number of variables in each set, generated for the respective threshold value based on Threshold Setting I. Consistent with this result, the highest number of variables (6) for a homogenous set given by a unique threshold (0.38) generates the highest KMO. The table also shows that the highest number of homogenous sets given by the thresholds is 4, an indication that there are four possible

Table 3.3: Dimensionality detection for Dataset 2

SN/ Threshold	No. of hom. sets	No. of var in hom. set	KMO
[1] 0.38	**2**	**6, 2**	**0.8242 0.5000**
[2] 0.39	2	5, 2	0.7942 0.5000
[3] 0.40	2	5, 2	0.7942 0.5000
[4] 0.41	3	4, 2, 2	0.8058 0.5000 0.5000
[13] 0.50	3	4, 2, 2	0.8058 0.5000 0.5000
[23] 0.60	3	4, 2, 2	0.8058 0.5000 0.5000
[24] 0.61	3	3, 2, 2	0.7202 0.5000 0.5000
[25] 0.62	3	3, 2, 2	0.7202 0.5000 0.5000

[26] 0.63	4	2, 2, 2, 2	0.50	0.50	0.50	0.50
[38] 0.75	4	2, 2, 2, 2	0.50	0.50	0.50	0.50

dimensions that underlie the data. The claim of an existing dimensionality in the data is further buttressed with CFA.

CONFIRMATORY TEST OF MODEL ADEQUACY FOR DATASET 2

Since the highest expected number of dimensions does not exceed four, the CFA is carried out for a maximum of four factor solutions in Dataset 2. Thus, we test the adequacy of one to four factor models equivalent to one to four dimensions. As indicated earlier, if a dataset generates an optimal threshold for dimensionality detection, then this threshold should automatically yield an optimal factor solution.

Table 3.4: Significance test of factor models for Dataset 2

Model	Chi-Square	Df	Sig.
1	41.949	27	0.033
2	18.505	19	0.489
3	10.144	12	0.603
4	2.584	6	0.859

From the table, Model 2 is the least-fitting factor solution since the *p*-value begins to get greater than 0.05 with a two-factor solution model. It also shows that factor solutions containing two factors or more are all suitable. However, since the unique threshold of 0.38 with the highest KMO identifies two homogenous sets, the 2-factor solution is the best.

IMPLEMENTATION OF ORDER STATISTICS ALGORITHM PROCEDURE

The algorithm generates the correlation matrix for the set of p variables, $X_1, X_2, \ldots, X_p$ and returns the order statistics for the variables $X_{(1)}, X_{(2)}, \ldots, X_{(p)}$. For the ordered variables, the correlation matrix is then generated for the application of the dimensionality detection algorithm. The implementation is carried using Dataset 1 and yields threshold values 0.31(0.01)0.94 for Threshold Setting I, for example.

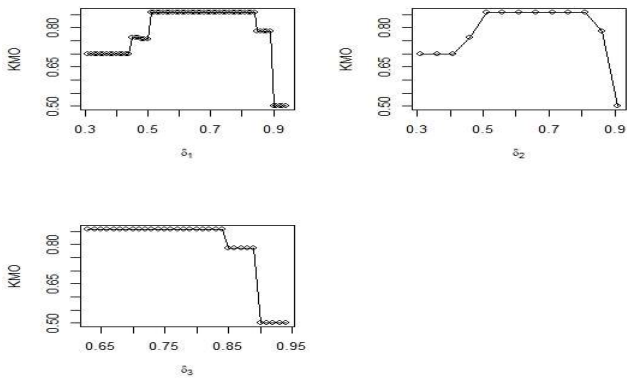

FIGURE 3.3: Plot of implementation FOR DATASET 1 BASED ON ORDER STATISTIC PROFILE

Figure 3.3 shows a plot of KMOs against the corresponding thresholds. It could be observed that there is no unique threshold, and that the highest saturation point for all three graphs lie within [0.6, 0.85]. Thus, any threshold within this range could be optimal for dimensionality detection for this dataset.

Table 3.5 shows specific thresholds with corresponding KMO values and the number of homogeneous sets generated for the respective threshold value based on Threshold Setting I.

Table 3.5: Dimensionality detection for Dataset 1

SN/ Threshold	No. of hom. sets	KMO
[1] 0.31	1	0.6995
[2] 0.32	1	0.6995
[14] 0.44	1	0.6995
[20] 0.50	1	0.7571
[21] 0.51	2	0.8587 0.5000
[54] 0.84	2	0.8587 0.5000
[57] 0.89	2	0.7848 0.5000
[64] 0.94	3	0.50 0.50 0.50

Similar results are obtained for Pearson's correlation. It is also observed that the same number of variables are obtained in each homogenous set for each cut-off for both the Pearson's correlation and that of order statistics. These results further confirm that there is no dimensionality in this dataset. The lack of a threshold value with unique highest KMO buttresses the other methods that the data lacks dimensionality.

4 Summary, Conclusion and Recommendation

Usually, a rather covert threshold has hitherto been used for dimensionality detection which may lead to misleading results. The study has presented automated threshold settings using an algorithm that generates a data-specific optimal threshold from the data structure for detecting the exact dimensionality of multivariate data for more accurate results. The proposed method would serve as the basis for the application of the well-known statistical tools which may now be useful for purposes of interpretation of the extracted reduced dimensions.

The result has shown that though a dataset could be identified to have dimensionality by exploratory methods, there may practically not be any underlying dimensionality. It is also clearly demonstrated that dimensionality detection is threshold sensitive. It is therefore reasonable to allow the data structure to generate its own optimal threshold suitable for determining its dimensions.

It is acknowledged that studies on the subject could be sensitive to the likely presence of extreme values. The study could therefore be extended to take care of these extremes in the data, an approach that is expected to further save computational time. Further threshold settings may also be examined to meet the challenge posed by negative correlations.

References

Anderson, T.W. (2003). *Introduction to Multivariate Statistical Analysis,* New Jersey: Prentice Hall

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

Johnson, R. A., & Wichern, D. W. (2014). *Applied Multivariate Statistical Analysis* (6th ed.). NJ: Pearson

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

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison Wesley.

Nkansah, B. K. (2018). On the Kaiser-Meier-Olkin's Measure of Sampling Adequacy. *Journal of Mathematical Theory and Modelling*, 8(7), 52-76

Nkansah, B. K. , Zakaria, A., & Howard, N. K. (2019). Effect of Measurement Scales on Results of Item Response Theory Models and Multivariate Techniques. *Journal of Informatics and Mathematical Sciences*, 11(1), 51-79

Rutledge D. N., Roger J-M, & Lesnoff, M. (2021). Different Methods for Determining the Dimensionality of Multivariate Models. *Front. Anal. Sci.* 1:754447. doi: 10.3389/frans.2021.754447

Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics* (5th ed.). Boston, Pearson

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

Zhang, M. (2016), Exploring dimensionality of scores for mixed-format tests, Unpublished Doctorial Thesis, University of Iowa

Zhang, J., & Stout, W. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. Psychometrika, 64, 129–152.

Zupluoglu, C. (2013), Assessed Dimensionality of Latent Structures Underlying Dichotomous Item Response Data with Imperfect Models, Unpublished Doctorial Thesis, University of Minnesota