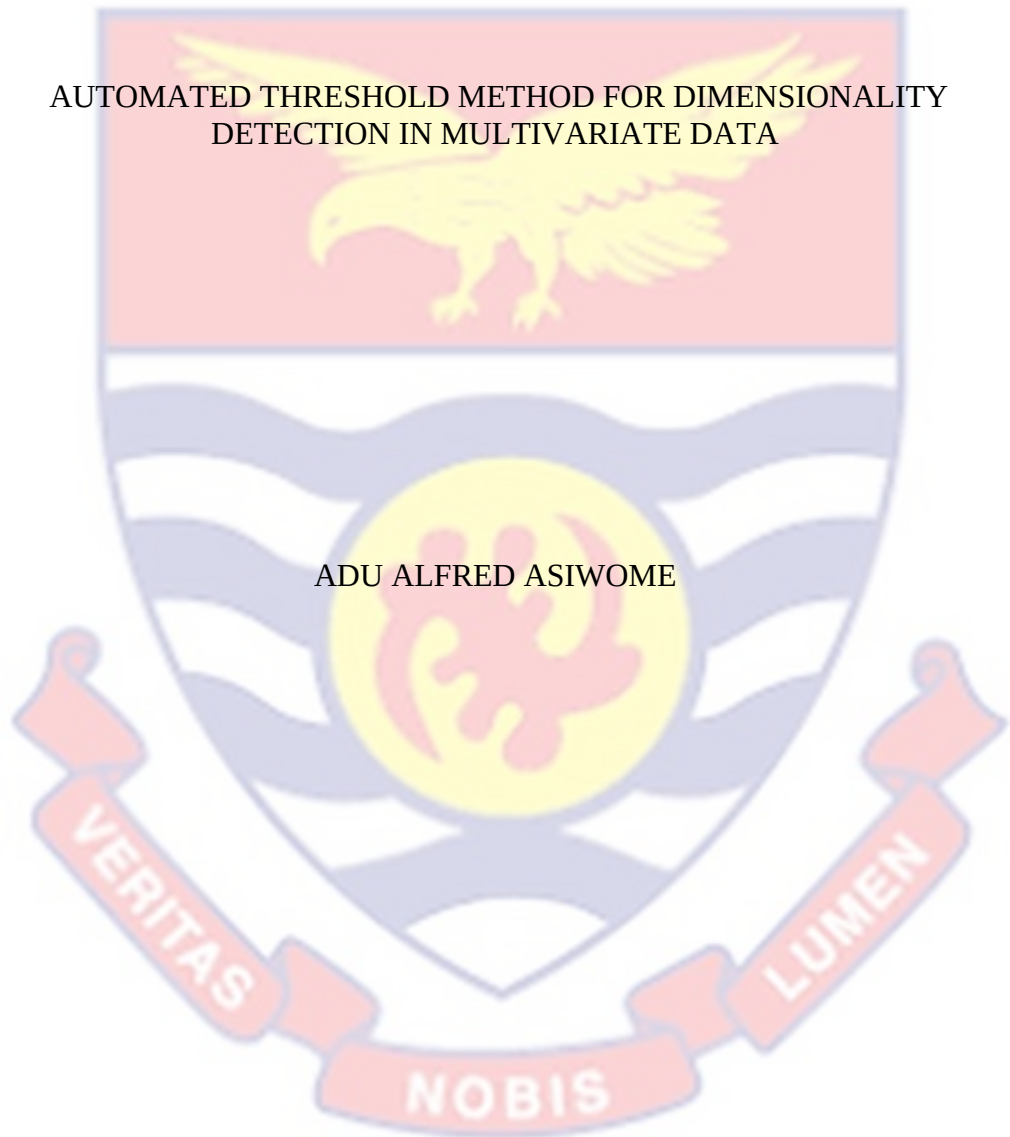


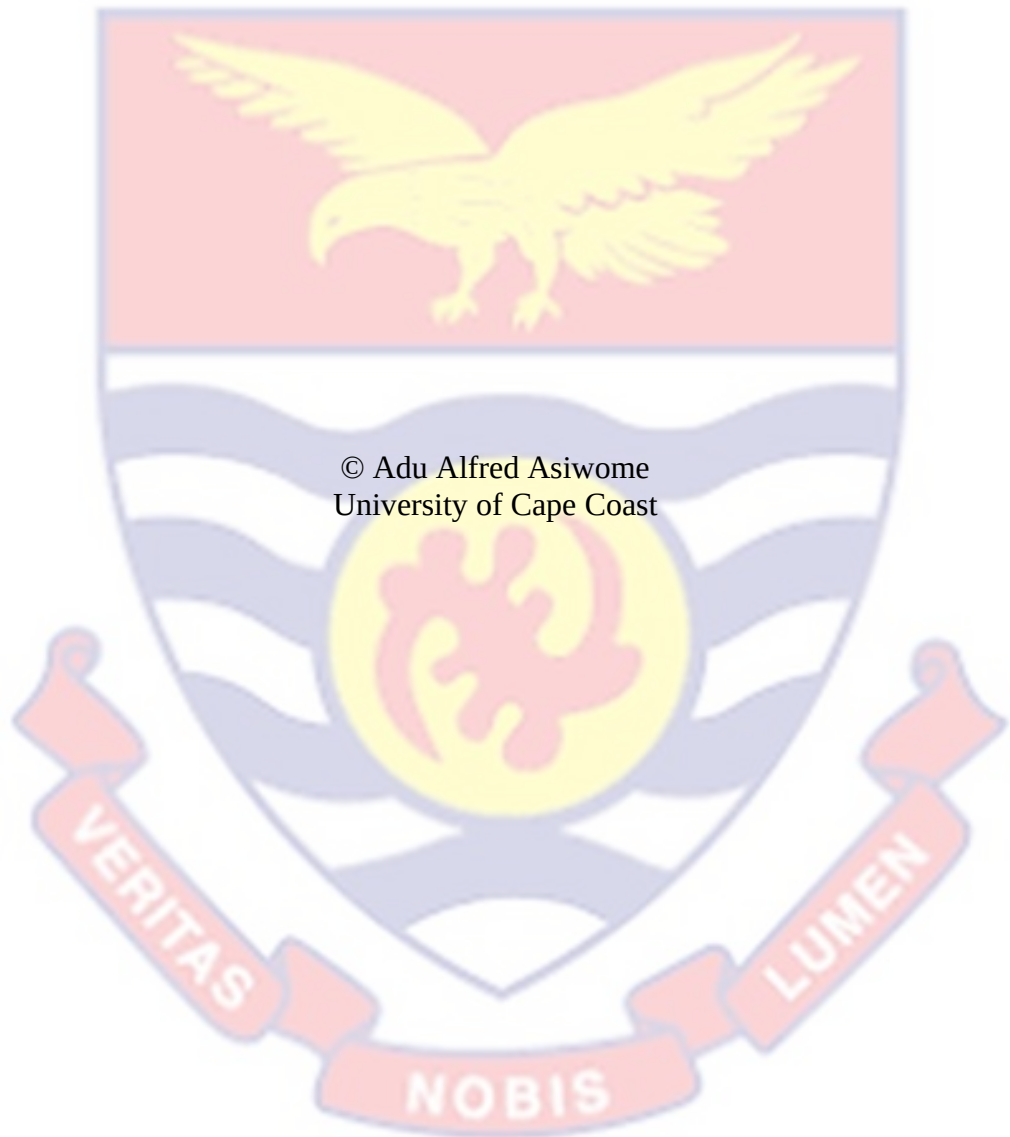
UNIVERSITY OF CAPE COAST

AUTOMATED THRESHOLD METHOD FOR DIMENSIONALITY
DETECTION IN MULTIVARIATE DATA

ADU ALFRED ASIWOME



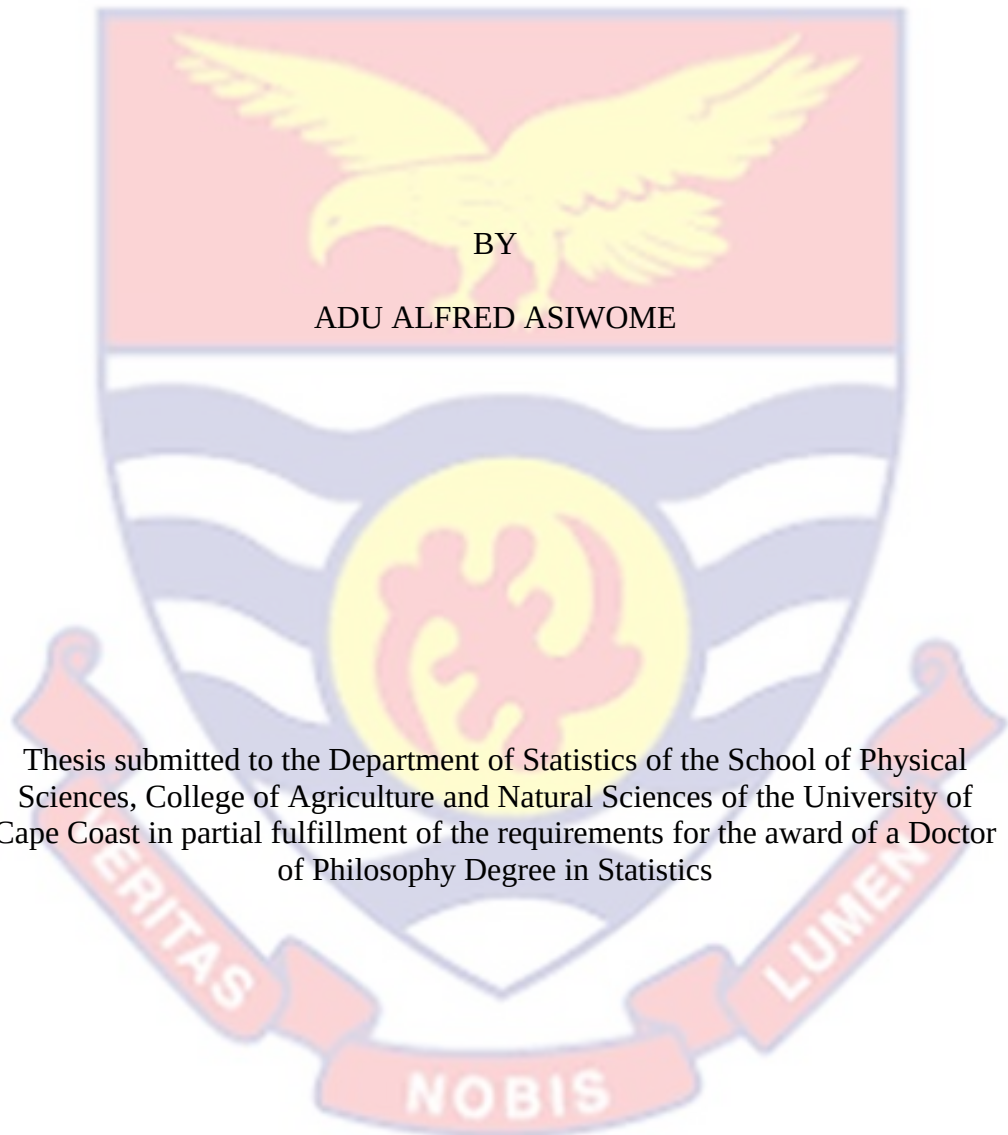
2022



© Adu Alfred Asiwome
University of Cape Coast

UNIVERSITY OF CAPE COAST

AUTOMATED THRESHOLD METHOD FOR DIMENSIONALITY
DETECTION IN MULTIVARIATE DATA



Thesis submitted to the Department of Statistics of the School of Physical Sciences, College of Agriculture and Natural Sciences of the University of Cape Coast in partial fulfillment of the requirements for the award of a Doctor of Philosophy Degree in Statistics

NOVEMBER 2022

DECLARATION

Candidate's Declaration

I hereby declare that this thesis is the result of my own original research and that no part of it has been presented for another degree in this university or elsewhere.

Candidate's Signature..... Date.....

Name: Adu Alfred Asiwome

Supervisors' Declaration

We hereby declare that the preparation of the thesis was supervised in accordance with the guidelines laid down by the University of Cape Coast.

Principal Supervisor's Signature Date.....

Name: Prof. Bismark Kwao Nkansah

Co- Supervisor's Signature Date.....

Name: Dr. David Mensah

ABSTRACT

Multivariate methods such as principal component analysis and factor analysis have been used to interpret multivariate data. However, these statistical applications are not able to determine prior to their application whether a dimension exists within the multivariate data set since it is possible to have a dimensionless multivariate dataset. In addition, these statistical applications are method dependent, it is therefore imperative to propose an independent technique for detecting dimensionality using automated threshold settings which are thresholds generated based on the structure of the data and not by the judgement of the researcher so that these statistical applications will be for purposes of interpretation or giving meaning to the data structure. Also, the formation of dimensionality in the well-known multivariate techniques is not analytically or computationally presented. They therefore offer a leave-or-take result with no understanding of the formation of the dimensions. This study therefore filled this gap by successfully proposing an independent dimensionality detection method using three automated threshold settings that generate data specific thresholds by allowing the data structure to generate the optimal threshold for detecting dimensionality of the multivariate data set for more accurate results. The study also established the robustness of the method using Pearson's correlation which hinges on the mean and another correlation profile that does not hinge on a statistic which is affected by extreme values, in this case order statistic which hinges on the median. The algorithm converged in all cases. Confirmatory factor analysis are carried out for confirmation of results. The proposed method completely removes the challenge of subjectivity associated with dimensionality detection, and hence is highly recommended.

KEY WORDS

Dimensionality

Homogenous Sets

Multidimensionlity

Non Homogenous Sets

Similarity Detector

Thresholds

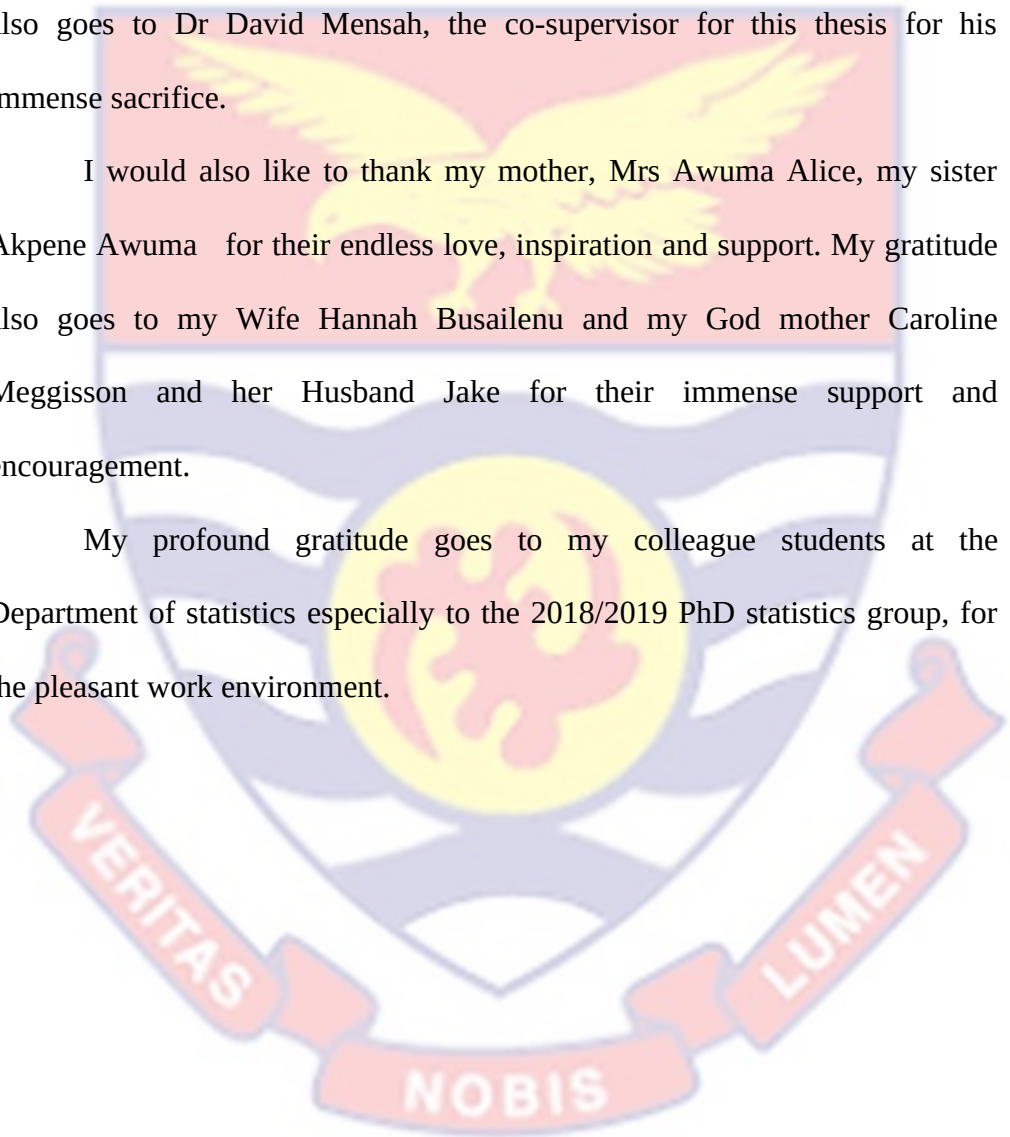


ACKNOWLEDGEMENTS

The work presented in this thesis could never be accomplished without the support of many people. Foremost, I would like to express my sincere gratitude to my main supervisor Prof. Bismark Kwao Nkansah for his invaluable guidance and support in all aspects of my thesis work. My gratitude also goes to Dr David Mensah, the co-supervisor for this thesis for his immense sacrifice.

I would also like to thank my mother, Mrs Awuma Alice, my sister Akpene Awuma for their endless love, inspiration and support. My gratitude also goes to my Wife Hannah Busailenu and my God mother Caroline Meggisson and her Husband Jake for their immense support and encouragement.

My profound gratitude goes to my colleague students at the Department of statistics especially to the 2018/2019 PhD statistics group, for the pleasant work environment.



DEDICATION

To my wife, children and entire family for their support.



TABLE OF CONTENTS

	Page
DECLARATION	ii
ABSTRACT	iii
KEY WORDS	iv
ACKNOWLEDGEMENTS	v
DEDICATION	vi
LIST OF TABLES	xi
LIST OF FIGURES	xii
LIST OF ABBREVIATIONS	xiii
CHAPTER ONE: INTRODUCTION	
Background to the Study	1
Statement of the Problem	6
Purpose of the Study	7
Objectives of the Study	8
Significance of the Study	8
Delimitation	9
Limitation of the Study	11
Definition of Terms	11
Organisation of the Study	12
CHAPTER TWO: LITERATURE REVIEW	
Introduction	14
Review of Studies on Dimensionality	14
Measurement Scale and Sample Size for Determination of Dimensionality	37
Exploratory versus Confirmatory Assessment	37

Review of Dimensionality Assessment Methods	39
Considerations That Could Influence Results of Factor Analysis	39
Illustrative Dataset and Cut-off Values in Dimensionality Detection	42
Chapter Summary	44
CHAPTER THREE: RESEARCH METHODS	
Introduction	45
Generalized Rule for Determining Expected Dimensions of Datasets	45
Dimensionality of a Dataset	48
Confirmatory Factor Analysis	49
Item Response Theory	52
Dimensionality in IRT	53
Generalized Partial Credit Model (GPCM)	54
Essential Dimensionality	56
Methods Using the IRT Definition of Dimensionality	56
Proposed Method on the Kaiser-Meier-Olkin's Measure of Sampling	58
Order Statistics Correlation Coefficient	60
Definition and Properties	60
Chapter Summary	61
CHAPTER FOUR: RESULTS AND DISCUSSION	
Introduction	62
Kaiser-Meyer-Olkin Measure of Sampling Adequacy based on	65
An Automated Dimensionality Detection	66
Generalization of the Modified Rule for Determining Expected Dimensions	67
Similarity Based Dimensionality Detector	68
Algorithm 2: Similarity-based Dimensionality Detector	69

Explanation of Dimensionality Detection Algorithm Procedure	70
Automated Threshold Setting	70
Mathematical Background	71
Automated Threshold Setting Algorithm Procedures	71
Automated Threshold Setting 1 ($\delta 1$)	71
Automated Threshold Setting Two ($\delta 2$)	72
Automated Threshold Setting Three ($\delta 3$)	72
Selecting an Optimal Threshold from the Saturation Point	78
Confirmatory Test of Model Adequacy for Dataset 1	84
Threshold Selection for Optimal Factor Solution	88
Confirmatory Test of Model Adequacy for Dataset 2	93
Dimensionality Detection Based on Order Statistics	94
Order Statistics Algorithm Procedure	94
Robustness of Method using a Reduced Dataset	98
Original Data Layout	98
Implementation (for simulated data)	112
Discussion	118
Chapter Summary	119
CHAPTER FIVE: SUMMARY, CONCLUSIONS AND RECOMMENDATIONS	
Summary	125
Conclusions	130
Recommendations	131
REFERENCES	132
APPENDICES	163

APPENDIX A: DIMENSIONALITY DETECTION CODES – PEARSON’S CORRELATION APPROACH	163
APPENDIX B: CODES FOR SIMULATING DATA BASED ON ITEM RESPONSE THEORY	242



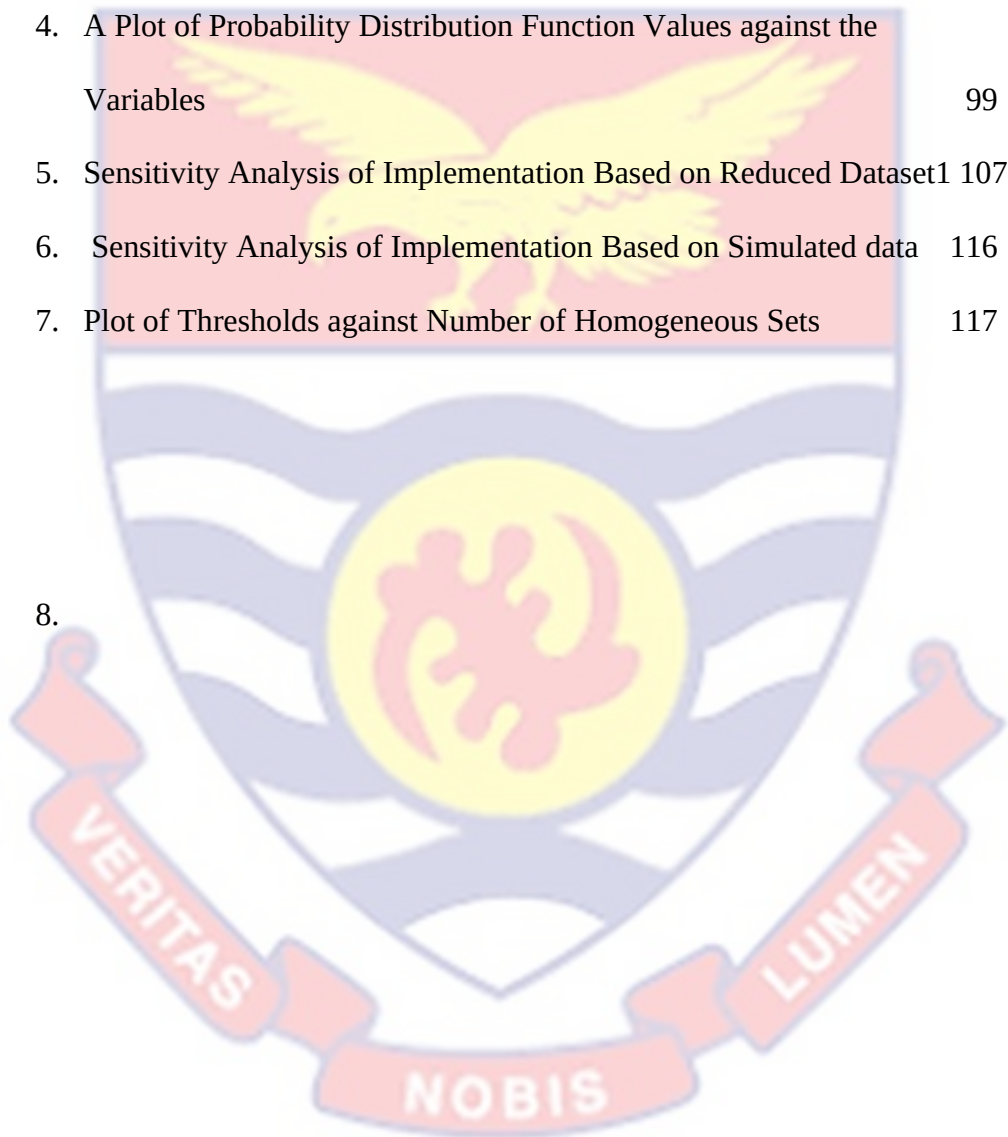
LIST OF TABLES

	Page
1. A Guide for Interpreting KMO Measure	59
2. Correlation Matrix for Dataset 1	62
3. KMO for Dataset 1 Based on Subgroupings	63
4. Correlation Matrix for Dataset 2	63
5. KMO Values for Dataset 2 from Literature	64
6. Dimensionality Detection for Dataset 1	73
7. Threshold Selection for Optimal Factor Solution	80
8. Significance test of Factor Solutions for Dataset 1	84
9. Dimensionality Detection for Dataset 2	85
10. Threshold Selection for Optimal Factor Solution Based on Dataset 2	88
11. Significance Test of Factor Solutions for Dataset 2	93
12. Order Statistic Implementation	95
13. Original Data Layout on p Variables	98
14. Data Layout of Extracted Feature $T(y)$	101
15. Correlation Matrix for Reduced Dataset 1	102
16. Dimensionality Detection for Reduced Dataset 1	103
17. Summary of Thresholds and Highest KMOs for Full and	108
18. Dimensionality Detection for Simulated data	113

LIST OF FIGURES

	Page
1. Sensitivity Analysis for Dataset 1	78
2. Sensitivity Analysis of Implementation Based on Dataset 2	87
3. Sensitivity Analysis for Dataset 1 Based on Order Statistic profile	97
4. A Plot of Probability Distribution Function Values against the Variables	99
5. Sensitivity Analysis of Implementation Based on Reduced Dataset1	107
6. Sensitivity Analysis of Implementation Based on Simulated data	116
7. Plot of Thresholds against Number of Homogeneous Sets	117

8.

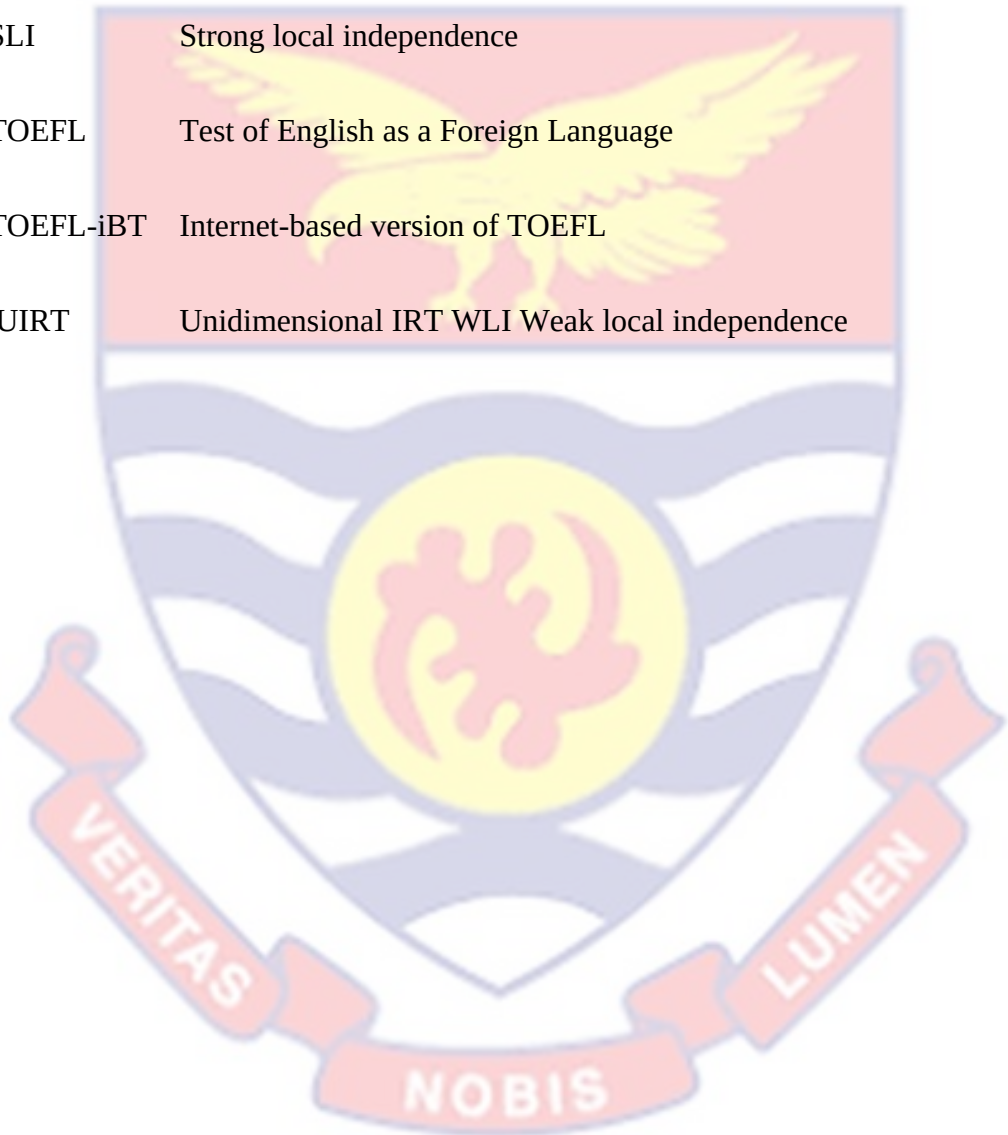


LIST OF ABBREVIATIONS

KMO	Kaiser Meir Olkins Measure of Sampling Adequacy
1PL	One-parameter logistic model in IRT
2PL	Two-parameter logistic model in IRT
3PL	Three-parameter logistic model in IRT
AP	College Board Advanced Placement Exams
CH	The Caliński and Harabasz (1974) index
CINEG	Common-item nonequivalent groups design in equating
CFA	Confirmatory factor analysis
CR	Constructed-response item format
DH	The Duda and Hart (1973) index
EFA	Exploratory factor analysis ES Common-item effect size
GR	The graded-response model in IRT
IRF	Item response function in IRT
IRT	Item response theory
LS	Least square estimation method
M3PL	Multidimensional extension of the 3PL model in IRT
MAP	Minimum average partial procedure
MAP-R2	MAP with Pearson partial correlations squared
MAP-R4	MAP with Pearson partial correlations to the fourth power

MAP-P2	MAP with polychoric partial correlations squared
MAP-P4	MAP with polychoric partial correlations to the fourth power
MC	Multiple-choice item format
MDIFF	Multidimensional item difficulty parameter in IRT
MDISC	Multidimensional item discrimination parameter in IRT
MGR	Multidimensional extension of the GR model in IRT
MIRT	Multidimensional IRT
ML	Maximum likelihood estimation method
O1	Assessment outcome 1: whether unidimensionality holds
O2	Assessment outcome 2: the number of dimensions
O3	Assessment outcome 3: specific dimensional structure
O4	Assessment outcome 4: explanations of dimensions
PA	Parallel analysis
PA-MCm	PA based on Monte Carlo simulation with a mean threshold
PA-MC95	PA based on Monte Carlo simulation with a 95th-percentile threshold
PA-Pm	PA based on permutations with a mean threshold
PA-P95	PA based on permutations with a 95th-percentile threshold
PCA	Principal component analysis
PD	Positive definite matrix

PFA	Principal factor analysis
RMSEA	Root mean square error of approximation calculated by Mplus
RMSR	Root mean square residual calculated by Mplus
SEM	Structural equation model
SLI	Strong local independence
TOEFL	Test of English as a Foreign Language
TOEFL-iBT	Internet-based version of TOEFL
UIRT	Unidimensional IRT WLI Weak local independence



CHAPTER ONE

INTRODUCTION

This chapter is made up of the background to the study, the statement of the problem, objectives of the study, outline of the thesis as well the limitation of the thesis.

Background to the Study

The dimensionality of a dataset is mostly defined 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 underlying the data.

A number of statistical applications come in handy to determine the number of dimensions underlying a multivariate dataset. However these techniques are not designed as preliminary techniques for dimensionality detection which is required before the application of these statistical methods for purposes of interpretation.

There is only one attempt (Nkansah, 2018) at determining an independent technique for dimensionality detection. However, the procedure is quite subjective as a key element of threshold setting is experimenter specific. It may be necessary therefore to review the known statistical applications and also attempt to propose an objective dimensionality detection method that can be used to determine whether a dimension exists within a multivariate dataset.

The idea that an instrument's test items all measure the same thing is one of the most important assumptions in measurement theory. The underlying latent variable of a composite score must be unidimensional in order to make

psychological sense whether sorting people on an attribute, characterizing individual differences, or grouping them by ability (Hattie, 1985).

One of the most significant goals of evaluating unidimensionality is to summarize the patterns of correlations between the observed variables (Tabachnick & Fidell, 2001). To account for the underlying phenomenon, this is frequently accomplished by reducing variables to the minimum number possible. The underlying phenomena is thought to be the fundamental cause of the observed variables' correlation in the first place. One or more dimensions may be reflected in the underlying phenomena. The structure of a phenomena is referred to as dimensionality (Pett, Lackey, & Sullivan, 2003). One dominant latent variable or phenomenon is referred to as unidimensionality. In the social and behavioral sciences, composite scale scores are frequently employed to make conclusions, and unidimensionality is assumed when employing these composite scores. A structural analysis of a set of observable variables can be performed using a variety of statistical approaches (e.g., factor analysis or multidimensional scaling). Finally, these approaches should produce a sufficient number of dimensions to support the usage of composite scores and to explain the pattern of correlations between observed variables. Dimensions (also known as latent variables) are built variables that appear before the observed variables. That example, if two test items are correlated, it is considered that they have something unseen in common.

Dimensionality Assessment Tools

There are a number of dimensionality assessment tools that have been developed for multivariate data. Popular methods for assessing dimensionality of multivariate dataset are principal components analysis (PCA) and factor analysis (FA). Both methods are linear models that reduce the data on fewer components or factors. The first step in either method is an eigenvalue/eigenvector decomposition of a square, symmetric matrix. One major difference between PCA and FA lies in the type of matrix that is decomposed. In PCA, the matrix that is traditionally decomposed is a correlation matrix, whereas the decomposed matrix in FA is a reduced correlation matrix (i.e., it contains communality estimates along the main diagonal instead of ones). Because of this difference, the complete set of principal components will account for the total amount of variance in the data, while the full set of factors will account for the common variance in the data. However, both sets (principal components and factors) will correspond to the calculated eigenvalues from their respective matrices, and will be in decreasing order. The first principal component (or factor) will have a corresponding eigenvector that indicates a direction in space that accounts for the most variance (or common variance) in the data, the second will account for the next largest amount of variance (or common variance), and so on. These principal components and factors show the underlying structure of the data. However, unlike FA, PCA is a mathematical identity, which orients the data space such that each dimension corresponds to orthogonal directions that account for the largest amount of variation in the data. Therefore, it is not possible to rotate the PCA solution and maintain this identity, whereas rotation of the solution is commonly seen in FA.

After the first step, one needs to decide on the number of dominant dimensions to retain by examining eigenvalues that correspond to each principal component or factor to reduce the data. The number of dominant eigenvalues that underlie a set of data indicates the dominant dimensions within the data. The idea is to choose the smallest number of dominant dimensions that still account for a significant amount of (total or common) variance in the data. Interestingly, determination of the number of dominant dimensions has typically been based on a PCA solution regardless of whether a FA solution is the ultimate goal. When choosing the dominant eigenvalues that underlie a set of data, one must use some decision criteria to justify the choice. Consequently, there are several proposed decision criteria used in PCA. For example, one could use Cattell's scree test (Cattell, 1996) or Kaiser's rule (eigenvalues greater than one; Kaiser, 1960).

One of the better performing methods is a bootstrapped version of Horn's parallel analysis procedure (Horn, 1965; Lambert et al., 1990). If only one dominant eigenvalue is retained from any of these decision methods, the data are assumed to be unidimensional. Any larger number of dominant factors would indicate a multidimensional dataset. Although FA and PCA are appropriate in many analytic situations, the two procedures do have their limitations. The issues with factor analysis and principal components analysis on dichotomous IRT response data have been well-documented (Bernstein & Teng, 1989; McDonald, 1981; Reise, 1999). One such problem is the existence of what has been called "difficulty factors". These difficulty factors occur because binary IRT data often violate the primary assumption of linearity in factor analysis—the assumption that there exists a linear

relationship between observed variables and the underlying latent trait(s). When assumptions of linearity are violated, spurious dominant factors, or difficulty factors, can appear because items with similar difficulty tend to form additional factors distinct from the true dominant underlying dimension(s), thus resulting in overestimation of the true dimensionality of the dataset.

It is important to remember that Principal Components Analysis (PCA) and Factor Analysis (FA) are two independent approaches that are sometimes confused. In other words, PCA has been suggested to be a type of FA (Fabrigar et al., 1999). Factor analysis (FA) is a statistical process that is used to discover which observable variables constitute individual subsets that eventually combine to produce dimensions from a set of latent variables. These variables are used to demonstrate the underlying phenomena that causes the observed variables to correlate (Tabachnick & Fidell, 2001). Exploratory factor analysis (EFA) is a technique for discovering the underlying unobserved dimension of a set of test items in order to get a hypothetical understanding of them (s). When only a minute or no prior knowledge on the data structure is provided, EFA's main goal is to explain the correlation between a group of observed variables (i.e., test items). As a result, EFA is viewed as a tool for developing hypotheses. An EFA is frequently used as a preliminary evaluation technique. When constructing or modifying a scale, an EFA is used to determine the validity of instruments in typical test development practice. For example, researchers might use an EFA to identify an instrument's dimensionality, then utilize that information to create merged scores for hypothesis testing or statistical inferences.

Statement of the Problem

In order to determine the number of dimensions underlying a multivariate dataset, statistical techniques such as principal component analysis, factor analysis and item response theory modelling have been utilized. However, these statistical techniques are not able to determine prior to their application whether dimensions exist within the multivariate dataset, as it is possible to have a dimensionless multivariate data. Thus, available techniques are not designed as preliminary techniques for dimensionality detection which is required before the application of these statistical methods. It appears therefore that there is no initial justification yet for the application of the well-known dimensionality-reduction statistical applications.

Additionally, for the same dataset, different techniques may yield different dimensionality. Even though the relative importance of the dimensions may differ from technique to technique, the basic number of dimensions should be the same, and the technique for finding this number is what appears to be missing. It may be necessary therefore to review these applications and also attempt to propose a dimensionality detection method that can be used to determine whether dimensionality exists within a multivariate dataset.

Studies on dimensionality detection are almost absent in literature. The work which is specific to dimensionality detection is one by Nkansah (2018), which observed some drawbacks. In particular, the study uses an experimenter specific threshold in KMO determination, a threshold based on the judgement of the experimenter. This approach is quite subjective. The goal of this research is to avoid the subjectivity that generally characterizes dimensionality

detection by proposing a-data specific threshold which is a threshold generated from the data structure. It is also observed that the procedure outlined in Nkansah (2018) is computationally expensive since the duration involved in calculating KMO using the original correlation structure far exceeds the duration for the computation based on a much smaller spanning set. This study also investigates the sensitivity and robustness of the method based on the correlation profile. Unlike the literature, the study would be sensitive to the likely presence of extreme values that may affect results by focusing on the use of only the highest contributors to homogeneity within a dimension. This approach will therefore be expected to save computational time and produce a more reliable result.

Also, the formulation of dimensionality in the well-known multivariate techniques is not analytically or computationally presented. They therefore offer a leave-or-take result with no understanding of the formation of the dimensions.

This study therefore seeks to fill this knowledge gap by proposing a dimensionality detection method that could be used to determine whether a dimension exists within a multivariate dataset.

Purpose of the Study

The purpose of this study is to propose a robust automated threshold dimensionality detection method which is not based on an experimenter specific threshold but on the structure of the data for more accurate results. Also, it is anticipated that the robustness of the method to correlation profile would lead to a computationally less expensive approach for calculating the homogeneity of a dataset.

Objectives of the Study

The main objective is to develop an automated threshold method for detecting dimensionality in multivariate datasets. To guide the study, the specific objectives are as follows:

1. Assess the standard statistical techniques for detecting the number of dimensions underlying a multivariate dataset.
2. To propose a dimensionality detection method in a multivariate dataset that would serve as a justification for the application of a dimensionality-reduction technique.
3. To determine a robust dimensionality detection method using the correlation profile of the multivariate data structure.

Significance of the Study

The study would serve as an independent well-structured methodology that comes handy for dimensionality detection. Taking into consideration the level of significance of the correlation coefficients, indicators that influence a dimension are identified by a data specific cut-off value for more reliable results. By data-specific cut-off value, the approach allows the data itself to identify its own threshold suitable for dimensionality detection as opposed to a threshold based on the judgment of the experimenter. By this cut-off value, variables may be considered to belong together if their pair-wise correlation coefficient is equal to or exceeds the cut-off. A data specific cut-off value could identify perhaps the exact dimensionality in the dataset prior to analysis by identifying and excluding those subsets variables that are likely to reduce the true measure of homogeneity in the data. In determining the sets of indicators that constitute the various dimensions, there are some indicators that

may not influence any of the dimensions. These indicators would constitute the non-homogeneous sets. In addition, it is possible to have a number of indicators that influence multiple dimensions. It is also observed that dimensionality could be affected by prevalence of negative correlations among the indicators. Unlike existing procedures that are clearly unstructured, the proposed technique takes into consideration all the afore-mentioned cases that are likely to influence the detection of the true dimensionality in a given data. This way, a good justification could be found for a more focused further application of dimensionality-reduction technique of the dataset.

Delimitation

The study considers two correlation profiles in developing the dimensionality detection technique. The Pearson's correlation which hinges on the mean and Order statistics which hinges on the median will be used to determine the robustness of the method to other correlation profiles. The methods are applied to both simulated and existing datasets. The study also compared the results of a correlation profile generated using the k highest contributors after controlling for outliers as opposed to the results of a correlation profile generated using all the variables in the original dataset.

Description of Datasets

A couple of datasets have been employed in this thesis to study dimensionality detection. Here, we describe the source of these datasets and comment on the reason for their selection for this research. These datasets have been numbered for easy identification in Chapter Four.

Dataset 1 (Performance of Sales Personnel)

The dataset contains evaluations on the performance of sales people employed by a marketing firm. The company seeks to assess the value of its sales people by devising a test, or a series of tests, that would reveal whether or not they have a productivity for high-quality sales performance. The company chose an arbitrary sample of 50 salespeople and evaluated them on three performance indicators: sales growth, sales profitability, and new account sales. These metrics have been converted to a scale of one to 100, with ten representing "average" performance.

Each of the 50 salespeople would take one of four tests, each of which appears to measure creativity, mechanical reasoning, abstract thinking, and mathematical abilities. The table contains a sample of 50 observations on $p = 7$ variables (Johnson & Wichern, 2007).

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

This encompasses an unpublished data which included marks graded out of 100% earned by 72 students in a senior high school on nine subjects. These modules 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.

Simulation of Datasets

The data is simulated on a seven-point polytomous scale with sample size of 200 and dimensionality of three on thirty variables. The simulation is

done using the *mirt* package in R software under the command: *simdata(a, d, N, itemtype)* (Chalmers, 2012), where argument *a* denotes a vector/matrix of discrimination parameter values, *d* vector/matrix of difficulty parameter values, *N* sample size and *itemtype* the underlying IRT model. These opinions are outlined to produce the anticipated dataset. The response datasets are simulated using the generalized partial credit model (gpcm). The generation of three-dimensional dataset necessitates a $k \times 3$ matrix of discrimination values (*a*). Regarding response format, the seven-point scale made use of $k \times 7$ matrix of difficulty values (*d*). The idea of item response theory is reviewed in the methodology in Chapter Three.

Limitation of the Study

As indicated in the objectives of the study, earlier studies on dimensionality detection did not investigate the robustness of the method to other correlation profiles as only Pearson's correlation which hinges on the mean was employed. Though our study considered a correlation profile which hinges on a statistic not affected affected by extreme values namely order statistics, the researcher desired to consider other additional correlation profiles but could not do this due to time constraints.

Definition of Terms

Correlation: A measure of the strength of the relationship between two variables.

KMO: A measure of homogeneity among variables.

Multivariate Data: Data collected on two or more variables.

Partial Correlation: Correlation between two variables controlling for the effect of other(s).

Organisation of the Study

The first chapter covers the general introduction of the study. It first considers the background to the study. In the background, the idea of using well known statistical applications such factor analysis principal, component analysis and IRT for dimensionality detection in multivariate data is introduced and the associated challenge with the use of these technique have been pointed out. This provides the motivation for the study which is provided in the statement of the problem. It is then followed by the objectives of the study. Finally in Chapter One, the description of the various datasets used in the study have been provided.

The review of relevant literature is done in Chapter Two. It focuses on works done by earlier authors on dimensionality. Chapter Three reviews important concepts and methods employed. It reviews the concepts of factor analysis, principal component analysis and Item Response Theory modelling and KMO. Chapter Four deals with analysis and results. It uses a number of datasets to generate results for the proposed dimensionality detection technique. In Chapter Five, the summary of the entire work is presented. Conclusions based on the results are drawn and relevant recommendations are made.

Chapter Summary

This chapter focused on issues and preconceptions regarding the area of study. It outlined a brief synopsis of the fundamentals of dimensionality and also shed some light on unidimensionality and multidimensionality. What the study sought to achieve was also outlined. The gap identified by the research and possible steps for filling this gap were also discussed. It was revealed that

for a multivariate Dataset, the original number of variables are assumed to constitute the number of dimensions underlying the dataset. However not all indicators may influence the phenomenon under study. It is therefore imperative to determine the minimum number of latent constructs (dimensionality) that may underlie the data. Consequently, it is expected that a dataset may have only one dimension or multiple dimensions underlying it. Other areas the research would have considered but for time constraints is also captured.



CHAPTER TWO

LITERATURE REVIEW

Introduction

This chapter reviews the works of researchers that are relevant to the study. It highlights mainly dimensionality studies by several authors and also points out the gaps in these studies.

Review of Studies on Dimensionality

Mengyao (2016) investigated the dimensionality of mixed-format test scores. They discovered that dimensionality assessment improves test developers' and consumers' knowledge of how test scores translate human talents into numbers. Dimensionality assessment addresses a variety of concerns, including (a) whether unidimensionality is true; (b) the number of dimensions that influence test scores; and (c) the linkages between items, underlying dimensions, and items and dimensions. Test developers and users can carefully validate explicit understandings and applications of test scores using the results of dimensionality assessments. The widespread use of mixed-format assessments muddles both theoretically and procedurally dimensionality assessment. The researchers initially suggested a methodology designed specifically for exploratory dimensionality assessment in mixed-format tests. This dissertation examined the performance of a number of widely used and promising dimensionality evaluation methods and approaches using data from three large-scale mixed-format examinations.

Alejandra (2018) investigated factor regression for dimensionality reduction and data integration strategies using cancer data. He noted that two major obstacles in modern statistical applications are the vast amount of data

recorded per individual and the fact that such data are frequently collected in batches rather than all at once, resulting in mean and variance distortions. They solved these problems by developing a new sparse latent factor regression model for integrating heterogeneous data. The model provides a tool for data exploration by reducing dimensionality, correcting so-called batch effects, and estimating sparse low-rank covariance matrices. They looked at how to learn the dimension of latent components using a variety of sparse priors, both local and non-local. Our model is fitted in a deterministic manner using an EM technique for which closed-form updates are derived; this contributes a novel scalable algorithm for non-local priors, which is of interest outside the scope of this thesis. They also demonstrated numerous applications, with a focus on bioinformatics. The findings largely indicated an improvement in the accuracy of low-dimensional data reconstructions, with non-local priors significantly enhancing factor cardinality and non-zero factor loadings inference. Furthermore, the batch effect correction significantly improved the recovery of latent variables. Overall, the thesis introduces a novel technique to latent factor regression that balances sparsity and sensitivity while still being computationally efficient, and it opens up new paths for future study on dimension-reduction-based data integration

Statistical inference in high-dimensional matrix models was investigated by (Löffler, 2020). Matrix models suggested, are common in current statistics. They're used in finance to analyze asset interdependence, genomics to impute missing data, and movie recommender systems to simulate the relationship between users and movie ratings, among other applications. High-dimensional models, in which the number of parameters

exceeds the number of data points by many orders of magnitudes, or nonparametric models, in which the quantity of interest is an infinite dimensional operator, are common.

This leads to novel techniques as well as new theoretical phenomena that can arise when estimating a parameter of interest or its functionals, or when building confidence sets. In this thesis, we will look at three of these matrix models as examples and establish statistical theory for them: Completion of matrices, Principal Component Analysis (PCA) with Gaussian data, and Markov chain transition operators. In the 'Bernoulli' and 'trace-regression' models, studies started with matrix completion and looked for adaptive confidence sets. When the variance of the errors is unknown, they showed that adaptive confidence sets do not exist in the 'Bernoulli' model, but they presented an explicit construction in the 'trace-regression' model. Finally, based on a testing argument, they demonstrated that adaptive confidence sets exist in the 'Bernoulli' model in the situation of known variance. Then they looked at PCA in a Gaussian observation model with complexity assessed by the effective rank, which is the reciprocal of the first principal component's percentage of variance explained. We look at how to estimate linear eigenvector functionals and prove Berry-Essen type constraints. We uncover a new phenomenon as a result of the problem's high dimensionality: The sample eigenvector-based plug-in estimator may have non-negligible bias and hence no longer be n -consistent. They demonstrated how to de-bias this estimator and provide precise matching minimax lower bounds by obtaining n convergence rates. Finally, they looked at nonparametric estimate of a Markov chain's transition operator and transition density. They expected that the

transition operator's unique values diminish exponentially. Discrete, low frequency observations of periodised, reversible stochastic differential equations, for example, satisfy this requirement. We build a new algorithm and demonstrate improved convergence rates using penalization techniques from low rank matrix estimation. Assessed Distributional Properties of High-Dimensional Data.

Mansoor (2013), a multivariate statistical analysis of high-dimensional data was the subject of this PhD. Hessionite Carlo simulations were used to study the increasing dimension asymptotic (IDA) qualities of a number of multivariate non-normality tests when the dimension grows proportionately with the amount of data. For circumstances when $p/n \rightarrow c$, a novel non-normality test based on principal components is proposed. Meaning the power and size of the test are examined through Monte Carlo Simulations. Monte Carlo simulations with various combinations of n and p are performed to investigate the test's power and size. The study looked into the relationship between a distribution's second central moment and its initial raw moment. To infer the systematic relationship between mean and standard deviation, a model with a slope parameter is proposed, and three different estimators of this parameter are developed, and their consistency demonstrated in the context of increasing the number of variables proportionally to the number of observations. To model the link between the mean and standard deviation of the excess return and test hypotheses about the parameter, a Bayesian regression approach was used. The data from the Stockholm exchange market were used in an empirical case. Finally, three novel approaches for testing

panel cointegration of high-dimensional data were incorporated in the error correction framework.

Zupluoglu (2013) used imperfect models to analyze the dimensionality of latent structures underlying dichotomous item response data by using both real and simulated data. The study explored the impact of model misspecification due to minor latent components on a range of dimensionality evaluation approaches described in the literature. The study took into account a variety of dimensionality evaluation processes based on eigenvalue inspection (i.e., parallel analysis), conditional covariances (i.e., DETECT), and model selection approaches (e.g., NOHARM and Mplus based chi-square statistics, RMSEA, GFI, AIC). Two studies were carried out. Using sample datasets chosen from a very large real item response dataset considered as the population, the average, standard deviation, and range of the number of dimensions indicated by different techniques were explored. Also a full simulation study was conducted, and the analytical methods' performances were assessed using the number of key dimensions in the true generating model as a benchmark. The current research yielded some intriguing and thought-provoking findings about the performance of some well-known and widely employed procedures under various conditions. The current study's findings suggest that most of the methods proposed in the literature and available to practitioners are not always useful tools in dimensionality assessment, especially when the goal of dimensionality assessment is to identify latent traits with major influences when the underlying factor structure is complex and minor factors are present. When the underlying latent structure was factorially complicated, the current investigation gave some insight into

the performance of alternative dimensionality evaluation methodologies with mis specified models.

Tian (2009) investigated dimensionality reduction for high-dimensional data categorization. The study looked at dimensionality reduction issues in classification for both multivariate and functional data with high dimensionality. High-dimensional data, according to the study, refers to data having a large number of variables, which is often greater than the number of observations. Engineering, biometrics, psychometrics, and neuroimaging are just a few of the fields that deal with high-dimensional data. Classifying these data is a tough task due to the large number of variables, which complicates traditional classification algorithms and makes many traditional procedures unfeasible. Adding a dimensionality reduction step before applying a classification approach is a natural solution. Two ways are proposed for dealing with multivariate data. The first is based on simulated annealing (SA), and the second is based on multivariate adaptive stochastic search (MASS). In each cycle, they both use stochastic search methods to select a small number of optimal transformation directions from a huge number of random possibilities. The proposed approaches have the advantage of being able to accurately project data onto very low-dimensional non-linear as well as linear domains. These methods are meant to resemble variable selection methods like the Lasso, or variable combination methods like PCA, or a method that combines the two. MASS, in particular, can modify the model complexity level adaptively, and so performs well when variable selection or variable combination methods fail. We compare the strengths of SA and MASS to various classical and modern categorization approaches in a variety of

simulated and real-world investigations. Problems with classification of functional data are also addressed. We present a functional adaptive classification (FAC) method that considers the functional response and generates extremely accurate and understandable results. FAC is similarly based on a stochastic search technique that is directed by the model complexity evaluation. This frequently leads to a straightforward link between functional variables and the reduced data, making the model more understandable. To demonstrate the efficiency of the suggested strategy, simulation studies and an fMRI time course study are included.

Thinesh (2018) used a variety of generalized hyperbolic distributions to investigate dimension reduction and grouping of high-dimensional data. Model-based clustering, according to the study, is a probabilistic strategy in which each cluster is viewed as a component in an appropriate mixture model. One of the most extensively used model-based strategies is the Gaussian mixture model. However, due to the over-parametrized solutions that develop in high-dimensional spaces, this model performs badly when clustering high-dimensional data. Instead, this study looked at how to combine dimension reduction approaches with clustering using a variety of generalized hyperbolic distributions. The techniques of dimension reduction, principal component analysis, and factor analysis, as well as their extensions, were examined. Then, using both simulated and real data sets, the aforementioned dimension reduction strategies were separately matched with a mixture of generalized hyperbolic distributions to demonstrate the clustering performance attained under each strategy. The clustering method based on principal component

analysis produced superior classification results for the majority of the data sets than the clustering method based on extending the factor analysis mode.

Janecek (2009) investigated efficient feature reduction and classification methods, as well as their applicability in drug discovery and E-mail Categorization. They claimed that as the dimensionality of the feature space grows, many types of data analysis and classification become significantly more difficult, and that data also become increasingly sparse in the space it occupies, posing significant challenges for both supervised and unsupervised learning. The curse of dimensionality is a phenomenon that arises from the fact that high-dimensional data is sometimes difficult to work with. When there are few observations (i.e., data samples) relative to the number of features, a large number of features can increase the noise in the data and hence the error of a learning system. Feature selection and dimensionality reduction methods (often referred to as feature reduction methods) are two strategies for addressing these issues by reducing the amount of features and consequently the data's dimensionality. Several studies have been conducted in recent years to improve feature selection and dimensionality reduction strategies, and significant progress has been made in terms of picking, extracting, and creating effective feature sets. However, due to the significant impact of different feature reduction approaches on classification accuracy, there are still a number of unanswered concerns in this subject.

Furthermore, as the number of possible features for diverse application areas grows, additional concerns arise. This thesis looked into some of these

unanswered concerns, such as the relationship between different feature reduction techniques and classification accuracy. The goal is to find a set of features that best mimic the original data while maintaining a high level of classification accuracy. The computational cost of feature reduction techniques is the basis for other difficulties. Due to the large amount of data, it is necessary to design computationally efficient feature reduction approaches that can be used in parallel. To solve this issue, the thesis investigated numerous ways for leveraging task and/or data parallelism in NMF, as well as introducing computationally efficient adaptations of current NMF algorithms. They researched innovative initialization strategies for NMF based on feature selection, as well as fast and effective classification methods based on NMF, to speed up the runtime of NMF even further. Furthermore, there are a number of issues to consider when evaluating the interpretability of dimension reduction strategies. The information about how much an original feature contributes is often lost when a linear combination of dimensionality reduction algorithms is used.

In this thesis, we look at how the improved interpretability of NMF factors due to non-negativity requirements may be used to keep the original data interpretable. Experiments are carried out on datasets from two very distinct application fields, each with its own set of research challenges: email categorization and in silico drug discovery screening.

Timmerman & Lorenzo-Seva (2011) used parallel analysis to measure the dimensionality of ordered polytomous elements. They discovered that parallel analysis (PA) is an often suggested method for determining the

dimensionality of a set of variables. PA comes in a variety of forms, each of which can produce different dimensionality indicators. To determine the number of common components underlying ordered polytomously scored variables, the authors used the most applicable PA technique.

Instead of the currently used principal component analysis (PCA) and primary axes factoring, they proposed minimal rank factor analysis (MRFA) as an extraction approach. Based on data containing major and minor components, simulation research revealed that all processes consistently point to the number of major common elements. Although a polychoric-based PA outperformed a Pearson-based PA by a small margin, convergence issues may limit its empirical application.

PA-MRFA with a 95% threshold based on polychoric correlations or, in the case of nonconvergence, Pearson correlations with mean thresholds appear to be a good choice for identifying the number of common variables in practical applications. The PA-MRFA technique, which is based on common factors, fared best in the simulation experiment. Second best is PA based on PCA with a 95% threshold, as this method performed well in the simulation experiment's empirically applicable conditions.

Kim (1994) investigated new methods for determining the dimensionality of standardized test data. The researchers discovered that a novel dimensionality index based on the conditional covariance of item scores given a latent variable is defined and studied in educational and psychological test data. By using cluster analysis, this index accurately detects the test dimensionality in terms of both identifying the number of dimensions present

in the test and identifying the items contributing to each dimension. Furthermore, this index accurately measures the test data's lack of unidimensionality, and its asymptotic behavior under unidimensionality gives theoretical support. To detect dimensional disagreement of item pairs, a new significance test based on a kernel smoothing technique is devised. A simulated evaluation of this method demonstrates a reasonable type 1 error rate in relation to its nominal level of significance, as well as great power performance, when compared to existing procedures. When data is not unidimensional, the unidimensional parametric item and ability calibration processes BILOG and LOGIST are checked to see what is truly being assessed as unidimensional ability. The accuracy of ability estimation is also examined in terms of average standard error. As their claimed unidimensional ability estimate, both BILOG and LOGIST appear to present a composite of underlying latent qualities. The average standard errors are relatively invariant to the degree of lack of unidimensionality as a result of this, but the direction of the composite being assessed best changes routinely and by a substantial amount with various degrees of multidimensionality.

Under nonparametric IRT models, Alexander et al. (2004) conducted a comparative evaluation of test data dimensionality assessment methodologies. The dimensionality of item response data can be determined using non-parametric item response theory approaches. MSP, DETECT, HCA/CCPROX, and DIMTEST were all considered. The methods were first compared on a theoretical level. Second, using the default parameters of each program, simulation research was conducted to examine the performance of MSP, DETECT, and HCA/CCPROX in detecting a simulated dimensional structure

of a matrix of item response data. The approaches that employ conditional covariances on the latent trait (DETECT and HCA/CCPROX) were superior in discovering the simulated structure in various design cells versus the method that used normed unconditional covariances (MSP). Third, based on the data used in DETECT and DIMTEST, the accuracy of the decision to accept or deny unidimensionality was examined. This decision did not always reflect the item pool's true dimensionality.

With small sample sizes and short test lengths, (Andre et al., 1998) evaluated the dimensionality of Item Response Matrices. To apply standard item response theory models legally, the assumption of unidimensionality must be met, according to the researchers. The extent to which it can be proven that the dimensional structure underlying a test is consistent with the blueprint determines the validity of score-based conclusions. In settings similar to those observed in small-volume administrations, little study has been done to examine the behavior of dimensionality assessment algorithms. The goal of this research was to look into empirical data. With data sets constructed to reflect brief tests and small samples, Type I error rates and rejection rates for 3-dimensionality evaluation techniques were calculated. With unidimensional data sets, the G₂; difference test from TESTFACT (Wilson, Wood, & Gibbons, 1991) and the LISREL8 (Jöreskog & Sörbom, 1993a) chi-square statistic had an inflated Type I error rate, whereas the approximate chi-square statistic from a NOHARM (Fraser & McDonald, 1988) analysis did not. All procedures have significant rejection rates when using simulated 2-dimensional data sets. The independent factors changed strongly altered the behavior of the G₂; difference test, which was not the case

for the approximation chi-square statistic. These findings are examined in terms of their relevance for small-volume administrations.

Bayesian dimensionality assessment for the multidimensional nominal response model was investigated by (Javier et al., 2017). For the multidimensional nominal response model, the work introduced Bayesian estimation and assessment methodologies. This paradigm is useful for performing nominal factor analysis on items with a finite number of unordered response categories. In contrast to standard factorial models, the key feature of this model is that each response category on the latent dimensions has a slope, rather than having slopes connected with the items. For estimation, the multidimensional nominal response model's extensive parameterization necessitates large samples. When the sample size is moderate or small, some of these factors may be difficult to empirically identify, causing the estimation process to fail. To estimate the parameters and number of dimensions underlying the multidimensional nominal response model, we present a Bayesian MCMC inferential approach.

The standardized generalized discrepancy measure, which needs resampling data and is computationally more demanding, was compared to two Bayesian techniques to model evaluation: discrepancy statistics (DIC, WAICC, and LOO), which provide an indication of the relative value of different models. The findings of a simulation research comparing these two approaches reveal that the standardized generalized discrepancy measure may be used to correctly predict the model's dimensionality, whereas the discrepancy statistics are suspect. The study also contains a real-world

example in which the model is used to perform an exploratory factor analysis of nominal data in the context of learning styles. In the disciplines of ability measurement, attitude scales, sample surveys, market research, and so on, nominal variables are commonly collected from a variety of item response formats. Multiple-choice questions, for example, have one correct answer and several distractors. When the data comes from multiple-choice items, the factorial analysis of nominal variables is frequently carried out by dichotomizing the data into correct and incorrect responses and then running the dichotomous data matrix through a categorical factor analysis technique. In some cases, however, dichotomization is not an option because the focus is on the relationship between latent dimensions and answer categories. Each category in a market research item, for example, could represent a buying choice, and there is no natural way to dichotomize the data.

The study discovered that the factorial analysis of responses with an implicit ordering, as well as their estimation and testing methodologies, have long been discussed in the psychometric literature (Christofferson, 1975; Bartholomew, 1980; Reckase, 2009). These models are based on a normal or logistic function that uses a vector of slopes to link observed responses and dimensions. Furthermore, the distribution of responses across the item's categories is determined by a set of intercept parameters (Mislevy, 1986). Because of the inherent challenges of the underlying psychometric paradigm, nominal variable component analysis is a more recent development. This model is a multidimensional expansion of (Bock's, 1972) nominal response model, which assumes things load in one dimension. The slopes of the nominal response model are parameters of the categories rather than

parameters of the items. The ordinal model has two thresholds and two slopes for an item with three response categories and measures two dimensions (say), whereas the nominal model has two thresholds and four slopes (one category has no parameters and the other categories have one slope in each dimension). In the psychometric literature, applications of constrained versions of the multidimensional nominal response model (MNRM) have been published. Hoskens et al. (2001), for example, used a restricted MNRM to assess cognitive components involved in item solving; in this model, parameter limitations are imposed to represent the components tested by the categories.

Another version of the MNRM created by (Johnson & Bolt, 2010) targeted at separating a general dimension of ability from subsidiary variables that describe test taking strategy. The MNRM will be used in its entirety in this article to undertake an exploratory factor analysis of nominal variables. Except when essential to identify the model, none of the parameters in the exploratory analysis are fixed to a constant value. The MNRM's extensive parameterization causes complications in parameter interpretation and estimation. (Thissen et al., 2010; Falk & Cai, 2016) introduced many parameterizations aiming at providing parameters with a clear meaning in terms of parameter interpretation. This paper focuses on the inferential parts of the problem, specifically the estimation of the number of dimensions. The MNRM's estimate issues arise because the response patterns' contingency table is often excessively sparse due to the vast number of response categories that must be modeled. Using computer algorithms like Latent GOLD,

maximum likelihood estimates can be generated (Vermunt & Magidson, 2016).

When the sample size is approximately a few hundred people, however, the maximum-likelihood estimation process may run into difficulties, resulting in significant standard errors. Convergence issues are most common for parameters in categories with a low response frequency, which can occur even when the sample size is rather large. For example, with a sample of 500 or more people, it's not uncommon to encounter categories with frequencies of less than 10, which means that reliable estimates for the many parameters that describe the category are impossible to come by. Apart from the issues of estimating, measuring the fit of the nominal model in the frequentist framework is problematic since goodness-of-fit statistics are based on asymptotic reasoning that rarely adhere to genuine model application conditions. By defining prior distributions for the parameters and shifting the inference to a Bayesian setting, the statistical difficulties of the nominal model can be solved. Bayesian inference combines information from the sample with information from prior distributions, resulting in estimates that are more stable, alleviate problems of lack of convergence for some parameters, and provide a method for simulating the posterior distribution of model evaluation statistics. The study presented a Bayesian inferential approach for determining the MNRM's latent dimensionality.

The suggested approach is based on Markov chain Monte Carlo (MCMC) procedures that use basic Bayesian estimating algorithms. In the framework of item response theory, Bayesian estimation has already been

applied to ordinal answers (Kieftenbeld & Natesan, 2012) and multidimensional models (Levy et al., 2009). By replicating the distribution of evaluation statistics, Bayesian processes have been successfully applied to testing model fit (Sinharay et al., 2006). The definition of model evaluation statistics for a nominal model, on the other hand, is a relatively new subject of study. We used two model evaluation statistics that were recently proposed in a Bayesian statistical context, the widely applicable information criterion (WAIC) and the leave-one-out cross-validation (LOO), both of which are based on information theory (Gelman et al., 2014) and have never been used in a psychometric context to our knowledge. The article also covers Levy et al's adaptation of the standardized generalized dimensionality discrepancy measure (SGDDM) to the nominal case. The SGDDM was created to evaluate the dichotomous item response model, but it was later expanded to ordinal factorial models.

The SGDDM gives useful information for dimensionality assessment of the nominal model, as demonstrated in this paper. The remainder of the article is divided into the sections below. The MNRM, as well as the restrictions for parameter identification and the rotation problem, are described in Section Multidimensional Nominal Response Model. The MCMC Bayesian estimation algorithm is described in the section's Bayesian Parameter Estimation, while the model evaluation statistics are described in the Section's Bayesian Model Evaluation. The simulation study in section's simulation study analyzes the Bayesian inferential procedure under actual situations. A real data research is presented in the context of a questionnaire of learning

styles, in which the response categories indicate several learning styles and there is no implicit order among them.

The effect of distributional differences on dimensionality assessment using DIMTEST was investigated by (Walker et al., 2006). Some people feel that most exams are multidimensional, meaning that they examine more than one underlying construct, according to the study. The fundamental goal of this research is to show how differences in the secondary ability distribution affect statistical dimensionality detection and to distinguish between substantive and statistical dimensionality. This study shows how altering the ability distributions influences the results generated from DIMTEST, a nonparametric statistical process based on the notion of essential unidimensionality, given dichotomous data simulated as multidimensional. As the mean of the secondary ability distribution approached the extremes and/or the standard deviation of the secondary ability distribution approached zero, the power of DIMTEST dropped. This has crucial ramifications for both academics and practitioners because, while a test may measure extra dimensions from a substantive standpoint, statistically, these dimensions may not be discovered.

Heating et al. (2010) investigated the optimization and uncertainty assessment of severely nonlinear groundwater models with a large number of parameters. Highly parameterized and CPU-intensive groundwater models are increasingly being utilized to explain and predict flow and transport through aquifers, according to the findings. Despite their widespread use, these models pose considerable hurdles to parameter estimation and predictive uncertainty analysis algorithms, especially global techniques that typically require a high

number of forward runs. In this paper, we provide a general methodology for parameter estimation and uncertainty analysis that can be used in these circumstances. Following the derivation of a surrogate model that mimics essential properties of a full process model, we evaluate and apply nullspace Monte Carlo (NSMC), a pragmatic uncertainty analysis tool that combines the capabilities of gradient-based search with parameter dimensionality reduction.

The results of NSMC are contrasted with a formal Bayesian approach employing the differential evolution adaptive metropolis algorithm as part of the surrogate model study. This kind of comparison has never been done before, especially with such high parameter dimensionality. Despite the inversion problem's highly nonlinear nature, the presence of several local minima, and the relatively large parameter dimensionality, both techniques performed well, and the results are comparable. The knowledge collected from the surrogate model study is then used to calibrate the full, highly parameterized, and CPU heavy groundwater model, as well as to investigate the predictive uncertainty of the model's predictions. The methodology described here can be used to any highly parameterized and CPU-intensive environmental model in which efficient methods like NSMC are the only viable way to do predictive uncertainty analysis.

The Bayesian assessment of dimensionality in reduced rank regression was investigated by (Jukka et al., 2010). In the multivariate reduced rank regression framework, which incorporates numerous models such as MANOVA, factor analysis, and cointegration models for multiple time series, the research investigated a Bayesian inference about dimensionality. A closed

form approximation to the posterior distribution of the dimensionality is derived using the fractional Bayes approach, and some asymptotic features of the approximation are established. Simulation is used to investigate finite sample properties, and the method is applied to growth curve data and cointegrated multivariate time series. According to the findings, a common scenario in multivariate analysis is the examination of relationships between sets of variables using explicit parametric models or descriptive methods like principal components and canonical correlations.

Although it was long recognized that these instances could be represented jointly in terms of multivariate regression with a reduced rank structure for certain parameters (see, for example, Anderson's pioneering work in 1951), the general statistical community has only recently fully appreciated this approach. The generality of the reduced rank regression (RRR) framework is one of its strongest features, as it incorporates various well-known models for multiple time series, including MANOVA, factor analysis, linear simultaneous equations models, and many others. In conventional full rank multivariate regression, the most common source of model uncertainty is the selection of suitable predictor variables. For the latter model selection problem, there are several plausible methods available (Brown et al., 1998; George & Foster, 2000). Producing reasonable conclusions on the dimensionality of the subspace of regression coefficients for a fixed set of predictor variables has been more difficult. To estimate dimensionality in RRR, (Geweke, 1996) suggests a computationally intensive approach, and (Kleibergen & Paap, 2002) employ extensive Monte Carlo simulation schemes to obtain the posterior distribution of the dimensionality.

The conveniently computable one-formula solutions without subjective input from the user, such as information theoretic criteria (Akaike, 1974), approximation logarithmic Bayes factor (Schwarz, 1978), or sequential tests, are the methods that tend to be employed in applications (Anderson, 1951; Izenman, 1980; Jo-hansen, 1995). Some recommendations within a narrow class of reduced rank models have also been made; see, for example, (Chao and Phillips, 1999) for a criterion adapted to cointegration models. Only the approach of (Schwarz, 1978) seeks to approximate the posterior distribution of dimensionality among these methods. This is significant because the posterior distribution is an appealing representation of the dimensionality inference's uncertainty. However, the Schwarz approximation is known to be a bit sloppy, and it frequently underestimates the underlying model dimension (Kass & Raftery, 1995). The approximate posterior distribution of the dimension of the parameter structure was calculated using O'Hagan's fractional marginal likelihood (FML) technique (O'Hagan, 1995, 1997). Their method produced an analytically tractable answer that may be used without the user's subjective input. Its qualities are studied both theoretically and by application to a variety of real and simulated data sets.

Mares (2016) investigated variable selection in the dimensionality. The researcher discovered that today's high-throughput technologies are resulting in a vast amount of data to be studied. The goal of the research was to develop mathematical and statistical approaches for extracting as much information as possible from the available data. However, the high dimensionality of the data, both in terms of sample size and the number of features or variables, creates significant obstacles. Increased computer power

and the usage of distributed computation technologies make it easier to deal with the enormous number of samples. The huge number of features or variables increases the risk of using the improper explanatory factors to explain variance in both noise and signal. One approach to overcoming this challenge is to select a smaller set of features from the original set that are most important given an assumed prediction model from the initial set. This method is known as variable or feature selection, and it entails making a bias or statistical assumption about which attributes are more important. Different statistical assumptions about the mathematical relationship between predicted and explanatory factors, as well as which explanatory variables should be deemed more relevant, are used in different feature selections. The initial contribution of the researcher is to combine the strength of several variable selection approaches based on various statistical assumptions. The researchers began by categorizing existing feature selection methods based on their assumptions and evaluating their scaling capacity for high-dimensional data, especially when the number of samples is substantially fewer than the number of features. The study introduced a new algorithm that combines the findings of many feature selection methods based on distinct assumptions about the function that generated the data, and we show that our method is more sensitive than using each method alone.

One of the most common simplifying assumptions is that the predicted variable and the explanatory variables have a linear relationship. The second contribution is to show that, even though the underlying function that created the data is not always linear, at least one feature selection procedure based on the linearity assumption is consistent. The study developed a new technique

based on these theoretical discoveries that offer superior results when the underlying function that created the data is at most partially linear. When given enough training data, neural networks, particularly deep learning architectures, have been found to fit very non-linear prediction models. They do not, however, include feature selecting tools. The study made a contribution by evaluating the performance of these models when given a large number of features and fewer samples, proposing a method for feature selection, and demonstrating that combining this feature selection method with deep learning architectures outperforms not using feature selection in certain situations. Several feature selection strategies, including the ones suggested in this thesis, rely on resampling techniques or the use of several algorithms for the same dataset. Their benefit is derived in part from the use of additional computational capacity. As a result, our final contribution is an efficient data distribution and load-balanced parallel calculation for re-sampling-based algorithms.

Measurement Scale and Sample Size for Determination of Dimensionality

One of the key findings of (Nkansah, Zakaria, & Howard, 2019) is the optimal size of data for detecting factors in IRT generated data. It is found in that study that a sample size of 150 is optimal for various types of scales with varying underlying dimensionality. However, it is also found that sample size of 200 could perform quite close to that of 150. Likert scale with more points are also found to perform much better allowing higher underlying dimensionality to be captured in factor analysis. Under optimal sample size, likert scale of five-points or higher is preferred. Particularly, seven-point scale is identified to be suitable for data with high underlying dimensions.

Exploratory versus Confirmatory Assessment

Both exploratory and confirmatory methods could be used to assess dimensionality (Reckase, 2009; Svetina & Levy, 2014). When there is no clear hypothesis or evidence on the dimensional structure of the given data, exploratory dimensionality assessment, which is the focus of this dissertation, is frequently used. It has been used in operational testing programs to check the alignment of real dimensionality with the desired dimensionality, either alone or in combination with confirmatory dimensionality assessment (e.g., Fu, Chung, & Wise, 2013; Jang & Roussos, 2007; Wilson, 2000; Zwick, 1987).

Before examining other psychometric processes, exploratory dimensionality evaluation is frequently used as part of a preliminary investigation (e.g., MIRT equating, see Brossman & Lee, 2013). In this thesis, more exploratory dimensionality evaluation methods are studied, while the insights from these methods may also be beneficial in some confirmatory cases. Dimensionality could also be described within the framework of IRT. IRT, according to proponents, enables for a clear and exact understanding of the ideas of unidimensionality and multidimensionality (Hattie, 1985; Nandakumar, 1991; Stout, 1987; Stout et al., 1996; Zhang & Stout, 1999a, 1999b). Stout (1990) proposed a classic IRT definition of dimensionality as the smallest number of latent features required for a locally independent and monotone model, which was further expanded on by (Nandakumar, 1991), (Stout et al., 1996), and (Zhang and Stout, 2000, 1999a, 1999b).

Local independence specifies whether item answers are mutually independent or pairwise uncorrelated after adjusting for underlying latent qualities, depending on whether strong local independence (SLI; Lord, 1980) or weak local independence (WLI; McDonald, 1981) is evaluated. The likelihood of properly answering an item change monotonically with the values of latent features, according to monotonicity. The data are called unidimensional when a single latent attribute is sufficient to generate such a model. The number of latent qualities necessary defines the number of dimensions if the data is not unidimensional. (Stout, 1987) defined essential dimensionality as the number of main or dominant latent features based on IRT, which has proven to be one of the most important notions in the development of nonparametric dimensionality evaluation processes.

Review of Dimensionality Assessment Methods

Factor Analysis

Factor analysis has long been used to investigate dimensionality in multivariate data. For an overview of the application of factor analytic approaches to dimensionality evaluation, see (Hattie, 1985), (Reckase, 2009), (Stone & Yeh, 2006), and (Velicer, Eaton, & Fava, 2000). EFA refers to a group of statistical approaches that are used to explain observed variances and covariances in a broad sense (Kline, 2010). EFA does not need the use of a theorized dimensional structure, unlike confirmatory factor analysis (CFA), which appears to be favorable for exploratory purposes. The number of dimensions equals the number of components or factors to keep, according to EFA. When only one component or element is preserved, the data is deemed unidimensional; otherwise, some degree of multidimensionality appears. The

data's dimensional structure correlates to a specific factor solution created by EFA.

Considerations That Could Influence Results of Factor Analysis

There are a number of considerations that could influence the result of factor analysis. (van der Eijka & Rose, 2015) found that, generally, factor analysis conducted on ordered categorical survey data is prone to over-dimensionalisation, irrespective of the mode of analysis. However, the risk when using some extraction methods such as the eigenvalue-greater-than-one rule (or K1 rule), are reduced for polychoric, rather than Pearson's correlations. The focus on data generated on categorical variables primarily violates the assumption of interval-level measurement and questions keep being raised regarding the circumstances under which this leads to substantially misleading results. The literature is not clear on the matter, and this is also the opinion of (van der Eijk and Rose, 2015). Their attempt in this regard estimated the risk of over-dimensionalisation when factor analysis is used on data generated on Likert-type data. They specifically stress that the data that may be factor suitable could be affected in some five main ways: (1) the nature of the underlying distribution; (2) the number of items; (3) the level of random noise; (4) the range of positions of the items on the underlying dimension; and (5) the skewness of the items. Based on their study, van der Eijk and Rose therefore recommend, among others, that the K1 should not be used, given available alternatives; and that the polychoric correlations are to be preferred to Pearson's correlations within the condition of smaller number of items.

The consequences of violating the assumptions are evident in inflated

probability chi-square tests of fit, lowered standard errors, and inflated error variances in confirmatory factor analysis (Finch & West, 1997). When item response scales have more scale points or categories, the repercussions are less severe. An item with an ordered seven-point response scale, for example, is more likely than a dichotomous item to nearly satisfy the assumption of factor analysis. When categorical variables approach a normal distribution, the number of categories has no effect on the chi-square test of fit between the model and the data, according to (Byrne ,2001).

Furthermore, factor loadings and factor coefficients are only slightly underestimated under these conditions. When responses to items follow an approximate normal distribution, research suggests that items with five or more ordered response categories do relatively well in confirmatory factor analysis (Byrne, 2001). (De Bruin, 2004) employed two ways to deal with the problem of non-normality and nonlinearity of items. These include (a) using item response theory measurement models and (b) using item parcels rather than individual items as the primary units of factor analysis.

Factor analysis is based on the correlations among items on which the data is generated. For Pearson's product-moment correlations to adequately reflect the relationships among the variables, observed variables must be measured on interval scale (e.g., MacCallum, 2009; Tompson, 2004). This condition is also required for the assumption of linearity of relationships among latent variables. However, Likert scale items, which are categorical in nature violate this condition. This has been the main concern in the literature. It is observed, for example, that the correlation between assumed underlying continuous variables in a Likert scale items is attenuated by the categorisation

(Olsson, 1979), though the extent of the attenuation is not uniform. The smaller the number of categorisation, the larger the attenuation, *ceteris paribus*. In addition, attenuation depends on the (observed) distribution of the scores: it is minimal when responses are approximately normally distributed with approximately equal means and is maximal for variable pairs that are skewed in opposite directions. Thus, (Flora, LaBrish & Chalmers, 2012) report a true population correlation coefficient of 0.75 observed as 0.25 when the continuous variables are categorised into five-point items; however, for other item pairs, the attenuation was much less severe.

These observations imply that observed product-moment correlations may be quite different from their underlying true values, and so is also the factor structure derived from the observed correlations. This would likely lead to over-dimensionalisation with factors discriminating between left and right skewed items (for example, in Gorsuch, 1983; and Van Schuur, 2003). Moreover, categorisation of true continua leads necessarily to violations of the linearity, adding to the inadequacy of the product-moment correlation to represent the relationship between items (Flora et al., 2012). For factor analysis of ordinal data, polychoric correlations are often recommended. Extensive discussion on polychoric correlation can be found in a number of texts (Uebersax, 2006)

Illustrative Dataset and Cut-off Values in Dimensionality Detection

As noted in Chapter One, the illustrative Dataset 1 on the performance of sales personnel is contained in several texts (Johnson & Wichern, 2007; Anderson, 2003; Mardia, Kent & Bibby, 1979). In these presentations, the data were used to illustrate some multivariate techniques, particularly factor

analysis. The data are one of several datasets that have been used in studies of problems associated with factor analysis by (Benyi, 2018) and on the Kaiser-Meier-Olkin's measure of sampling adequacy (Nkansah, 2018). These studies have found that although the data is suitable for factor analysis, it surprisingly does not yield a reasonable factor solution. A study of dimensionality in the data made use of a cut-off value of 0.5 and identified only one dimension underlying the data. The factor analysis reveals challenges of interpretation of factors. Thus, there is theoretically one dimension and hence the data is suitable for factor extraction. However, it statistically has no 'significant' dimension. This implies that a factor model is not suitable for the data. The problems identified with this data cover contrasting factors and one-indicator factor solutions, which are inconsistent with the features of the variance-covariance matrix of the data. It is apparent therefore that the variance-covariance structure of this data makes it difficult for determination of its dimensionality.

The data is therefore very suitable as a test data for the implementation of the methodology developed in this thesis. The study will therefore explain more clearly the nature of the data structure that makes its dimensionality difficult to detect.

Dataset 1 has also been studied (Apanyin, 2021) with canonical correlation analysis technique. It has been demonstrated that the first three columns of the data on seven variables could constitute one subset variables whilst the remaining four constitutes the second subset vector. This way, canonical correlations could be found for pairs of canonical variables from the two sets.

Dataset 2 on student performance on nine subjects has also been studied in Benyi (2018) using a cut-off value of 0.5. Two homogeneous sets have been identified in this data indicating a dimensionality of two. A cut-off value as low as 0.2 has been identified by subjective choice to support identification of homogeneous groupings in data. The choice of a cut-off value is clearly dependent on the data structure and a good choice of cut-off value is required to identify appropriate dimensionality.

Chapter Summary

The literature has focused on the relevance of the data used, the cut off value, some methods that were used in dimensionality detection in a multivariate dataset. A multivariate dataset is either characterised as unidimensional or multidimensional. It is clear from the literature that no specific study has focused on a structured and rigorous presentation of dimensionality detection in multivariate data. Our study therefore seeks to fill this gap by developing an objective and robust dimensionality detection method.

CHAPTER THREE

RESEARCH METHODS

Introduction

As noted in the introductory chapter, this work is mainly motivated by the work of Nkansah (2018) on the computation of the KMO. We review the generalized rule as presented for determining the expected dimensions in multivariate data and point out in Remark (3.1) the main point of contention in the rule that motivates this study. The underlying concepts of the KMO are orders zero and one correlation coefficients. The chapter will examine these concepts and point out the perspectives taken on them by the study. It will review relevant multivariate techniques that have dimensionality detection embedded in them and which will be applied in the study.

Generalized Rule for Determining Expected Dimensions of Datasets

Suppose a multivariate dataset is generated on a set of p variables (x_1, x_2, \dots, x_p) with correlation coefficients that are generally significant. 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 $(X_i, X_j), i, j \in I = (1, 2, \dots, p)$ with the highest correlation coefficient. Let this pair be (x_u, x_v) , and label the set as $S_1 = (x_u, x_v)$ 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 $i \in I\check{1}$, then we obtain a final first

homogeneous set $S_1 = \check{1}\check{1}$ with index set

$$I_1 = \{i_1, i_2, \dots, i_{g1}\} \subset I.$$

We will form a new set S_2 from the elements $x_i \notin S_1$, $i \in I\check{1}$. Denote

$T_1 = I\check{1}$. Consider the pair (x_i, x_j) , $i, j \in I\check{1}$ with the highest correlation

coefficient that meets the cut-off value τ . This pair is (x_{l_1}, x_{l_2}) . Thus, we

obtain the second set $S_2 = \{x_{l_1}, x_{l_2}\}$, and an index set $I_2 = \{l_1, l_2\}$. Now, if the

correlation coefficients $r_{x_k, x_i} > \tau$, $\forall k \in I_2$, $i \in I\check{2}$, 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 some $i \in I\check{2}$ then we obtain a final second homogeneous set

$$S_2 = \{x_{l_1}, x_{l_2}, \dots, x_{l_{g2}}\} \text{ with index set } I_2 = \{i_1, i_2, \dots, i_{g2}\} \subset I.$$

Consider all elements $x_i \notin (S_1 \cup S_2)$, $i \in I\check{1}\check{1}\check{1}$. Denote

$T_2 = I\check{1}\check{1}\check{1}$. 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}) . Thus, we obtain the third set $S_3 = \check{1}\check{1}$, and an index set

$$I_3 = \{t_1, t_2\}.$$

If the correlation coefficients $r_{x_k, x_i} > \tau$, $\forall k \in I_3$, $i \in I\check{3}$, 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 some $i \in I \setminus I_3$ then we obtain the final third homogeneous set $S_3 = \{x_{t_1}, x_{t_2}, \dots, x_{t_{g_2}}\}$ with index set $I_3 = \{t_1, t_2, \dots, t_{g_3}\} \subset I$.

We attempt to form the q th 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

$S_q = \{x_{d_1}, x_{d_2}, \dots, x_{d_{g_q}}\}$, and the index set $I_q = \{d_1, d_2, \dots, d_{g_q}\} \subset I$. 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 q th homogeneous set $S_q = \{x_{d_1}, x_{d_2}, \dots, x_{d_{g_q}}\}$ with index set $I_q = \{d_1, d_2, \dots, d_{g_q}\} \subset I$.

If for some set S_{l+1} and index set I_{l+1} , and for $x_i \notin \left(\bigcup_{k=1}^l S_k \right)$, $r_{x_i, x_j} < \tau$, for all

$i, j \in I \setminus \bigcup_{k=1}^l I_k$, then S_{l+1} 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 3.1

It can be observed that the procedure described for identifying homogeneous sets is obviously influenced by a fixed cut-off value of τ for which variables may be considered to belong together if their pair-wise correlation coefficient exceeds τ . Since the value of τ is set by the experimenter, it is highly subjective even though it is based on a general assessment of the size of correlations coefficients among the variables.

It is already acknowledged (Nkansah, 2018) that there are some variables which are likely to contaminate the measure of homogeneity of the data. These are elements identified in the indexed set T_l of elements that are not found in any of the homogeneous sets. An automated procedure would therefore include a process that screens the variables to include only those that are identified with a particular homogeneous group.

Dimensionality of a Dataset

As they choose one or more techniques to examine their own data, researchers make a hazy decision on how they interpret dimensionality. In the extant literature, the term dimensionality has been employed in a variety of ways, both as a property of a test and as a characteristic of test scores (Reckase, 2009). For the purpose of this research, dimensionality of a dataset may be described broadly in two ways: unidimensionality and multidimensionality.

A given multivariate dataset could be of unidimensionality and multidimensionality. We attempt to establish working definitions of the two types of dimensionality in relation to two statistical applications that deal with dimensionality of dataset. These are Item Response Theory (IRT) and Factor Analysis (FA). A unidimensional test is one that has one latent trait underlying the data (Hattie, 1985). In relation to IRT, a multivariate dataset is said to be unidimensional if it is possible to find a vector of values $\varphi = (\varphi_i)$ such that the probability of correctly answering an item g is $\pi_{ig} = f_g(\varphi_i)$ and local dependence holds for each value of φ . In factor analysis, if only one factor explains a phenomenon, then the data is essentially unidimensional. A

multidimensional test however has two or more distinct latent traits underlying the data. These two concepts will be discussed in detail later in relation to the dimensionality assessment methods.

In factor analysis, latent variables represent unobserved constructs and are referred to as factors or dimensions. In this thesis, only the Confirmatory Factor Analysis (CFA) would be relevant and is therefore reviewed briefly.

Confirmatory Factor Analysis

Suppose that an exploratory factor analysis of data on the indicators X_1, X_2, \dots, X_p yields an m -factor solution given by

$$x_i = \sum_{j=1}^m l_{ij} f_j + \varepsilon_i, \quad i=1, 2, \dots, p \quad (3.1)$$

In Equation (3.1), $m \leq p$ and f_j are the factors specific to the individual indicator x_i with loading l_{ij} on the j th factor. Usually, indicator variables with loading higher than 0.5 are considered influential in the formation of the factor. The m factors model in Equation (3.1) could also be represented as

$$X = \Lambda F + E \quad (3.2)$$

where Λ is the matrix of loadings and F is the vector of specific factors. The correlation matrix R could then be approximated as

$$R = \Lambda \Lambda' + \Psi \quad (3.3)$$

The matrices $\Lambda \Lambda'$ and Ψ are respectively, the reproduced correlation matrix based on the m -factor model and diagonal matrix of specific variances

whose elements are given by $\psi_i = 1 - \sum_{j=1}^m l_{ij}^2, \quad i=1, 2, \dots, p$. Equation (3.3) is

the fundamental factor analysis equation that provides the principle of

hypothesis in Confirmatory Factor Analysis (CFA).

In factor analysis, the number of factors that can be extracted is the same as the number of variables. Each factor i explains a certain amount (λ_i) of overall variance in the observed variables, and the factors are always listed in the order of how much variation they explain. Thus, $\lambda_1 > \lambda_2 > \dots > \lambda_p$.

The test of adequacy of the m -factor model is equivalent to the test of the hypothesis

$$H_0: \rho = \Lambda \Lambda' + \Psi \quad \text{against} \quad H_a: \rho \neq \Lambda \Lambda' + \Psi$$

$p \times p$ $p \times m$ $m \times p$ $p \times p$
 $p \times p$ $p \times m$ $m \times p$ $p \times p$

The null hypothesis means that the m factors are adequate in approximating the original correlation matrix. If H_0 is rejected, it means that the factor model does not significantly represent the underlying dimensions of the correlation matrix. Thus, the alternative hypothesis means that ρ is any other positive definite matrix that cannot be factorized as under H_0 . Under H_0 , the maximum of the likelihood function, with $\hat{\mu} = \bar{x}$ and $\hat{\Sigma} = \hat{\Lambda} \hat{\Lambda}' + \hat{\Psi}$, where $\hat{\Lambda}$ and $\hat{\Psi}$ are the maximum respective likelihood estimates of Λ and Ψ , is proportional to



where
$$S_n = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})(x_j - \bar{x})'$$

By the general likelihood method,

$$\begin{aligned}
 -2 \ln \Lambda &= -2 \ln \left(\frac{\max_{\theta \in \Theta_0} L(\theta)}{\max_{\theta \in \Theta} L(\theta)} \right)^{-\frac{n}{2}} \\
 &= -2 \ln \left(\frac{|\hat{\Sigma}|}{|S_n|} \right)^{-\frac{n}{2}} + n \left[\text{tr}(\hat{\Sigma}^{-1} S_n) - p \right]
 \end{aligned}$$

with degrees of freedom $\nu - \nu_0 = \frac{1}{2}[(p-m)^2 - p - m]$.

Under a maximum likelihood estimate of the parameters in H_0 , $\text{tr}(\hat{R}^{-1} R_n) - p = 0$. Thus, the statistic becomes

$$-2 \ln \Lambda = n \ln \left(\frac{|\hat{\Sigma}|}{|S_n|} \right)^{-\frac{n}{2}} \tag{3.4}$$

Bartlett (1954) shows that the chi-square approximation to the sampling distribution of $-2 \ln \Lambda$ may be improved by replacing n in Equation (3.4) with the multiplicative factor $\left[n - 1 - \frac{1}{6}(2p + 4m + 5) \right]$. The hypothesis H_0 is thus rejected at α level of significance if

$$\left[n - 1 - \frac{1}{6}(2p + 4m + 5) \right] \ln \frac{|\hat{\Lambda} \hat{\Lambda}' + \hat{\psi}|}{|S_n|} > \chi_{\frac{1}{2}[(p-m)^2 - p - m]}^2(\alpha) \tag{3.5}$$

provided n and $(n-p)$ are large.

Remarks 3.2

In this study, it will be observed that several factor models may be significant (i.e., H_0 will not be rejected) for a number of values of m . In this case, it will be sufficient to use the smallest value of m . On the other hand, it is also possible that there will be no value of m at which the H_0 would be

rejected. This means that there are no significant underlying factors, even though exploratory techniques are able to identify factor.

Item Response Theory

The item response theory (IRT), also known as the unobserved response theory, is a collection of mathematical models that attempt to explain the relationship between latent constructs (unobservable characteristics or attributes) and their results (i.e. observed outcomes, responses or performance). They create a link between the characteristics of items on an instrument, the responses of individuals to these items, and the underlying property being measured. The unobserved construct (e.g. stress, knowledge, attitudes) and measure items are organized in a latent continuum, according to IRT.

As a result, its primary application is to determine an individual's place on that continuum. The label item response theory refers to the theory's focus on the item, as opposed to conventional test theory's test-level focus. Meaning, IRT simulates each examinee's response to each test item for a given ability. The term is broad, encompassing a wide range of instructive materials. Multiple choice questions with correct and incorrect answers, popular statements on questionnaires that allow respondents to express their level of agreement (a rating or Likert scale), patient symptoms rated as present or absent, and diagnostic information in complicated systems are examples of these. The premise behind IRT is that the chance of a correct/keyed response to an item is a mathematical function of the individual and item factors. A single unobserved concept or dimension is understood as the person parameter.

The difficulty of an item (known as "location" for its location on the difficulty range); discrimination (slope or correlation), which represents how steeply an individual's rate of success changes with their ability; and a pseudo guessing parameter, which represents the asymptote at which even the least able persons will score due to guessing (for instance, 25 percent probability on a multiple choice item with four possible responses).

In addition, the goal of IRT is to provide a framework for evaluating how effectively exams and particular items on assessments work. IRT is most commonly employed in education, where psychometricians use it to conceive and design examinations, manage groupings of items for examinations, and link item problems for subsequent editions of examinations. Unobserved trait models are another name for IRT models. The word "unobserved" is used to stress that item responses that do not allow fractions are considered observable displays of hypothesized traits or attributes that cannot be tested directly but must be inferred from the visible replies.

Dimensionality in IRT

The IRT viewpoint is important in research because it provides clear stipulations of the correlation between item score variable X_j and the unobserved construct $\theta = [\theta_1 \dots \theta_d]^T$. When $d=1$, the dataset is unidimensional; on the other hand, when $d>1$, the dataset is multidimensional. Both parametric and non-parametric methods have been developed to determine dimensionality-inherent item clusters and evaluate relationships inter and intra these clusters.

Generalized Partial Credit Model (GPCM)

In the context of IRT, parametric dimensionality assessment methods depend on certain Unidimensional IRT models and Multidimensional IRT models for dichotomously scored MC items and polytomously scored CR items. We attempt to describe the models studied in this research. For dichotomous item score variables, the most generic UIRT model in prevalent operative use is the three-parameter logistic (3PL) model (Lord, 1980), whose Item Response Function is given by

$$p_j = Pr(X_j = 1 | \theta) = c_j + (1 - c_j) \exp \left(\frac{a_j \theta - b_j}{1 + \exp(a_j \theta + d_j)} \right) \quad (3.6)$$

In the equation 3.6, θ is the single unobserved trait; a_j , b_j , and c_j denote the item discrimination, difficulty, and pseudo-guessing parameters for item j , respectively. The 3-Parameter model becomes the two-parameter logistic (2PL) model by setting $c_j = 0$, and further becomes the one-parameter logistic (1PL) model by setting $a_j = 1$ and $c_j = 0$. A multidimensional modification of the 3PL (M3PL) model (Reckase, 2009) is given by

$$p_j = Pr(X_j = 1 | \theta) = c_j + (1 - c_j) \frac{\exp(a_j \theta + d_j)}{1 + \exp(a_j \theta + d_j)} \quad (3.7)$$

where multiple unobserved constructs contained in θ together determine how probable a randomly chosen examinee answers item j correctly, a_j is a transposed vector of slope parameters, d_j is the intercept, and c_j is the pseudo-guessing parameter. According to Reckase (2009), the multidimensional discrimination (MDISC) for item j is defined as

$$MDISC_j = \sqrt{a_j a_j} \quad (3.8)$$

which is equivalent to a_j in Equation (3.6). The multidimensional difficulty (MDIFF), which is analogous to b_j in Equation (3.6), is given by

$$MDIFF_j = \frac{-d_j}{MDISC_j} = \frac{-d_j}{\sqrt{a_j a_j}} \quad (3.9)$$

Equation (3.7) further highlights the model's compensatory nature: a low value on one unseen construct could be offset by high value(s) on at least one other unobserved construct. Compensatory models have been frequently employed in the dimensionality assessment field because they appear to better mimic the genuine intellectual processes observed in many educational examinations and do not provide significant theoretical or computational challenges (e.g., Hattie, 1984; Mroch & Bolt, 2006; Nandakumar, 1994; van Abswoude et al., 2004). The graded response (GR) model (Samejima, 1969) for the unidimensional case begins with the outline of the cumulative category response functions for polytomous item score variables.

$$p_{j0}^{\zeta} = Pr(X_j \geq 0 | \theta) = \zeta \quad p_{jk}^{\zeta} = Pr(X_j \geq k | \theta) = \zeta \exp(-\zeta b_{jk}), \quad k = 1, \dots, K_j - 1 \quad (3.91)$$

The parameters in Equation (3.9.1)) are like those in the 2PL model, the only difference being the difficulty parameter b_{jk} is assigned to response categories from 1 to $(K_j - 1)$. The IRF of the GR model is defined as the disparity amongst cumulative category response functions:

$$p_{jk}^{\square} = Pr(X_j = k | \theta) = \zeta (p_{jk}^{\zeta} - p_{j(k+1)}^{\zeta}), \quad k = 0, \dots, K_j - 2$$

$$p_{j(K_j-1)} = Pr(X_j = K_j - 1 | \theta) = p_{j(K_j-1)}^{\zeta}, \quad (3.92)$$

Essential Dimensionality

In this study, the non-parametric IRT approach was used in accordance with the concept of essential dimensionality. The definition of critical dimensionality for dichotomous variables was discussed informally by Stout (1987). In Stout, this was formalized (1990). Junker (1991) expanded the notion to include polytomous variables. Similar ideas have been presented discreetly from an EFA perspective; for example, factors responsible for the majority of the observed association are preserved while the other minor factors are ignored (Stout, 1990). "A construct vector is dominating if the residual covariances among the items are minimal in anticipated value after conditioning on," to put it another way (Junker, 1991, p. 258). The number of dominating dimensions is the emphasis of essential dimensionality. If only one dominant dimension is visible, data is believed to be essentially unidimensional.

Methods Using the IRT Definition of Dimensionality

Poly-DIMTEST (Nandakumar et al., 1998) is an extension of DIMTEST (Stout, 1987; Nandakumar & Stout, 1993) to study unidimensionality of polytomous data. This approach tests H_0 of the essential unidimensionality, given by

$$H_0 : d_E = 1 \text{ versus } H_1 : d_E > 1$$

where d_E is the number of dominant dimensions defined in Stout (1987, 1990). Because it assumes the IRT notion of dimensionality (Stout, 1987, 1990) as mentioned previously, the Poly-DIMTEST approach was recognized as an IRT approach in this investigation.

Poly DIMTEST captures numerous phases and is quite similar to DIMTEST with a few minor differences (for technical details, refer to Nandakumar et al., 1998). Examinees are given a test of items based on the notation described earlier. AT1, AT2, and PT are the three categories that the initial test is broken into. M items are found in both AT1 and AT2, while the remaining (J-2M) items cover PT, where M is a tiny number. According to Nandakumar and Stout (1993)'s DIMTEST recommendations, four or more AT components are required for reliable estimations, however it is better when it goes below J/4.

To give adequate power, the number of PT pieces should be at least 15. (Stout et al., 1996). AT1 items are expected to measure the same unobserved construct as AT2 items, and AT2 items are expected to have an item difficulty distribution similar to AT1 items. AT1 items could be identified at random or using EFA, indicating two Poly-DIMTEST modes: confirmatory and exploratory. Items measuring the same content subdomain or having the same item format could be designated as AT1 items in a confirmatory mode. An algorithm proposed by Nandakumar and Stout (1993) could be used to choose AT1 items in an exploratory mode. They suggested that on the second factor, items with substantial absolute loadings (e.g., more than 0.15) be used. When running a Poly-DIMTEST exploratory model, it is also a good idea to divide the initial sample into two categories. EFA is conducted in the first category, and the Stout's T statistic is computed in the second. The first category should have a minimum of 500 people (Nandakumar, 1994). When computing the T statistic, the examinees' category may be divided into many subcategories based on their PT item scores. Under H_0 , the T statistic approximates the

conventional normal distribution (Nandakumar et al., 1998). If the P-value falls below a predetermined significant level, the basic unidimensionality assumption is invalidated.

Proposed Method on the Kaiser-Meier-Olkin's Measure of Sampling Adequacy

In this thesis, we offer a computationally robust methodology for employing Kaiser-Meier Olkin's Measure of Sampling Adequacy (KMO) as a dimensionality identification method in a multivariate dataset. It first investigates a systematic strategy for determining the dataset's initial dimensionality. It then divides the variables into two groups: those that do not contribute to any dimension (non-homogeneous sets) and those that contribute to many dimensions (multidimensional sets). According to the literature, a KMO value of 0.6–1.0 is a natural good measure (Rencher, 2002; Nkansah, 2018).

KMO values less than 0.6, on the other hand, indicate that the sample is unsuitable, and that corrective action should be performed. The suitability of a sample is traditionally determined by four factors.

1. The sample's representativeness
2. Sample size
3. Variability in the population
4. Estimation precision that is desired

KMO, which is a measure of similarity between variables, is used to determine the acceptability of a sample rather than looking into each of the features individually. To determine the suitability of a dataset for dimensionality detection, a number of approaches are used. The Kaiser-Meier-Measure Olkin's of Sampling Adequacy (abbreviated KMO) is a commonly used method. It's a detection metric for determining the degree to which a dimension's indicators are homogeneous.

A low KMO score indicates that the connection between the two variables cannot be explained by a well-defined unseen factor, and hence dimensionality detection may not be appropriate.

Table 1: A Guide for Interpreting KMO Measure

KMO Measure	Recommendation
≥ 0.90	Marvellous
0.80+	Meritorious
0.70+	Meddling
0.60+	Mediocre
0.50+	Miserable
≤ 0.50	Unacceptable

Source: Nkansah (2018)

According to the criterion in Table 1, the overall KMO measure should be 0.8 or higher to achieve acceptable results. Although a value of greater than 0.6 is permissible, this rule of thumb appears to have gained widespread acceptance (Rencher, 2002). The index allows for a comparison of the size of observed correlation coefficients versus partial correlation coefficients. The KMO can be calculated using the following equation.

$$KMO = \frac{\sum_{i < j} r_{ij}^2}{\sum_{i < j} r_{ij}^2 + \sum_{i < j} pr_{ij}^2} \tag{3.93}$$

where r_{ij}^2 is the square of the correlation coefficient between anypairs 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

Order Statistics Correlation Coefficient

Definition and Properties

Let $(x_{ij}, x_{kj}), j=1, 2, \dots, n; i, k=1, 2, \dots, p$ be n observations on any two variables from the set (X_1, X_2, \dots, X_p) . By rearranging pairwise 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 $x_{i(1)} \leq x_{i(2)} \leq \dots \leq x_{i(n)}$ are the order-statistics of x_i and $x_{k[1]}, x_{k[2]}, \dots, x_{k[n]}$ are the associated concomitants of x_k . Re-versing the roles of x and y , we also define the order statistics of y and the corresponding concomitants which are denoted by y_{1, \dots, y_N} and x_{1, \dots, x_N} respectively. The order statistics correlation coefficient can be defined as

$$r_x(x, y) \triangleq \sum_{i=1}^N i i i \tag{3.94}$$

The order statistics correlation coefficient has the basic properties of a correlation coefficient, as follows

$$1. -1 \leq r_x \leq 1$$

2. $r_x(x, y)$ attains $+1(-1)$, where $x \wedge y$ are in strict increasing(decreasing) relationship

3. $r_x(x', y') = r_x(x, y)$ for $x' = k_x x + const_x \wedge y' = k_y y + const_y, k_x > 0 \wedge k_y > 0$

4. If x and y are mutually independent and each is independent identically distributed (IID), the expectation $E[r_x(x', y')] = 0$ as $N \rightarrow \infty$

Chapter Summary

In this chapter, we reviewed the known statistical techniques that are used to interpret multivariate data. Notably, factor analysis, Principal Component Analysis and Item Response theory Modelling. The general orthogonal factor model and underlying basic conditions and assumptions were discussed.

The methods of principal components and maximum likelihood have been discussed and their relative desirable properties have been pointed out. The 3-parameter IRT model has been critically examined. Confirmatory factor analysis which uses the Chi square statistic to assess model fit has been discussed with the purpose of using it to assess model fit in chapter four. Unidimensionality and multidimensionality have also been discussed. The generalisation of the dimensionality detection algorithm and a measure of homogeneity, KMO have been thoroughly explored.

CHAPTER FOUR

RESULTS AND DISCUSSION

Introduction

This chapter presents the development and implementation of the algorithms based on methodology for addressing the detection of dimensionality in data. Implementation is carried out on two existing datasets described in Chapter One and one simulated data. The proposed dimensionality detection methods are similarity measures which hinges 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. A brief preview of the structure and key findings for each of Datasets 1 and 2 obtained in Nkansah (2018) is presented as follows:

Table 2: Correlation Matrix for Dataset 1

	x_1	x_2	x_3	x_4	x_5	x_6	
x_2		0.926					
x_3		0.844	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

Source: Nkansah (2018)

Using an experimenter specific threshold of 0.5, only one dimension is detected to underlie the dataset since only one homogenous set is found constituted by five variables given by $S_1 = \{x_1, x_2, x_3, x_5, x_7\}$. The KMO values for Dataset 1 based on the methodology described in Chapter Three are summarised in Table 3.

Table 3: KMO for Dataset 1 Based on Subgroupings

SN	Groupings	KMO Value
1	All	0.6161
2	S_1 only	0.6413

Source: Nkansah (2018)

Table 4: Correlation Matrix for 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

Source: Nkansah (2018)

The study detected dimensionality for this data set using an experimenter specific threshold. It identified that two dimensions underlie the data set since there were two homogenous sets given by

$S_1 = \{x_3, x_7, x_6\} \wedge S_2 = \{x_2, x_9, x_4\}$. The KMO values for dataset 2 are summarised in the table below.

Table 5: KMO Values for Dataset 2 from Literature

SN	Groupings	KMO
1	All	0.8222
2	S_1 only	0.8503
3	S_2 only	0.8365

Source: Nkansah (2018)

Remarks 4.1

The drawbacks in the study described above are specified as follows:

1. Calculation of KMO dwelling on the original correlation structure and not a spanning set may lead to abuse of information and misleading results since KMO depends on the original correlation structure.
2. KMO determination is based on experimenter specific thresholds and not thresholds based on the data structure.
3. Computation is expensive since the duration involved in calculating KMO using the original correlation structure far outweighs the duration regarding the computation based on a spanning set.
4. The method does not investigate the sensitivity and robustness based on the correlation profile as only the Pearson's correlation is used.

In this study, the proposal to address the drawbacks identified above are as follows:

1. Develop an approach to determine KMO based on a threshold not influenced by the choice of the experimenter but by the structure of the data
2. Develop an approach for calculating KMO based on a spanning set, not the original correlation structure, to provide reliable results.
3. Develop an approach that is robust to correlation structure

It is anticipated that the implementation of these proposals would lead to a computationally less expensive approach for calculating a measure of homogeneity of a dataset.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy based on Pearson's Correlation Profile

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. Given a pair of random variables (X, Y) , the Pearson correlation coefficient, ρ , is given by

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Pearson's correlation coefficient, when applied to a sample, is commonly represented by r_{xy} .

For a given paired data $(x_1, y_1), \dots, (x_n, y_n)$,

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

(4.1)



The Kaiser-Meyer-Olkin Measure of Sampling Adequacy (or simply KMO) is a diagnostic measure for assessing the extent to which the indicators of a dimension belong together. The KMO is given by

$$KMO = \frac{\sum_{i < j} r_{ij}^2}{\sum_{i < j} r_{ij}^2 + \sum_{i < j} pr_{ij}^2} \quad (4.2)$$

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.

For the purpose of the approach adopted in this study, the KMO may be expressed in the form

$$KMO = \frac{1}{1 + \frac{\sum_{i < j} pr_{ij}^2}{\sum_{i < j} r_{ij}^2}} \quad (4.3)$$

In an attempt to examine the ratio of the partial correlation to zero order correlation. For KMO to be large, the ratio of partial correlation to zero order correlation must be small. We need to form the homogeneous set in such a way that the variables in a set correlate highly with each other but not with members of the other set. For this homogenous set we expect the ratio to be small for any pair of elements controlling for all others

An Automated Dimensionality Detection

In this section, an updated version of the generalized rule (Nkansah, 2018) is described that is based on an automated threshold. Subsequently, a description of methods for determining the threshold is provided.

Generalization of the Modified Rule for Determining Expected Dimensions of Datasets

Suppose the dataset is generated on a set of p variables y_1, y_2, \dots, y_p with correlation coefficients that are generally significant. We generate a data specific threshold δ_0 using an automated threshold setting for which variables may be considered to belong together if their pairwise correlation coefficients exceed δ_0 . First take the pair (i, j) , $i, j \in I = (1, 2, \dots, p)$ with the highest correlation coefficient. Let this pair be (y_m, y_n) and label the set as $S_1 = (y_m, y_n)$ and the index as $I_1 = (m, n)$. If the correlation coefficient $r_{y_k, y_i} \geq \delta_0$, $\forall k \in I_1, i \in I \setminus I_1$, then $y_i \in S_1$, otherwise $y_i \notin S_1$. The sets S_1 and I_1 are updated each time. Now if $r_{y_k, y_i} < \delta_0$ for some $k \in I_1$ and some $i \in I \setminus I_1$, then we obtain the first homogeneous set $S_1 = (y_{i_1}, y_{i_2}, \dots, y_{i_{g_1}})$ with index $I_1 = (i_1, i_2, \dots, i_{g_1}) \subset I$.

We would form a new set S_2 from the elements $y_i \notin S_1, i \in I \setminus I_1$. Denote $T_1 = I \setminus I_1$. Consider the pair (i, j) , $i, j \in T_1$ with the highest correlation that meets the cut off value δ_0 . This pair is (y_{i_1}, y_{i_2}) . Thus we obtain the second pair $S_2 = (y_{i_1}, y_{i_2})$ and an index set $I_2 = (I_1, I_2)$. Now if the correlation coefficient $r_{y_k, y_f} \geq \delta_0 \quad \forall k \in I_2, i \in I \setminus I_2$, then $y_i \in S_1$, otherwise $y_i \notin S_1$. The sets S_2 and I_2 are updated each time. Now if $r_{y_k, y_f} < \delta_0$ for some $k \in I_2$ and some $i \in I \setminus I_2$ then we obtain a final second homogeneous set $S_1 = (y_{i_1}, y_{i_2}, \dots, y_{i_{g_2}})$ with index $I_1 = (i_1, i_2, \dots, i_{g_2}) \subset I$.

Consider all elements $y_i \notin (S_1 \cup S_2), i \in I \setminus (I_1 \cup I_2)$. Denote $T_2 = I \setminus (I_1 \cup I_2)$. To form a new set, take the pair (i, j) , $i, j \in T_2$ with the highest

correlation coefficient that meets the cut off, δ_0 . Let the pair be (i, j) . Thus, we obtain the third set $S_3 = (y_{i_1}, y_{i_2}, \dots, y_{i_{g_3}})$ and an index $I_3 = (i_1, i_2, \dots, i_{g_3}) \subset I$.

If the correlation coefficient $r_{y_k, y_l} \geq \delta_0$, $\forall k \in I_3, l \in I \setminus I_3$, then $y_l \in S_3$, otherwise $y_l \notin S_3$. The sets S_3 and I_3 are updated each time. Now if $r_{y_k, y_l} < \delta_0$ for some $k \in I_3$ and some $l \in I \setminus I_3$, then we obtain the third homogenous set $S_3 = (y_{i_1}, y_{i_2}, \dots, y_{i_{g_3}})$ with index $I_3 = (i_1, i_2, \dots, i_{g_3}) \subset I$.

We attempt to form the q th set S_q from the elements $y_i \notin (i, k=1 \dots q-1 S_k)$, $i \in I \setminus (i, k=1 \dots q-1 I_k)$. Denote $T_{q-1} = (i, k=1 \dots q-1 I_k)$. Take the pair $(i, j), i, j \in T_{q-1}$ with the highest correlation coefficient that meets the threshold value δ_0 .

Thus, we obtain $S_q = (y_{d_1}, y_{d_2}, \dots, y_{d_{g_q}})$ and an index set $I_q = (d_1, d_2, \dots, d_{g_q}) \subset I$. Now if $r_{y_k, y_l} < \delta_0$ for some $k \in I_q$ and some $l \in I \setminus I_q$, then we obtain the q th

homogenous set $S_q = (y_{d_1}, y_{d_2}, \dots, y_{d_{g_q}})$ with index $I_q = (d_1, d_2, \dots, d_{g_q}) \subset I$.

If for some set S_{i+1} and index set I_{i+1} and for $y_i \notin (i, k=1 \dots l S_k)$,

$r_{y_k, y_l} < \delta_0$ for all $i, j \in I \setminus (i, k=1 \dots l I_k)$ then S_{i+1} is the last set of variables in the original set of p variables and there are a total of l dimensions underlying the correlation matrix.

Similarity Based Dimensionality Detector

Consider a p variate random variable, $Y = (Y_1, Y_2, \dots, Y_p)$ on which an $n \times p$ data set is observed on a given multivariate system. Let C_Y denote a $p \times p$ matrix of pairwise similarity measure based on Y . We assume that without knowledge of the appropriate number of dimensions underlying the data

under consideration, an appropriate guess of the number of dimensions will be at most the number of variables defining the data. It is also possible that the variables are inter-related in some sense which may be simple or complex. This relationship may be informative about the intrinsic dimension underlying the data and thus, an appealing source for building a dimensionality detection scheme for detecting dimension in such data. The similarity-based algorithm for detecting dimension is outlined in algorithm1, 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 $Y_i, i = 1, \dots, p$.

Algorithm 2: Similarity-based Dimensionality Detector

Initialization: Data: $Y = Y_1, Y_2, \dots, Y_p$. Set threshold, $\delta = \delta_0$

Compute similarity matrix, $C_Y = \tau(Y) = \tau(Y_1, \dots, Y_p)$.

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. $D_{Y_k, S_f} = D_{Y_k, Y_i}, D_{Y_k, Y_j}, Y_k \in Y^*$

2. if $D_{Y_k, S_f} \geq \delta_0$

- $S_f = \{(Y_i, Y_j, Y_k)\}$

- $m_f = N(S_f)$

- $H_s = S_f$

- $Y^* = Y^* \setminus Y_k$

- $n^* = p - m_f$

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

Explanation of Dimensionality Detection Algorithm Procedure

The algorithm initializes by generating the variance-covariance matrix for p variables. Since Pearson's correlation is symmetric, the lower triangular matrix is used. The highest pairwise correlation in the variance-covariance matrix is selected and the associated variables constitute the spanning set S_f . All variables in the reduced matrix whose pairwise correlations are at least the threshold δ_0 are used to update the spanning set giving us the first homogeneous set S_1 . The process is repeated until all possible homogenous

sets are formed. The algorithm terminates when the number of variables in the reduced dataset n^* is at most two. The number of homogeneous sets gives us the number of dimensions underlying the dataset for each threshold.

Automated Threshold Setting

The use of threshold is primal in dimensionality detection as 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 highly likely that a single threshold may generate multiple homogeneous sets.

Mathematical Background

$$\delta_1 = [i, \dots, a_n]$$

$$\delta_2 = [i, \dots, a_n]$$

$$\delta_3 = [\delta_1 \geq \tilde{\delta}_1]$$

$$\alpha_2 = \frac{a_n - a_1}{k\delta}$$

$$\text{where } \alpha_1 = 0.01, \alpha_2 = \frac{a_n - a_1}{k\delta}, a_1 = \min(D_Y)$$

$$a_n = \max(D_Y)$$

$$\tilde{\delta}_1 = \text{median of } \delta_1$$

We set $k\delta = 12$

Automated Threshold Setting Algorithm Procedures

Automated Threshold Setting 1 (δ_1)

The algorithm picks the lowest pairwise correlation in the variance - covariance matrix, generates series of thresholds using a step value of 0.01 until all thresholds in the variance -covariance matrix are accommodated. This

generates a total of 80 thresholds. The dimensionality detection algorithm is then used to generate homogeneous sets for each of these 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. Sensitivity analysis is then carried out for these thresholds in an attempt to pick the optimal threshold suitable for dimensionality detection for the dataset.

Automated Threshold Setting Two (δ_2)

The correlation profile used is the Pearson's correlation which is normally distributed. Statistically majority (about 97%) of the data points lie 3 standard deviations about the mean. This gives us 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 variance-covariance matrix to the resultant standard deviation to generate series of thresholds. This generated 38 thresholds. The dimensionality detection algorithm is then used to generate homogeneous sets for each of these thresholds. Since multidimensionality is expected some thresholds yielded more than one homogeneous set. The KMO values are then calculated for each homogeneous set that corresponds to each threshold. Sensitivity analysis is then carried out for these thresholds in an attempt to pick the optimal threshold suitable for dimensionality detection for the data set.

Automated Threshold Setting Three (δ_3)

This procedure is based on Threshold Setting 1. Statistically the variance-covariance matrix used which hinges on Pearson's correlation is

symmetric. This makes the data points normally distributed. In view of this the algorithm determines the median for thresholds generated using automated threshold setting 1 and selects those thresholds that are at least the median. This is similar to the usage of the lower triangular matrix of the variance covariance matrix. The dimensionality detection algorithm is then used to generate homogeneous sets for each of these thresholds. Since multidimensionality is expected, some thresholds yield more than one homogeneous set. The KMO values are then calculated for each homogeneous set that corresponds to each threshold. Sensitivity analysis is then carried out for these thresholds in an attempt to pick the optimal threshold suitable for dimensionality detection for the dataset.

Table 6: Dimensionality Detection for Dataset 1

SN/Threshold	No of hom. Sets	SN/KMO
[1] 0.15	1	[1] 0.6995
[2] 0.16	1	[[2]] 0.6995
[3] 0.17	1	[[3]] 0.6995
[4] 0.18	1	[[4]] 0.6995
[5] 0.19	1	[[5]] 0.6995
[6] 0.20	1	[[6]] 0.6995
[7] 0.21	1	[[7]] 0.6995
[8] 0.22	1	[[8]] 0.6995
[9] 0.23	1	[[9]] 0.6995
[10] 0.24	1	[[10]]0.6995
[11] 0.25	1	[[11]] 0.6995
[12] 0.26	1	[[12]] 0.6995
[13] 0.27	1	[[13]] 0.6995
[14] 0.28	1	[[14]] 0.6995
[15] 0.29	1	[[15]]] 0.6995
[16] 0.30	1	[[16]] 0.6995
[17] 0.31	1	[[17]]0.6995

[18] 0.32	1	[[18]] 0.6995
[19] 0.33	1	[[19]] 0.6995
[20] 0.34	1	[[20]] 0.6995
[21] 0.35	1	[[21]] 0.6995
[22] 0.36	1	[[22]] 0.6995
[23] 0.37	1	[[23]] 0.6995
[24] 0.38	1	[[24]] 0.6995
[25] 0.39	1	[[25]] 0.6995
[26] 0.40	1	[[26]] 0.6995
[27] 0.41	1	[[27]] 0.6995
[28] 0.42	1	[[28]] 0.7571
[29] 0.43	1	[[29]] 0.7571
[30] 0.44	1	[[30]] 0.7571
[31] 0.45	1	[[31]] 0.7571
[32] 0.46	1	[[32]] 0.7571
[33] 0.47	1	[[33]] 0.7571
[34] 0.48	1	[[34]] 0.7571
[35] 0.49	1	[[35]] 0.7571
[36] 0.50	1	[[36]] 0.7571
[37] 0.51	1	[[37]] 0.7571
[38] 0.52	1	[[38]] 0.7571
[39] 0.53	1	[[39]] 0.7571
[40] 0.54	1	[[40]] 0.7571
[41] 0.55	1	[[41]] 0.7571
[42] 0.56	1	[[42]] 0.7571

[43] 0.57	1	[[43]] 0.7571	
[44] 0.58	2	[[44]] 0.8587	0.5000
[45] 0.59	2	[[45]] 0.8587	0.5000
[46] 0.60	2	[[46]] 0.8587	0.5000
[47] 0.61	2	[[47]] 0.8587	0.5000
[48] 0.62	2	[[48]] 0.8587	0.5000
[49,] 0.63	2	[[49]] 0.8587	0.5000
[50] 0.64	2	[[50]]0.8587	0.5000
[51] 0.65	2	[[51]] 0.8587	0.5000
[52] 0.66	2	[[52]] 0.8587	0.5000
[53] 0.67	2	[[53]] 0.8587	0.5000
[54] 0.68	2	[[54]] 0.8587	0.5000
[55] 0.69	2	[[55]] 0.8587	0.5000
[56] 0.70	2	[[56]] 0.8587	0.5000
[57] 0.71	2	[[57]] 0.8587	0.5000
[58] 0.72	2	[[58]] 0.8587	0.5000
[59] 0.73	2	[[59]] 0.8587	0.5000
[60] 0.74	2	[[60]] 0.8587	0.5000
[61] 0.75	2	[[61]] 0.8587	0.5000
[62] 0.76	2	[[62]]0.8587	0.5000
[63] 0.77	2	[[63]]] 0.8587	0.5000
[64] 0.78	2	[[64]] 0.8587	0.5000
[65] 0.79	2	[[65]] 0.8587	0.5000
[66] 0.80	2	[[66]] 0.8587	0.5000
[67] 0.81	2	75[[67]] 0.8587	0.5000
[68] 0.82	2	[[68]]0.8587	0.5000
[69] 0.83	2	[[69]] 0.8587	0.5000

Source: Author's Construct (2022)

From Table 6, it is observed that the automated threshold setting which allows the dataset to generate its own threshold is used to generate series of thresholds, thus 80 in this case. Though some thresholds may appear small, it is not in our power to determine the thresholds generated by the algorithm. Our mandate is to allow the algorithm detect the optimal threshold that is suitable for dimensionality detection for this dataset. Subsequently the dimensionality detection algorithm generates homogenous set(s) for each threshold. It is possible to have either a single homogeneous set giving one dimension, two homogeneous sets giving us two dimensions and so on as we outlined early on in Chapter One that it is possible for a multivariate dataset to be unidimensional meaning we have only one latent construct underlying the data and also multidimensional meaning we have two or more latent traits explaining the phenomenon being studied . The modified automated KMO algorithm is then used to calculate the KMO for each homogeneous set. It is observed that if there are multiple homogenous sets for a particular threshold, the algorithm calculates the KMO for each homogenous set. For instance, for threshold 0.92, resulting in two homogeneous sets, the algorithm calculates the KMO for each homogeneous set namely 0.7848 and 0.5 respectively. Since KMO is a measure of homogeneity the higher it is the better. So in a case where we have multiple KMO we pick the highest .We then graph the KMO against the thresholds to determine either a unique optimal threshold or a saturation point if any.

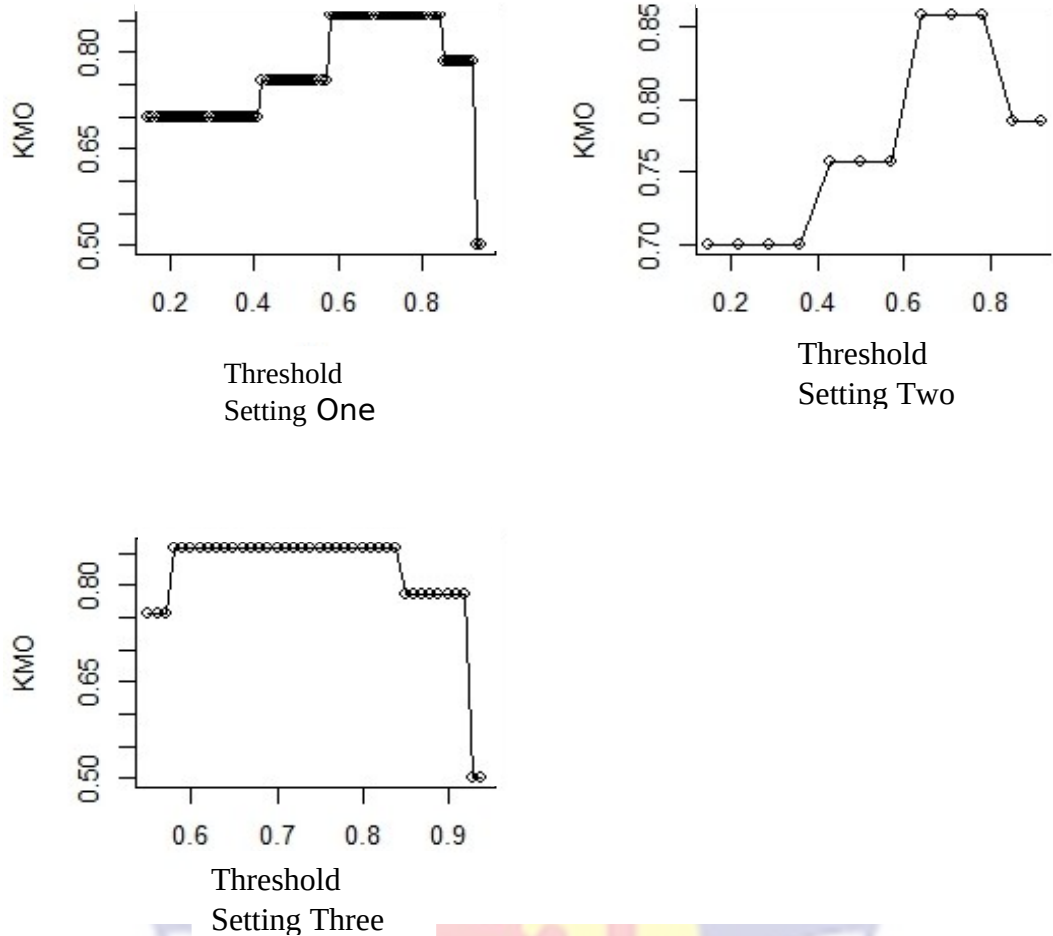


Figure 1: Sensitivity Analysis for Dataset 1

The graphs in Figure 1 depict the plot of KMOs against the various thresholds. Though we have a couple of saturation points, our interest is the one that corresponds to the highest KMO. It could be observed that the saturation points for all three graphs corresponding to the highest KMO lie within 0.6 and 0.85 inclusive. This means any threshold within this range could be the optimal threshold for dimensionality detection for this dataset.

Selecting an Optimal Threshold from the Saturation Point

A resultant saturation point indicates that any threshold within this range is suitable for dimensionality detection for the given dataset. However, each threshold within this range could give different factor solutions when the

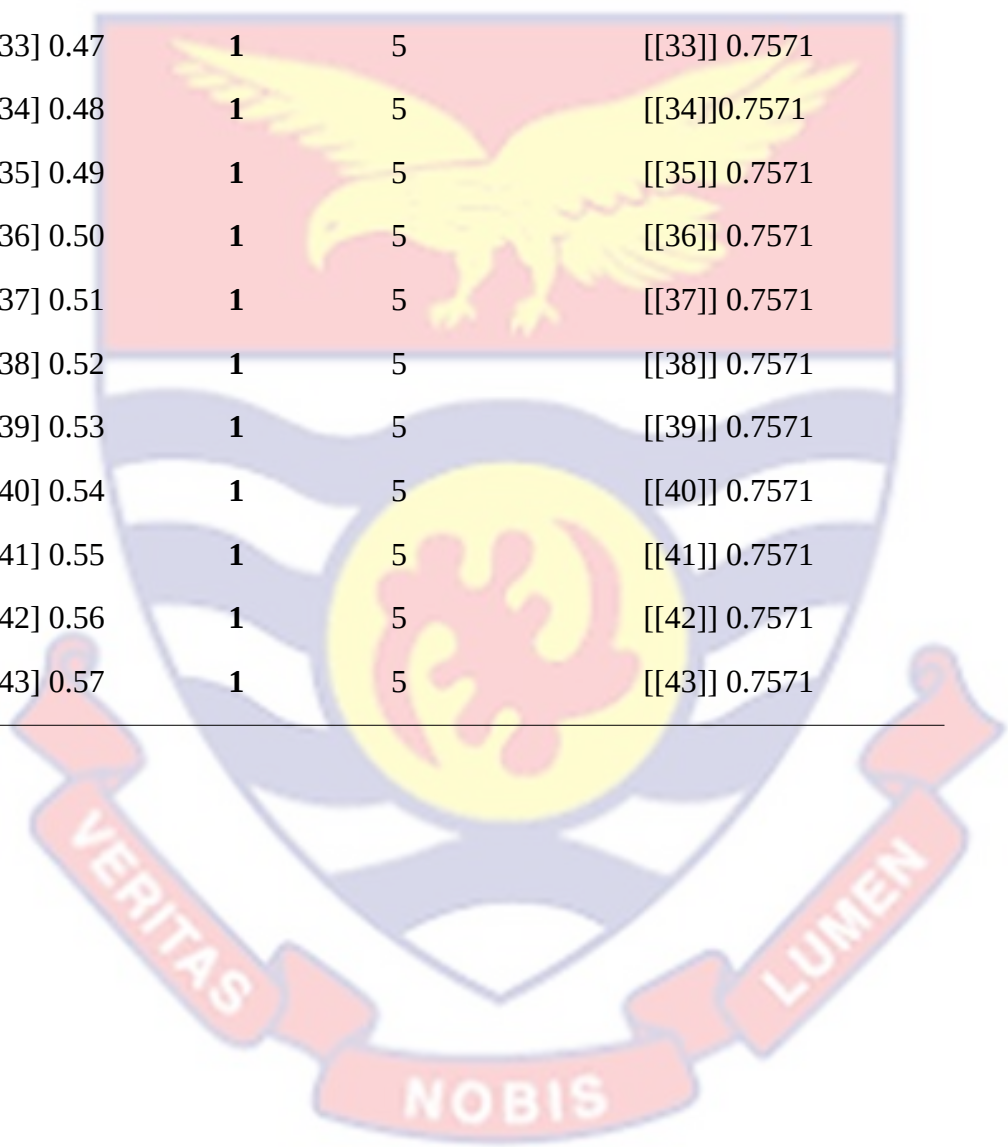
multivariate dataset is subjected to factor analysis. Consequently we should be able to select a threshold that generates the optimal factor solution. The dimensionality detection outlined in this research hinges on KMO which is calculated for each homogeneous set. We observed that the larger the number of variables in a homogeneous set, the higher the KMO. This suggests that the threshold within the saturation point which accounts for the homogeneous set with the highest number of variables and a corresponding highest KMO should be the threshold that generates the optimal factor solution. We attempt to investigate this in the next table.



Table 7: Threshold Selection for Optimal Factor Solution

SN/Threshold	No. of hom. Sets	No. of variables in hom. Set	KMO
[1] 0.15	1	6	[1] 0.6995
[2] 0.16	1	6	[[2]] 0.6995
[3] 0.17	1	6	[[3]] 0.6995
[4] 0.18	1	6	[[4]] 0.6995
[5] 0.19	1	6	[[5]] 0.6995
[6] 0.20	1	6	[[6]] 0.6995
[7] 0.21	1	6	[[7]] 0.6995
[8] 0.22	1	6	[[8]] 0.6995
[9] 0.23	1	6	[[9]] 0.6995
[10] 0.24	1	6	[[10]]0.6995
[11] 0.25	1	6	[[11]] 0.6995
[12] 0.26	1	6	[[12]] 0.6995
[13] 0.27	1	6	[[13]] 0.6995
[14] 0.28	1	6	[[14]] 0.6995
[15] 0.29	1	6	[[15]]] 0.6995
[16] 0.30	1	6	[[16]] 0.6995
[17] 0.31	1	6	[[17]]]0.6995
[18] 0.32	1	6	[[18]] 0.6995
[19] 0.33	1	6	[[19]] 0.6995
[20] 0.34	1	6	[[20]] 0.6995
[21] 0.35	1	6	[[21]] 0.6995
[22] 0.36	1	6	[[22]] 0.6995
[23] 0.37	1	6	[[23]] 0.6995
[24] 0.38	1	6	[[24]] 0.6995
[25] 0.39	1	6	[[25]] 0.6995
[26] 0.40	1	6	[[26]] 0.6995

[27] 0.41	1	5	[[27]] 0.6995
[28] 0.42	1	5	[[28]] 0.7571
[29] 0.43	1	5	[[29]] 0.7571
[30] 0.44	1	5	[[30]] 0.7571
[31] 0.45	1	5	[[31]] 0.7571
[32] 0.46	1	5	[[32]] 0.7571
[33] 0.47	1	5	[[33]] 0.7571
[34] 0.48	1	5	[[34]] 0.7571
[35] 0.49	1	5	[[35]] 0.7571
[36] 0.50	1	5	[[36]] 0.7571
[37] 0.51	1	5	[[37]] 0.7571
[38] 0.52	1	5	[[38]] 0.7571
[39] 0.53	1	5	[[39]] 0.7571
[40] 0.54	1	5	[[40]] 0.7571
[41] 0.55	1	5	[[41]] 0.7571
[42] 0.56	1	5	[[42]] 0.7571
[43] 0.57	1	5	[[43]] 0.7571



©University of Cape Coast <https://ir.ucc.edu.gh/xmlui>

[44]	0.58	2	4, 2	[[44]]	0.8587	0.5000
[45]	0.59	2	4, 2	[[45]]	0.8587	0.5000
[46]	0.60	2	4, 2	[[46]]	0.8587	0.5000
[47]	0.61	2	4, 2	[[47]]	0.8587	0.5000
[48]	0.62	2	4, 2	[[48]]	0.8587	0.5000
[49,]	0.63	2	4, 2	[[49]]	0.8587	0.5000
[50]	0.64	2	4, 2	[[50]]	0.8587	0.5000
[51]	0.65	2	4, 2	[[51]]	0.8587	0.5000
[52]	0.66	2	4, 2	[[52]]	0.8587	0.5000
[53]	0.67	2	4, 2	[[53]]	0.8587	0.5000
[54]	0.68	2	4, 2	[[54]]	0.8587	0.5000
[55]	0.69	2	4, 2	[[55]]	0.8587	0.5000
[56]	0.70	2	4, 2	[[56]]	0.8587	0.5000
[57]	0.71	2	4, 2	[[57]]	0.8587	0.5000
[58]	0.72	2	4, 2	[[58]]	0.8587	0.5000
[59]	0.73	2	4, 2	[[59]]	0.8587	0.5000
[60]	0.74	2	4, 2	[[60]]	0.8587	0.5000
[61]	0.75	2	4, 2	[[61]]	0.8587	0.5000
[62]	0.76	2	4, 2	[[62]]	0.8587	0.5000
[63]	0.77	2	4, 2	[[63]]	0.8587	0.5000
[64]	0.78	2	4, 2	[[64]]	0.8587	0.5000
[65]	0.79	2	4, 2	[[65]]	0.8587	0.5000
[66]	0.80	2	4, 2	[[66]]	0.8587	0.5000
[67]	0.81	2	4, 2	[[67]]	0.8587	0.5000
[68]	0.82	2	4, 2	[[68]]	0.8587	0.5000
[69]	0.83	2	4, 2	[[69]]	0.8587	0.5000
[70]	0.84	2	3, 2	[[70]]	0.8587	0.5000
[71]	0.85	2	3, 2	[[71]]	0.7848	0.5000
[72]	0.86	2	3, 2	[[72]]	0.7848	0.5000
[73]	0.87	2	3, 2	[[73]]	0.7848	0.5000



Source: Author's Construct (2022)

As opposed to the earlier assertion that the threshold that accounts for the optimal factor solution should be the threshold with the homogeneous set that has the highest number of variables and a corresponding highest KMO, from the Table 7, it could be observed that all thresholds within the saturation point have homogeneous sets with the same number of variables but not a unique highest KMO. This suggests that it is possible that none of these thresholds could generate an optimal factor solution. An attempt is subsequently made to show whether it is possible to use any of these thresholds to generate an optimal factor solution for these data by carrying out a confirmatory factor analysis.

Confirmatory Test of Model Adequacy for Dataset 1

In this dataset, we first test the adequacy of the one-, two- and three factor models equivalent to one, two and three dimensions to determine which is the most suitable. As indicated earlier by our assessment, we suspect that each threshold within the saturation point could yield different factor solutions. Our aim here is to conduct a confirmatory factor analysis to justify our assertions.

Table 8: Significance test of Factor Solutions 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

Source: Author's Construct (2022)

Since our highest number of dimensions for the data does not exceed three, it suggests a 3-factor solution. The confirmatory factor analysis is therefore carried out for a maximum of three factors. In Table 8, no specific factor solution is seen to fit the model due to the small *p*-values. 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 these data may not be practically suitable for factor extraction.

Table 9: Dimensionality Detection for Dataset 2

Threshold	No. of hom. Sets	KMOS
0.38	2	[[1]] 0.8242 0.5000
0.39	2	[[2]] 0.7942 0.5000
0.40	2	[[3]] 0.7942 0.5000
0.41	3	[[4]] 0.8058 0.5000 0.5000
0.42	3	[[5]] 0.8058 0.5000 0.5000
0.43	3	[[6]] 0.8058 0.5000 0.5000
0.44	3	[[7]] 0.8058 0.5000 0.5000
0.45	3	[[8]] 0.8058 0.5000 0.5000
0.46	3	[[9]] 0.8058 0.5000 0.5000
0.47	3	[[10]] 0.8058 0.5000 0.5000
0.48	3	[[11]] 0.8058 0.5000 0.5000
0.49	3	[[12]] 0.8058 0.5000 0.5000
0.50	3	[[13]] 0.8058 0.5000 0.5000
0.51	3	[[14]] 0.8058 0.5000 0.5000
0.52	3	[[15]] 0.8058 0.5000 0.5000
0.53	3	[[16]] 0.8058 0.5000 0.5000
0.54	3	[[17]] 0.8058 0.5000 0.5000
0.55	3	[[18]] 0.8058 0.5000 0.5000
0.56	3	[[19]] 0.8058 0.5000 0.5000
0.57	3	[[20]] 0.8058 0.5000 0.5000
0.58	3	[[21]] 0.8058 0.5000 0.5000

0.59	3	[[22]] 0.8058 0.5000 0.5000
0.60	3	[[23]] 0.8058 0.5000 0.5000
0.61	3	[[24]] 0.7202 0.5000 0.5000
0.62	3	[[25]] 0.7202 0.5000 0.5000
0.63	4	[[26]] 0.5 0.5 0.5 0.5
0.64	4	[[27]] 0.5 0.5 0.5 0.5
0.65	4	[[28]] 0.5 0.5 0.5 0.5
0.66	4	[[29]] 0.5 0.5 0.5 0.5
0.67	4	[[30]] 0.5 0.5 0.5 0.5
0.68	4	[[31]] 0.5 0.5 0.5 0.5
0.69	4	[[32]] 0.5 0.5 0.5 0.5
0.70	4	[[33]] 0.5 0.5 0.5 0.5
0.71	4	[[34]] 0.5 0.5 0.5 0.5
0.72	4	[[35]] 0.5 0.5 0.5 0.5
0.73	4	[[36]] 0.5 0.5 0.5 0.5
0.74	4	[[37]] 0.5 0.5 0.5 0.5
0.75	4	[[38]] 0.5 0.5 0.5 0.5

Author's Contract (2022)

From Table 9 it is observed that the automated threshold setting generates series of thresholds, 38 in this case. The task is to allow the algorithm to detect the optimal threshold that is suitable for dimensionality detection for this dataset. Subsequently, the dimensionality detection algorithm generates homogenous set(s) for each threshold. The modified

automated KMO algorithm is then used to calculate the KMO for each homogeneous set. It is observed that if there are multiple homogenous sets for a particular threshold, the algorithm calculates the KMO for each homogenous set. For instance, for threshold 0.62, resulting in three homogeneous sets the algorithm calculates the KMO for each homogeneous set, namely, 0.7202, 0.5 and 0.5, respectively. Since KMO is a measure of homogeneity the higher it is the better. So in a case where we have multiple KMOs, we pick the highest.

We then generate a graph of the KMOs against the thresholds to determine either a unique optimal threshold or a saturation point, if any.

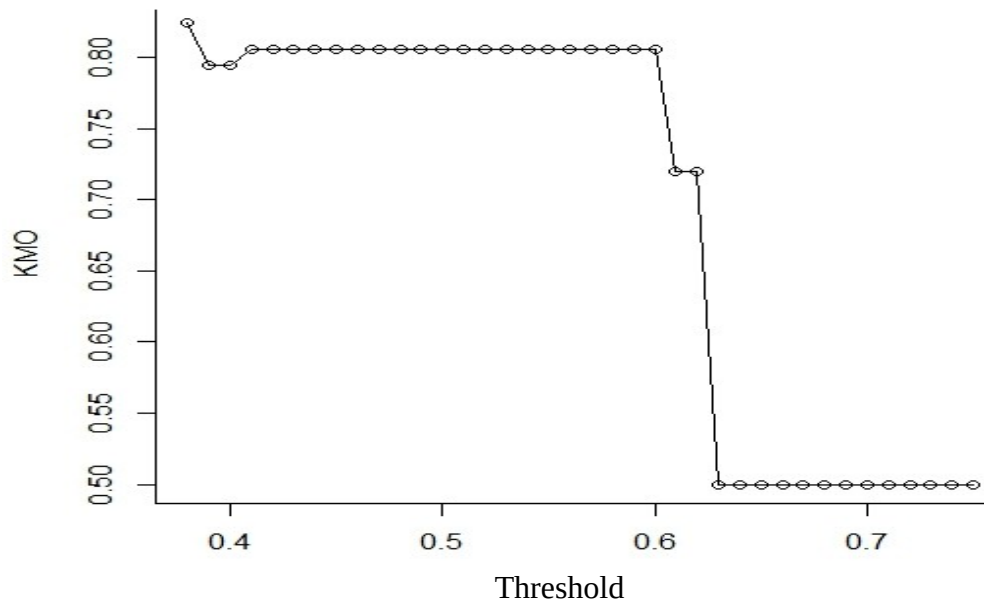


Figure 2: Sensitivity Analysis of Implementation Based on Dataset 2

The graph in Figure 2 shows that there is a saturation point between 0.42 and 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 for dimensionality detection for this dataset since it corresponds to the highest KMO (0.8242).

Threshold Selection for Optimal Factor Solution

If there is a unique optimal threshold for dimensionality detection for a dataset as demonstrated in Dataset 2, then this threshold should automatically give the optimal factor solution. We proceed to confirm this and also perform a confirmatory factor analysis. We attempt to show that the highest KMO is the one that corresponds to the homogeneous set with the highest number of variables and that the threshold for this particular homogeneous set should be the threshold that generates the optimal factor solution. In Table 10, the specific number of variables in each homogenous set is provided along with the respective KMO.

Table 10: Threshold Selection for Optimal Factor Solution Based on Dataset 2

Threshold	No. of hom. Sets	No. of variables in hom. Sets	KMO
0.38	2	6	0.8242
		2	0.5000
0.39	2	5	0.7942
		2	0.5000
0.40	2	5	0.7942
		2	0.5000
0.41	3	4	0.8058
		2	0.5000
		2	0.5000
0.42	3	4	0.8058
		2	0.5000
		2	0.5000
0.43	3	4	0.8058
		2	0.5000

		2	0.5000
0.44		4	0.8058
	3	2	0.5000
		2	0.5000
0.45		4	0.8058
	3	2	0.5000
0.46		2	0.5000
		4	0.8058
	3	2	0.5000
		2	0.5000
0.47		4	0.8058
	3	2	0.5000
0.48		2	0.5000
		4	0.8058
	3	2	0.5000
		2	0.5000
0.49		4	0.8058
	3	2	0.5000
		2	0.5000
0.50		4	0.8058
	3	2	0.5000
		2	0.5000
0.51		4	0.8058
	3	2	0.5000
		2	0.5000
0.52		4	0.8058
	3	2	0.5000
		2	0.5000
0.53		4	0.8058
	3	2	0.5000
		2	0.5000

0.54		4	0.8058
	3	2	0.5000
		2	0.5000
0.55		4	0.8058
	3	2	0.5000
		2	0.5000
0.56		4	0.8058
	3	2	0.5000
		2	0.5000
0.57		4	0.8058
	3	2	0.5000
		2	0.5000
0.58		4	0.8058
	3	2	0.5000
		2	0.5000
0.59		4	0.8058
	3	2	0.5000
		2	0.5000
0.60		4	0.8058
	3	2	0.5000
		2	0.5000
0.61		3	0.7202
	3	2	0.5000
		2	0.5000
0.62		3	0.7202
	3	2	0.5000
		2	0.5000
0.63		2	0.5000
	4	2	0.5000
		2	0.5000

		2	0.5000
		2	0.5000
0.64		2	0.5000
	4	2	0.5000
		2	0.5000
		2	0.5000
0.65	4	2	0.5
		2	0.5
		2	0.5
		2	0.5
0.66	4	2	0.5
		2	0.5
		2	0.5
0.67	4	2	0.5
		2	0.5
		2	0.5
		2	0.5
0.68	4	2	0.5
		2	0.5
		2	0.5
		2	0.5
0.69	4	2	0.5
		2	0.5
		2	0.5
0.70	4	2	0.5
		2	0.5
		2	0.5
		2	0.5

0.71	4	2	0.5
		2	0.5
		2	0.5
		2	0.5
0.72	4	2	0.5
		2	0.5
		2	0.5
0.73	4	2	0.5
		2	0.5
		2	0.5
0.74	4	2	0.5
		2	0.5
		2	0.5
0.75	4	2	0.5
		2	0.5
		2	0.5

Source: Author's Construct (2022)

Table 10 displays the various thresholds, their corresponding homogeneous sets, number of variables in each homogenous set and corresponding KMOs.

From Table 10 it could be observed that a threshold of 0.38, corresponding to a homogenous set with the highest number of variables (6) has the highest KMO making it the threshold that generates the optimal factor solution for Dataset 2 as suspected earlier. Though the two homogeneous sets

both have the same threshold, the KMO for the entire data set (0.822) shows that the required homogeneous set is the one whose KMO is equal to or greater than the full KMO.

Confirmatory Test of Model Adequacy for Dataset 2

It is indicated earlier that the second confirmatory test to verify our assertion that if a dataset generates an optimal threshold for dimensionality detection, then this threshold should automatically yield an optimal factor solution based on a confirmatory factor analysis test.

In this dataset, we first test the adequacy of the one-, two, three and four factor models equivalent to one, two, three and four dimensions to determine which is the most suitable.

Table 11: Significance Test of Factor Solutions 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

Source: Author's Construct (2022)

Table 11 shows the best possible fitting factor solutions that can be obtained for Dataset 2. 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. The question now is: which factor solution is optimal? The results based on our algorithm (**Table 10**) indicates that the two factor solution would be the best

since it contains a homogenous set with the highest number of variables. This is in line with our earlier results.

Dimensionality Detection based on Order Statistics

It is intimated in our objective that works done in earlier research do not investigate the robustness of their method to other correlation profiles as only Pearson's Correlation is used. Statistically the variance-covariance matrix of the Pearson's correlation hinges on the mean which is affected by extreme values. So, an attempt is made to test the robustness of the algorithm using another correlation profile which hinges on a statistic which is not affected by extreme values. In this study we make use of the order statistic, which hinges on the median. Here the values of the p variables are ordered. A correlation matrix is then generated for these ordered variables.

Order Statistics Algorithm Procedure

The algorithm generates the correlation matrix for p variables, $x_1, x_2, x_3, \dots, x_p$ and returns the order statistics for the variables $x_{(1)}, x_{(2)}, x_{(3)}, \dots, x_{(p)}$. Meaning the sample values placed in ascending order. $x_{(1)}, x_{(2)}, x_{(3)}, \dots, x_{(p)}$ is the set of ordered values form the original sample values. The correlation matrix is then generated for these ordered variables. The dimensionality detection algorithm is then applied to detect dimensionality for the dataset.

Table 12: Order Statistic Implementation



[1] 0.31	1	[1] 0.6995
[2] 0.32	1	[[2] 0.6995
[3] 0.33	1	[[3] 0.6995
[4] 0.34	1	[4] 0.6995
[5] 0.35	1	[[5] 0.6995
[6] 0.36	1	[[6] 0.6995
[7] 0.37	1	[[7]] 0.6995
[8] 0.38	1	[8]] 0.6995
[9] 0.39	1	[9]] 0.6995
[10] 0.40	1	[10] 0.6995
[11] 0.41	1	[11] 0.6995
[12] 0.42	1	[12] 0.6995
[13] 0.43	1	[13] 0.6995
[14] 0.44	1	[14] 0.6995
[15] 0.45	1	[15] 0.7614
[16] 0.46	1	[16] 0.7614
[17] 0.47	1	[17] 0.7614
[18] 0.48	1	[18] 0.7571
[19] 0.49	1	[19] 0.7571
[20] 0.50	1	[[20]] 0.7571
[21] 0.51	2	[[21]] 0.8587 0.5000
[22] 0.52	2	[[22] 0.8587 0.5000
[23] 0.53	2	[[23] 0.8587 0.5000
[24] 0.54	2	[[24]] 0.8587 0.5000
[25] 0.55	2	[[25]] 0.8587 0.5000
[26] 0.56	2	[[26]] 0.8587 0.5000
[27] 0.57	2	[[27]] 0.8587 0.5000
[28] 0.58	2	[[28]] 0.8587 0.5000
[29] 0.59	2	[[29]] 0.8587 0.5000
[30] 0.60	2	[30] 0.8587 0.5000
[31] 0.61	2	[31] 0.8587 0.5000
[32] 0.62	2	[32] 0.8587 0.5000
[33] 0.63	2	[33] 0.8587 0.5000
[34] 0.64	2	[34] 0.8587 0.5000
[35] 0.65	2	[35] 0.8587 0.5000
[36] 0.66	2	[36] 0.8587 0.5000
[37] 0.67	2	[37] 0.8587 0.5000
[38] 0.68	2	[38] 0.8587 0.5000
[39] 0.69	2	[39] 0.8587 0.5000
[40] 0.70	2	[40] 0.8587 0.5000
[41] 0.71	2	[41] 0.8587 0.5000
[42] 0.72	2	[42] 0.8587 0.5000
[43] 0.73	2	[43] 0.8587 0.5000
[44] 0.74	2	[44] 0.8587 0.5000
[45] 0.75	2	95 [45] 0.8587 0.5000
[46] 0.76	2	[46] 0.8587 0.5000
[47] 0.77	2	[47] 0.8587 0.5000
[48] 0.78	2	[48] 0.8587 0.5000
[49] 0.79	2	[49] 0.8587 0.5000

Source: Author’s Construct (2022)

Table 12 shows series of threshold generated using the automated threshold setting outlined early on. The Dimensionality detection algorithm generates homogeneous set(s) for each threshold. The KMO algorithm calculates the KMO for each homogeneous set. In case a threshold generates multiple homogeneous sets and corresponding multiple KMOs we choose the highest. This is because KMO is measure of homogeneity so the higher it is the better it is. We graph the KMO’S against the thresholds to determine either a unique optimal threshold or the saturation point if any.

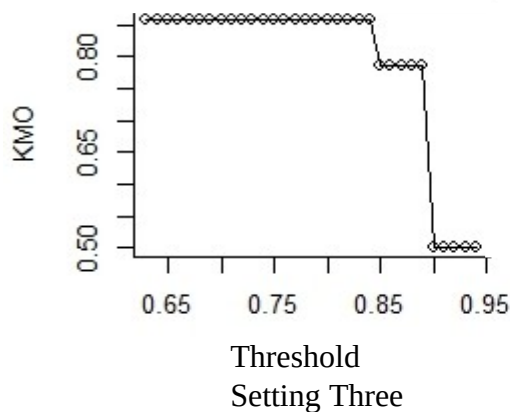
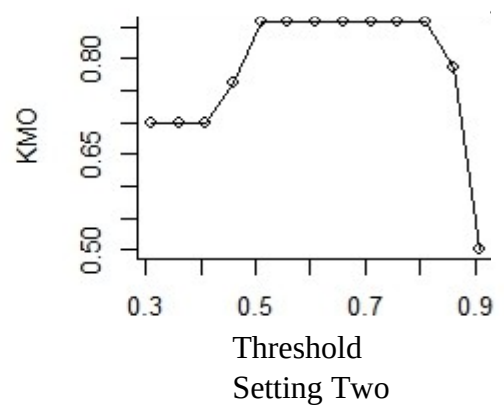
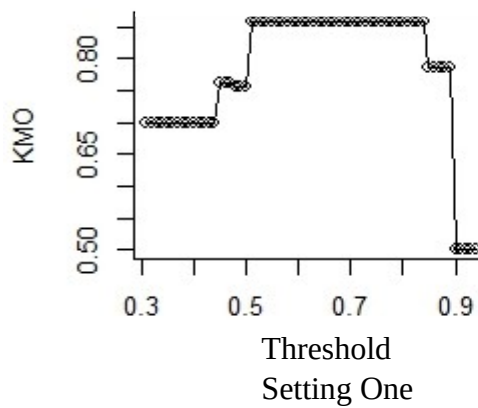


Figure 3: Sensitivity Analysis for Dataset 1 Based on Order Statistic profile

The graphs in Figure 3 show a plot of KMOs against the correspond thresholds. It could be observed that the saturation point for all three graphs lie within 0.6 and 0.85 inclusive meaning any threshold within this range could be the optimal threshold for dimensionality detection for this dataset. This is also supported by the highest KMOs.

Since similar results were obtained for Pearson's correlation, this establishes the robustness of the method to other correlation profiles that hinge on the median. It is also observed that the results show that the same number of variables in each homogenous set for each cut-off for both the Pearson's correlation and order statistics. This also establishes that the method is not sensitive to the correlation profile used.

Robustness of Method using a Reduced Dataset

The dimensionality detection method outlined early on used a correlation profile which hinges on all the original variables in the dataset. Meaning we generated the correlation matrix using all the variables in the original dataset. For a dataset with extreme values, some extreme values may not contribute prominently towards explaining the phenomenon under study. So it is important to control for those extreme values by selecting the k highest contributors. We then generate a correlation matrix for this reduced dataset for dimensionality detection. The idea is to compare the results for using a correlation profile that hinges on all the original variables in the dataset to the results of a correlation profile that hinges on the k highest contributors after controlling for extreme values.

Original Data Layout

Consider a set of p -variables and n observations. The data layout is given below.

Table 13: Original Data Layout on p Variables

y_1	y_2	y_p
y_{11}	y_{21}	y_{p1}
y_{12}	y_{22}	y_{p2}
\cdot	\cdot	\cdot
\cdot	\cdot	\cdot
y_{1n}	y_{2n}	y_{pn}

To determine the contributions of each of the p variables, we extract a feature from these variables that automatically controls for outliers based on probability distributions since this automatically controls for outlier.

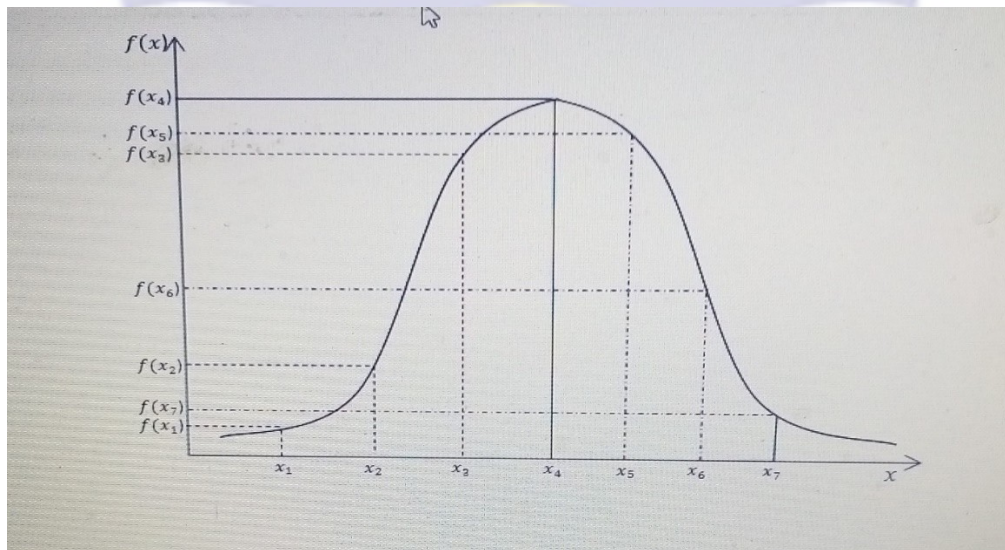


Figure 4: A Plot of Probability Distribution Function Values against the Variables

Statistically, whenever we plot the probability distribution function values against the observations, we get the density plot as shown in Figure 4. Regardless of the nature of the distribution, the variables are ordered automatically so the mode is always found at the peak. Also, the median

would divide the density in two equal halves. This means that values close to the tail of the distribution have smaller weights. From this, the variable that happens to be the median assumes the highest probability distribution value, in this case the highest probability density function value. Now it is observed from the diagram that $y_{11}f(y_{11}), y_{12}f(y_{12}), \dots, y_{1n}f(y_{1n})$ denote the mean of $y_{11}, y_{12}, \dots, y_{1n}$ which is their individual contributions towards the common center. Reason being statistically the k th moment about the origin is given by

$$E(Y^k) = \begin{cases} \int y^k f(y) dy, & y \text{ continuous} \\ \sum_y y^k f(y), & y \text{ discrete} \end{cases} \quad (4.4)$$

We set $k=1$ to get the first moment about the origin given by

$$E(Y) = \int y f(y) dy, \text{ for } y \text{ continuous} \quad (4.5)$$

This gives the mean of the distribution which gives information about the center of the distribution. Thus, $yf(y)$ indicates the contribution of each variable towards the center. Also, the closer a variable is to the center, the higher its contribution towards the common center $E(Y)$. It is therefore possible to identify the variables that contribute more based on their probability distribution values and then use these variables to generate the correlation profile as opposed to the use of all the original p variables. The Kernel Smoothing package in R is used to generate the probability distribution values.

It is clear from Figure 4 that each y has its own density value. We also observe that when the y is closer to the tail of the distribution, its density value is smaller but as it tends towards the center its density value increases. Based

on these arguments we extract a feature $T(y) = yf(y)$ where $T(y)$ is the statistics and $yf(y)$ is the contribution of each y towards the common center. The $T(y)$ extracted is shown in Table14.



Table14: Data Layout of Extracted Feature T(y)

$T(y_1)$	$T(y_2)$...	$T(y_p)$
$t(y_{11})$	$t(y_{21})$		$t(y_{p1})$
$t(y_{12})$	$t(y_{22})$		$t(y_{pn})$
\vdots	\vdots		\vdots
$t(y_{1n})$	$t(y_{2n})$		$t(y_{pn})$

Source: Author’s Construct (2022)

where $t(y_{ij}) = y_{ij} f(y_{ij})$, $i = 1, \dots, p; j = 1, \dots, n$

Now we order the columns of T(y) in order to select the k highest contributors based on the strength of their contribution and use the 70% training rule in Machine learning to select the k highest contributors. Mahanatesh (2020), intimated that in Machine learning 70% of the dataset should be used for training (to model) and 30% for testing (assessing the predictive performance of the model). Thus, the value of k is 0.7 times the size of the data. We then compute the correlation matrix for the reduced dataset, detect dimensionality for this reduced dataset and compare results with that of the full dataset.

Table 15: Correlation Matrix for Reduced Dataset 1

	X_1	X_2	X_3	X_4	X_5	X_6	
x_2		0.921					
x_3		0.881	0.834				
x_4		0.594	0.541	0.702			
x_5		0.774	0.767	0.695	0.595		
x_6		0.669	0.456	0.673	0.217	0.508	
x_7		0.917	0.957	0.848	0.429	0.650	0.533

Source: Author’s Construct (2022)

Comparing the pairwise correlations in the reduced data set after controlling for extreme values to the correlation matrix for all the original variables in Table 15 it observed that some pairwise correlations increased while others decreased. For example in the original correlation matrix (x_2, x_1) $\rho = 0.921$, while in the reduced dataset, correlation matrix ρ . We suspect these changes may be due to the control for outliers. The benefits may not be evident here but rather in the final results for dimensionality detection using this reduced correlation matrix. We proceed to detect dimensionality using the ‘reduced correlation matrix’.

Table 16: Dimensionality Detection for Reduced Dataset 1

SN/Threshold	No. of hom. Sets	SN/KMO
[1,] 0.22	1	[[1]] 0.7045
[2,] 0.23	1	[[2]] 0.7045
[3,] 0.24	1	[[3]] 0.7045
[4,] 0.25	1	[[4]] 0.7045
[5,] 0.26	1	[[5]] 0.7045
[6,] 0.27	1	[[6]] 0.7045
[7,] 0.28	1	[[7]] 0.7045
[8,] 0.29	1	[[8]] 0.7045
[9,] 0.30	1	[[9]] 0.7045
[10,] 0.31	1	[[10]] 0.7045
[11,] 0.32	1	[[11]] 0.7045
[12,] 0.33	1	[[12]] 0.7045
[13,] 0.34	1	[[13]] 0.7045
[14,] 0.35	1	[[14]] 0.7045
[15,] 0.36	1	[[15]] 0.7045
[16,] 0.37	1	[[16]] 0.7045
[17,] 0.38	1	[[17]] 0.7045
[18,] 0.39	1	[[18]] 0.7045
[19,] 0.40	1	[[19]] 0.7045
[20,] 0.41	1	[[20]] 0.7045
[21,] 0.42	1	[[21]] 0.7045
[22,] 0.43	1	[[22]] 0.7712
[23,] 0.44	1	[[23]] 0.7712
[24,] 0.45	1	[[24]] 0.7712
[25,] 0.46	1	[[25]] 0.7826

[26,] 0.47	1	[[26]] 0.7826
[27,] 0.48	1	[[27]] 0.7826
[28,] 0.49	1	[[28]] 0.7571
[29,] 0.50	1	[[29]] 0.7571
[30,] 0.51	1	[[30]]0.7571
[31,] 0.52	1	[[31]] 0.7571
[32,] 0.53	1	[[32]] 0.7571
[33,] 0.54	1	[[33]] 0.7571
[34,] 0.55	1	[[34]] 0.7571
[35,] 0.56	1	[[35]] 0.7571
[36,] 0.57	1	[[36]] 0.7571
[37,] 0.58	1	[[37]] 0.7571
[38,] 0.59	1	[[38]] 0.7571
[39,] 0.60	1	[[39]] 0.7571
[40,] 0.61	1	[[40]] 0.7571
[41,] 0.62	1	[[41]] 0.7571
[42,] 0.63	1	[[42]] 0.7571
[43,] 0.64	1	[[43]] 0.7826
[44,] 0.65	1	[[44]] 0.7826
[45,] 0.66	2	[[45]]0.8360 0.5000
[46,] 0.67	2	[[46]] 0.8360 0.5000
[47,] 0.68	2	[[47]] 0.8360 0.5000
[48,] 0.69	2	[[48]] 0.8360 0.5000

[49,] 0.70	2	[[49]] 0.8360 0.5000
[50,] 0.71	2	[[50]] 0.8360 0.5000
[51,] 0.72	2	[[51]] 0.8360 0.5000
[52,] 0.73	2	[[52]] 0.8360 0.5000
[53,] 0.74	2	[[53]] 0.8360 0.5000
[54,] 0.75	2	[[54]] 0.8360 0.5000
[55,] 0.76	2	[[55]] 0.8360 0.5000
[56,] 0.77	2	[[56]] 0.8360 0.5000
[57,] 0.78	2	[[57]] 0.8360 0.5000
[58,] 0.79	2	[[58]] 0.8360 0.5000
[59,] 0.80	2	[[59]] 0.8360 0.5000
[60,] 0.81	2	[[60]] 0.8360 0.5000
[61,] 0.82	2	[[61]] 0.8360 0.5000
[62,] 0.83	2	[[62]] 0.8360 0.5000
[63,] 0.84	2	[[63]] 0.7731 0.5000
[64,] 0.85	2	[[64]] 0.7731 0.5000
[65,] 0.86	2	[[65]] 0.7731 0.5000
[66,] 0.87	2	[[66]] 0.7731 0.5000
[67,] 0.88	2	[[67]] 0.7731 0.5000
[68,] 0.89	2	[[68]] 0.7731 0.5000
[69,] 0.90	2	[[69]] 0.7731 0.5000

[70,] 0.91	2	[[70]] 0.7731 0.5000
[71,] 0.92	3	[[71]] 0.5 0.5 0.5
[72,] 0.93	3	[[72]]0.5 0.5 0.5
[73,] 0.94	3	[[73]] 0.5 0.5 0.5
[74,] 0.95	3	[[74]] 0.5 0.5 0.5
[75,] 0.96	3	[[75]] 0.5 0.5 0.5

Source: Author’s Construct (2022)

Table 16 shows series of thresholds generated using the automated threshold setting outlined early on. Now comparing the KMOs for the reduced dataset in Table 16 to that of the full dataset in Table 6 indicate higher KMOs for the reduced dataset. For instance, a threshold of 0.22 in the full data set has a KMO of 0.6995 while the same threshold of 0.22 for the reduced dataset has a KMO of 0.7045. Similar patterns are observed for other thresholds. This gives a glimpse of a more superior result ahead. We graph the KMO against the thresholds to determine either a unique optimal threshold or the saturation point, if any.

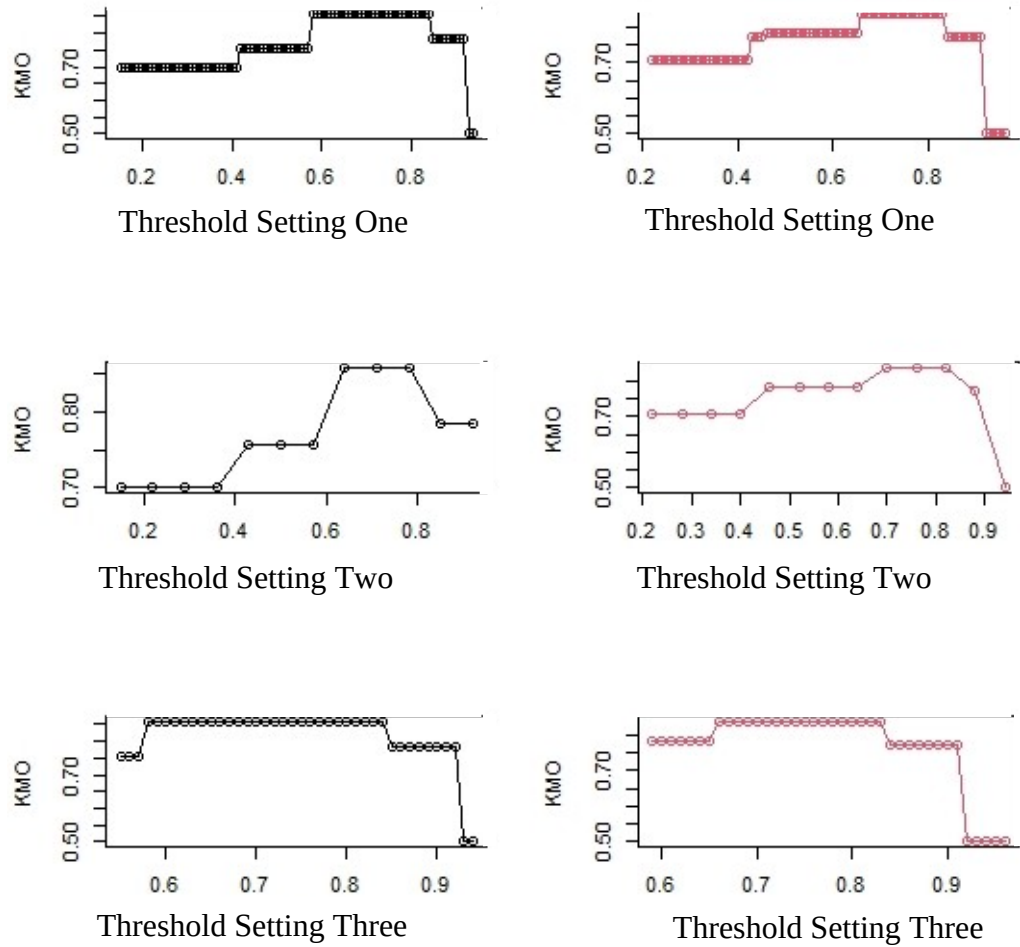


Figure 5: Sensitivity Analysis of Implementation Based on Reduced Dataset 1

The graph to the right shows the plot of KMOs against the corresponding thresholds for the reduced dataset. It could be observed that the saturation points for all three graphs lie within 0.65 and 0.85 inclusive meaning any threshold within this range could be the optimal threshold for dimensionality detection for this dataset. This is also supported by the highest KMOs. Though the number of dimensions generated by these thresholds remain the same for both the reduced and the full datasets, the interval of 0.65 and 0.85 for the reduced dataset (Graph to the right in Fig.5) is shorter as opposed to 0.6 and 0.85 (Graph to the left in Figure 5) inclusive for the full dataset. This makes the use of reduced dataset after controlling for extreme

values when detecting dimensionality computationally less expensive as opposed to the full dataset since it takes a shorter time to arrive at a shorter interval. We attempt to summarise the thresholds and highest KMO for both the full and reduced datasets in Table 17 to paint a clearer picture of the assertion.

Table 17: Summary of Thresholds and Highest KMOs for Full and Reduced Datasets.

Threshold for Full Dataset	Threshold for Reduced Dataset	Highest KMOs for Full Dataset	Highest KMOs for Reduced Dataset
[1] 0.15	[1] 0.22	[1] 0.6995	[1] 0.7045
[2] 0.16	[2] 0.23	[2] 0.6995	[2] 0.7045
[3] 0.17	[3] 0.24	[3] 0.6995	[3] 0.7045
[4] 0.18	[4] 0.25	[4] 0.6995	[4] 0.7045
[5] 0.19	[5] 0.26	[5] 0.6995	[5] 0.7045
[6] 0.20	[6] 0.27	[6] 0.6995	[6] 0.7045
[7] 0.21	[7] 0.28	[7] 0.6995	[7] 0.7045
[8] 0.22	[8] 0.29	[8] 0.6995	[8] 0.7045
[9] 0.23	[9] 0.30	[9] 0.6995	[9] 0.7045
[10] 0.24	[10] 0.31	[10] 0.6995	[10] 0.7045
[11] 0.25	[11] 0.32	[11] 0.6995	[11] 0.7045
[12] 0.26	[12] 0.33	[12] 0.6995	[12] 0.7045
[13] 0.27	[13] 0.34	[13] 0.6995	[13] 0.7045
[14] 0.28	[14] 0.35	[14] 0.6995	[14] 0.7045
[15] 0.29	[15] 0.36	[15] 0.6995	[15,] 0.7045

[16] 0.30	[16] 0.37	[16] 0.6995	[16] 0.7045
[17] 0.31	[17] 0.38	[17] 0.6995	[17] 0.7045
[18] 0.32	[18] 0.39	[18] 0.6995	[18] 0.7045
[19] 0.33	[19] 0.40	[19] 0.6995	[19] 0.7045
[20] 0.34	[20] 0.41	[20] 0.6995	[20] 0.7045
[21] 0.35	[21] 0.42	[21] 0.6995	[21] 0.7045
[22] 0.36	[22] 0.43	[22] 0.6995	[22] 0.7712
[23] 0.37	[23] 0.44	[23] 0.6995	[23] 0.7712
[24] 0.38	[24] 0.45	[24] 0.6995	[24] 0.7712
[25] 0.39	[25] 0.46	[25] 0.6995	[25] 0.7826
[26] 0.40	[26] 0.47	[26] 0.6995	[26] 0.7826
[27] 0.41	[27] 0.48	[27] 0.6995	[27] 0.7826
[28] 0.42	[28] 0.49	[28] 0.7571	[28] 0.7826
[29] 0.43	[29] 0.50	[29] 0.7571	[29] 0.7826
[30] 0.44	[30] 0.51	[30] 0.7571	[30] 0.7826
[31] 0.45	[31] 0.52	[31] 0.7571	[31] 0.7826
[32] 0.46	[32] 0.53	[32] 0.7571	[32] 0.7826
[33] 0.47	[33] 0.54	[33] 0.7571	[33] 0.7826
[34] 0.48	[34] 0.55	[34] 0.7571	[34] 0.7826
[35] 0.49	[35] 0.56	[35] 0.7571	[35] 0.7826
[36] 0.50	[36] 0.57	[36] 0.7571	[36] 0.7826
[37] 0.51	[37] 0.58	[37] 0.7571	[37] 0.7826
[38] 0.52	[38] 0.59	[38] 0.7571	[38] 0.7826
[39] 0.53	[39] 0.60	[39] 0.7571	[39] 0.7826

[40] 0.54	[40] 0.61	[40] 0.7571	[40] 0.7826
[41] 0.55	[41] 0.62	[41] 0.7571	[41] 0.7826
[42] 0.56	[42] 0.63	[42] 0.7571	[42] 0.7826
[43] 0.57	[43] 0.64	[43] 0.7571	[43] 0.7826
[44] 0.58	[44] 0.65	[44] 0.8587	[44] 0.7826
[45] 0.59	[45] 0.66	[45] 0.8587	[45] 0.8360
[46] 0.60	[46] 0.67	[46] 0.8587	[46] 0.8360
[47] 0.61	[47] 0.68	[47] 0.8587	[47] 0.8360
[48] 0.62	[48] 0.69	[48] 0.8587	[48] 0.8360
[49,] 0.63	[49] 0.70	[49] 0.8587	[49] 0.8360
[50] 0.64	[50] 0.71	[50] 0.8587	[50] 0.8360
[51] 0.65	[51,] 0.72	[51] 0.8587	[51] 0.8360
[52] 0.66	[52,] 0.73	[52] 0.8587	[52] 0.8360
[53] 0.67	[53] 0.74	[53] 0.8587	[53] 0.8360
[54] 0.68	[54] 0.75	[54] 0.8587	[54] 0.8360
[55] 0.69	[55] 0.76	[55] 0.8587	[55] 0.8360
[56] 0.70	[56] 0.77	[56] 0.8587	[56] 0.8360
[57] 0.71	[57] 0.78	[57] 0.8587	[57] 0.8360
[58] 0.72	[58] 0.79	[58] 0.8587	[58] 0.8360
[59] 0.73	[59] 0.80	[59] 0.8587	[59] 0.8360
[60] 0.74	[60] 0.81	[60] 0.8587	[60] 0.8360
[61] 0.75	[61] 0.82	[61] 0.8587	[61] 0.8360
[62] 0.76	[62] 0.83	[62] 0.8587	[62] 0.8360
[63] 0.77	[63] 0.84	[63] 0.8587	[63] 0.7731

[64] 0.78	[64] 0.85	[64] 0.8587	[64] 0.7731
[65] 0.79	[65] 0.86	[65] 0.8587	[65] 0.7731
[66] 0.80	[66] 0.87	[66] 0.8587	[66] 0.7731
[67] 0.81	[67] 0.88	[67] 0.8587	[67] 0.7731
[68] 0.82	[68] 0.89	[68] 0.8587	[68] 0.7731
[69] 0.83	[69] 0.90	[69] 0.8587	[69] 0.7731
[70] 0.84	[70] 0.91	[70] 0.8587	[70] 0.7731
[71] 0.85	[71] 0.92	[71] 0.7848	[71] 0.5000
[72] 0.86	[72] 0.93	[72,] 0.7848	[72] 0.5000
[73] 0.87	[73] 0.94	[73] 0.7848	[73] 0.5000
[74] 0.88	[74] 0.95	[74] 0.7848	[74] 0.5000
[75] 0.89	[75] 0.96	[75] 0.7848	[75] 0.5000
[76] 0.90		[76] 0.7848	
[77] 0.91		[77] 0.7848	
[78] 0.92		[78] 0.7848	
[79] 0.93		[79] 0.5000	
[80] 0.94		[80] 0.5000	

Source: Author's Construct (2022)

It could be observed from the table that the reduced dataset generated a smaller number of thresholds that is 75 as opposed to 80 thresholds for the full dataset. Also, the reduced dataset has higher KMOs than that of the full dataset. This is as a result of the selection of the k highest contributors rather than all variables in the full dataset. These observations resulted in a shorter saturation point for the reduced data set as opposed to a longer saturation point

for a full dataset owing mainly to the control of extreme values in the reduced dataset.

Implementation (for simulated data)

Table 18 gives the result of the implementation of the procedure in the simulated data using Threshold Setting 1. The results show that there is a highest KMO of 0.9643 corresponding to thresholds of 0.03 and 0.04. This shows that a highest KMO value is obtained at a very small cut-off value. It is however not clear the uniqueness of the highest KMO value. It requires an examination of the actual number of variables in the two homogeneous sets. In a large dataset such as this, this is quite cumbersome to present. It is quite clear that since multiple thresholds yield the same KMO value, the uniqueness of the value is not clearly determined.

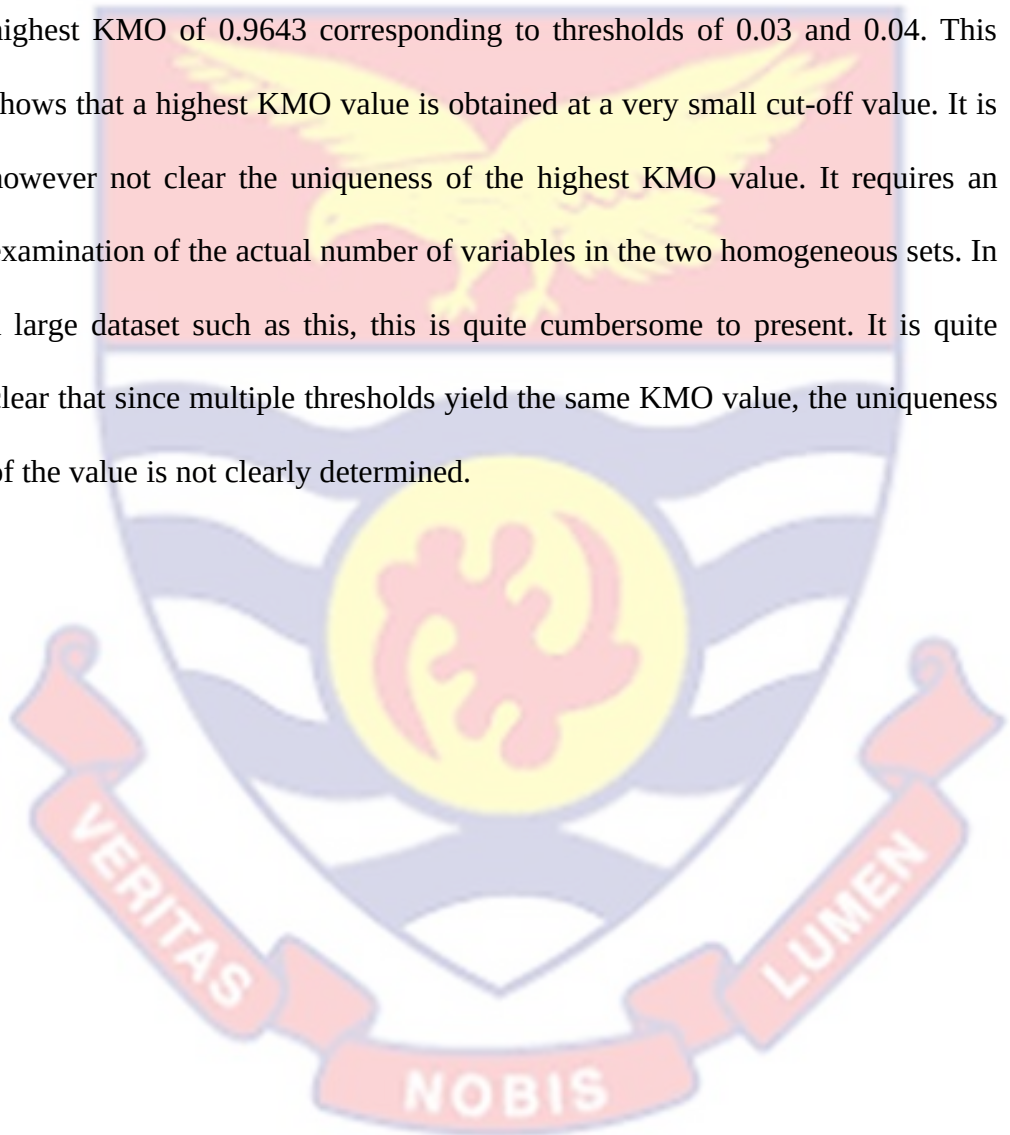


Table 18: Dimensionality Detection for Simulated data

Sn/Threshold	No. of hom. Sets	KMO
[1] 0.01	2	[1] 0.9582671
[2] 0.02	2	[2] 0.9618571
[3] 0.03	2	[3] 0.9642579
[4] 0.04	2	[4] 0.9642579
[5] 0.05	2	[5] 0.9627946
[6] 0.06	2	[6] 0.9627946
[7] 0.07	2	[7] 0.9627946
[8] 0.08	2	[8] 0.9627946
[9] 0.09	2	[9] 0.9619099
[10] 0.10	2	[10] 0.9615578
[11] 0.11	3	[11] 0.9604512
[12] 0.12	3	[12] 0.9604512
[13] 0.13	3	[13] 0.9603667
[14] 0.14	3	[14] 0.9591313
[15] 0.15	3	[15] 0.9582347
[16] 0.16	3	[16] 0.9582347
[17] 0.17	3	[17] 0.9582347
[18] 0.18	3	[18] 0.9565589
[19] 0.19	3	[19] 0.9565589
[20] 0.20	3	[20] 0.9565589
[21] 0.21	3	[21] 0.9505914
[22] 0.22	4	[22] 0.9518408

[23] 0.23	4	[23] 0.9518408
[24] 0.24	4	[24] 0.9405952
[25] 0.25	4	[25] 0.9408441
[26] 0.26	4	[26] 0.9408441
[27] 0.27	4	[27] 0.9408441
[28] 0.28	4	[28] 0.9408441
[29] 0.29	4	[29] 0.9324921
[30] 0.38	4	[30] 0.9132894
[31] 0.39	4	[31] 0.9132894
[32] 0.40	5	[32] 0.9132894
[33] 0.41	5	[33] 0.9132894
[34] 0.42	5	[34] 0.9132894
[35] 0.43	5	[35] 0.9132894
[36] 0.44	5	[36] 0.9132894
[37] 0.45	5	[37] 0.9132894
[38] 0.46	5	[38] 0.9132894
[39] 0.47	5	[39] 0.9132894
[40] 0.48	5	[40] 0.9132894
[41] 0.49	5	[41] 0.9132894
[42] 0.50	2	[42] 0.9132894
[43] 0.51	2	[43] 0.9132894
[44] 0.52	2	[44] 0.9132894
[45] 0.53	2	[45] 0.9132894
[46] 0.60	5	[46] 0.9132894

[47]	0.61	5	[47]	0.9132894
[48]	0.62	5	[48]	0.9132894
[49]	0.63	5	[49]	0.9132894
[50]	0.64	5	[50]	0.9132894
[51]	0.65	5	[51]	0.9132894
[52]	0.66	5	[52]	0.9132894
[53]	0.67	5	[53]	0.9132894
[54]	0.68	5	[54]	0.9132894
[55]	0.69	5	[55]	0.9132894
[56]	0.70	2	[56]	0.9132894
[57]	0.71	2	[57]	0.9132894
[58]	0.72	2	[58]	0.9132894
[59]	0.73	2	[59]	0.9132894
[60]	0.74	3	[60]	0.9132894
[61]	0.75	6	[61]	0.9617736
[62]	0.76	6	[62]	0.9607194
[63]	0.77	6	[63]	0.9607194
[64]	0.78	6	[64]	0.9607194
[65]	0.79	7	[65]	0.9607194
[66]	0.80	8	[66]	0.9607194
[67]	0.81	10	[67]	0.9536611
[68]	0.82	12	[68]	0.8691330
[69]	0.83	12	[69]	0.8691330
[70]	0.84	13	[70]	0.7693248

[71] 0.85 9 [71] 0.5000000

Source: Author's Construct (2022)

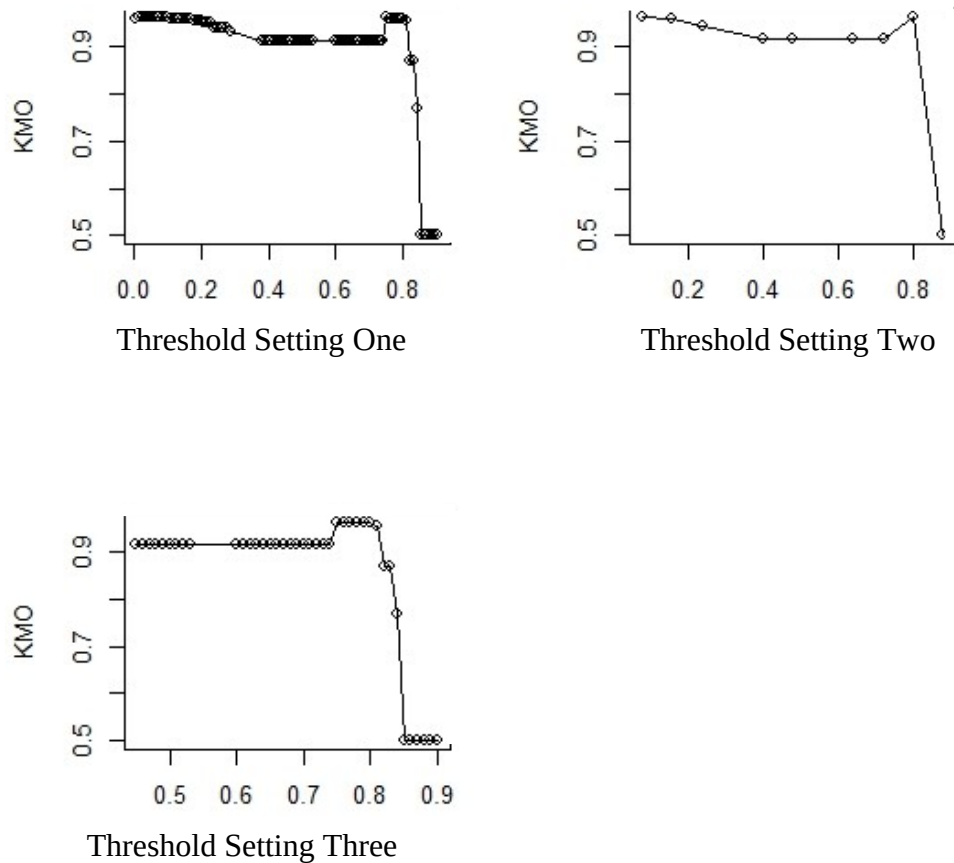


Figure 6: Sensitivity Analysis of Implementation Based on Simulated data

Figure 6 shows that there is a saturation point between 0.75 and 0.80 thresholds for all three threshold settings. However, it also clear that for the first two settings, there is another interval of much smaller thresholds that could also be examined for the highest KMO value.

As anticipated, the simulated data present some complexity in the identification of the optimal threshold. It is therefore necessary to observe the number of variables in each of the homogenous sets for each threshold. In

Table 18, it is observed that thresholds of 0.03 and 0.04 have the highest KMO of 0.9643 and given by Threshold Setting 1. This is a threshold that produces only two homogeneous sets. For Threshold Setting 3, the highest KMO (0.9618) corresponds to threshold value of 0.75 and produces six homogeneous sets. A plot of thresholds for the three settings against the number of homogeneous sets is given in Figure 7.

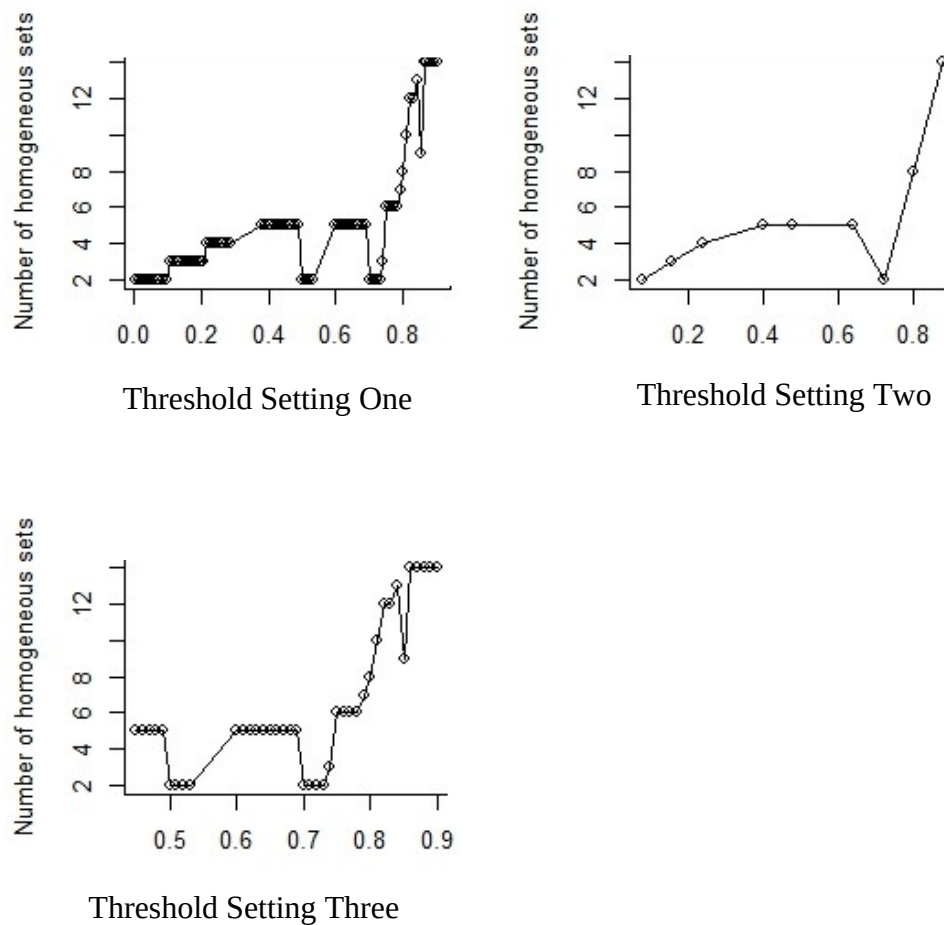


Figure 7: Plot of thresholds against number of homogeneous sets

Since the data involve large number of variables (30), it is clumsy to generate the actual number of variables for each homogeneous set. The result from these data reveals that an optimal threshold could be as small as 0.03, a

value that may be practically impossible to be thought of by a subjective consideration.

Remarks 4.2

It is observed with the simulation that there are some threshold values from the settings that could not be used for the implementation as those values run into errors with the generation of the KMO from the associated correlation matrices. For example, with Threshold Setting 1, the following values has issues associated with correlation matrix: 0.30, 0.31, 0.32, 0.33, 0.34 ,0.35, 0.36, 0.37, 0.54, 0.55, 0.56, 0.57, 0.58, 0.59; with Threshold Setting 2, the following values have issues with the correlation matrix: 0.32, 0.56; and with Threshold Setting 3, the following values have issues: 0.54, 0.55, 0.56, 0.57, 0.58, 0.59.

Remarks 4.3

The simulated dataset is obtained 9by Item Response Theory with underlying dimension of three. The approach presented here rather identifies two dimensions. This is not surprising as homogeneity of groupings in a simulated data may not be very well-defined as in real data. The difiiculty in constructing clear homogenous sets in such data might account for some thresholds that could not yield homogenous sets.

Discussion

The findings of this study shed light on some results in the literature. In the study (Benyi, 2018; Nkansah, 2018) of dimensionality on the Dataset 1 and that carried out in this study, it is found that there is actually no significant dimension underlying the correlation matrix, even though the dataset is

presented (Johnson & Wichern, 2009) apparently to demonstrate the concept of factor analysis. This result is initially observed from confirmatory factor analysis. In this study, it is further explained that the lack of significant dimension in the data is as a result of the fact that there is no homogeneous set with a unique highest KMO. Even though the use of KMO is not a new concept, it is only by a structured approach such as the one presented in this study that could unravel the detailed effect of the data structure on its true dimensionality.

The study on the simulated data has made interesting findings. It is observed that some threshold values may not generate homogeneous sets. This is an observation that is not noted in the literature. This means that it could be quite impractical to set a subjective cut-off value for certain datasets for the purpose of detecting dimensions. The study therefore affirms that to detect dimension in multivariate data, the way to go is to allow the data structure itself to determine its own threshold.

Chapter Summary

The chapter has focused on the implementation of the proposed dimensionality detection approach. It outlines a procedure for identifying the initial dimensionality in the data. The study of the formation of homogeneous groups in the dataset enables us to obtain preliminary understanding of the correlation structure of the data. The method starts with the determination of a data specific threshold as opposed to an experimenter-specific one, and then identifies a pair of indicators with the highest correlation that is at least equal to the threshold. This pair of variables forms the spanning set. All variables in the reduced dataset whose pairwise correlations with the variables in the

spanning set are at least the thresholds are used to update the spanning set until there is left no such variable. This forms the first homogeneous set. We begin another grouping by identifying a pair of indicators with highest pairwise correlation that is at least the threshold that is not found in the previous group. Any variable is added to this set in a manner described earlier. We continue this way until no such grouping can be formed. The groupings so formed are expected to be homogeneous within each group. The number of homogeneous sets then constitute the expected minimum number of dimensions that underlie the data.

Dimensionality detection is threshold sensitive, in view of this the study identifies three major threshold settings which allow the data to call its optimal threshold suitable for dimensionality detection.

The study proposes three automated threshold settings using an automated algorithm that generates a data-specific threshold by allowing the data structure to generate the optimal threshold for detecting dimensionality of the multivariate data for more accurate results.

For automated threshold setting 1, the algorithm picks the lowest pairwise correlation in the variance -covariance matrix, generates series of thresholds using a step value of 0.01 until all thresholds in the variance -covariance matrix are accommodated. The dimensionality detection algorithm is then used to generate homogeneous sets for each of these thresholds. Since multidimensionality is expected some thresholds yielded more than one homogeneous set. The KMO values are then calculated for each homogeneous set that corresponds to each threshold. Sensitivity analysis is then carried out

for these thresholds in an attempt to pick the optimal threshold suitable for dimensionality detection for the data set.

Also, for automated threshold setting two, the correlation profile used is the Pearson's correlation which is normally distributed. Statistically majority (about 97%) of the data points lie 3 standard deviations about the mean. This gives us 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 variance-covariance matrix to the resultant standard deviation to generate series of thresholds. This generated 38 thresholds. The dimensionality detection algorithm is then used to generate homogeneous sets for each of these thresholds. Since multidimensionality is expected some thresholds yielded more than one homogeneous set. The KMO values are then calculated for each homogeneous set that corresponds to each threshold. Sensitivity analysis is then carried out for these thresholds in an attempt to pick the optimal threshold suitable for dimensionality detection for the data set.

In addition, for automated threshold setting three, the procedure is based on Threshold setting 1. Statistically the variance-covariance matrix used which hinges on Pearson's correlation is symmetric. This makes the data points normally distributed. In view of this the algorithm determines the median for thresholds generated using automated Threshold Setting 1 and selects those thresholds that are at least the median. This is similar to the use of the lower triangular matrix of the variance-covariance matrix. The dimensionality detection algorithm is then used to generate homogeneous sets for each of these thresholds. Since multidimensionality is expected, some

thresholds yield more than one homogeneous set. The KMO values are then calculated for each homogeneous set that corresponds to each threshold. Sensitivity analysis is then carried out for these thresholds in an attempt to pick the optimal threshold suitable for dimensionality detection for the data.

The study discovers that for Dataset 1, any threshold between 0.6 and 0.85 could be used to detect dimensionality for this data since there is a resultant saturation point. Also, for Dataset 2, a unique threshold of 0.38 is discovered as being the optimal threshold for dimensionality detection for this data set.

The study establishes that for data set 1, there were two homogeneous sets indicating that two dimensions underlie the data. Also, for data set two, there were two homogeneous sets indicating that two dimensions underlie the data. The study also established that Data set 1 could not be practically suitable for factor extraction since the confirmatory factor analysis test for Model adequacy did not indicate any model fit for up to three factors equivalent to three dimensions. These findings were based on Pearson's correlation.

The study observed that for any dataset with an optimal unique threshold suitable for dimensionality detection, this threshold generates a unique homogeneous set with the highest number of indicators and the highest KMO. This is demonstrated using dataset 2. On the other hand, if the dataset is not able to generate a unique homogeneous set with the highest KMO and the highest number of indicators, it is likely that the determination of the dimensionality of this dataset could be a challenge. This is demonstrated in dataset 1.

Also, for a dataset with extreme values, some extreme values may not contribute prominently towards explaining the phenomenon under study. So, it is important to control for those extreme values by selecting the k highest contributors. Kernel smoothing package in R is used to exclude extreme values in the dataset before dimensionality detection. The density function in R computes the values of the kernel density estimate, in our case the probability density values. Applying the plot function to an object created by density, in our case the observations will plot the estimate. This generates the probability distribution curve that reveals outliers.

From the arguments above, the dimensionality detection results for the reduced data set generated a smaller number of thresholds that is 75 as opposed to 80 thresholds for the full dataset. Also, the reduced dataset has higher KMOs than that of the full. This is as a result of the selection of the k highest contributors rather than all variables in the full dataset. These observations resulted in a shorter saturation point of 0.65 and 0.85 inclusive for the reduced data set as opposed to a longer saturation point of about 0.6 and 0.85 for the full dataset owing mainly to the control of extreme values in a reduced dataset. This renders the use of reduced dataset after controlling for extreme values when detecting dimensionality computationally less expensive as opposed to the use of the full dataset since it takes a shorter time to arrive at a shorter interval.

The study investigated the robustness of the proposed dimensionality detection method by using order statistic correlation profile which hinges on the median as opposed to Pearson's correlation which hinges on the mean. The

algorithm converged in both cases since similar results for Pearson's correlation were obtained.



CHAPTER FIVE

SUMMARY CONCLUSIONS AND RECOMMENDATIONS

This chapter presents the summary of the entire work. It highlights the main findings in all the preceding chapters. It then presents the conclusion to the study and prescribes recommendations based on the key findings of the study.

Summary

The purpose of this study is to propose an automated threshold method for detecting dimensionality in a multivariate dataset which would serve as the basis for the application of the well-known statistical tools for purposes of interpreting a multivariate dataset. The underlisted are the major findings from the study.

Automated Threshold Setting for Dimensionality Detection

Earlier researchers conducted research that explored a systematic approach that determines the initial dimensionality of the dataset. However, these researchers used an experimenter specific threshold which is a threshold based on the judgement of the experimenter for their studies which may lead to misleading results. Our study proposed three automated threshold setting using an automated algorithm that generates a data specific threshold by allowing the data structure to generate the optimal threshold for detecting dimensionality of the multivariate data set for more accurate results.

For automated threshold setting 1, the algorithm picks the lowest pairwise correlation in the variance covariance matrix, generates series of thresholds using a step value of 0.01 until all thresholds in the variance

covariance matrix are accommodated. The dimensionality detection algorithm is then used to generate homogeneous sets for each of these thresholds. Since multidimensionality is expected some thresholds yielded more than one homogeneous set. The KMO values are then calculated for each homogeneous set that corresponds to each threshold. Sensitivity analysis is then carried out for these thresholds in an attempt to pick the optimal threshold suitable for dimensionality detection for the data set.

Also, for automated threshold setting two, the correlation profile used is the Pearson's correlation which is normally distributed. Statistically majority (about 97%) of the data points lie 3 standard deviations about the mean. This gives us 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 variance-covariance matrix to the resultant standard deviation to generate series of thresholds. This generated 38 thresholds. The dimensionality detection algorithm is then used to generate homogeneous sets for each of these thresholds. Since multidimensionality is expected some thresholds yielded more than one homogeneous set. The KMO values are then calculated for each homogeneous set that corresponds to each threshold. Sensitivity analysis is then carried out for these thresholds in an attempt to pick the optimal threshold suitable for dimensionality detection for the data set.

In addition, for automated threshold setting three, the procedure is based on Threshold setting 1. Statistically the variance-covariance matrix used which hinges on Pearson's correlation is symmetric. This makes the data points normally distributed. In view of this the algorithm determines the

median for thresholds generated using automated threshold setting 1 and selects those thresholds that are at least the median. This is similar to the usage of the lower triangular matrix of the variance covariance matrix. The dimensionality detection algorithm is then used to generate homogeneous sets for each of these thresholds. Since multidimensionality is expected some thresholds yielded more than one homogeneous set. The KMO values are then calculated for each homogeneous set that corresponds to each threshold. Sensitivity analysis is then carried out for these thresholds in an attempt to pick the optimal threshold suitable for dimensionality detection for the data set.

The study discovered that for data set 1, any threshold between 0.6 and 0.85 could be used to detect dimensionality for this data set since there was a resultant saturation point. Also, for data set two, a unique threshold of 0.38 was discovered as being the optimal threshold for dimensionality detection for this data set.

Dimensionality Detection Method

Statistical applications such as factor analysis, principal component analysis are method dependent applications used by statisticians to interpret multivariate data. However, these applications are not able to determine whether data is dimensionless or not prior to their application. Our study proposed an automated method independent dimensionality detection approach that could be used by statisticians to have a prior knowledge of the dimensionality of a dataset before subsequent applications of these statistical tools for interpretation. Our approach also helps the researcher to generate multiple homogeneous sets for a threshold that results in multiple

homogeneous making it suitable for both unidimensional and multidimensional multivariate data.

The study established that for Dataset 1, there were two homogeneous sets indicating that two dimensions underlie the data. Also, for dataset two, there were two homogeneous sets indicating that two dimensions underlie the data. The study also established that Dataset 1 could not be practically suitable for factor extraction since the confirmatory factor analysis test for Model adequacy did not indicate any model fit for up to three factors equivalent to three dimensions.

Robustness of the Dimensionality Detection method to other Correlation Profiles

Previous researchers who attempted dimensionality detection did not investigate the robustness of the method to other correlation profiles since only Pearson's correlation which hinges on the mean was employed. Our study filled this gap by applying the Algorithm to other correlation profiles that hinge on the median specifically order statistic. The Algorithm converged in all cases indicating the robustness of the method to another correlation profile that hinges on the median specifically order statistic.

The study discovered that for dataset 1, any threshold between 0.6 and 0.85 could be used to detect dimensionality for this data set since there was a resultant saturation point. Also, for dataset two, a unique threshold of 0.38 was discovered as being the optimal threshold for dimensionality detection for this dataset.

Robustness of the Dimensionality Detection method using a reduced Dataset

Previous results of the implementation of the dimensionality detection method used a correlation profile which hinges on all the original variables in the data set. Meaning the correlation matrix generated used all the variables in the original dataset. For a dataset with extreme values, some extreme values may not contribute prominently towards explaining the phenomenon under study. So, it is important to control for those extreme values by selecting the k highest contributors, then generate a correlation matrix for this reduced dataset for dimensionality detection. The idea is to compare the results for using a correlation profile that hinges on all the original variables in the data set to the results of a correlation profile that hinges on the k highest contributors after controlling for extreme values. For our work, using, the dimensionality detection results for the reduced dataset generated a smaller number of thresholds that is 75 as opposed to 80 thresholds for the full dataset. Also, the reduced dataset has higher KMOs than that of the full. This is as a result of the selection of the k highest contributors rather than all variables in the full dataset. These observations resulted in a shorter saturation point of 0.65 and 0.85 inclusive for the reduced data set as opposed to a longer saturation point of about 0.6 and 0.85 for the full dataset owing mainly to the control of extreme values in a reduced data set.

This renders the use of reduced dataset after controlling for extreme values when detecting dimensionality computationally less expensive as opposed to the use of the full dataset since it takes a shorter time to arrive at a shorter interval.

Automated Modified KMO Algorithm

Our study modified and automated the KMO algorithm that allows the researcher to examine the ratio of partial correlations to zero order correlations. This algorithm also has the capacity to calculate the KMO for multiple homogeneous sets. This comes in handy for statisticians who wish to examine these relations. Also, our technique allows the researcher to resolve a KMO value outside the stipulated range by examining the features of those variables in the homogenous set whose KMO returned values outside the range.

Conclusions

Multivariate methods such as principal component analysis and factor Analysis have been used to interpret multivariate data. However, these statistical applications are not able to determine prior to their application whether a dimension exist within the multivariate data set since it is possible to have a dimensionless multivariate dataset. In addition, these statistical applications are method dependent, so it imperative to propose a method independent technique for detecting dimensionality using automated threshold settings which are thresholds generated based on the structure of the data and not the judgement of the researcher so that these statistical applications will be for purposes of interpretation or giving meaning to the data structure. Also, the formation of dimensionality in the well-known multivariate techniques is not analytically or computationally presented. They therefore offer a leave-or-take result with no understanding of the formation of the dimensions. This study therefore filled this gap by successfully proposing a method independent

dimensionality detection method using three automated threshold settings that generate data specific thresholds by allowing the data structure to generate the optimal threshold for detecting dimensionality of the multivariate data set for more accurate results. The study also established the robustness of the method using Pearson's correlation which hinges on the mean and another correlation profile that does not hinge on a statistic which is affected by extreme values, in this case order statistic which hinges on the median. The algorithm converged in all cases. Confirmatory factor analysis are carried out for confirmation of results.

Recommendations

The proposed approach for dimensionality detection shows it is threshold sensitive. It is therefore reasonable to allow the data structure to generate its own optimal threshold suitable for dimensionality detection. This will guide a reasonable application of relevant multivariate techniques on the data.

Also, the proposed method completely removes the challenge of subjectivity associated with dimensionality detection, and hence is highly recommended.

REFERENCES

- Abdi, H. (2003). Rotations. In M. S. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *Encyclopedia of Social Sciences Research Methods* (pp. 979–983). Thousand Oaks, CA: SAGE Publications, Inc.
- Achenbach, T. M. (1978). Psychopathology of childhood: Research problems and issues. *Journal of Consulting & Clinical Psychology*, 46, 759-776.
- Achenbach, T. M. (1991). *Manual for the Youth Self-Report and 1991 Profile*. Burlington VT: University of Vermont, Department of Psychiatry.
- Achenbach, T. M. (1995). Empirically based assessment and taxonomy: Applications to clinical research. *Psychological Assessment*, 7, 261-274.
- Ackerman, T. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement*, 20, 311–329.
- Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items are measuring. *Applied Measurement in Education*, 7, 255–278.
- Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practices*, 22, 37–51.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.

- Albert, J. H. (1992). Bayesian estimation of the polychoric correlation coefficient. *Journal of Statistical Computation and Simulation*, 44, 47–61.
- Alejandra Avalos-Pacheco (2018). *Factor Regression for Dimensionality Reduction and Data Integration Techniques with Applications to Cancer Data*. [Unpublished dissertation].
- Alexander, A. & Andres, V. (2004). A Comparative Study of Test Data Dimensionality Assessment Procedures Under Nonparametric IRT Models. *Applied Psychological Measurement*.28(1): 3-24.
- Allen, S. J., & Hubbard, R. (1986). Regression equations for the latent roots of random data correlation matrices with unities on the diagonal. *Multivariate Behavioral Research*, 2, 393–398.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.
- Anderson, T.W. (2003). *Introduction to multivariate Statistical Analysis*, New Jersey: Prentice Hall.
- André De Champlain & Marc E. Gessaroli (1998) *Assessing the Dimensionality of Item Response Matrices with Small Sample Sizes and Short Test Lengths*, *Applied Measurement in Education*, 11:3, 231-253.
- Apanyin. A (2021). *Canonical Correlation Techniques*. [Unpublished masters dissertation].

- Avalos Pacheco, Alejandra (2018) assessed Factor regression for dimensionality reduction, *Unpublished Doctorial Thesis, University of Warwick*.
- Ayotte, V., Saucier, J.F., Bowen, F., Laurendeau, M.C., Fournier, M., & Blais, J.G. (in press). Teaching multiethnic urban adolescents how to enhance their competencies: Effects of a middle school primary Prevention program on adaptation. *The Journal of Primary Prevention*.
- Baker, F. B., & Hubert, L. J. (1975). Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, 70, 31–38.
- Beale, E. M. L. (1969). Cluster analysis. *London: Scientific Control Systems*.
- Beck, A. T., Steer, R. A., Epstein, N., & Brown, G. (1990). *Beck Self-Concept Test. Psychological Assessment*, 2, 191-197.
- Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement*, 28, 77–92.
- Bennett, R. E., Rock, D. A., Braun, H. I., Frye, D., Spohrer, J. C., & Soloway, E. (1990). The relationship of expert-system scored constrained free-response items to multiplechoice and open-ended items. *Applied Psychological Measurement*, 14, 151–162.
- Bentler, P. M. (1990). *Comparative fit indices in structural models. Psychological Bulletin*, 107, 238-246.
- Bentler, P. M. (1990). *Comparative fit indices in structural models. Psychological Bulletin*, 107, 238-246.

Benyi, J. (2018). Some Problems Associated with Factor Analysis. *MPhil Thesis Submitted to the Univeristy of Cape Coast.*

Bernstein, I. H., & Teng, G. (1989). *Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. Psychological Bulletin, 105, 467-477.*

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: *Application of an EM algorithm. Psychometrika, 46, 443-459.*

Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). TESTFACT: Test scoring and full information item factor analysis (Version 4.0) [Computer software]. Lincolnwood, IL: *Scientific Software International.*

Boyle, G. J. (1994). Self-Description Questionnaire II: A review. *Test Critiques, 10, 632-643.*

Bridgeman, B., & Rock, D. A. (1993). Relationships among multiple-choice and openended analytical questions. *Journal of Educational Measurement, 30, 313-329.*

Brossman, B. G., & Lee, W. (2013). Observed score and true score equating procedures for multidimensional item response theory. *Applied Psychological Measurement, 37, 460-481.*

Brown, J. S., & Achenbach, T. M. (1995). Bibliography of published studies using the Child Behavior Checklist and related materials: 1995 edition. Burlington, VT: *University of Vermont, Department of Psychiatry.*

- Brown, P.J., Vannucci, M. and Fearn, T. (1998) *Multivariate Bayesian variable selection and prediction. Journal of the Royal Statistical Society, Series B*, 60, 627–641.
- Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, 36, 111–150.
- Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research*, 27, 509–540.
- Byrne, B. M. (1996). Measuring self-concept across the life span: Issues and instrumentation. Washington, DC: *American Psychological Association*.
- Byrne, B. M. (2001). *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Cai, L. (2013). flexMIRT (Version 2): Flexible multilevel multidimensional item analysis and test scoring [Computer program]. Chapel Hill, NC: Vector Psychometric Group. *Communications in Statistics*, 3, 1–27.
- Cattell, R. B. (1956). Validation and intensification of the Sixteen Personality Factor Questionnaire. *Journal of Clinical Psychology*, 12, 205-214.
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245–276.
- Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29.

- Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61, 1-36.
- Cho, S-J., Li, F., & Bandalos, D. (2009). Accuracy of the parallel analysis procedure with polychoric correlations. *Educational and Psychological Measurement*, 69, 748–759.
- Christofferson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, 40(1), 5–32.
- Cliff, N. (1988). The eigenvalues-greater-than-one rule and the reliability of components. *Psychological Bulletin*, 103, 276–279.
- Comfrey, A. L. & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Cook, L. L., Dorans, N. J., & Eignor, D. R. (1988). An assessment of the dimensionality of three SAT-Verbal test editions. *Journal of Educational Statistics*, 13, 19–43.
- Crain, R. M. (1996). The influence of age, race, and gender on child and adolescent multidimensional self-concept. In B. A. Bracken (Ed.), *Handbook of self-concept: Developmental, social, and clinical considerations* (pp. 395-420). New York: Wiley.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- D. J. Bartholomew (1980). *Factor Analysis for Categorical Data*.
- David, B., Cathy L.& Chalmers R.(2012).*Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis*. Front Psychology, 2012.

- De Ayala, R. J., & Hertzog, M. A. (1991). The assessment of dimensionality for use in item response theory. *Multivariate Behavioral Research*, 26, 765–792.
- Dorans, N. J., & Lawrence, I. M. (1999). The role of the unit of analysis in dimensionality assessment (*ETS Research Report RR-99-14*). Princeton, NJ: Educational Testing Service.
- Downing, S. M. (2006). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287–302). Lawrence Erlbaum Associates, Mahwah, NJ.
- Duda, R. O., & Hart, P. E. (1973). Pattern classification and scene analysis. *New York: Wiley*.
- Duong et al : kernel density estimation and kernel discriminant analysis for multivariate data in r'. *Journal of statistical software*, vol, 21, no..pp. 1-16,2007.
- Falk, F.& Cai, L. (2016). *A Flexible Full-Information Approach to the Modeling of Response Styles .Psychological Methods* 21(3)-(2015).
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). *New York: American Council on Education and Macmillan*.
- Finch, J.F. and West, S.G. (1997). *The Investigation of Personality Structure: Statistical Models. Journal of Research in Personality*, 31, 439-485.
- Fox, J. (2010). polycor: Polychoric and Polyserial Correlations. *R package version 0.7-8*. <http://CRAN.R-project.org/package=polycor>.

- Fraser, C. (1998). NOHARM: A Fortran program for fitting unidimensional and multidimensional normal ogive models in latent trait theory [Computer program]. *The University of New England, Center for Behavioral Studies, Armidale, Australia.*
- Frazer, C. & Mcdonald, R. (1988). Least Squares Item Factor Analysis. *Multivariate Behavioural Research*, 1988, 23, 267-269. Computer Software.
- Fu, J., Chung, S., & Wise, M. (2013). Dimensionality analysis of CBAL writing tests (ETS Research Report RR-13-10). *Princeton, NJ: Educational Testing Service.*
- Furr, R. M., & Bacharach, V. R. (2013). *Psychometrics: An introduction* (2nd ed.). Thousand Oaks, CA: *SAGE Publications, Inc.*
- Garrido, L. E., Abad, F. J., & Ponsoda, V. (2011). Performance of Velicer's minimum average partial factor retention method with categorical variables. *Educational and Psychological Measurement*, 71, 551–570.
- Garrido, L. E., Abad, F. J., & Ponsoda, V. (2013). A new look at Horn's parallel analysis with ordinal variables. *Psychological Methods*, 18, 454–474.
- Gelman, A. , Jonh, B. , Hal, S. &Donald, B. (2014). Bayesian Data Analysis *Journal of the American Statistical Association* 45(2).
- Gideon P. De Bruin (2004). *Problems with the factor analysis of items: Solutions based on item response theory and item parcelling. SA Journal of Industrial Psychology* | Vol 30, No 4 | a172 |

- Gierl, M. J., Tan, X., & Wang, C. (2005). Identifying content and cognitive dimensions on the SAT® (*College Board Research Report No. 2005-11*). New York, NY: The College Board.
- Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement*, 55, 377–393.
- Goldstein, H., & Wood, R. (1989). Five decades of item response modeling. *British Journal of Mathematical and Statistical Psychology*, 42, 139–167.
- Gorsuch, R. (1983). *Factor Analysis*. Hillsdale, NJ: L. Erlbaum Associates.
- Graham, J. W., & Hoffer, S. M. (2000). Multiple imputation in multivariate research. In T. D. Little, K. U. Schnable & J. Baumert (Eds), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 201-218). Erlbaum: Mahwah, NJ.
- Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, 19, 149–161.
- Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: American Council on Education and Praeger Publishers.
- Hakstian, A. R., & Muller, V. J. (1973). Some notes on the number of factors problem. *Multivariate Behavioral Research*, 8, 461–475.
- Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement*, 10, 287–302.

- Harter, S. (1998). The developmental of self-representations. In Damon (Series Ed.) and Eisenberg (Vol. Ed.) *Handbook of Child Psychology: Vol. 3, Social, Emotional and Personality Development*, (5th ed., pp 553-617). New York: Wiley.
- Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research*, 19, 49–78.
- Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9, 139–164.
- Hattie, J. (1992). *Self-concept*. Hillsdale, NJ: Erlbaum.
- Hattie, J. A. (1981). Decision criteria for determining unidimensionality (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (ProQuest document ID 303051687).
- Hay, I. (2000). *Gender self-concept profiles of adolescents suspended from high school*.
- Heating, K., Doherty, J., Jasper A.&Qinjuin K. (2010). *Optimization and uncertainty assessment of severely nonlinear groundwater models*. *Water Resources research*.
- Hendrickson, A. E., & White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *The British Journal of Statistical Psychology*, 17, 65–70.
- Hohensin, C., & Kubinger, K. D. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement*, 71, 732–746.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179–185.

- Houts, C. R., & Cai, L. (2013). flexMIRT® User's Manual Version 2: Flexible Multilevel Multidimensional Item Analysis and Test Scoring. Chapel Hill, NC: Vector Psychometric Group.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55.
- Hubert, L. J., & Levin, J. R. (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*, 83, 1072–1080.
- Jang, E. E., & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement*, 44, 1–21.
- Javier, R. & Carmen, E. (2017). *Bayesian Dimensionality Assessment for the Multidimensional Nominal Response Model Front. Psychol.*, 16 June 2017 Sec. Quantitative Psychology and Measurement.
- Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Upper Saddle River, NJ: Pearson.
- Joreskog, K. G., & Sorbom, D. (1993). LISREL 8: Structural equation modeling with the SIMPLIS command language. Chicago: *Scientific Software International. Journal of the Royal Statistical Society. Series B (Methodological)*.
- Jukka, C. & Mathas, V. (2010). *Bayesian assessment of dimensionality in reduced rank regression. Statistica Neerlandica*, 58(3): 255-270.
- Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika*, 56, 255–278.

- Kaiser, H. 1974. An index of factor simplicity. *Psychometrika* 39: 31–36.
- Kaiser, H. F. & Rice, J. (1974): "Little Jiffy Mark IV", *Psychometrika* 35 (December), 401 – 415.
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20, 141–151.
- Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, 15, 136–153.
- Keeling, K. B. (2000). A regression equation for determining the dimensionality of data. *Multivariate Behavioral Research*, 35, 457–468.
- Kieftenbeld, V., & Natesan, P. (2012). *Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. Applied Psychological Measurement*, 36(5), 399–419.
- Kim Rae – Him(2011): *New techniques for the dimensionality assessment of standardized test data*. [Unpublished doctoral thesis, University.
- Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford.
- Kling, K. C., Hyde, J. S., Showers, C. J. & Buswell, B. N. (1999). Gender differences in self-esteem: A meta-analysis. *Psychological Bulletin*, 125, 470-500.
- Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research*, 26, 457–477.

- Kolen, M. J., & Brennan, R. L. (2014). Test equating, scaling, and linking: *Methods and practices* (3rd ed.). New York, NY: Springer Science+Business Media.
- Kolen, M. J., & Lee, W. (2011). Introduction and overview. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 1)*. (CASMA Monograph Number 2.1). Iowa City, IA: CASMA, The University of Iowa.
- Kolen, M. J., & Lee, W. (Eds.). (2014). *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 3)*. (CASMA Monograph Number 2.3). Iowa City, IA: CASMA, The University of Iowa.
- Krus, D. J. (1975). *Order analysis of binary data matrices*. Los Angeles, CA: Theta Press.
- Krus, D. J., & Weiss, D. J. (1976). Empirical comparison of factor and order analysis on prestructured and random data. *Multivariate Behavioral Research*, 11, 95–104.
- Lancaster, H. O., & Hamdan, M. A. (1964). Estimation of the correlation coefficient in contingency tables with possibly nonmetrical characters. *Psychometrika*, 29, 383–391.
- Lautenschlager, G. J., Lance, C. E., & Flaherty, V. L. (1989). Parallel analysis criteria: Revised regression equations for estimating the latent roots of random data correlation matrices. *Educational and Psychological Measurement*, 49, 339–345.
- Leadbeater, B. J., Kupermine, G. P., Blatt, S. J., & Hertzog, C. (1999). A multivariate model of gender differences in adolescents' internalizing

- and externalizing problems. *Developmental Psychology*, 35, 1268-1282.
- Lee, S-Y. (1985). Maximum likelihood estimation of polychoric correlations in $r \times s \times t$ contingency tables. *Journal of Statistical Computation and Simulation*, 23, 53–67.
- Lee, S-Y., & Poon, W-Y. (1987). Two-step estimation of multivariate polychoric correlation. *Communications in Statistics — Theory and Methods*, 16, 307–320.
- Lee, S-Y., Poon, W-Y., & Bentler, P. M. (1990). Full maximum likelihood analyses of structural equation models with polytomous variables. *Statistics & Probability Letters*, 9, 91–97.
- Levy, J.& McManus, F. (2009). *Advancing the Bayesian Approach for Multidimensional Polytomous and Nominal IRT Models*. *Appl Psychol Meas*, 2017 Jan; 41(1): 3–16.
- Lewinsohn, P. M., Gotlib, I.H., & Seeley, J. R. (1997). Depressing-related psychosocial variables: Are they specific to depression in adolescents? *Journal of Abnormal Psychology*, 106, 365-375.
- Li, H-H., & Stout, W. F. (1995). Assessment of unidimensionality for mixed polytomous and dichotomous item data: Refinements of Poly-DIMTEST. *Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco*.
- Lissitz, R. W., Hou, X., & Slater, S. C. (2012). The contribution of constructed response items to large scale assessment: Measuring and understanding their impact. *Journal of Applied Testing Technology*, 13, 1–50.

- Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.
- Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, 18, 285–300.
- Löffler, M. (2020) explored Statistical inference in high-dimensional matrix models, *Unpublished Doctorial thesis* , University of Cambridge.
- Longman, R. S., Cota, A. A., Holden, R. R., & Fekken, G. C. (1989). A regression equation for the parallel analysis criterion in principal components analysis: Mean and 95th percentile eigenvalues. *Multivariate Behavioral Research*, 24, 59–69.
- Lord, F. M. (1970). Item characteristic curves estimated without knowledge of their mathematical form—A confrontation of Birnbaum's logistic model. *Psychometrika*, 35, 43–50.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.
- Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement*, 31, 234–250.
- M.D. Recase (2009). *Multidimensional Item Response Theory (Statistics for Social and Behavioral Sciences)*. *Psychometrika* volume 76, pages 504–506 (2011).

- MacCallum, R. C. (2009). *Factor analysis*. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 123–147). Sage Publications Ltd.
- MacCallum, R.C., Widaman, K.F., Zhang, S.B., & Hong, S.H. (1999). Sample size in factor analysis. *Psychological Methods*, 4(1), 84–99.
- Macteld, H. & Boeck, P. (2001) Multidimensional Componential Item Response Theory Models for Polytomous Items. *Applied Psychological Measurement* 25(1):19-37.
- Mag, A. J. (2009). *Efficient Feature Reduction and class and classification Methods*. Unpublished desertation, Univesitat Wien.
- Manhart, J. J. (1996). Factor analytic methods for determining whether multiple-choice and constructed-response tests measure the same construct. *Paper presented at the Annual Meeting of the National Council of Measurement in Education, New York, NY*.
- Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979) *Multivariate Analysis*. Academic Press, London.
- Marsh, H. W. & Hocevar, D. (1988). A new procedure for analysis of multitrait-multimethod data: An application of second order confirmatory factor analysis. *Journal of Applied Psychology*, 73, 107-111.
- Marsh, H. W. (1989). Age and sex effects in multiple dimensions of self-concept: Preadolescence to Early-adulthood. *Journal of Educational Psychology*, 81, 417-430.

- Marsh, H. W. (1990). A multidimensional, hierarchical model of self-concept: Theoretical and empirical justification. *Educational Psychology Review*, 2, 77-172.
- Marsh, H. W. (1992). *Self-Description Questionnaire II: Manual*. Sydney: University of Western Sydney, SELF Research Centre.
- Marsh, H. W. (1993). Academic self-concept: Theory measurement and research. In J. Suls (Ed.), *Psychological perspectives on the self* (Vol. 4, pp. 59-98). Hillsdale, NJ: Erlbaum.
- Marsh, H. W. (1997). The measurement of physical self-concept: A construct validation approach. In K. Fox (Ed.), *The physical self: From motivation to well-being* (pp. 27-58). Champaign, IL: Human Kinetics.
- Marsh, H. W., & Craven, R.G. (1997). Academic self-concept: Beyond the dustbowl. In G. Phye (Ed.) *Handbook of classroom assessment: Learning, achievement and adjustment*. Orlando, FL: Academic Press.
- Marsh, H. W., & Hattie, J (1996). Theoretical perspectives on the structure of self-concept. In B. A. Bracken (Ed.), *Handbook of self-concept* (pp 38-90). New York: Wiley.
- Marsh, H. W., & Hocevar, D. (1985). The application of confirmatory factor analysis to the study of self-concept: First and higher order factor structures and their invariance across age groups. **Psychological Bulletin**, 97, 562-582.
- Marsh, H. W., & O'Neill, R. (1984). Self Description Questionnaire III: The construct validity of multidimensional self-concept ratings by late adolescents. *Journal of Educational Measurement*, 21, 153-174.

- Marsh, H. W., & Peart, N. (1988). Competitive and cooperative physical fitness training programs for girls: Effects on physical fitness and on multidimensional self-concepts. *Journal of Sport and Exercise Psychology, 10*, 390-407.
- Marsh, H. W., & Shavelson, R. (1985). Self-concept: Its multifaceted, hierarchical structure. *Educational Psychologist, 20*, 107-125.
- Marsh, H. W., Balla, J. R., & Hau, K. T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical processes. In G. A. Marcoulides & R. E.
- Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indices in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 102*, 391-410.
- Marsh, H. W., Byrne, B. M., & Shavelson, R. (1988). A multifaceted academic self-concept: Its hierarchical structure and its relation to academic achievement. *Journal of Educational Psychology, 80*, 366-380.
- Marsh, H. W., Parada, R. H., Yeung, A. S. & Healey, J. (2001). Aggressive School Troublemakers and Victims: A Longitudinal Model Examining the Pivotal Role of Selfconcept. *Journal of Educational Psychology, 93*, 411-419.
- Marsh, H. W., Richards, G., & Barnes, J (1986). Multidimensional self-concepts: A long term follow-up of the effect of participation in an Outward Bound program. *Personality and Social Psychology Bulletin, 12*, 475-492.

- Martinson, E. O., & Hamdan, M. A. (1972). Maximum likelihood and some other asymptotically efficient estimators of correlation in two way contingency tables. *Journal of Statistical Computation and Simulation*, 1, 45–54.
- McDonald, R. P. (1981). *The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology*, 34, 100-117.
- McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, 34, 100–117.
- McDonald, R. P. (1985). Factor analysis and related methods. Hillsdale, NJ: Erlbaum.
- McDonald, R. P. (1994). Testing for approximate dimensionality. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), *Modern theories in measurement: Problems and issues* (pp. 31–61). Ottawa, Canada: Edumetrics Research Group, University of Ottawa.
- Mengyao Zhang (2016), Exploring dimensionality of scores for mixedformat tests, *Unpublihed Doctorial Thesis, University of Iowa*
- Miller, T. R., & Hirsch, T. M. (1992). Cluster analysis of angular data in applications of multidimensional item-response theory. *Applied Measurement in Education*, 5, 193– 211.
- Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50, 159–179.

- Montanelli, R. G., & Humphreys, L. G. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: A Monte Carlo study. *Psychometrika*, 41, 341–348.
- Mroch, A. A., & Bolt, D. M. (2006). A simulation comparison of parametric and nonparametric dimensionality detection procedures. *Applied Measurement in Education*, 19, 67–91.
- Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, 19, 73–90.
- Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, 43, 551–560.
- Muthén, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics*, 22, 43–65.
- Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49, 115–132.
- Muthén, B., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. *Unpublished technical report*. Retrieved from http://www.statmodel.com/bmuthen/articles/Article_075.pdf.
- Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus User's Guide*. Seventh Edition. Los Angeles, CA: Muthén & Muthén.
- Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement*, 28, 99–117.

- Nandakumar, R. (1994). Assessing dimensionality of a set of item responses: comparison of different approaches. *Journal of Educational Measurement*, 31, 17–35.
- Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational and Behavioral Statistics*, 18, 41–68.
- Nandakumar, R., Yu, F., Li, H-H., & Stout, W. (1998). Assessing unidimensionality of polytomous data. *Applied Psychological Measurement*, 22, 99–115.
- Nkansah, B. K. (2018). On the Kaiser-Meier-Olkin's Measure of Sampling Adequacy. *Journal of Mathematical Theory and Modeling*, 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.
- O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, 32, 396–402.
- Olson, C. L. (1979). *Practical considerations in choosing a MANOVA test statistic: A rejoinder to Stevens*. *Psychological Bulletin*, 86(6), 1350–1352.
- Olsson, U. (1979a). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44, 443–460.

- Olsson, U. (1979b). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research*, 14, 485–500.
- Pearson, K. (1901). Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 195, 1–47+405.
- Perkhounkova, Y., & Dunbar, S. B. (1999). Influences of item content and format on the dimensionality of tests combining multiple-choice and open-response items: An application of the Poly-DIMTEST procedure. *Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada.*
- Pett, M.A., Lackey, N.R. and Sullivan, J.J. (2003) *Making Sense of Factor Analysis: The Use of Factor Analysis for Instrument Development in Health Care Research*. SAGE Publications, Thousand Oaks. <http://dx.doi.org/10.4135/9781412984898>.
- R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.Rproject.org/>.
- Raïche, G., Walls, T. A., Magis, D., Riopel, M., & Blais, J-G. (2013). Non-graphical solutions for Cattell's scree test. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 9, 23–29.

- Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. (1965). Generalizability of stratified parallel tests. *Psychometrika*, 30, 39–56.
- Rashid, Moonson (2013). *Assessing Distributions Properties of High Dimensional Data*. [Unpublished Doctoral Dissertation], John Hopkins University.
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4, 207–230.
- Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.
- Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). *Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance*. *Psychological Bulletin*, 114, 552-566.
- Reise, V. (199). *Effects of Estimation Methods on Making Trait-Level Inferences from Ordered Categorical Items for Assessing Psychopathology*. *American psychological Assessment*.
- Rencher, A.C. (2002). *Methods of Multivariate Analysis: (2nd Ed.)*. New Jersey: John Wiley & Sons.
- Robinson, N. S., Garber, J., & Hilsman, R. (1995). *Cognitions and stress: Direct and moderating effects of depressive versus externalizing symptoms during the junior high transition*. *Journal of Abnormal Psychology*, 104, 453-463.
- Rodriguez, M. (2003). Construct equivalence of multiple-choice and constructed response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, 40, 163–184.

- Sajjan, Mahantesh. (2020). Re: *70% training and 30% testing spit method* ,<https://www.researchgate.net>.
- Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.
- Sass, D. A., & Schmitt, T. A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research*, 45, 73–103.
- Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL internet-based test. *Language Testing*, 26, 5–30.
- Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–353). Westport, CT: American Council on Education and Praeger Publishers.
- Schumacker (Eds.), *Advanced structural equation modeling techniques* (pp. 315-353). Hillsdale, NJ: Erlbaum.
- Schwarz, G. (1978). Estimation the dimension of a model. *Annals of Statistics*, 6, 461– 464.
- Segall, M. H., Lonner, W. J., & Berry, J. W. (1998). Cross-cultural psychology as a scholarly discipline: On the flowering of culture in behavioral research. *American Psychologist*, 53, 1101-1110.
- Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Validation of construct interpretations. *Review of Educational Research*, 46, 407-441.
- Singelis, T. M. (2000). Some thoughts on the future of cross-cultural social psychology. *Journal of Cross-Cultural Psychology*, 31, 76-91.

- Sinharray, s. & Russel, A. (2006). *Assessing Fit of Cognitive Diagnostic Models. Educational and Psychological Measurement* 67(2):239-257.
- Song, L-Y, Singh, J., & Singer, M. (1994). The Youth Self Report inventory: A study of its measurement fidelity. *Psychological Assessment*, 6, 236-245.
- SPSS (1999). *SPSS for Windows (Release 10.0.5)*. Chicago: SPSS.
- Stone, C. A., & Yeh, C-C. (2006). Assessing the dimensionality and factor structure of multiple-choice exams: An empirical comparison of methods using the Multistate Bar Examination. *Educational and Psychological Measurement*, 66, 193–214.
- Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52, 589–617.
- Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, 55, 293–325.
- Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, 20, 331–354.
- Sue, S. (1999). Science, ethnicity and bias: Where have we gone wrong? *American Psychologist*, 54., 1070-1077
- Svetina, D., & Levy, R. (2012). An overview of software for conducting dimensionality assessment in multidimensional models. *Applied Psychological Measurement*, 36, 659–669.

- Svetina, D., & Levy, R. (2014). A framework for dimensionality assessment for multidimensional item response models. *Educational Assessment*, 19, 35–57.
- Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.) *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp. 82– 98). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.
- T. Siva Tian (2009) explored Dimensionality reduction for classification with high-dimensional data, *Unpublished Doctorial Thesis, University of Southern California*.
- Tabachnick, B.G. and Fidell, L.S. (2001) *Using Multivariate Statistics. 4th Edition, Allyn and Bacon, Boston*.
- Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52, 393–408.
- Tallis, G. M. (1962). The maximum likelihood estimation of correlation from contingency tables. *Biometrics*, 18, 342–353.
- Tate, R. (2002). Test dimensionality. In J. Tindal & T. M. Haladyna (Eds.), *Large scale assessment programs for all students: Development, implementation, and analysis* (pp. 155–181). Mahwah, NJ: Lawrence Erlbaum.
- Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, 27, 159–203.

- Tatsuoka, M. M. (1971). *Multivariate analysis: Techniques for educational and psychological research*. New York: Wiley.
- Thinesh, Pathmanathan (2018). *Dimension reduction and clustering of high dimensional data using a mixture of generalized hyperbolic distributions*. [*Unpublished Masters Thesis*], University of McMaster.
- Thissen, D., Wainer, H., & Wang, X. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement*, 31, 113–123.
- Thissen, M., Darion, D.&Van, O. (2010). *Integration and Convergence in Regional Europe: European Regional Trade Flows from 2000 to 2010*. Netherlands Environmental Assessment Agency. PBL publication number: 1036.
- Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association. (International Standard Book Number: 1-59147-093-5).
- Thurstone, L. L. (1935). *The vectors of mind: Multiple-factor analysis for the isolation of primary traits*. Chicago, IL: The University of Chicago Press.
- Thurstone, L. L. (1947). *Multiple-factor analysis: A development and expansion of The Vectors of Minds*. Chicago, IL: The University of Chicago Press.

- Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, 16, 209–220.
- Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 29–44). Lawrence Erlbaum Associates, Hillsdale, NJ.
- Traub, R. E., & Fisher, C. W. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement*, 1, 355–369.
- Traub, R. E., & MacRury, K. (1990). Antwort-auswahl- vs freie-antwort-aufgaben bei lernerfolgs-tests [Multiple-choice vs. free-response in the testing of scholastic achievement]. In K. Ingenkamp & R. S. Jäger (Eds.), *Tests und trends 8: Jahrbuch der pädagogischen diagnostic* (pp. 128–159). Weinheim, Germany: Beltz Verlag.
- Uebersax JS. *Factor analysis and SEM with tetrachoric and polychoric correlations*. *Statistical Methods for Rater Agreement web site*. 2006. Available at: <http://john-uebersax.com/stat/sem.htm>.
- Van Abswoude, A. A. H., van der Ark, L. A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement*, 28, 3–24.

- Van, E., Cees, R., Jonathan (2015) : *Factor analysis of survey data – assessing the probability of incorrect dimensionalisation. PLoS ONE*,10 (3).
0118900/1-0118900/31. ISSN 1932-6203.
- Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41, 321–327.
- Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). *Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components*. In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 41–71). Boston, MA: Kluwer Academic Publishers.
- Vermont, K. &Magidson, J. (2016). *Latent Class Cluster Analyses. Applied Latent Class Analysis* Vol. 42, No. 3 (1980), pp. 293-321 (29 pages)
Volume 46.Issue 10.
- Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24, 185–201.
- Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement*, 27, 1–14.
- Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practices*, 15, 22–29.
- Walker. C.M. (2006). *Statistical Versus Substantive Dimensionality: The Effect of Distributional Differences on Dimensionality Assessment Using*

DIMTEST. Educational and Psychological Measurement 2006 / 10

Vol. 66; Iss. 5.

- Waller, N. G. (1995). *MicroFACT 1.0: A microcomputer factor analysis program for ordered polytomous data and mainframe sized problems*. St. Paul, MN: Assessment Systems Corporation.
- Weng, L., & Cheng, C. (2005). Parallel analysis with unidimensional binary data. *Educational and Psychological Measurement*, 65, 697–716.
- Wijbrandt van Schuur, 2010, "Replication data for: Mokken Scale Analysis: A Nonparametric Version of Guttman Scaling for Survey Research", <https://doi.org/10.7910/DVN/8VWE6A>, Harvard Dataverse, V1.
- Wilson, K. M. (2000). An exploratory dimensionality assessment of the TOEIC test (*ETS Research Report RR-00-14*). Princeton, NJ: Educational Testing Service.
- Wilson, K. M., & Graves, K. (1999). Validity of the secondary level English proficiency test at Temple University — *Japan (ETS Research Report RR-99-11)*. Princeton, NJ: Educational Testing Service.
- Wise, S. L. (1983). Comparisons of order analysis and factor analysis in assessing the dimensionality of binary data. *Applied Psychological Measurement*, 7, 311–321.
- Wothke, W. (1993). Nonpositive definite matrices in structural modeling. In: K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 256–293). Newbury Park, CA: Sage.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187–213.

- Yen, W. M., & Fitzpatrick, A. R. (2006). Item Response Theory. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 111–153). Westport, CT: American Council on Education and Praeger Publishers.
- Zhang, J., & Stout, W. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika*, 64, 129–152.
- Zhang, J., & Stout, W. (1999b). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, 64, 213–249.
- Zhang, M., Kolen, M. J., & Lee, W. (2014). A comparison of test dimensionality assessment approaches for mixed-format tests. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (volume 3). (CASMA Monograph Number 2.3). Iowa City, IA: CASMA, The University of Iowa.
- Zupluoglu Cengiz (2013), assessed Dimensionality of Latent Structures Underlying Dichotomous Item Response Data with Imperfect Models, *Unpublished Doctorial Thesis, University of Minnesota* .
- Zwick, R. J. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement*, 24, 293–308.

APPENDICES

APPENDIX A

DIMENSIONALITY DETECTION CODES – PEARSON'S

CORRELATION APPROACH

```
#=====
#Dimension detection in
#Multivariate datasets
#=====
library(mvtnorm)
library(MASS)
library(pscl)
library(Matrix)
library(foreign)

Dim_Detector<-function(Mdata,thold=0.5){

#Function to compute correlation matrix
Corrv<-function(mdata){

Cormat<-matrix(0,dim(mdata)[2],dim(mdata)[2])

for(i in 1: dim(Cormat)[1]){

for(j in 1:dim(Cormat)[2]){

Cormat[i,j]<-cor(mdata[,i],mdata[,j])
```

```
    }  
  }  
  Cormat  
  }  
  #=====  
  #Function to select first  
  #Spanning set based on  
  #Correlation Coefficient  
  #=====  
  Span_set<-function(xmat){  
    max_val<-max(xmat)  
  
    stvl<-function(x,max_v){  
      ifelse(x==max_v,1,0)  
    }  
  
    Rid<-lapply(seq_len(dim(xmat)[1]),function(i){  
      stv<-stvl(xmat[i,],max_val)  
      ld<-which(stv==1)  
      c(i,ld)  
    })  
  
    #Sr<-NULL  
  
    ls_len<-unlist(lapply(seq_len(length(Rid)),function(j){  
      Vecx<-as.vector(Rid[[j]])
```



```
sl<-length(Vecx)
sl}))
#if(any(ls_len==2)){
S<-Rid[[which(ls_len==2)]]
S<-c(S[2],S[1])
#}else{
#break
#}
S}
#-----
Vbind<-function(X,y){
#=====
#Function to combine vectors of
#different lengths
#=====
mbind<-function(x,y){
slab<-NULL
a<-dim(x)[1];b<-length(y)
if(a==b){slab<-cbind(x,y)}
if(a>b){slab<-
cbind(x, y=c(y, rep(NA,(a-b))))
}
```



```
slab
}

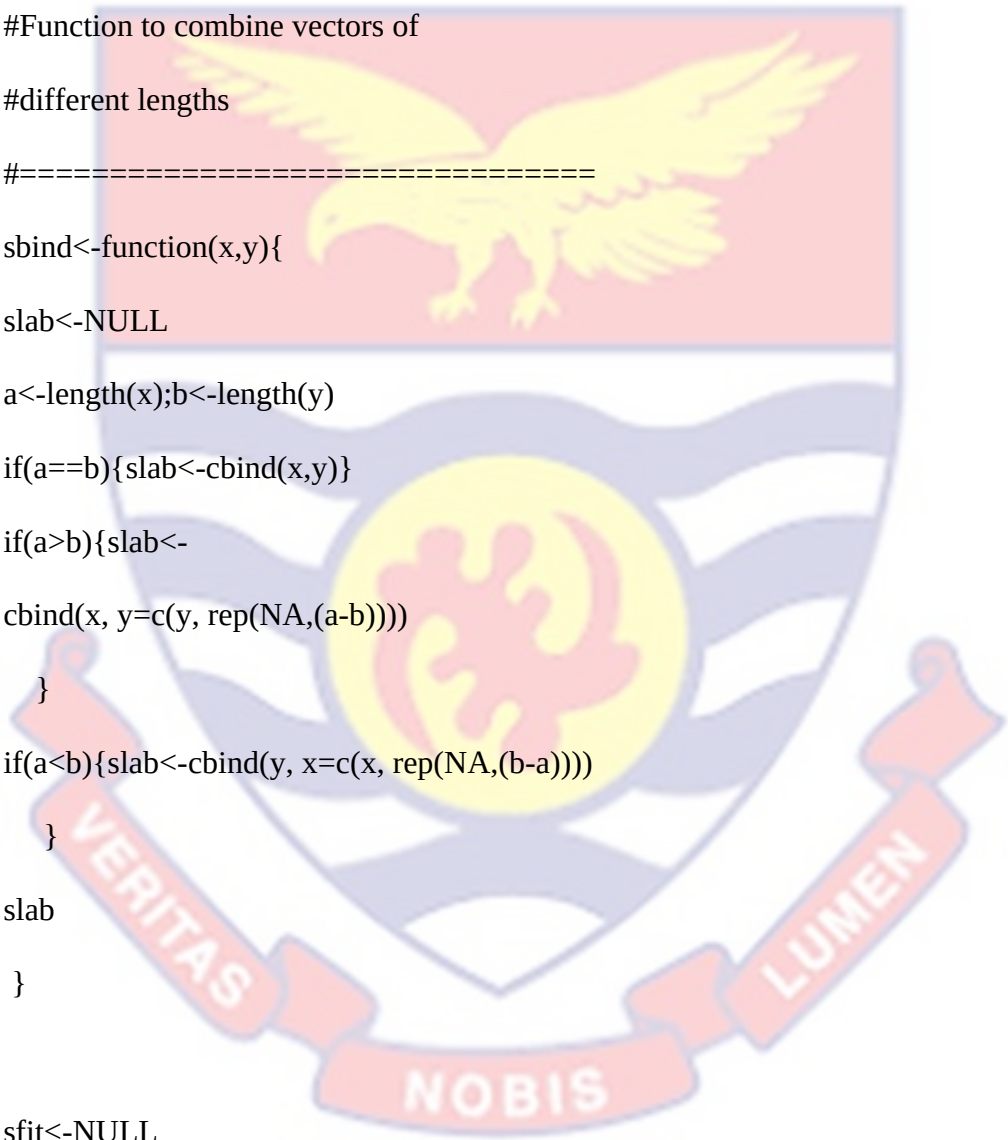
#=====

#Function to combine vectors of
#different lengths
#=====

sbind<-function(x,y){
slab<-NULL
a<-length(x);b<-length(y)
if(a==b){slab<-cbind(x,y)}
if(a>b){slab<-
cbind(x, y=c(y, rep(NA,(a-b))))
}
if(a<b){slab<-cbind(y, x=c(x, rep(NA,(b-a))))
}
slab
}

sfit<-NULL

if(length(X)==0){sfit<-sbind(X,y)}else{
if(length(X)>1 & is.matrix(X)=="TRUE"){
sfit<-mbind(X,y)
```



```
}  
}  
sfit }  
  
Foutput<-function(rmat){  
  Dresult<-sapply(seq_len(dim(rmat)[2]),function(i){  
    rmat[,i])  
  
    Dresult}  
  
#=====  
#Updating spanning function  
#Based on pairwise Correlation  
#=====  
CompwS<-function(S_index,Cor_Mat,thold){  
  
  Sxupdator<-function(Sx_set,x_dex,Cor_mat,thold){  
    pwcr<-unlist(lapply(seq_len(length(Sx_set)),function(i){  
      Crr<-c(Cor_mat[Sx_set[i],x_dex],Cor_mat[x_dex,Sx_set[i]])  
  
      rr<-which(Crr==0)  
  
      Drr<-Crr[-rr]  
  
      Drr  
    })))  
  
    if( all(pwcr>=thold)=="TRUE"){  
  
      Sx_set<-c(Sx_set,x_dex)}else{Sx_set<-Sx_set}
```

```
Sx_set}
```

```
N<-dim(Cor_Mat)[2]
```

```
m<-seq_len(N)[-S_index]
```

```
New_set<-S_index
```

```
for(i in 1:length(m)){
```

```
NS<-Sxupdater(New_set,m[i],Cor_Mat,thold)
```

```
New_set<-NS
```

```
}
```

```
Hset<-New_set;Nhset<-seq_len(N)[-Hset]
```

```
list(Hset=Hset,Nhset=Nhset)}
```

```
#Start sequential updating
```

```
tol<-2
```

```
clab<-colnames(data.frame(Mdata))
```

```
Srecord<-NULL
```

```
count<-0
```

```
Cmat<-Corrv(Mdata)
```

```
if(any(Cmat<0)=="TRUE"){
```

```
Cmat<-abs(Cmat)}else{
```

```
Cmat<-Cmat}
```

```
crmat<-as.matrix(tril(Cmat));diag(crmat)<-0
```

```
d<-dim(crmat)[2]
```

```
lhset<-d
```

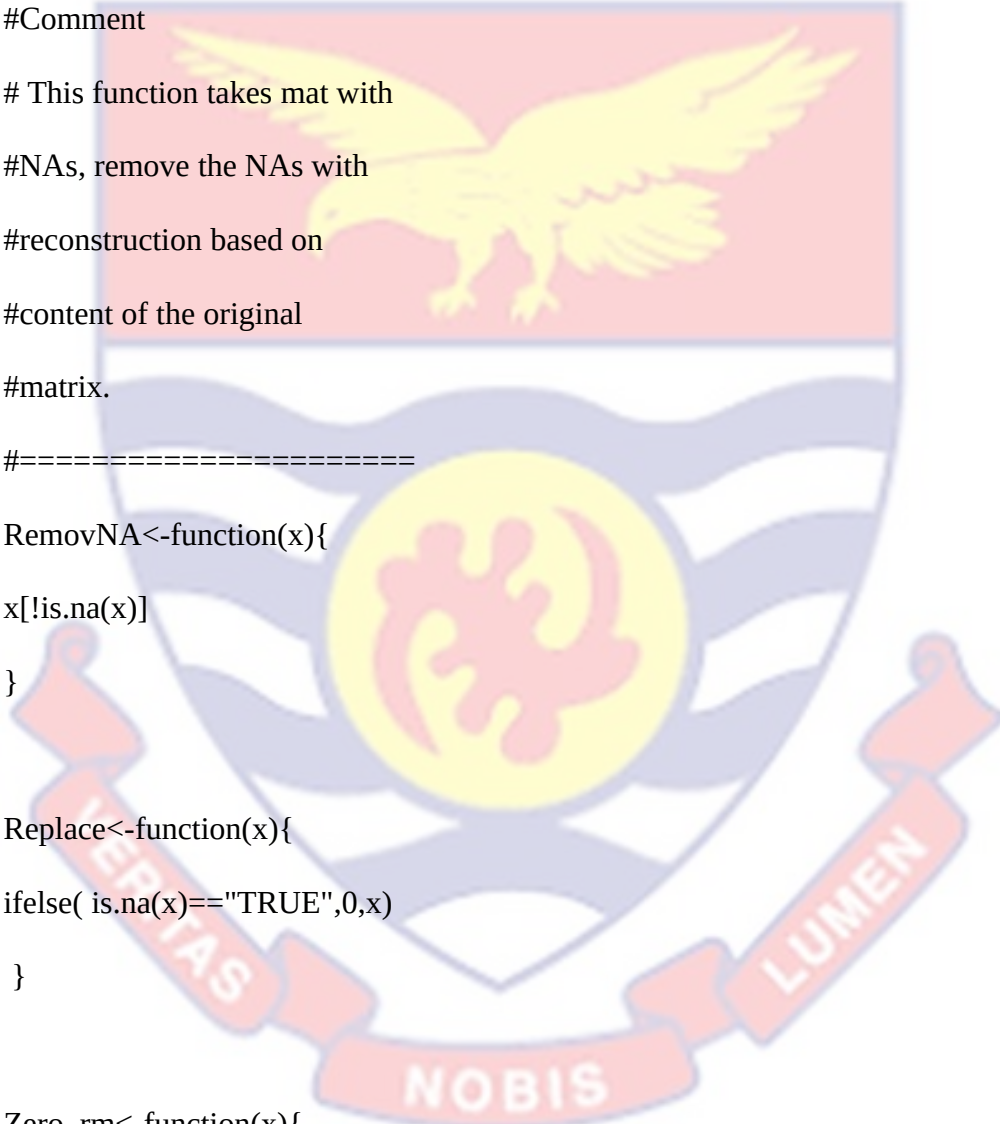
```
Nlab<-seq_len(d)
```

```
S<-Span_set(crmat)
```

```
while(lnhset>tol){  
  count<-count+1  
  fit<-CompwS(S,crmat,thold)  
  hset<-fit$Hset  
  nhset<-fit$Nhset  
  mlab<-clab[hset]  
  rlab<-clab[-hset]  
  Srecord<-Vbind(Srecord,mlab)  
  Rcrmat<-crmat[-hset,-hset]  
  NHset<-rlab  
  if(is.matrix(Rcrmat)=="FALSE"){  
    break  
  }  
  S_new<-Span_set(Rcrmat)  
  S<-S_new;crmat<-Rcrmat  
  clab<-rlab  
  lnhset<-length(nhset)  
  cat(count,lnhset,S,mlab,"\n")  
}  
  
if(count>1 & lnhset<tol){  
  cat("Algorithm failed to converge","\n")  
  return(NULL)  
}else{cat("Algorithm covered","\n")  
  list(Srecord=Srecord,NHset=NHset)}
```



```
}  
  
#-----  
  
Screen_result<-function(mat_result){  
#=====
```



```
#Comment  
# This function takes mat with  
#NAs, remove the NAs with  
#reconstruction based on  
#content of the original  
#matrix.  
#=====
```

```
RemovNA<-function(x){  
x[!is.na(x)]  
}  
  
Replace<-function(x){  
ifelse( is.na(x)=="TRUE",0,x)  
}  
  
Zero_rm<-function(x){  
x[-which(x==0)]  
}  
  
MatRed<-function(mat){
```

```
fx<-function(x){  
  ifelse(all(x==0)=="TRUE",1,0)}  
nd<-as.vector(apply(mat,2,fx))  
Nmat<-mat[,-which(nd==1)]  
Nmat}
```

```
Rlist<-function(Mmat){  
  
  Ttf<-function(ylist){  
    vv<-NULL  
    if(any(ylist==0)){vv<-Zero_rm(ylist)}else{vv<-ylist[]}  
    vv  
  }  
  if(is.matrix(Mmat)=="TRUE"){  
    rlist<-apply(Mmat,2,list)  
    fresult<-lapply(seq_len(length(rlist)),function(j){  
      vv<-NULL  
      tts<-unlist(rlist[[j]])  
      if(any(tts==0)){vv<-Zero_rm(tts)}else{vv<-tts[]}  
    })  
  }  
}
```

```
if(is.vector(Mmat)=="TRUE"){
```

```
fresult<-Ttf(Mmat)
```

```
}  
  
fresult}  
  
#-----  
  
#Call  
#-----  
smat<-Replace(mat_result)  
mmat<-MatRed(smat)  
  
Fresult<-Rlist(mmat)  
Fresult}  
#-----  
  
Data.names<-function(Data_frame,rlist){  
  
Data_frame<-if(is.data.frame(Data_frame)=="FALSE"){  
Data_frame<-data.frame(Data_frame)  
}  
}  
}  
}  
  
n.set<-names(Data_frame)  
  
xsame<-function(x,y){  
which(x==y)}  
  
smvec<-function(xdata,ystand){  
  
ufit<-unlist(lapply(seq_len(length(xdata)),function(i){
```

```
xsame(ystand,xdata[i])
)))
ufit
}
data_var<-lapply(seq_len(length(rlist)),function(j){
smvec(rlist[[j]],n.set)
)
Hdata<-sapply(seq_len(length(data_var)),function(t){
Data_frame[,data_var[[t]]
})
list(data_var=data_var,Hdata=Hdata)}
#-----
#Algorithm to compute KMO of a Multivariate dataset
#-----
KMO_Val<-function(mdata){
#=====
#Comments:
#This function computes
#the KMO of a multivariate
#data
#-----
```

```
R<-cor(as.matrix(mdata))  
Qmat<-function(Rm){  
  RI<-solve(Rm)  
  Dm<-sqrt(diag(diag(RI)))  
  Dr<-solve(Dm)  
  Q<-(Dr%*%RI)%*%Dr  
  Q}  
Q<-do.call(Qmat,list(R))  
  
#Function to compute  
Sr_sq<-function(CMat){  
  rsq<-NULL  
  for(i in 1:dim(CMat)[1]){  
    for(j in 1:dim(CMat)[2]){  
      if(i<j){  
        rsq<-cbind(rsq,CMat[i,j])  
      }  
    }  
  }  
  Rsq<-sum(as.vector(rsq)^2)  
  Rsq  
}
```



```
Sum_Rsq<-Sr_sq(R)
Sum_Pr_sq<-Sr_sq(Q)

KMO<-1/(1+(Sum_Pr_sq/Sum_Rsq))
KMO}

#list(KMO=KMO,Sum_Rsq=Sum_Rsq,Sum_Pr_sq=Sum_Pr_sq,Q=Q,R=R
)})
#=====END OF ALGORITHM=====

VData_KMO<-function(Mdata,result_list){

kmo_oneH<-function(xdata){
km<-KMO_Val(xdata)
km}
kmo_twomore<-function(hdata_list){
lapply(hdata_list,KMO_Val)
}

dkmo<-lapply(seq_len(length(result_list)),function(i){
Rmat<-Screen_result(result_list[[i]]$Srecord)
HData_set<-Data.names(Mdata,Rmat)
if(is.list(Rmat)=="TRUE"){kmos<-kmo_twomore(HData_set$Hdata)
}else{kmos<-kmo_oneH(HData_set$Hdata)
}
}
```

```
kmos
```

```
})
```

```
dkmo
```

```
}
```

```
#Sensitivity analysis
```

```
#=====
```

```
SenAnalysis<-function(Mdata,set_thold){
```

```
m<-length(set_thold)
```

```
Sresult<-lapply(seq_len(m),function(i){
```

```
Dim_Detector(Mdata,set_thold[i])
```

```
})
```

```
Sresult}
```

```
#=====
```

```
#Data driven threshold setting
```

```
#=====
```

```
strehod<-function(Mdata){
```

```
CMat<-Corrv(Mdata)
```

```
Lmat<-as.matrix(tril(CMat));diag(Lmat)<-0
```

```
Nmat<-as.vector(Lmat)
```

```
NCrmat<-Nmat[-which(Nmat==0)]
```

```
crange<-round(range(NCrmat),2)
```

```
sth<-seq(crange[1],crange[2],0.01)
a<-((crange[2]-crange[1])/12)
b<-round(a,2)
st2<-seq(crange[1],crange[2],b)
st3<-sth[which(sth>=median(sth))]
list(crange=crange,sth=sth,a=a,st2=st2,st3=st3)}

#Example
#Import data into R
Datta<-read.spss("Speformance.sav", use.value.label=TRUE,
to.data.frame=TRUE)
mdata<-data.matrix(Datta)[-c(8,9)]
colnames(mdata)<-c("X1","X2","X3","X4","X5","X6","X7")
Result<-Dim_Detector(mdata,thold=0.5)
Rmat<-Screen_result(Result$Srecord)
HData_set<-Data.names(mdata,Rmat)
kmo_data<-KMO_Val(mdata)
kmo_Hset1<-KMO_Val(HData_set$Hdata[[1]])
kmo_Hset2<-KMO_Val(HData_set$Hdata[[2]])

#Example 2
#Import data into R
Datta<-read.spss("concreteStrength.sav", use.value.label=TRUE,
to.data.frame=TRUE)
mdata<-data.matrix(Datta)
```

```
colnames(mdata)<-c("X1","X2","X3","X4","X5","X6","X7")
```

```
Result<-Dim_Detector(mdata,thold=0.)
```

```
#Example 2: Dataset 3
```

```
#Import data into R
```

```
Datta<-read.spss("FATHERdata2.sav", use.value.label=TRUE,  
to.data.frame=TRUE)
```

```
mdata<-data.matrix(Datta)[,c(5:24)]
```

```
colnames(mdata)<-
```

```
c("X1","X2","X3","X4","X5","X6","X7","X8","9","X10","X11","X12","X  
13","X14","X15","X16","X17","X18","X19","X20")
```

```
Result<-Dim_Detector(mdata,thold=0.34)
```

```
#Sensitivity analysis
```

```
#-----
```

```
#Example
```

```
#Import data into R
```

```
Datta<-read.spss("Speformance.sav", use.value.label=TRUE,  
to.data.frame=TRUE)
```

```
mdata<-data.matrix(Datta)[,-c(8,9)]
```

```
colnames(mdata)<-c("X1","X2","X3","X4","X5","X6","X7")
```

```
sthod1<-strehod(mdata)$sth
```

```
sthod2<-strehod(mdata)$st2
```

```
sthod3<-strehod(mdata)$st3
```

```
Sfit<-SenAnalysis(mdata,sthod)
Sfit2<-SenAnalysis(mdata,sthod2)
Sfit3<-SenAnalysis(mdata,sthod3)
```

```
kmo_val<-VData_KMO(mdata,Sfit3)
```

```
KMO_LS<-unlist(lapply(seq_len(length(kmo_val)),function(i){
lc<-length(kmo_val[i])
kkmo<-ifelse(lc>1,max(unlist(kmo_val[i])),unlist(kmo_val[i]))
kkmo}))
```

Dimensionality Detection Codes: Order Statistics Approach

Algorithm 3: Order statistics correlation approach

```
#=====
#Order Statistic correlation approach to
#Dimensionality Detection
#=====
#Dimension detection in
#Multivariate datasets
#=====
library(mvtnorm)
library(MASS)
library(pscl)
library(Matrix)
library(foreign)
library(corrplot) # For correlation plot (Correlogram)
```



```
#=====
#Dimension detection in
#Multivariate datasets
#=====

library(mvtnorm)
library(MASS)
library(pscl)
library(Matrix)
library(foreign)

Dim_Detector<-function(Mdata,thold=0.5){
#-----
#Order statistics correlation
#====Function=====
OStat_Cor<-function(mydata){
mat <- as.matrix(mydata)
n <- ncol(mat)
cor.mat<- matrix(NA, n, n)
corr.c<-function(x,y){
Dnum.x<-Num.x<-numeric(length(x))
x.new<-x[order(x,decreasing=FALSE)]
y.new<-y[order(y,decreasing=FALSE)]
for(j in 1:length(x.new)){
Num.x[j]<-(x.new[j]-x[length(x.new)-j+1])*y[order(x,decreasing=FALSE)][j]
Dnum.x[j]<-(x.new[j]-x[length(x.new)-j+1])*y.new[j]}
cr.x<-sum(Num.x)/sum(Dnum.x)
```

```
rx<-round(cr.x,5)
return(rx)}

Cvmat<-function(n,mat,cor.mat){
for (i in 1:n) {
  for (j in 1:n) {
    cor.mat[i, j] <-corr.c(mat[,i],mat[,j])
  }
}
return(cor.mat)}

cormat<-Cvmat(n,mat,cor.mat)
#colnames(cormat)<-rownames(cormat) <-colnames(mydata)
return(cormat)}

#=====
#Function to select first
#Spanning set based on
#Correlation Coefficient
#=====
Span_set<-function(xmat){
max_val<-max(xmat)

stvl<-function(x,max_v){
ifelse(x==max_v,1,0)
}
}
```

```
Rid<-lapply(seq_len(dim(xmat)[1]),function(i){
  stv<-stvl(xmat[i,],max_val)
  ld<-which(stv==1)
  c(i,ld)
})
#Sr<-NULL
ls_len<-unlist(lapply(seq_len(length(Rid)),function(j){
  Vecx<-as.vector(Rid[[j]])
  sl<-length(Vecx)
  sl}))
#if(any(ls_len==2)){
  S<-Rid[[which(ls_len==2)]]
  S<-c(S[2],S[1])
  #}else{
  #break
  #}
S}
#-----
Vbind<-function(X,y){
#=====
#Function to combine vectors of
#different lengths
#=====
mbind<-function(x,y){
slab<-NULL
```

```
a<-dim(x)[1];b<-length(y)
if(a==b){slab<-cbind(x,y)}
if(a>b){slab<-
cbind(x, y=c(y, rep(NA,(a-b))))
}
slab
}
#=====
#Function to combine vectors of
#different lengths
#=====
sbind<-function(x,y){
slab<-NULL
a<-length(x);b<-length(y)
if(a==b){slab<-cbind(x,y)}
if(a>b){slab<-
cbind(x, y=c(y, rep(NA,(a-b))))
}
if(a<b){slab<-cbind(y, x=c(x, rep(NA,(b-a))))
}
slab
}
sfit<-NULL
if(length(X)==0){sfit<-sbind(X,y)}else{
if(length(X)>1 & is.matrix(X)=="TRUE"){
```

```
sfit<-mbind(X,y)

}

}

sfit }

Foutput<-function(rmat){

Dresult<-sapply(seq_len(dim(rmat)[2]),function(i){

rmat[,i]})

Dresult}

#=====

#Updating spanning function

#Based on pairwise Correlation

#=====

CompwS<-function(S_index,Cor_Mat,thold){

Sxupdator<-function(Sx_set,x_dex,Cor_mat,thold){

pwcr<-unlist(lapply(seq_len(length(Sx_set)),function(i){

Crr<-c(Cor_mat[Sx_set[i],x_dex],Cor_mat[x_dex,Sx_set[i]])

rr<-which(Crr==0)
```



```
Drr<-Crr[-rr]

Drr

)))

if( all(pwcr>=thold)==TRUE){

Sx_set<-c(Sx_set,x_dex)}else{Sx_set<-Sx_set}

Sx_set}

N<-dim(Cor_Mat)[2]

m<-seq_len(N)[-S_index]

New_set<-S_index

for(i in 1:length(m)){

NS<-Sxupdator(New_set,m[i],Cor_Mat,thold)

New_set<-NS

}

Hset<-New_set;Nhset<-seq_len(N)[-Hset]

list(Hset=Hset,Nhset=Nhset)}

#Start sequential updating

tol<-2

clab<-colnames(data.frame(Mdata))

Srecord<-NULL
```

```
count<-0
Cmat<-OStat_Cor(Mdata)
if(any(Cmat<0)== "TRUE"){
Cmat<-abs(Cmat)}else{
Cmat<-Cmat}
crmat<-as.matrix(tril(Cmat));diag(crmat)<-0
d<-dim(crmat)[2]
lnhset<-d
Nlab<-seq_len(d)
S<-Span_set(crmat)
while(lnhset>tol){
count<-count+1
fit<-CompwS(S,crmat,thold)
hset<-fit$Hset
nhset<-fit$Nhset
mlab<-clab[hset]
rlab<-clab[-hset]
Srecord<-Vbind(Srecord,mlab)
Rcrmat<-crmat[-hset,-hset]
NHset<-rlab
if(is.matrix(Rcrmat)== "FALSE"){
break
}
S_new<-Span_set(Rcrmat)
S<-S_new;crmat<-Rcrmat
```

```
clab<-rlab
Inhset<-length(nhset)
cat(count,Inhset,S,mlab,"\n")
}
if(count>1 & Inhset<tol){
cat("Algorithm failed to converge","\n")
return(NULL)
}else{cat("Algorithm covered","\n")
list(Srecord=Srecord,NHset=NHset)}
}
```

```
#-----
Screen_result<-function(mat_result){
#=====
#Comment
# This function takes mat with
#NAs, remove the NAs with
#reconstruction based on
#content of the original
#matrix.
#=====
RemovNA<-function(x){
x[!is.na(x)]
}
Replace<-function(x){
```

```
ifelse( is.na(x)=="TRUE",0,x)  
}
```

```
Zero_rm<-function(x){  
x[-which(x==0)]  
}
```

```
MatRed<-function(mat){
```

```
fx<-function(x){
```

```
ifelse(all(x==0)=="TRUE",1,0)}
```

```
nd<-as.vector(apply(mat,2,fx))
```

```
Nmat<-mat[,-which(nd==1)]
```

```
Nmat}
```

```
Rlist<-function(Mmat){
```

```
Ttf<-function(ylist){
```

```
vv<-NULL
```

```
if(any(ylist==0)){vv<-Zero_rm(ylist)}else{vv<-ylist[]}
```

```
vv
```

```
}
```

```
if(is.matrix(Mmat)== "TRUE"){  
  
rlist<-apply(Mmat,2,list)  
  
fresult<-lapply(seq_len(length(rlist)),function(j){  
  
vv<-NULL  
  
tts<-unlist(rlist[[j]])  
  
if(any(tts==0)){vv<-Zero_rm(tts)}else{vv<-tts[]}  
  
})  
  
}  
  
if(is.vector(Mmat)== "TRUE"){  
  
fresult<-Ttf(Mmat)  
  
}  
  
fresult}  
  
#-----  
  
#Call  
  
#-----  
  
smat<-Replace(mat_result)  
  
mmat<-MatRed(smat)
```




```
Fresult<-Rlist(mmat)

Fresult}

#-----

Data.names<-function(Data_frame,rlist){
Data_frame<-if(is.data.frame(Data_frame)=="FALSE"){
Data_frame<-data.frame(Data_frame)
}else{Data_frame<-Data_frame}

n.set<-names(Data_frame)

xsame<-function(x,y){
which(x==y)}

smvec<-function(xdata,ystand){
ufit<-unlist(lapply(seq_len(length(xdata)),function(i){
xsame(ystand,xdata[i])
}))
ufit
}

data_var<-lapply(seq_len(length(rlist)),function(j){
smvec(rlist[[j]],n.set)
})

Hdata<-sapply(seq_len(length(data_var)),function(t){

Data_frame[,data_var[[t]]]
```

```
})  
  
list(data_var=data_var,Hdata=Hdata)}  
  
#-----  
  
#Algorithm to compute KMO of a Multivariate dataset
```

```
#-----  
  
KMO_Val<-function(mdata){  
#=====
```

```
#Comments:  
  
#This function computes  
#the KMO of a multivariate  
#data  
#-----
```

```
R<-cor(as.matrix(mdata))  
  
Qmat<-function(Rm){
```

```
RI<-solve(Rm)  
  
Dm<-sqrt(diag(diag(RI)))  
  
Dr<-solve(Dm)
```

```
Q<-(Dr%*%RI)%*%Dr
```

```
Q}
```

```
Q<-do.call(Qmat,list(R))
```

```
#Function to compute
```

```
Sr_sq<-function(CMat){
```

```
rsq<-NULL
```

```
for(i in 1:dim(CMat)[1]){
```

```
for(j in 1:dim(CMat)[2]){
```

```
if(i<j){
```

```
rsq<-cbind(rsq,CMat[i,j])
```

```
}
```

```
}}
```

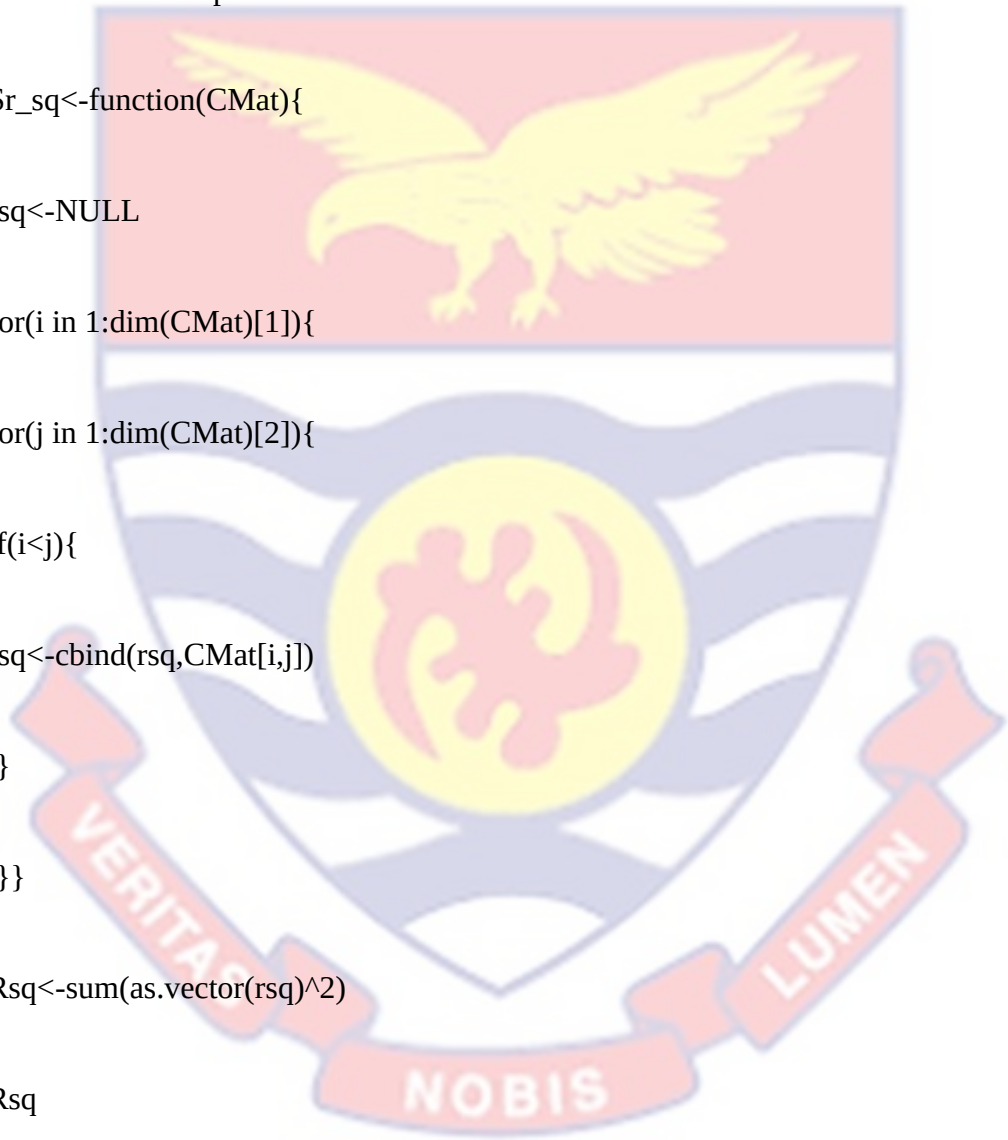
```
Rsq<-sum(as.vector(rsq)^2)
```

```
Rsq
```

```
}
```

```
Sum_Rsq<-Sr_sq(R)
```

```
Sum_Pr_sq<-Sr_sq(Q)
```



```
KMO<-1/(1+(Sum_Pr_sq/Sum_Rsq))

KMO}

#list(KMO=KMO,Sum_Rsq=Sum_Rsq,Sum_Pr_sq=Sum_Pr_sq,Q=Q,R=R)}

#=====END OF ALGORITHM=====

VData_KMO<-function(Mdata,result_list){

kmo_oneH<-function(xdata){

km<-KMO_Val(xdata)

km}

kmo_twomore<-function(hdata_list){

lapply(hdata_list,KMO_Val)

}

dkmo<-lapply(seq_len(length(result_list)),function(i){

Rmat<-Screen_result(result_list[[i]]$Srecord)

HData_set<-Data.names(Mdata,Rmat)

if(is.list(Rmat)=="TRUE"){kmos<-kmo_twomore(HData_set$Hdata)
```

```
}else{kmos<-kmo_oneH(HData_set$Hdata)

}

kmos

})

dkmo

}

#Sensitivity analysis

#=====

SenAnalysis<-function(Mdata,set_thold){

m<-length(set_thold)

Sresult<-lapply(seq_len(m),function(i){

Dim_Detector(Mdata,set_thold[i])

})

Sresult}

Correlogram<-function(mydata){

invisible(Ttr<-order_stat_cor(mydata))

mcol<- colorRampPalette(c("red", "white", "blue"))(20)
```



```
corrplot(Ttr,method="square",order="hclust",tl.col="black",tl.srt=45,bg="light  
blue",col=mcol)
```

```
invisible(return(Ttr))}
```

```
#=====
```

```
#-----
```

```
#Example
```

```
#-----
```

```
#Import data into R
```

```
Datta<-read.spss("Speformance.sav", use.value.label=TRUE,  
to.data.frame=TRUE)
```

```
mdata<-data.matrix(Datta)[-c(8,9)]
```

```
colnames(mdata)<-c("X1","X2","X3","X4","X5","X6","X7")
```

```
Result<-Dim_Detector(mdata,thold=0.7)
```

```
Rmat<-Screen_result(Result$Srecord)
```

```
HData_set<-Data.names(mdata,Rmat)
```

```
kmo_data<-KMO_Val(mdata)
```

```
kmo_Hset1<-KMO_Val(HData_set$Hdata[[1]])
```

```
kmo_Hset2<-KMO_Val(HData_set$Hdata[[2]])
```

```
#-----
```

#Example 2

#-----

#Import data into R

#-----

```
Datta<-read.spss("concreteStrength.sav", use.value.label=TRUE,  
to.data.frame=TRUE)
```

```
mdata<-data.matrix(Datta)
```

```
colnames(mdata)<-c("X1","X2","X3","X4","X5","X6","X7")
```

```
Result<-Dim_Detector(mdata,thold=0.)
```

```
#Sensitivity analysis
```

```
#-----
```

```
sthod<-seq(0.2,0.9,by=0.01)
```

```
Sfit<-SenAnalysis(mdata,sthod)
```

```
kmo_val<-VData_KMO(mdata,Sfit)
```

```
KMO_LS<-unlist(lapply(seq_len(length(kmo_val)),function(i){
```

```
lc<-length(kmo_val[i])
```

```
kkmo<-ifelse(lc>1,max(unlist(kmo_val[i])),unlist(kmo_val[i]))
```

```
kkmo}))
```

```
max(kkmo)
```

KMO Algorithm

Modified Algorithm to Compute KMO of a Multivariate dataset

```
KMOD<-function(cvMat,Sx){
```

```
#=====
```

```
#Comments:
```

```
#This function computes
```

```
#the KMO of a multivariate
```

```
#data
```

```
#-----
```

```
Qmat<-function(Rm){
```

```
RI<-solve(Rm)
```

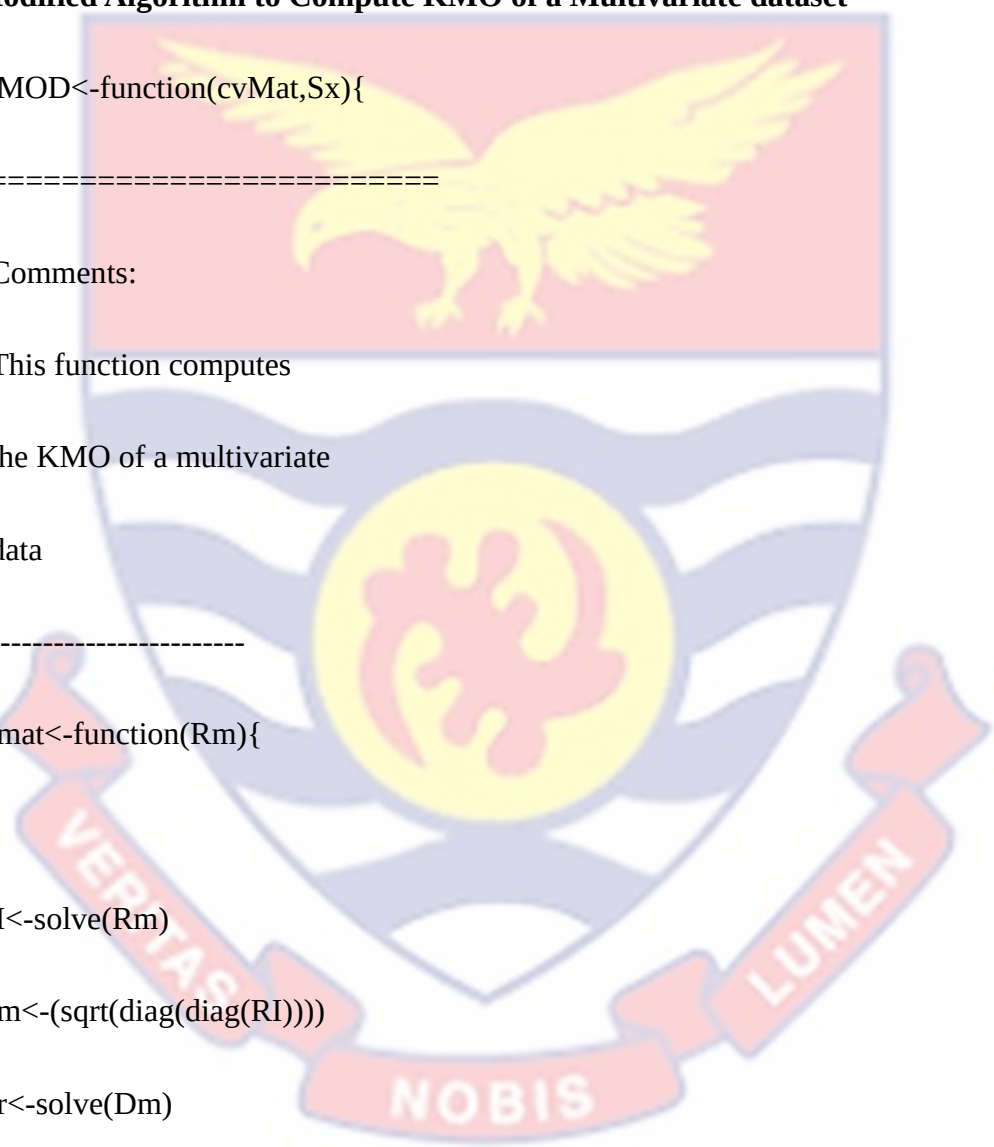
```
Dm<-sqrt(diag(diag(RI)))
```

```
Dr<-solve(Dm)
```

```
Q<-(Dr**RI)**Dr
```

```
Q}
```

```
#Function to compute
```



```
Sr_sq<-function(CMat){  
  
rsq<-NULL  
  
for(i in 1:dim(CMat)[1]){  
  
for(j in 1:dim(CMat)[2]){  
  
if(i<j){  
rsq<-cbind(rsq,CMat[i,j])  
}  
}}  
  
Rsq<-sum(as.vector(rsq)^2)  
  
Rsq  
  
}  
  
R<-cvMat  
Rs<-R[c(Sx),c(Sx)]  
Q<-do.call(Qmat,list(Rs))  
Qr<-Q;diag(Qr)<-0  
Sum_Rsq<-Sr_sq(R)  
Sum_Pr_sq<-Sr_sq(Qr)  
  
KMO<-1/(1+(Sum_Pr_sq/Sum_Rsq))
```

```
list(KMO=KMO,Sum_Rsq=Sum_Rsq,Sum_Pr_sq=Sum_Pr_sq,Q=Q,R=R,Qr  
=Qr)}
```

```
#=====END OF ALGORITHM=====
```

Dimensionality Detection Codes: Reduced Dataset Approach

```
#=====
```

```
#Dimension detection in
```

```
#Multivariate datasets
```

```
#=====
```

```
library(mvtnorm)
```

```
library(MASS)
```

```
library(pscl)
```

```
library(Matrix)
```

```
library(foreign)
```

```
#=====
```

```
#Data-based tuning schemes
```

```
#=====
```

```
#Load library ks for kernel density estimation
```

```
#=====
```

```
library(ks)#For kernel density estimation
```



```
library(latex2exp)

#Package for hi, hc

#hcl<-hlscv(mdata) # Not applicable because data contain duplicated values

hi<-hpi(mdata) #plug in estimator

hc1<-hscv(mdata) #Smoothed Cross validation

hn<-hns(mdata)

#=====

#Euclidean distance metric

#=====

Dxx<-function(x1,x2){
sqrt(sum((x1-x2)^2))}

#=====

#=====

#S2 and S1 statistic functions

#=====

S2x<-function(W2xstat){

s2x<-unlist(lapply(seq_len(length(W2xstat)),function(i){

Dxx(W2xstat[i],W2xstat[-i])

}))
```

```
s2x}  
  
#-----  
  
S1x<-function(W1xstat){  
  
dxabs<-function(x1,x2){  
sqrt(sum(abs(x1-x2)))  
}  
  
s1x<-unlist(lapply(seq_len(length(W1xstat)),function(i){  
dxabs(W1xstat[i],W1xstat[-i])  
}))  
s1x}  
  
#-----  
  
#-----  
  
#Function to compute  
#W2i statistic in h2 and h2i  
  
#-----  
  
W2stat<-function(y,a0=2){  
  
m<-length(y)  
  
Dxx<-function(x1,x2){sum((x1-x2)^a0)}  
  
sim_vec<-unlist(lapply(seq_len(length(y)),function(j){
```

```
(1/(m-1))*Dxx(y[-j],y[j])
```

```
)))
```

```
sim_vec}
```

```
#=====
```

```
#Function to compute
```

```
#Average KNN distances
```

```
#of observations
```

```
#=====
```

```
KNNy<-function(y,K=10){
```

```
Edist<-function(x1,x2){
```

```
sqrt(sum((x1-x2)^2))}
```

```
uspar<-function(X,K){
```

```
SDmat<-function(x){
```

```
Evec<-function(x1,x){
```

```
unlist(lapply(seq_len(length(x)),function(j){
```

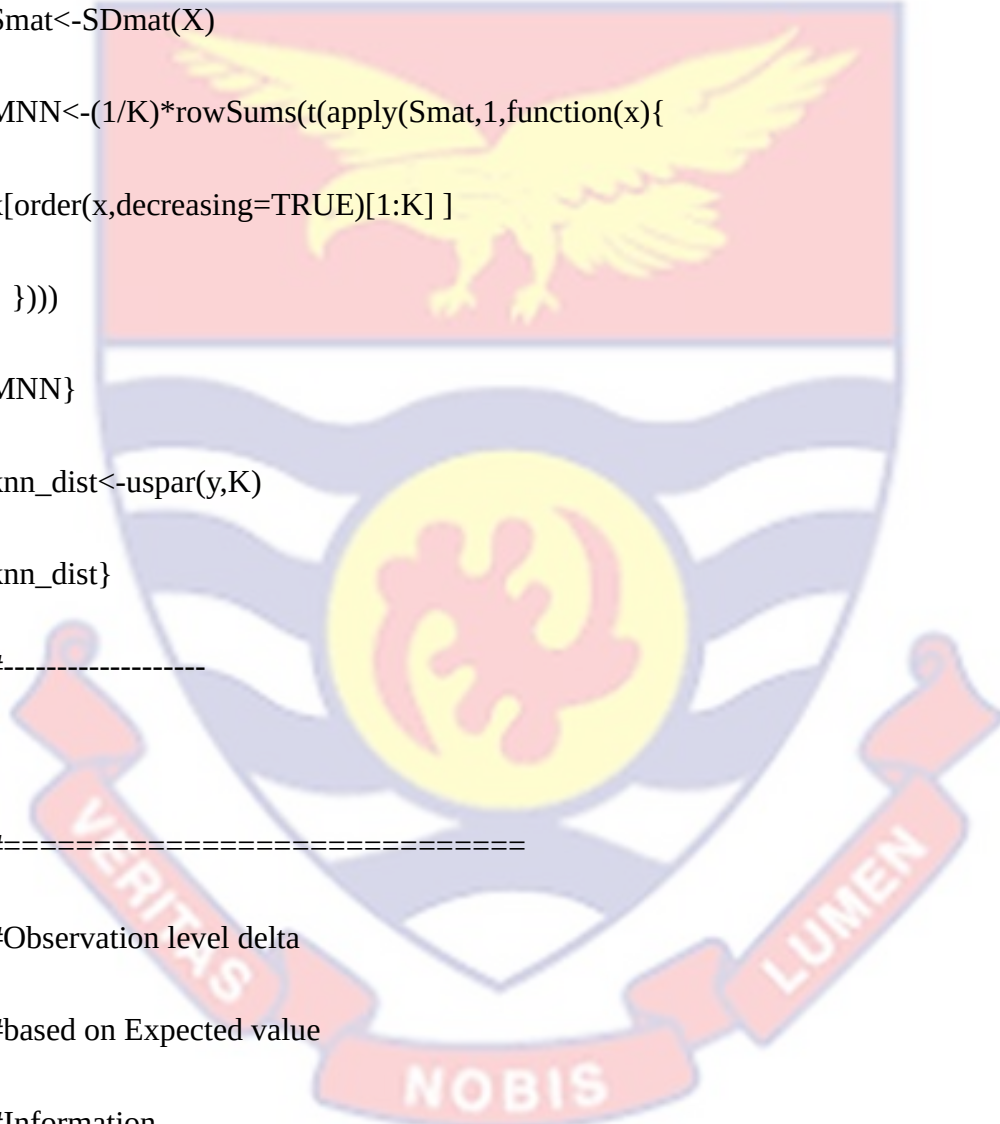
```
Edist(x1,x[j])
```

```
}))
```

```
}
```

```
Emat<-t(sapply(seq_len(length(x)),function(j){
```

```
Evec(x[j],x[-j]))))  
  
Emat[which(Emat<0)]<-0  
  
Emat  
  
}  
  
Smat<-SDmat(X)  
  
MNN<-(1/K)*rowSums(t(apply(Smat,1,function(x){  
x[order(x,decreasing=TRUE)[1:K] ]  
})))  
MNN}  
  
knn_dist<-uspar(y,K)  
  
knn_dist}  
  
#-----  
  
#=====
```

The watermark is the official crest of the University of Cape Coast. It features a shield with a yellow eagle at the top, a central yellow circle containing a red stylized figure, and a red banner at the bottom with the Latin motto "NOBIS VERITAS LUMEN". The shield is set against a background of blue and white wavy lines.

```
#Observation level delta  
#based on Expected value  
#Information  
  
#=====
```

```
delt_ExptVi<-function(xdata){  
  
hr<-(range(xdata)[2]-range(xdata)[1])/6
```

```
Fit<-kde(xdata,eval.points=xdata,hr)
```

```
fx<-Fit$estimate
```

```
Wx=fx*xdata
```

```
list(fx=fx,Wx=Wx)}
```

```
#-----
```

```
#-----
```

```
#Common weight proposals
```

```
#-----
```

```
#=====
```

```
#delta_g function
```

```
#=====
```

```
delt_g<-function(a,b,stepp=0.01){
```

```
x_int<-seq(a,b,by=stepp)
```

```
x_int
```

```
}
```

```
#-----
```

```
#=====
```

```
#Function for delta 1
```

```
#=====
```

```
delt1_fun<-function(S2stat,alph){
```



```
Sm2<-min(S2stat)
```

```
1/sqrt((alplt*Sm2))
```

```
}
```

```
#=====
```

```
#-----
```

```
#Delta 2 function
```

```
#=====
```

```
delt2_fun<-function(S2stat,alplt){
```

```
Sm2<-min(S2stat)
```

```
1/(alplt*sqrt(Sm2))
```

```
}
```

```
#-----
```

```
#-----
```

```
#Delta 3 function
```

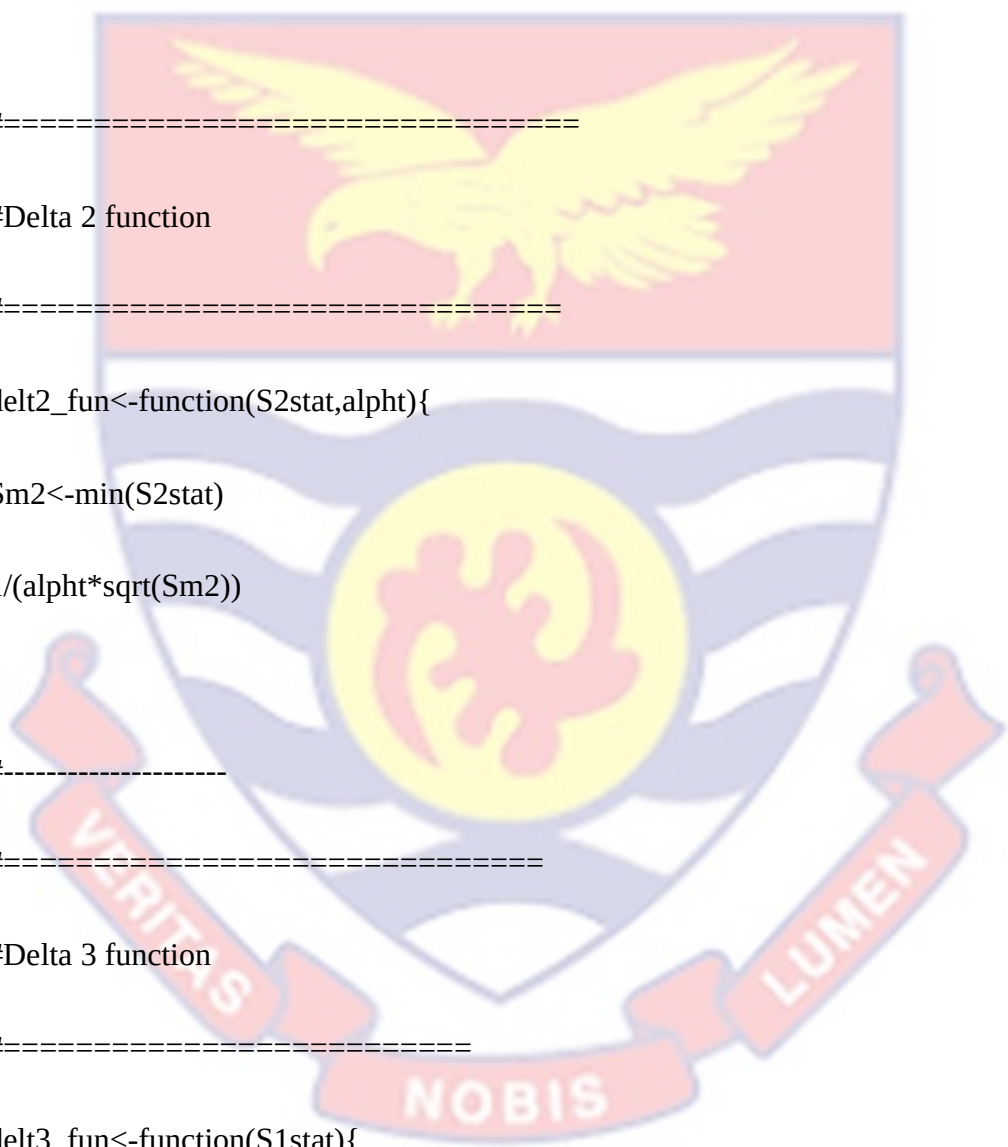
```
#=====
```

```
delt3_fun<-function(S1stat){
```

```
S1m<-min(S1stat)
```

```
1/S1m
```

```
}
```



```
#-----  
#=====
```

#Delta 4 function

```
#=====
```

```
delt4_fun<-function(S2stat){  
  hs<-XOGK(S2stat)  
  dl4<-1/sqrt(hs)  
  dl4}  
#-----  
#=====
```

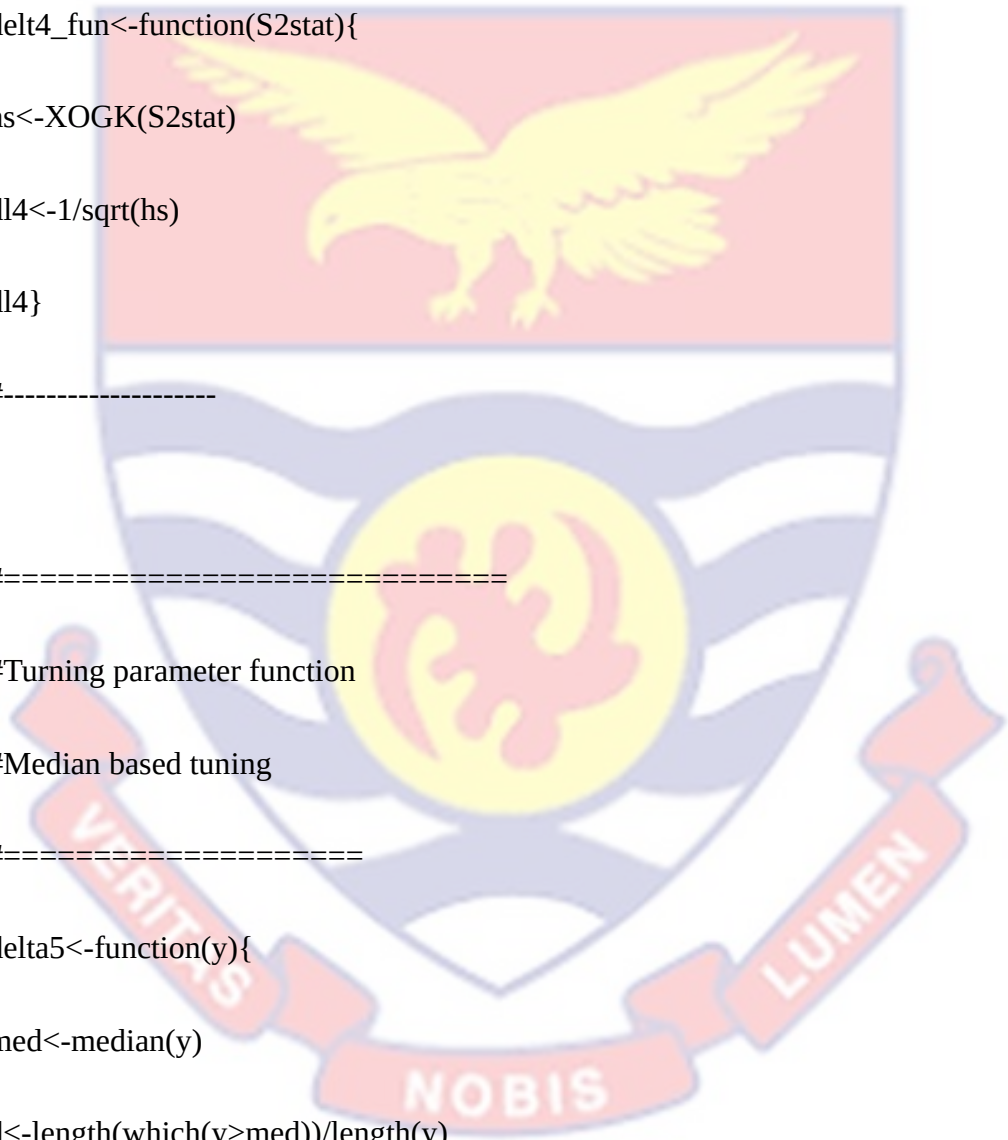
#Turning parameter function

#Median based tuning

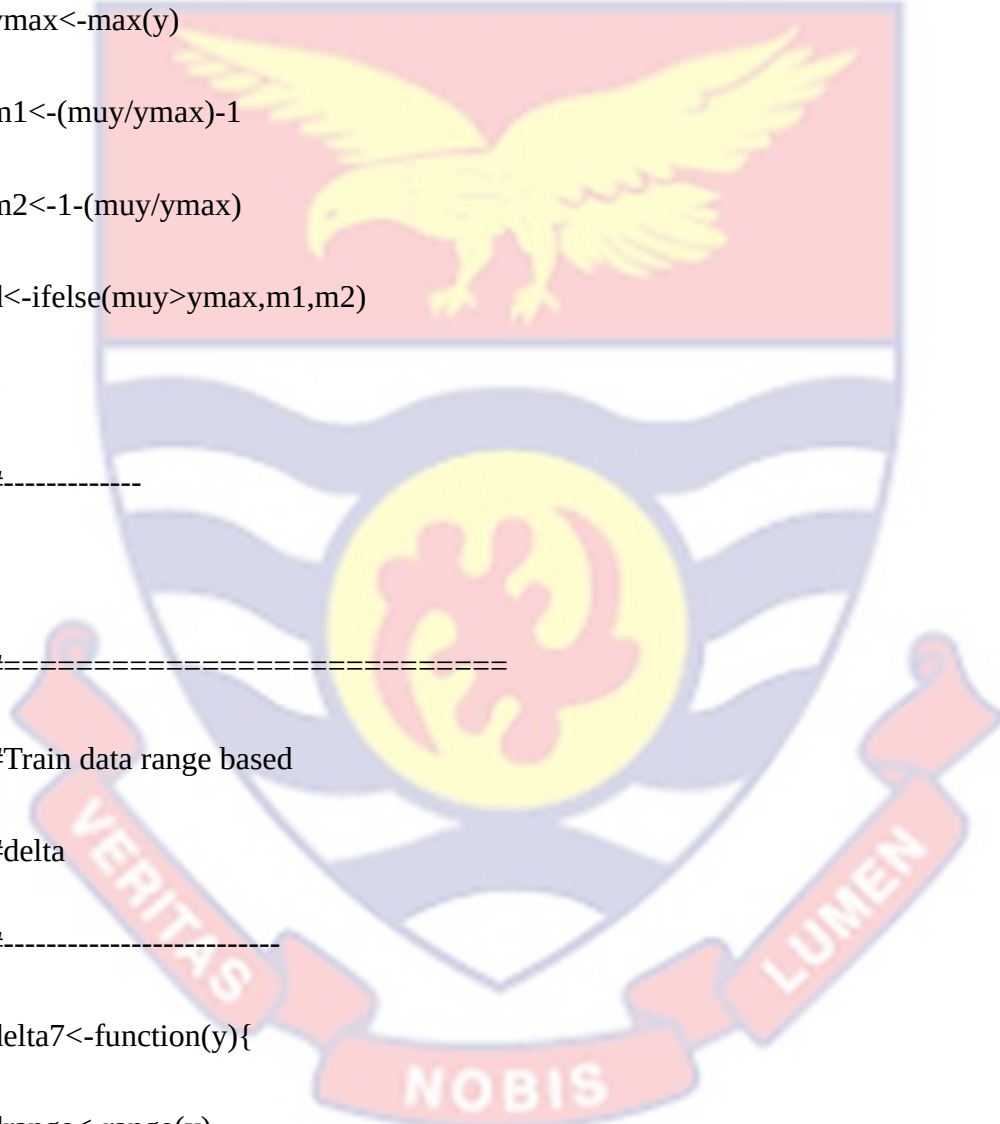
```
#-----
```

```
delta5<-function(y){  
  med<-median(y)  
  d<-length(which(y>med))/length(y)  
  d}  
#=====
```

#Full training data-based



```
#tuning function  
  
#-----  
  
delta6<-function(y){  
  
muy<-median(y)  
  
ymax<-max(y)  
  
m1<-(muy/ymax)-1  
  
m2<-1-(muy/ymax)  
  
d<-ifelse(muy>ymax,m1,m2)  
  
}  
  
#-----  
  
#=====
```



```
#Train data range based  
#delta  
  
#-----  
  
delta7<-function(y){  
  
drange<-range(y)  
  
eps<-(drange[2]-drange[1])/6  
  
delt7=1/sqrt(eps)  
  
delt7}
```

```
#-----
```

```
#=====
```

```
#Common weight based Expected
```

```
#===delta 8
```

```
#-----
```

```
delta8<-function(xdata){
```

```
dfit<-delt_ExptVi(xdata)
```

```
Wx<-dfit$Wx
```

```
umx=median(Wx)
```

```
mx=max(Wx)
```

```
mcx<-c(umx,mx)
```

```
delt_rat<-min(mcx)/max(mcx)
```

```
delt_rat}
```

```
#-----
```

```
#=====
```

```
#Common weight based Expected
```

```
#=== delta 9
```

```
#=====

delta9<-function(xdata){

dfit<-delt_ExptVi(xdata)

Wx<-dfit$Wx

mxni<-min(Wx)

mx=max(Wx)

delt<-1/sqrt((mx-mxni))

delt}

#-----

#==== End of common weight====

#-----

#Varying deltas

#-----

delt1i<-function(delt,S2stat){

m<-length(S2stat)-1

dm2<-sapply(seq_len(length(delt)),function(i){

1/sqrt((m*delt[i]*S2stat))

})

dm2}
```



```
delta2i<-function(delt,S2stat){  
  
m<-length(S2stat)-1  
  
dm1<-sapply(seq_len(length(delt)),function(i){  
  
1/(delt[i]*sqrt((m*S2stat)))  
  
})  
dm1}  
  
delta3i<-function(S1_W2istat,delt){  
  
m<-length(S1_W2istat)-1  
  
dmx<-sapply(seq_len(length(delt)),function(i){  
  
1/(delt[i]*sqrt((m*S1_W2istat)))  
  
})  
dmx}  
  
delta4i<-function(S1_W2istat,delt){  
  
m<-length(S1_W2istat)-1  
  
dmn<-sapply(seq_len(length(delt)),function(i){  
  
1/sqrt((m*delt[i]*S1_W2istat))  
  
})  
  
dmn}
```

```
delt5i<-function(S2_W1istat){
```

```
1/(S2_W1istat)
```

```
}
```

```
delt6i<-function(S2_W1istat){
```

```
1/sqrt(S2_W1istat)
```

```
}
```

```
delt7i<-function(S2stat){
```

```
1/sqrt(S2stat)
```

```
}
```

```
delt8i<-function(y){
```

```
rnum<-unlist(lapply(seq_len(length(y)),function(i){
```

```
length(which(y[-i]>y[i]))
```

```
}))
```

```
deltr=1-(rnum/length(y))
```

```
deltr
```

```
}
```

```
delt9i<-function(y){
```

```
lnum<-unlist(lapply(seq_len(length(y)),function(i){
```

```
length(which(y[-i]<y[i]))  
  
  )))
```

```
deltl<-1-(lnum/length(y))
```

```
deltl}
```

```
delt10i<-function(y){
```

```
deltr<-delt6i(y)
```

```
deltl<-delt7i(y)
```

```
deltm<-0.5*(deltl+deltr)
```

```
deltm}
```

```
#-----
```

```
#=====
```

```
#delta 11i function
```

```
#=====
```

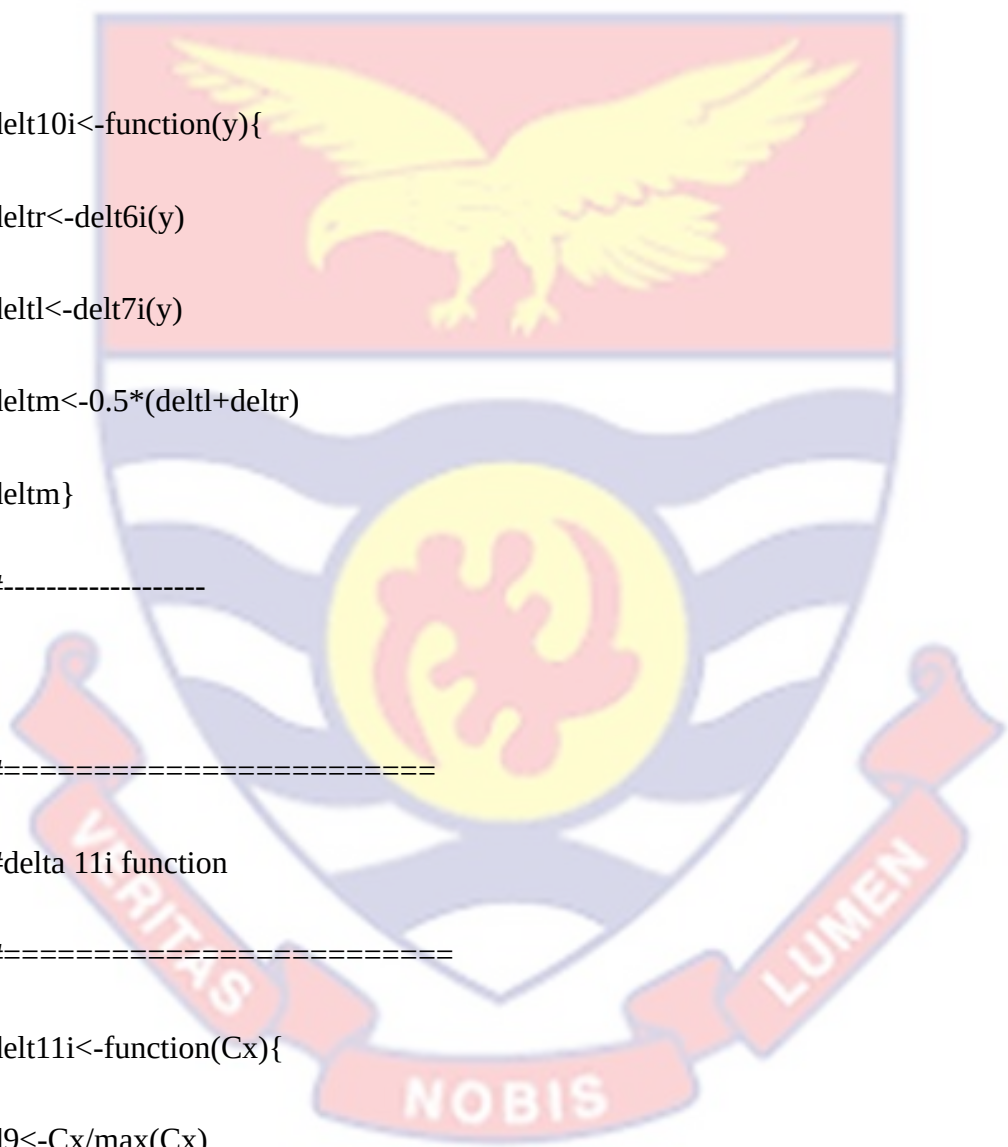
```
delt11i<-function(Cx){
```

```
d9<-Cx/max(Cx)
```

```
d9}
```

```
#-----
```

```
#=====
```



```
# delta 12i function

#=====

delt12i<-function(Cx,W2val){

uct<-median(Cx)

drat<-W2val/uct

d11=1/sqrt(drat)

d11}

#-----

#=====

#delta 13i function

#-----

delt13i<-function(Cx,alphs){

m<-length(Cx)

delts<-sapply(seq_len(length(alphs)),function(i){

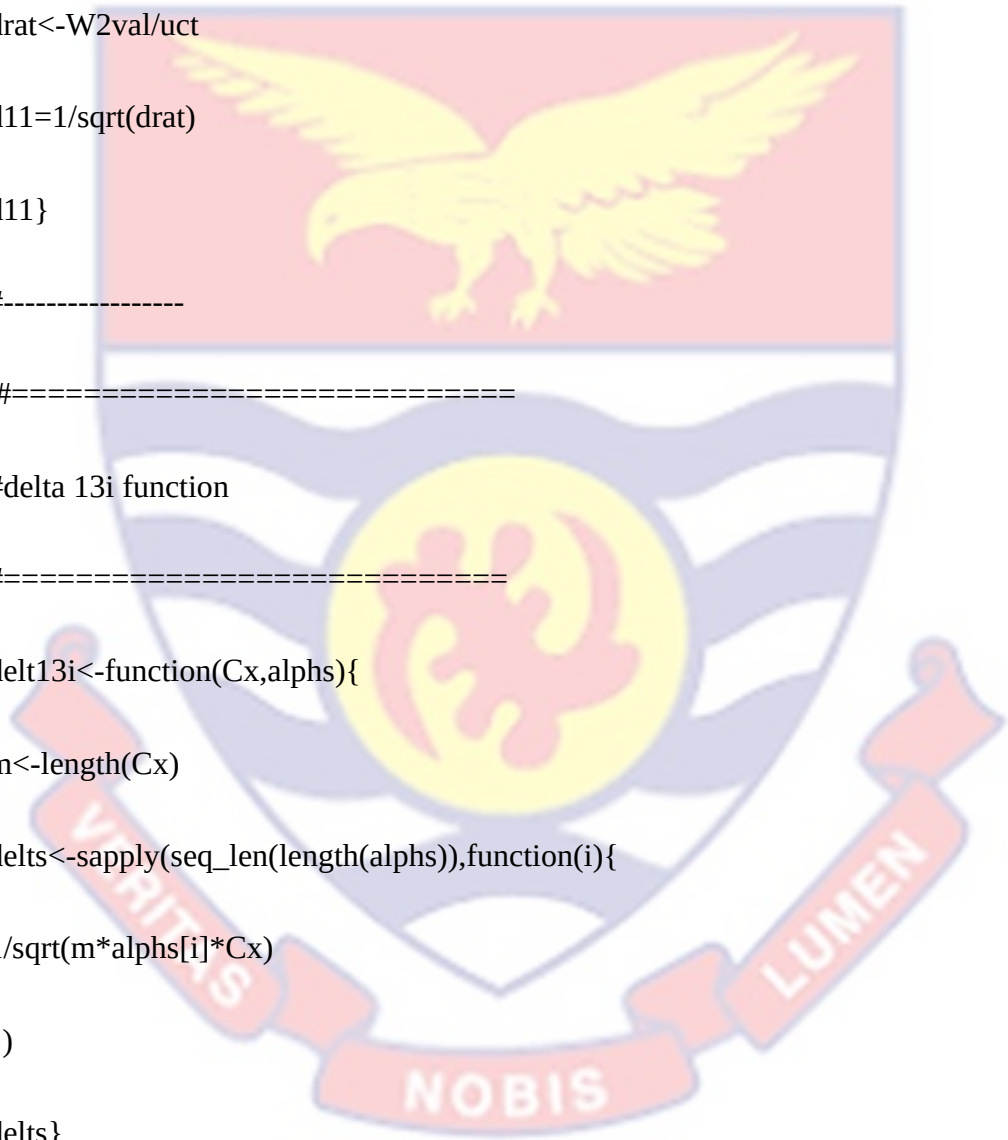
1/sqrt(m*alphs[i]*Cx)

})

delts}

#-----

#=====
```



```
#delta 14i function

#=====

delt14i<-function(S2Cx_stat,alphs){

m<-length(S2Cx_stat)

delts<-sapply(seq_len(length(alphs)),function(i){

1/sqrt(m*alphs[i]*S2Cx_stat)

})

delts}

#-----

#-----

#====hc1 function

#-----

hc1_func<-function(smdist,deltn){

hc1<-sqrt(mean(deltn*smdist))

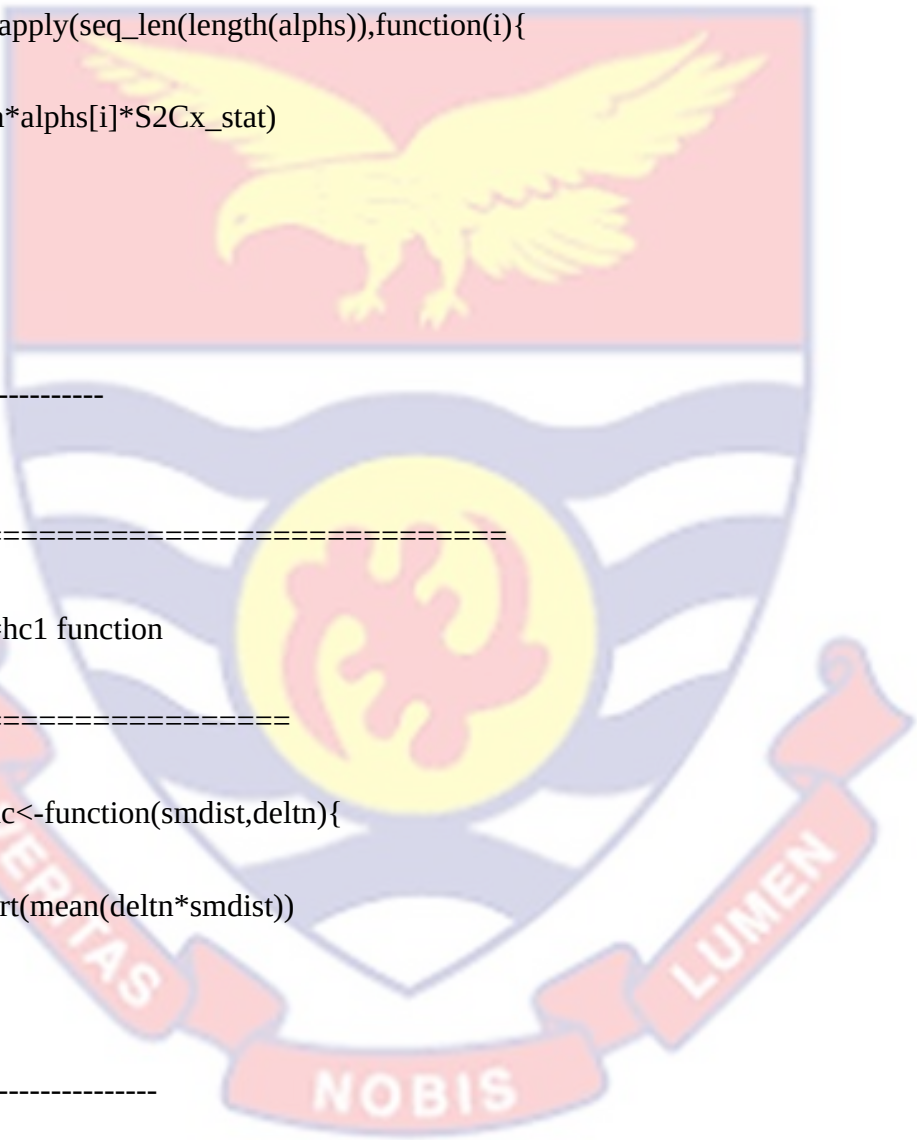
hc1}

#-----

#-----

#=== hc2 function

#=====
```




```
#-----  
#=====
```

#====h1v function

```
#=====
```

```
h1v<-function(Sim_KNN,deltv){  
  result<-Sim_KNN  
  hw<-deltv*(result)  
  h1v<-mean(hw)  
h1v}  
#-----  
#=====
```

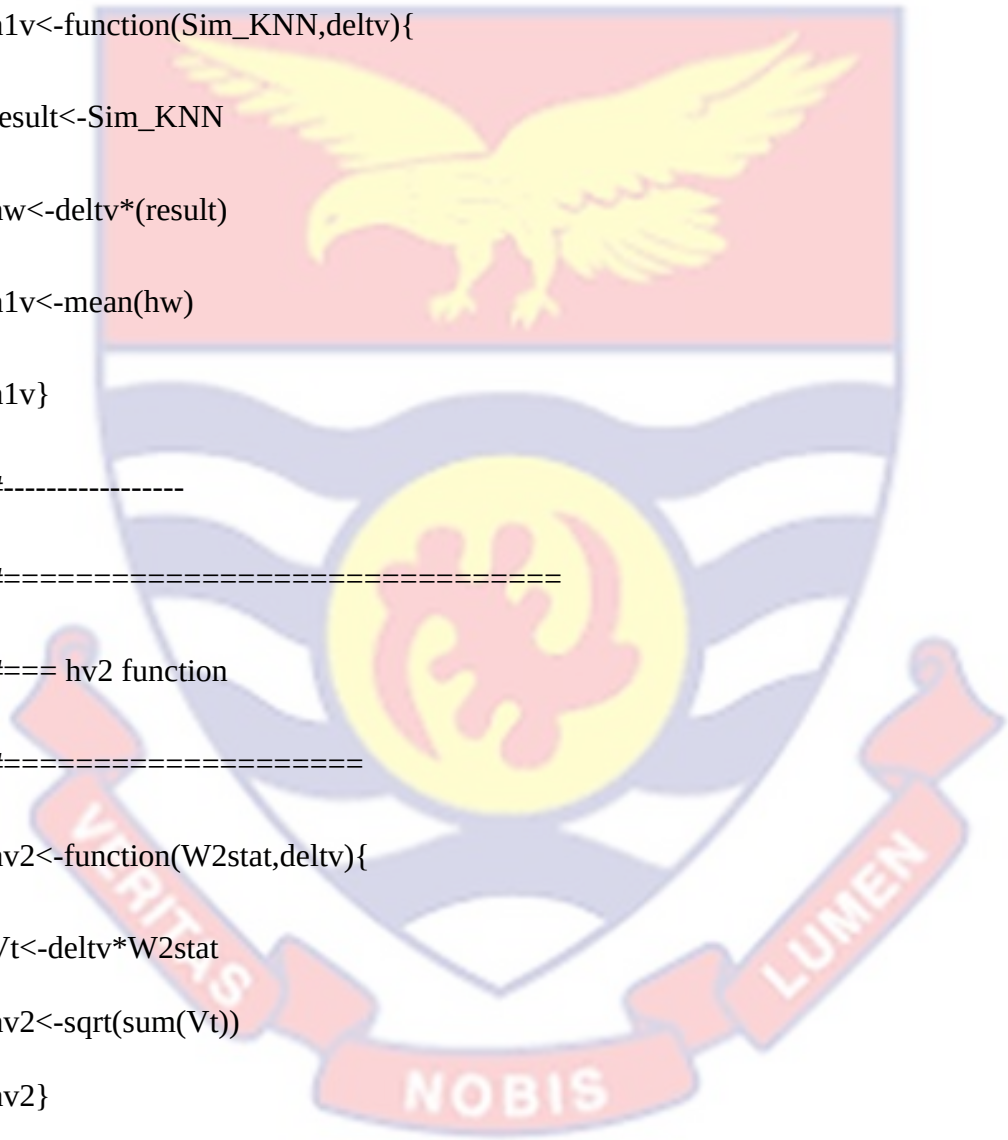
#=== hv2 function

```
#=====
```

```
hv2<-function(W2stat,deltv){  
  Vt<-deltv*W2stat  
  hv2<-sqrt(sum(Vt))  
hv2}  
#-----  
#=====
```

#====hv3 function

```
#=====
```



```
hv3<-function(S1_W2istat,deltv){  
  result<-S1_W2istat  
  lambda<-sqrt(deltv)  
  hx<-lambda*result  
  h3v<-sqrt(mean(hx))  
  h3v}
```

```
#-----
```

```
#=====
```

```
#===== hv5 function
```

```
#=====
```

```
hv4<-function(S2stat,deltv3){
```

```
Vt<-deltv3*S2stat
```

```
h4<-sqrt(XOGK(Vt))
```

```
h4}
```

```
#-----
```

```
#Function to compute correlation matrix
```

```
Corrv<-function(mdata){
```

```
Cormat<-matrix(0,dim(mdata)[2],dim(mdata)[2])
```

```
for(i in 1: dim(Cormat)[1]){
```

```
for(j in 1:dim(Cormat)[2]){
```

```
Cormat[i,j]<-cor(mdata[,i],mdata[,j])
```

```
  }
```

```
}
```

```
Cormat
```

```
}
```

```
PDF_Dim_Detector<-function(Mdata,thold=0.5){
```

```
#=====
```

```
#Function to select first
```

```
#Spanning set based on
```

```
#Correlation Coefficient
```

```
#=====
```

```
Span_set<-function(xmat){
```

```
max_val<-max(xmat)
```

```
stvl<-function(x,max_v){
```

```
ifelse(x==max_v,1,0)
```

```
}
```

```
Rid<-lapply(seq_len(dim(xmat)[1]),function(i){
```

```
stvl<-stvl(xmat[i,],max_val)
```

```
ld<-which(stv==1)
```

```
c(i,ld)
```

```
})
```

```
#Sr<-NULL
```

```
ls_len<-unlist(lapply(seq_len(length(Rid)),function(j){
```

```
Vecx<-as.vector(Rid[[j]])
sl<-length(Vecx)
sl}))
#if(any(ls_len==2)){
S<-Rid[[which(ls_len==2)]]
S<-c(S[2],S[1])
#}else{
#break
#}
S}
#-----
Vbind<-function(X,y){
#=====
#Function to combine vectors of
#different lengths
#=====
mbind<-function(x,y){
slab<-NULL
a<-dim(x)[1];b<-length(y)
if(a==b){slab<-cbind(x,y)}
if(a>b){slab<-
cbind(x, y=c(y, rep(NA,(a-b))))}
```



```
}  
slab  
}  
  
#=====  
#Function to combine vectors of  
#different lengths  
#-----  
sbind<-function(x,y){  
slab<-NULL  
a<-length(x);b<-length(y)  
if(a==b){slab<-cbind(x,y)}  
if(a>b){slab<-  
cbind(x, y=c(y, rep(NA,(a-b))))  
}  
if(a<b){slab<-cbind(y, x=c(x, rep(NA,(b-a))))  
}  
slab  
}  
  
sfit<-NULL  
  
if(length(X)==0){sfit<-sbind(X,y)}else{  
  
if(length(X)>1 & is.matrix(X)=="TRUE"){  
  
sfit<-mbind(X,y)  
  
}
```

```

}

sfit }

Foutput<-function(rmat){

Dresult<-sapply(seq_len(dim(rmat)[2]),function(i){

  rmat[,i]))

Dresult}

#=====
#Updating spanning function
#Based on pairwise Correlation
#=====

CompwS<-function(S_index,Cor_Mat,thold){

Sxupdator<-function(Sx_set,x_dex,Cor_mat,thold){
pwcr<-unlist(lapply(seq_len(length(Sx_set)),function(i){
Crr<-c(Cor_mat[Sx_set[i],x_dex],Cor_mat[x_dex,Sx_set[i]])
rr<-which(Crr==0)
Drr<-Crr[-rr]
Drr
}))
if( all(pwcr>=thold)=="TRUE"){
Sx_set<-c(Sx_set,x_dex)}else{Sx_set<-Sx_set}
Sx_set}

N<-dim(Cor_Mat)[2]

```

```
m<-seq_len(N)[-S_index]
New_set<-S_index
for(i in 1:length(m)){
NS<-Sxupdator(New_set,m[i],Cor_Mat,thold)
New_set<-NS
}
Hset<-New_set;Nhset<-seq_len(N)[-Hset]
list(Hset=Hset,Nhset=Nhset)}

#Start sequential updating
tol<-2
clab<-colnames(data.frame(Mdata))
Srecord<-NULL
count<-0
Cmat<-Corrv(Mdata)
if(any(Cmat<0)== "TRUE"){
Cmat<-abs(Cmat)}else{
Cmat<-Cmat}

crmat<-as.matrix(tril(Cmat));diag(crmat)<-0
d<-dim(crmat)[2]
Inhset<-d
```



```
Nlab<-seq_len(d)

S<-Span_set(crmat)

while(lnhset>tol){

count<-count+1

fit<-CompwS(S,crmat,thold)

hset<-fit$Hset

nhset<-fit$Nhset

mlab<-clab[hset]

rlab<-clab[-hset]

Srecord<-Vbind(Srecord,mlab)

Rcrmat<-crmat[-hset,-hset]

NHset<-rlab

if(is.matrix(Rcrmat)==FALSE){

break

}

S_new<-Span_set(Rcrmat)

S<-S_new;crmat<-Rcrmat

clab<-rlab

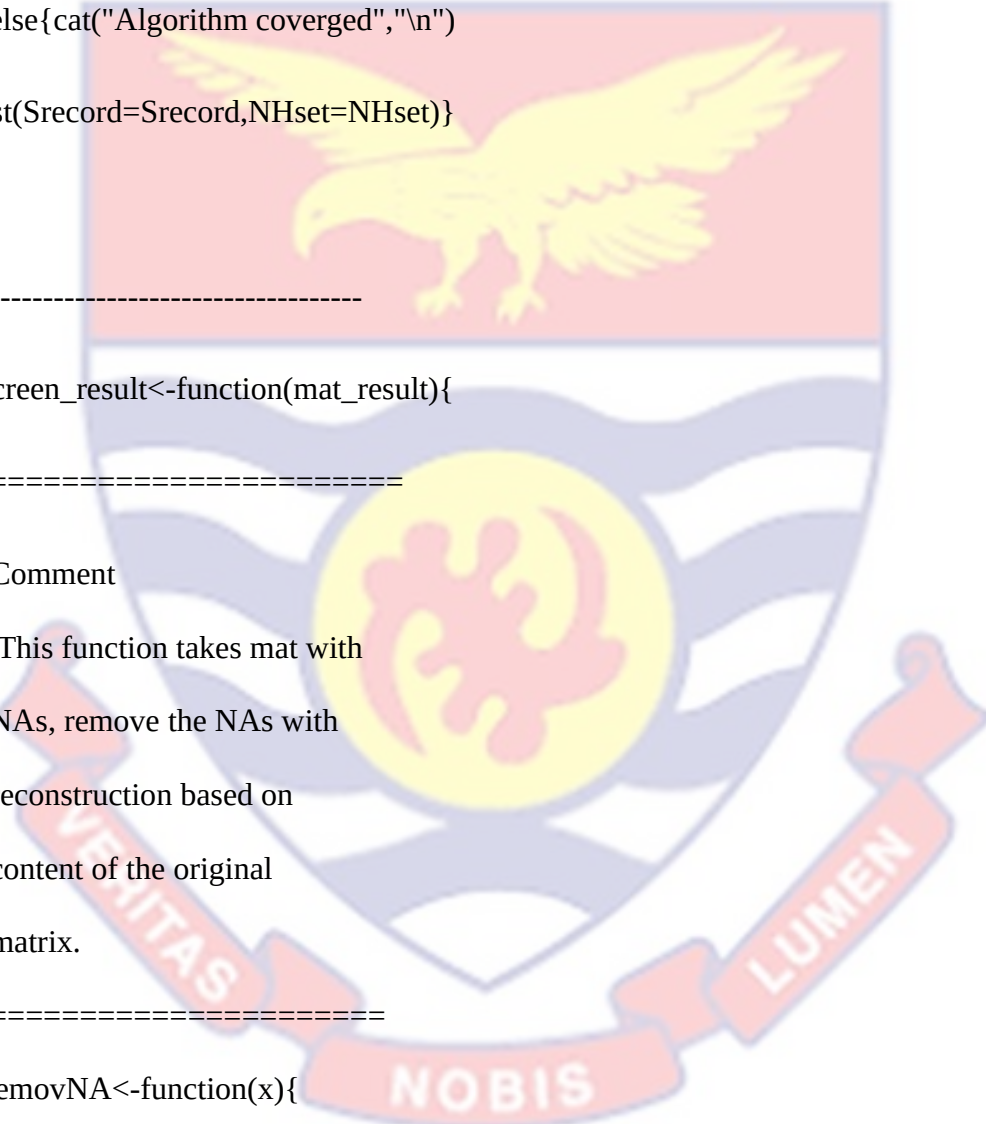
lnhset<-length(nhset)

cat(count,lnhset,S,mlab,"n")
```



```
}  
  
if(count>1 & lnhset<tol){  
  
cat("Algorithm failed to converge","\n")  
  
return(NULL)  
  
}else{cat("Algorithm covered","\n")  
list(Srecord=Srecord,NHset=NHset)}  
}
```

```
#-----  
  
Screen_result<-function(mat_result){  
  
#-----  
  
#Comment  
# This function takes mat with  
#NAs, remove the NAs with  
#reconstruction based on  
#content of the original  
#matrix.  
  
#=====
```



```
RemovNA<-function(x){  
x[!is.na(x)]  
}
```

```
Replace<-function(x){
```



```
ifelse( is.na(x)=="TRUE",0,x)  
}
```

```
Zero_rm<-function(x){  
x[-which(x==0)]  
}
```

```
MatRed<-function(mat){  
fx<-function(x){  
ifelse(all(x==0)=="TRUE",1,0)}  
nd<-as.vector(apply(mat,2,fx))  
Nmat<-mat[,-which(nd==1)]  
Nmat}
```

```
Rlist<-function(Mmat){  
Ttf<-function(ylist){  
vv<-NULL  
if(any(ylist==0)){vv<-Zero_rm(ylist)}else{vv<-ylist[]}  
vv  
}
```

```
if(is.matrix(Mmat)=="TRUE"){  
rlist<-apply(Mmat,2,list)  
fresult<-lapply(seq_len(length(rlist)),function(j){  
vv<-NULL
```

```
tts<-unlist(rlist[[j]])  
if(any(tts==0)){vv<-Zero_rm(tts)}else{vv<-tts[]}  
})  
}  
  
if(is.vector(Mmat)=="TRUE"){  
fresult<-Ttf(Mmat)  
}  
fresult}  
#-----  
  
#Call  
#-----  
smat<-Replace(mat_result)  
mmat<-MatRed(smat)  
  
Fresult<-Rlist(mmat)  
Fresult}  
#-----  
  
Data.names<-function(Data_frame,rlist){  
  
Data_frame<-if(is.data.frame(Data_frame)=="FALSE"){  
Data_frame<-data.frame(Data_frame)  
}  
}else{Data_frame<-Data_frame}
```

```
n.set<-names(Data_frame)

xsame<-function(x,y){
  which(x==y)}

smvec<-function(xdata,ystand){
  ufit<-unlist(lapply(seq_len(length(xdata)),function(i){
  xsame(ystand,xdata[i])
  )))
  ufit
}

data_var<-lapply(seq_len(length(rlist)),function(j){
  smvec(rlist[[j]],n.set)
})

Hdata<-sapply(seq_len(length(data_var)),function(t){
  Data_frame[,data_var[[t]]
})

list(data_var=data_var,Hdata=Hdata)}

#-----
#Algorithm to compute KMO of a Multivariate dataset
#-----

KMO_Val<-function(mdata){
```

```
#=====
```

```
#Comments:
```

```
#This function computes
```

```
#the KMO of a multivariate
```

```
#data
```

```
#-----
```

```
R<-cor(as.matrix(mdata))
```

```
Qmat<-function(Rm){
```

```
RI<-solve(Rm)
```

```
Dm<-sqrt(diag(diag(RI)))
```

```
Dr<-solve(Dm)
```

```
Q<-(Dr%*%RI)%*%Dr
```

```
Q}
```

```
Q<-do.call(Qmat,list(R))
```

```
#Function to compute
```

```
Sr_sq<-function(CMat){
```

```
rsq<-NULL
```

```
for(i in 1:dim(CMat)[1]){
```

```
for(j in 1:dim(CMat)[2]){
```

```
if(i<j){
```

```
rsq<-cbind(rsq,CMat[i,j])
```

```
}  
}}  
Rsqr<-sum(as.vector(rsqr)^2)  
Rsqr  
}  
  
Sum_Rsqr<-Sr_sq(R)  
Sum_Pr_sq<-Sr_sq(Q)  
  
KMO<-1/(1+(Sum_Pr_sq/Sum_Rsqr))  
KMO}  
  
#list(KMO=KMO,Sum_Rsqr=Sum_Rsqr,Sum_Pr_sq=Sum_Pr_sq,Q=Q,R=R)}  
  
#=====END OF ALGORITHM=====  
  
VData_KMO<-function(Mdata,result_list){  
  
Muti_Homoser<-function(Datta,rmat){  
if(is.list(rmat)!="TRUE"){  
Udata<-lapply(seq_len(length(rmat)),function(j){  
Data.names(Datta,rmat[[j]])$Hdata  
})}else{  
Udata<-Data.names(Datta,rmat)$Hdata  
}  
Udata}
```



```
kmo_oneH<-function(xdata){  
  km<-KMO_Val(xdata)  
  km}  
kmo_twomore<-function(hdata_list){  
  m1<-length(hdata_list)  
  ut<-unlist(lapply(seq_len(m1),function(i){  
    KMO_Val(hdata_list[[i]])  
  })))  
  ut }  
  
dkmo<-lapply(seq_len(length(result_list)),function(i){  
  Rmat<-Screen_result(result_list[[i]]$Srecord)  
  HData_set<-Muti_Homoset(Mdata,Rmat)  
  
  if(is.list(Rmat)=="TRUE"){kmos<-kmo_twomore(HData_set)  
    }else{kmos<-kmo_oneH(HData_set)  
    }  
  kmos  
})  
  
dkmo  
}  
  
#=====
```

```
#Sensitivity analysis
#=====

SenAnalysis<-function(Mdata,set_thold){
m<-length(set_thold)
Sresult<-lapply(seq_len(m),function(i){
PDF_Dim_Detector(Mdata,set_thold[i]
)})
Sresult}
#=====

#Data driven threshold setting
#=====

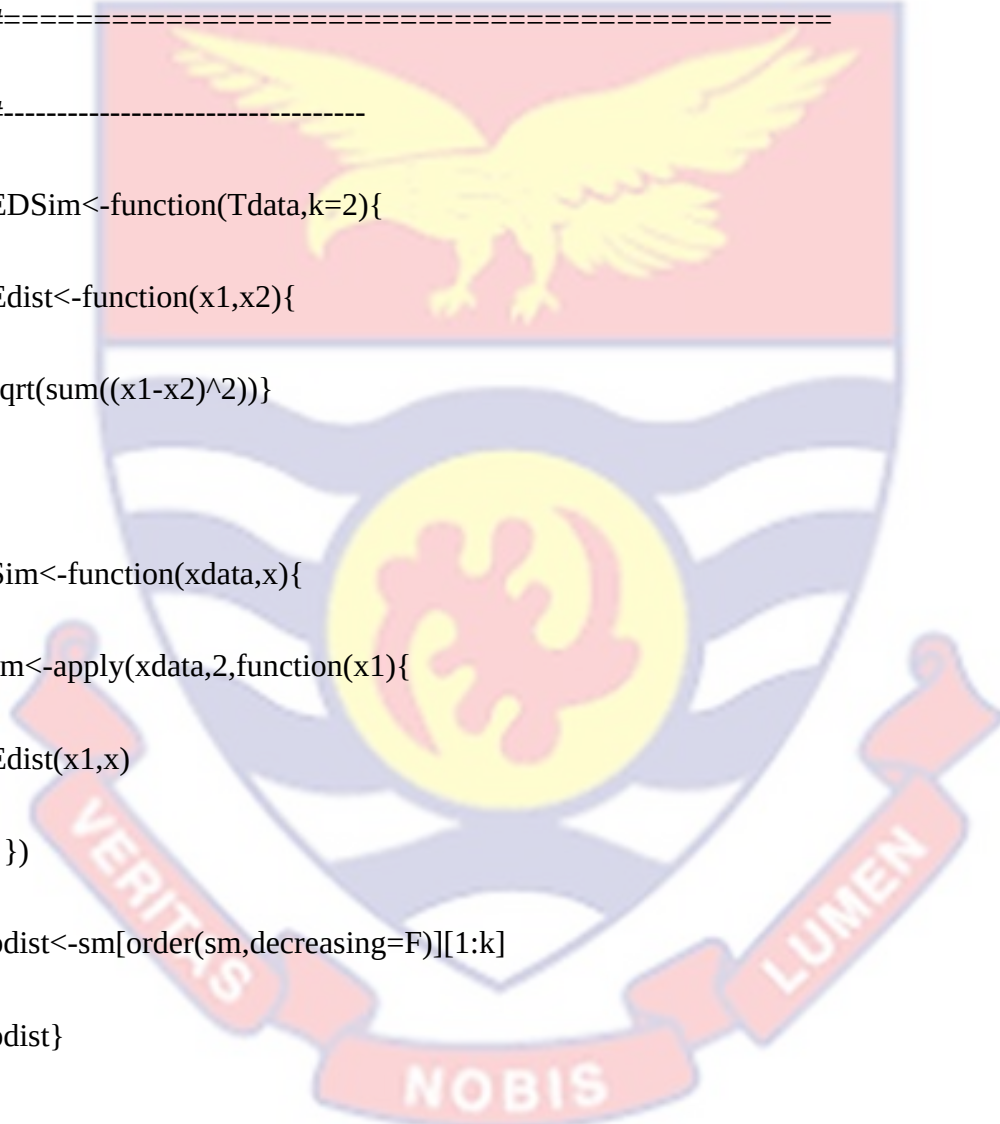
strehod<-function(Mdata){
CMat<-Corrv(Mdata)
Lmat<-as.matrix(tril(CMat));diag(Lmat)<-0
Nmat<-abs(as.vector(Lmat))
NCrmat<-Nmat[-which(Nmat==0)]
crange<-round(range(NCrmat),2)
sth<-seq(crange[1],crange[2],0.01)

a<-((crange[2]-crange[1])/12)

b<-round(a,2)

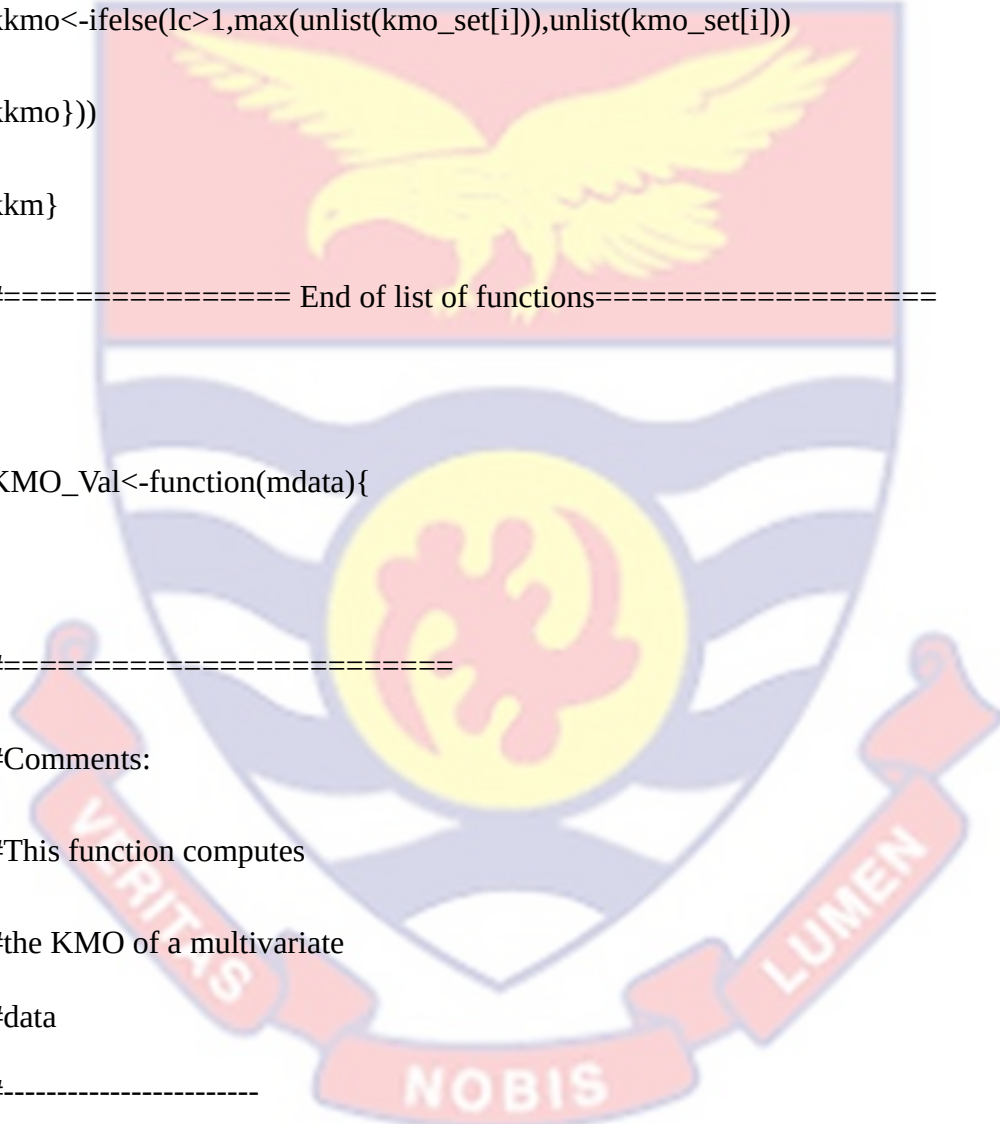
st2<-seq(crange[1],crange[2],b)
```

```
st3<-sth[which(sth>=median(sth))]  
  
list(crange=crange,sth=sth,a=a,st2=st2,st3=st3)}  
  
#=====
```



```
#Alternative approaches to threshold setting  
  
#-----  
#-----  
EDSim<-function(Tdata,k=2){  
  Edist<-function(x1,x2){  
    sqrt(sum((x1-x2)^2))  
  }  
  Sim<-function(xdata,x){  
    sm<-apply(xdata,2,function(x1){  
      Edist(x1,x)  
    })  
    bdist<-sm[order(sm,decreasing=F)][1:k]  
    bdist}  
  
  smat<-t(sapply(seq_len(dim(Tdata)[2]),function(i){  
    Sim(Tdata[,-i],Tdata[,i])  
  })))
```

```
smat}  
  
KMO_Set<-function(kmo_set){  
  
kkm<-unlist(lapply(seq_len(length(kmo_set)),function(i){  
  
lc<-length(kmo_set[i])  
  
kkmo<-ifelse(lc>1,max(unlist(kmo_set[i])),unlist(kmo_set[i]))  
kkmo}))  
kkm}  
  
#===== End of list of functions=====
```

The logo of the University of Cape Coast is a watermark in the background. It features a shield with a yellow eagle at the top, a central yellow circle with a red figure, and a red banner at the bottom with the Latin motto "VERITAS LUMEN NOBIS".

```
KMO_Val<-function(mdata){  
  
#=====
```

#Comments:
#This function computes
#the KMO of a multivariate
#data
#-----

```
R<-cor(as.matrix(mdata))  
  
Qmat<-function(Rm){  
  
RI<-solve(Rm)
```

```
Dm<-(sqrt(diag(diag(RI))))
```

```
Dr<-solve(Dm)
```

```
Q<-(Dr%*%RI)%*%Dr
```

```
Q}
```

```
Q<-do.call(Qmat,list(R))
```

```
#Function to compute
```

```
Sr_sq<-function(CMat){
```

```
rsq<-NULL
```

```
for(i in 1:dim(CMat)[1]){
```

```
for(j in 1:dim(CMat)[2]){
```

```
if(i<j){
```

```
rsq<-cbind(rsq,CMat[i,j])
```

```
}
```

```
}}
```

```
Rsq<-sum(as.vector(rsq)^2)
```

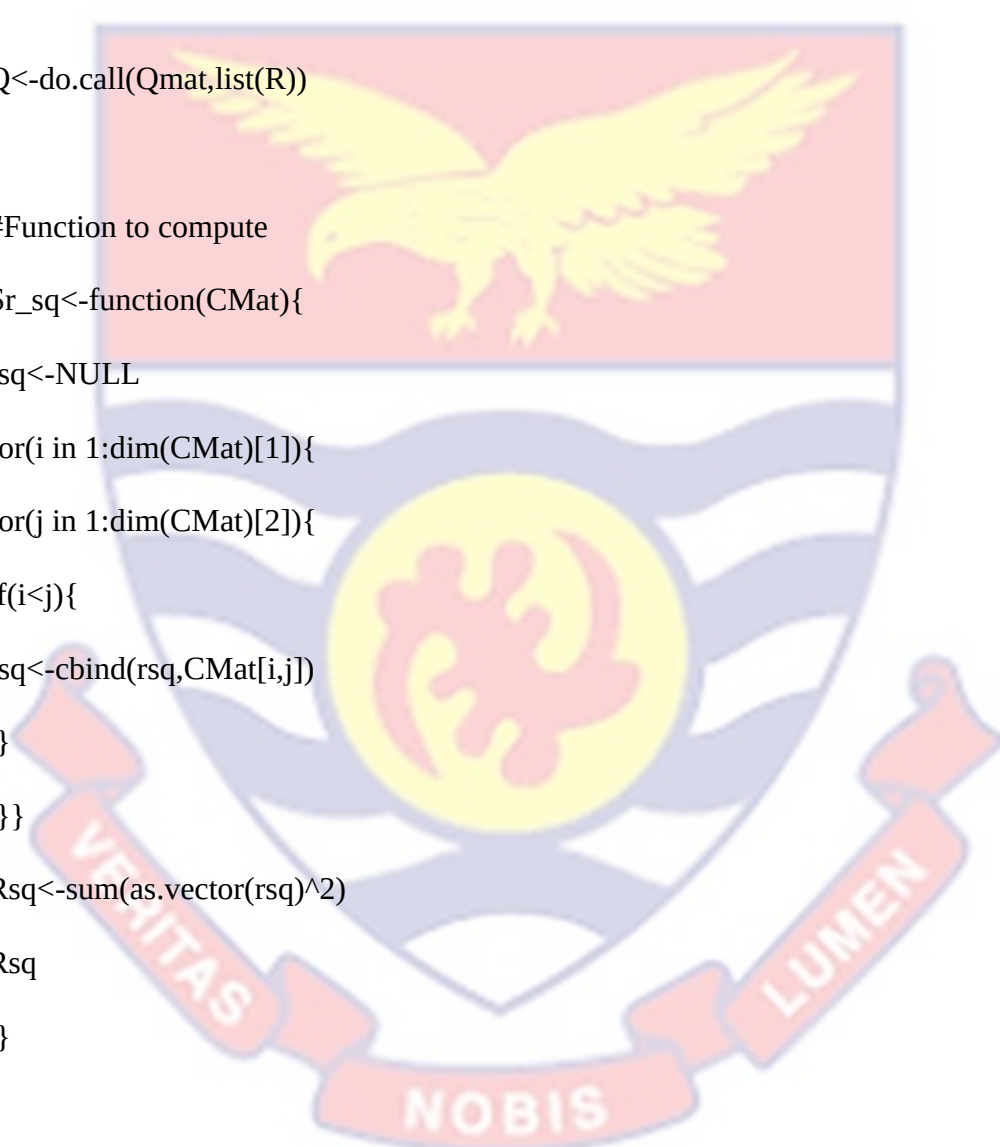
```
Rsq
```

```
}
```

```
Sum_Rsq<-Sr_sq(R)
```

```
Sum_Pr_sq<-Sr_sq(Q)
```

```
KMO<-1/(1+(Sum_Pr_sq/Sum_Rsq))
```




```
KMO}  
  
#=====  
  
#====Probability density function  
  
#====based features  
  
#=====  
  
DensFeat<-function(Datta,sm_par,dprop=0.7){  
  dens_feat<-function(ydatta,smooth_par,dat_prop=0.7){  
    Fit<-kde(ydatta,h=smooth_par,eval.points=ydatta)  
    fy<-Fit$estimate  
    Ty=fy*ydatta  
    Uy=mean(Ty)  
    Zstat<-((Ty-mean(Ty))/sqrt(var(Ty)))  
    zr<-order(Ty,decreasing=T)  
    K=round(length(ydatta)*dat_prop,0)  
    y_select<-ydatta[1:K]  
    sprob=Ty/max(Ty)  
    list(fy=fy,Ty=Ty,Uy=Uy,Zstat=Zstat,y_select=y_select,sprob=sprob)}  
  }  
  
Dfit<-lapply(seq_len(dim(Datta)[2]),function(j){
```

```
dens_feat(Datta[,j],sm_par[j],dprop)

})

fymat<-sapply(seq_len(length(Dfit)),function(i){

Dfit[[i]]$fy

})

Tymat<-sapply(seq_len(length(Dfit)),function(i){

Dfit[[i]]$Ty

})

Zystat<-sapply(seq_len(length(Dfit)),function(i){

Dfit[[i]]$Zstat

})

fymat<-sapply(seq_len(length(Dfit)),function(i){

Dfit[[i]]$fy

})

yselect<-sapply(seq_len(length(Dfit)),function(i){

Dfit[[i]]$y_select

})

list(Dfit=Dfit,fymat=fymat,Tymat=Tymat,Zystat=Zystat,yselect=yselect)}

#Import data into R
```

```
Datta<-read.spss("Speformance.sav", use.value.label=TRUE,
to.data.frame=TRUE)

mdata<-data.matrix(Datta)[-c(8,9)]

colnames(mdata)<-c("X1","X2","X3","X4","X5","X6","X7")

sm_par<-unlist(lapply(seq_len(dim(mdata)[2]),function(i){
hscv(mdata[,i])
}))

perform_data_fit<-DensFeat(mdata,sm_par,dprop=0.7)

yselect<-perform_data_fit$yselect
fymat<-perform_data_fit$fymat
Tymat<-perform_data_fit$Tymat
Zymat<-perform_data_fit$Zymat
CVMatP<-Corrv(mdata)
CVMatP_rdata<-Corrv(yselect)
CVMatP_Tymat<-Corrv(Tymat)
#CVMatO<-OStat_Cor(mdata)

sthod1_rdata<-strehod(yselect)$sth
sthod2_rdata<-strehod(yselect)$st2
sthod3_rdata<-strehod(yselect)$st3

Ffit1_rdata<-SenAnalysis(yselect,sthod1_rdata)
```

```
Ffit2_rdata<-SenAnalysis(yselect,sthod2_rdata)

Ffit3_rdata<-SenAnalysis(yselect,sthod3_rdata)

kmo_val3_rdata<-VData_KMO(yselect,Ffit3_rdata)

kmo_val2_rdata<-VData_KMO(yselect,Ffit2_rdata)

kmo_val1_rdata<-VData_KMO(yselect,Ffit1_rdata)

KMO1_rdata<-KMO_Set(kmo_val1_rdata)

KMO2_rdata<-KMO_Set(kmo_val2_rdata)

KMO3_rdata<-KMO_Set(kmo_val3_rdata)

par(mfrow=c(3,2))

plot(mdata[,1],fymat[,1],xlab=expression(y[1]),ylab=expression(f(y[1])))

plot(mdata[,2],fymat[,2],xlab=expression(y[2]),ylab=expression(f(y[2])))

plot(mdata[,3],fymat[,3],xlab=expression(y[3]),ylab=expression(f(y[3])))

plot(mdata[,4],fymat[,4],xlab=expression(y[4]),ylab=expression(f(y[4])))

plot(mdata[,5],fymat[,2],xlab=expression(y[5]),ylab=expression(f(y[5])))

par(mfrow=c(3,2))

plot(fymat[,1],Tymat[,1],xlab=expression(f(y[1])),ylab=expression(T(y[1])))

plot(fymat[,2],Tymat[,2],xlab=expression(f(y[2])),ylab=expression(T(y[2])))

plot(fymat[,3],Tymat[,3],xlab=expression(f(y[3])),ylab=expression(T(y[3])))

plot(fymat[,4],Tymat[,4],xlab=expression(f(y[4])),ylab=expression(T(y[4])))

plot(fymat[,5],Tymat[,2],xlab=expression(f(y[5])),ylab=expression(T(y[5])))
```

```
#=====

#Using Original data

#=====

sthod1<-strehod(mdata)$sth

sthod2<-strehod(mdata)$st2

sthod3<-strehod(mdata)$st3

Ffit1<-SenAnalysis(mdata,sthod1)

Ffit2<-SenAnalysis(mdata,sthod2)

Ffit3<-SenAnalysis(mdata,sthod3)

kmo_val3<-VData_KMO(mdata,Ffit3)

kmo_val2<-VData_KMO(mdata,Ffit2)

kmo_val1<-VData_KMO(mdata,Ffit1)

KMO1<-KMO_Set(kmo_val1)

KMO2<-KMO_Set(kmo_val2)

KMO3<-KMO_Set(kmo_val3)

par(mfrow=c(3,2))

plot(sthod1,KMO1,type="o",xlab=expression(delta[1]),ylab="KMO",col=1)

plot(sthod1_rdata,KMO1_rdata,type="o",xlab=expression(delta[1]),ylab="K
MO",col=2)

plot(sthod2,KMO2,type="o",xlab=expression(delta[2]),ylab="KMO",col=1)
```



```
plot(sthod2_rdata,KMO2_rdata,type="o",xlab=expression(delta[2]),ylab="K  
MO",col=2)
```

```
plot(sthod3,KMO3,type="o",xlab=expression(delta[3]),ylab="KMO",col=1)
```

```
plot(sthod3_rdata,KMO3_rdata,type="o",xlab=expression(delta[3]),ylab="K  
MO",col=2)
```

```
#Example 2
```

```
#Import data into R
```

```
#Dataset3 : Concrete strength
```

```
#=====
```

```
Datta<-read.spss("concreteStrength.sav", use.value.label=TRUE,  
to.data.frame=TRUE)
```

```
Mdata<-data.matrix(Datta)[-c(8,9)]
```

```
colnames(Mdata)<-c("X1","X2","X3","X4","X5","X6","X7")
```

```
#hi<-hpi(mdata)
```

```
smooth_par<-unlist(lapply(seq_len(dim(Mdata)[2]),function(i){
```

```
hscv(Mdata[,i])
```

```
}))
```

```
#hn<-hns(mdata)

Fitd<-DensFeat(Mdata,sm_par,dprop=0.7)

CVMat<-Corrv(mdata2)

sthod1<-strehod(mdata2)$sth

sthod2<-strehod(mdata2)$st2

sthod3<-strehod(mdata2)$st3

Sfit1<-SenAnalysis(mdata2,sthod1)

Sfit2<-SenAnalysis(mdata2,sthod2)

Sfit3<-SenAnalysis(mdata2,sthod3)

kmo_val3<-VData_KMO(mdata2,Sfit3)

kmo_val2<-VData_KMO(mdata2,Sfit2)

kmo_val1<-VData_KMO(mdata2,Sfit1)

time1<-system.time(kmo_val1<-VData_KMO(mdata2,Sfit1))

time2<-system.time(kmo_val2<-VData_KMO(mdata2,Sfit2))

time3<-system.time(kmo_val3<-VData_KMO(mdata2,Sfit3))

KMO1<-KMO_Set(kmo_val1)

KMO2<-KMO_Set(kmo_val2)

KMO3<-KMO_Set(kmo_val3)
```

```
par(mfrow=c(2,2))  
  
plot(sthod1,KMO1,type="o",xlab=expression(delta[1]),ylab="KMO")  
  
plot(sthod2,KMO2,type="o",xlab=expression(delta[2]),ylab="KMO")  
  
plot(sthod3,KMO3,type="o",xlab=expression(delta[3]),ylab="KMO")
```

APPENDIX B

CODES FOR SIMULATING DATA BASED ON ITEM RESPONSE THEORY

SIMULATION ON SEVEN-POINT SCALE

```
library(mirt)  
library(ltm)  
library(psych)  
library(polycor)  
  
d=matrix(c(  
0.357,0.714,1.071,1.428,1.785,2.142,2.5,  
0.331,0.662,0.993,1.324,1.655,1.986,2.32,  
0.283,0.566,0.849,1.132,1.415,1.698,1.98,  
-3.0,-2.574,-2.145,-1.716,-1.289,-0.858,-0.429,  
-2.5,-2.142,-1.785,-1.428,-1.071,-0.714,0.357,  
-2.0,-1.716,-1.430,-1.144,-0.858,-0.572,-0.286,  
0.300,0.600,0.900,1.200,1.500,1.800,1.830,2.1,  
-2.5,-2.142,-1.785,-1.428,-1.071,-0.714,0.357,  
0.286,0.572,0.858,1.144,1.430,1.716,2.0,
```

0.376,0.752,1.128,1.504,1.880,2.256,2.63,
0.321,0.642,0.963,1.284,1.605,1.926,2.25,
-1.7,-1.458,-1.215,-0.972,-0.729,-0.486,-0.243,
-2.30,-1.974,-1.645,-1.316,-0.987,-0.658,-0.329,
0.357,0.714,1.071,1.428,1.785,2.142,2.5,
-2.7,-2.316,-1.930,-1.544,-1.158,-0.772,-0.386,
0.283,0.566,0.849,1.132,1.415,1.698,1.98,
-2.30,-1.974,-1.645,-1.316,-0.987,-0.658,-0.329,
0.343,0.686,1.029,1.379,1.715,2.058,2.400,
0.450,0.900,1.350,1.800,2.250,2.700,3.15,
-3.18,-2.724,-2.270,-1.816,-1.362,-0.908,-0.454,
0.557,1.114,1.671,2.228,2.785,3.342,3.9,
0.386,0.772,1.158,1.544,1.930,2.316,2.7,
0.304,0.608,0.912,1.216,1.520,1.824,2.13,
0.286,0.572,0.858,1.144,1.430,1.716,2.0,
0.367,0.734,1.101,1.468,1.835,2.202,2.57,
-0.84,-0.72,-0.600,-0.480,-0.360,-0.240,-0.120,
-0.20,-0.174,-0.145,-0.116,-0.087,-0.058,-0.029,
-0.36,-0.306,-0.255,-0.204,-0.153,-0.102,-0.051,
-0.63,-0.540,-0.450,-0.360,-0.270,-0.180,-0.090,
0.059,0.118,0.117,0.236,0.295,0.354,0.41,
0.029,0.058,0.087,0.116,0.145,0.174,0.20,
0.100,0.200,0.300,0.400,0.500,0.600,0.70,
0.101,0.202,0.303,0.404,0.505,0.606,0.71,
-0.51,-0.365,-0.292,-0.219,-0.146,-0.073,

0.127,0.254,0.381,0.508,0.635,0.762,0.89,
-0.95,-0.816,-0.680,-0.544,-0.408,-0.272,-0.163,
-0.65,-0.558,-0.465,-0.372,-0.279,-0.186,-0.093,
-0.4,-0.342,-0.285,-0.228,-0.171,-0.114,-0.057,
0.071,0.142,0.213,0.284,0.355,0.426,0.5,
-0.60,-0.516,-0.430,-0.344,-0.258,-0.172,-0.086,
0.129,0.258,0.387,0.516,0.645,0.774,0.900,
0.086,0.172,0.258,0.344,0.430,0.516,0.600,
0.057,0.114,0.171,0.228,0.285,0.342,0.400,
0.021,0.042,0.063,0.084,0.105,0.126,0.15,
0.043,0.086,0.129,0.172,0.215,0.258,0.300,
0.026,0.052,0.078,0.104,0.130,0.156,0.180,
-0.76,-0.654,-0.545,-0.436,-0.327,-0.218,-0.109,
0.033,0.066,0.099,0.132,0.165,0.198,0.230,
-0.19,-0.162,-0.135,-0.108,-0.081,-0.054,-0.027,
-0.4,-0.342,-0.285,-0.228,-0.171,-0.114,-0.057),ncol=7,byrow=TRUE)

#Difficulty parameter

```
d40=matrix(c(  
d[c(1:20,26:45),1],d[c(1:20,26:45),2],d[c(1:20,26:45),3],  
d[c(1:20,26:45),4],d[c(1:20,26:45),5],d[c(1:20,26:45),6],  
d[c(1:20,26:45),7]),ncol=7,byrow=FALSE) #40 Variables
```

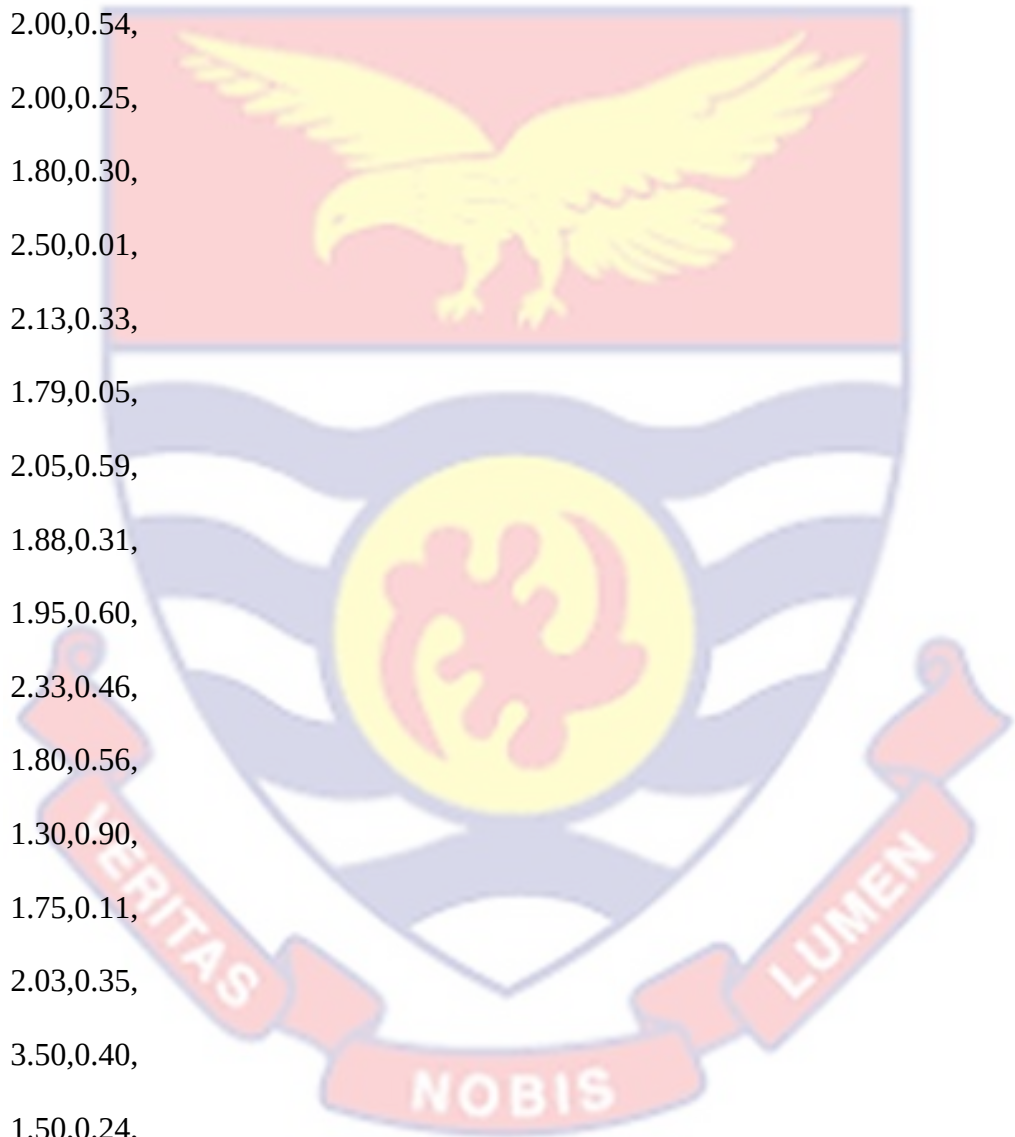
```
d30=matrix(c(  
d[c(1:15,26:40),1],d[c(1:15,26:40),2],d[c(1:15,26:40),3],
```



```
d[c(1:15,26:40),4],d[c(1:15,26:40),5],d[c(1:15,26:40),6],  
d[c(1:15,26:40),7]),ncol=7,byrow=FALSE) #30 Variables
```

```
#Two-dimensional Dataset
```

```
a2=matrix(c(  
2.00,0.54,  
2.00,0.25,  
1.80,0.30,  
2.50,0.01,  
2.13,0.33,  
1.79,0.05,  
2.05,0.59,  
1.88,0.31,  
1.95,0.60,  
2.33,0.46,  
1.80,0.56,  
1.30,0.90,  
1.75,0.11,  
2.03,0.35,  
3.50,0.40,  
1.50,0.24,  
2.06,0.15,  
1.95,0.05,  
2.65,0.19,  
2.93,0.13,
```



3.15,0.25,

2.17,0.20,

1.95,0.16,

2.15,0.09,

1.89,0.18,

0.54,2.00,

0.25,2.00,

0.30,1.80,

0.01,2.50,

0.33,2.13,

0.05,1.79,

0.59,2.05,

0.31,1.88,

0.60,1.95,

0.46,2.33,

0.56,1.80,

0.90,1.30,

0.11,1.75,

0.35,2.03,

0.40,3.50,

0.24,0.50,

0.15,2.06,

0.05,1.95,

0.19,2.65,

0.13,2.93,



2.00,0.54,0.56,

2.00,0.25,0.90,

1.80,0.63,0.11,

2.50,0.41,0.35,

2.13,0.33,0.40,

1.79,0.05,0.24,

2.05,0.59,0.15,

1.88,0.31,0.05,

1.95,0.60,0.19,

2.33,0.46,0.13,

1.80,0.56,0.25,

1.30,0.90,0.20,

1.75,0.11,0.16,

2.03,0.35,0.09,

3.50,0.40,0.18,

1.50,0.24,0.54,

2.06,0.15,0.25,

1.95,0.05,0.30,

2.65,0.19,0.01,

2.93,0.13,0.33,

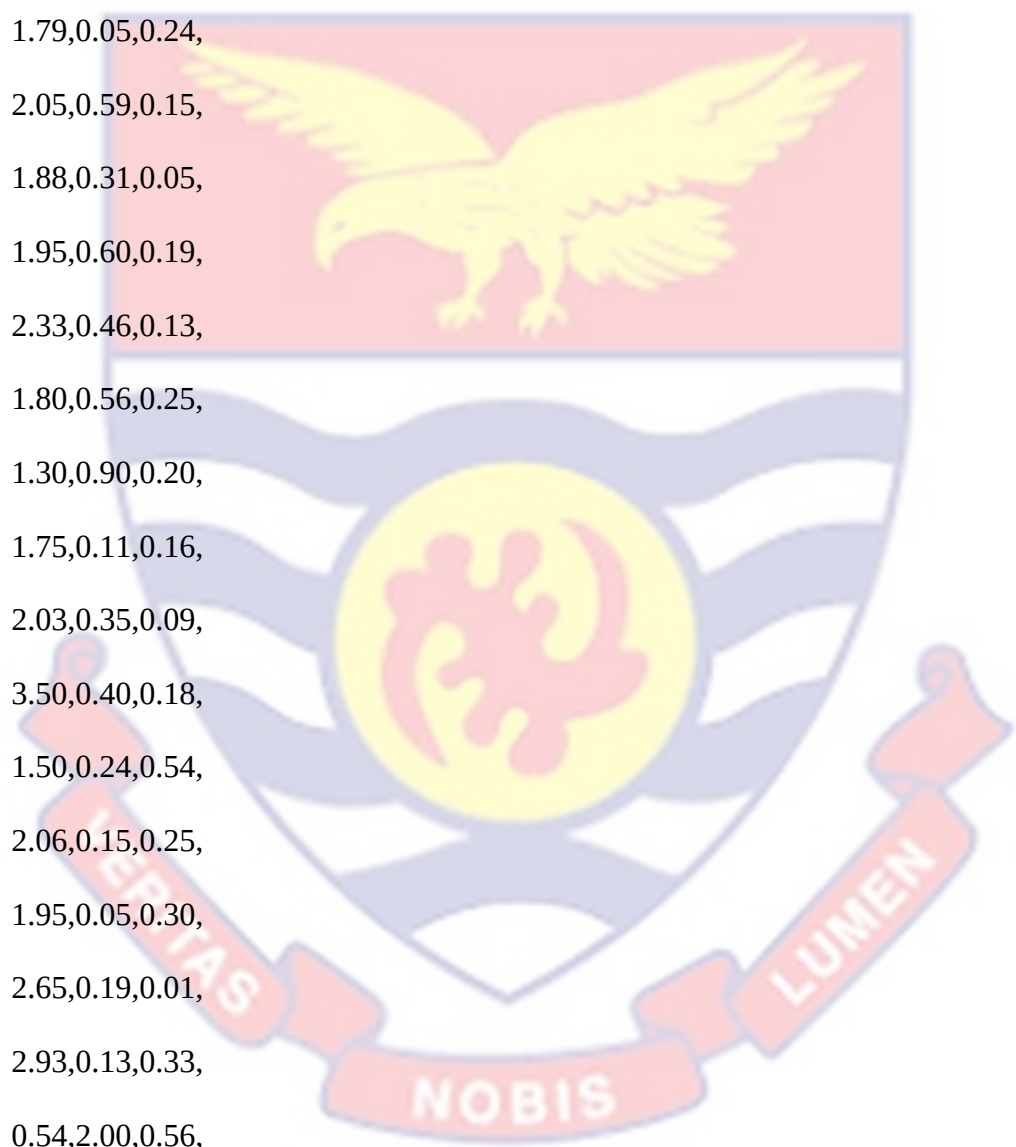
0.54,2.00,0.56,

0.25,2.00,0.90,

0.30,1.80,0.11,

0.01,2.50,0.35,

0.33,2.13,0.40,

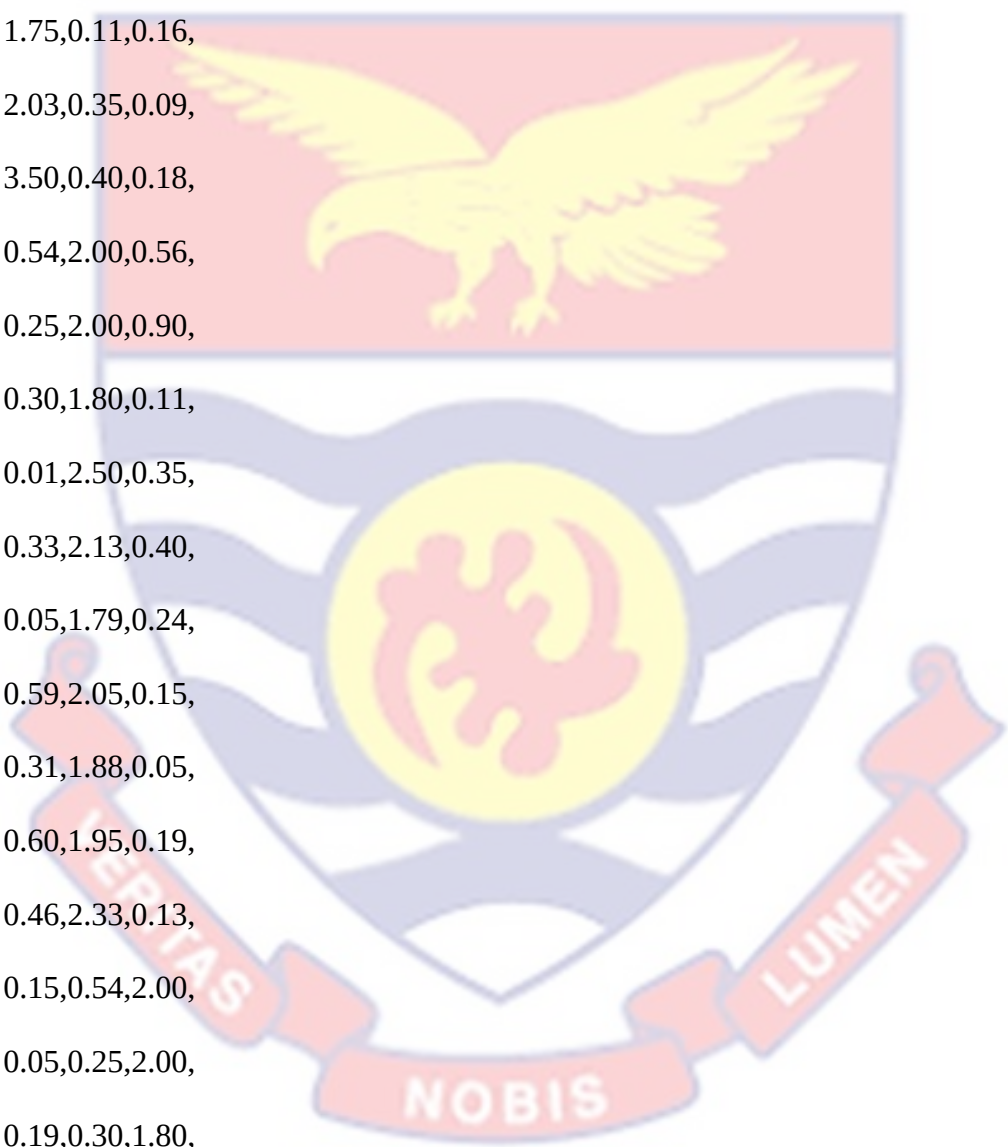


```
0.05,1.79,0.24,  
0.59,2.05,0.15,  
0.31,1.88,0.05,  
0.60,1.95,0.19,  
0.46,2.33,0.13,  
0.56,1.80,0.25,  
0.90,1.30,0.20,  
0.11,1.75,0.16,  
0.15,0.54,2.00,  
0.05,0.25,2.00,  
0.19,0.30,1.80,  
0.13,0.01,2.50,  
0.25,0.05,2.13,  
0.20,0.59,1.79,  
0.16,0.31,2.05  
) ,ncol=3,byrow=TRUE) #40 Variables  
  
a3.30=matrix(c(  
2.00,0.54,0.56,  
2.00,0.25,0.90,  
1.80,0.63,0.11,  
2.50,0.41,0.35,  
2.13,0.33,0.40,  
1.79,0.05,0.24,  
2.05,0.59,0.15,
```



1.88,0.31,0.05,
1.95,0.60,0.19,
2.33,0.46,0.13,
1.80,0.56,0.25,
1.30,0.90,0.20,
1.75,0.11,0.16,
2.03,0.35,0.09,
3.50,0.40,0.18,
0.54,2.00,0.56,
0.25,2.00,0.90,
0.30,1.80,0.11,
0.01,2.50,0.35,
0.33,2.13,0.40,
0.05,1.79,0.24,
0.59,2.05,0.15,
0.31,1.88,0.05,
0.60,1.95,0.19,
0.46,2.33,0.13,
0.15,0.54,2.00,
0.05,0.25,2.00,
0.19,0.30,1.80,
0.13,0.01,2.50,
0.25,0.05,2.13

),ncol=3,byrow=TRUE) #30 Variables



```
set.seed(2001);data2=simdata(a=a3.30,d=d30,N=200,itemtype="gpcm")  
  
#Data simulation  
  
set.seed(2001);data3=simdata(a=a3.40,d=d40,N=200,itemtype="gpcm")  
  
#Factor Analyses  
  
FA30<-fa(r=data2,nfactors = 3,n.obs = 200,rotate = "varimax",fm="pa",cor =  
"poly")  
r=FA30$r  
print(r,digits=3,max=2000)  
print(FA30$loadings,digits = 3)  
print(FA30$fit)  
  
FA40<-fa(r=data3,nfactors = 3,n.obs = 200,rotate = "varimax",fm="pa",cor =  
"poly")  
r=FA40$r  
print(r,digits=3,max=2000)  
print(FA40$loadings,digits = 3)  
print(FA40$fit)
```

