KABIRU BALA

COMPARATIVE ANALYSIS OF CLASSIC ALGORITHMS AND METHODS IN EXPLORATORY DATA ANALYSIS: A RETROSPECTIVE COHORT STUDY OF ANTIRETROVIRAL PATIENTS, VISITING FEDERAL TEACHING HOSPITAL GOMBE STATE, NIGERIA

PhD THESIS

YEAR 2023

NEAR EAST UNIVERSITY

INSTITUTE OF GRADUATE STUDIES

DEPARTMENT OF BIOSTATISTICS

COMPARATIVE ANALYSIS OF CLASSIC ALGORITHMS AND METHODS IN EXPLORATORY DATA ANALYSIS: A RETROSPECTIVE COHORT STUDY OF ANTIRETROVIRAL PATIENTS, VISITING FEDERAL TEACHING HOSPITAL GOMBE STATE, NIGERIA

PhD. THESIS

Kabiru BALA

Nicosia

September, 2023

NEAR EAST UNIVERSITY
INSTITUTE OF GRADUATE STUDIES
DEPARTMENT OF BIOSTATISTICS

COMPARATIVE ANALYSIS OF CLASSIC ALGORITHMS AND METHODS IN EXPLORATORY DATA ANALYSIS: A RETROSPECTIVE COHORT STUDY OF ANTIRETROVIRAL PATIENTS, VISITING FEDERAL TEACHING HOSPITAL GOMBE STATE, NIGERIA

PhD. THESIS

Kabiru BALA

Supervisor
Prof. Dr. İlker ETİKAN

Nicosia
September, 2023

Approval

We, the undersigned, affirm the thesis titled "Comparative Analysis of Classic Algorithms and Methods in Exploratory Data Analysis: A Retrospective Cohort Study of Antiretroviral Patients at Federal Teaching Hospital, Gombe State, Nigeria" by Kabiru Bala meets the standards for the degree of Doctor of Philosophy in Biostatistics.

Thesis Committee;

Chair of the committee: **Pro. Dr. İlker ETİKAN (Advisor)**

Near East University

Sign: .....

Member: **Prof. Dr. Beyza ŞAHİN**

Pamukkale University

Sign: .....

Member: **Prof. Selim YAVUZ SANİSOĞLU**

Ankara Yıldırım Beyazıt University

Sign: .....

Member: **Assoc. Prof. Dr. Özgür TOSUN**

Near East University

Sign: .....

Member: **Assoc. Prof. Dr. Prof. UĞUR BİLGE**

Akdeniz University

Sign: .....

Approved by: **Prof. Dr. K. Hüsnü Can BAŞER**

Director of Institute of Graduate Studies

Near East University

Sign: .....

Dedication

I, Kabiru Bala, declare that all material, documents, analyses, and results in this thesis were obtained and presented following the academic regulations and ethical principles of the Health Science Institute at Near East University. As required by these rules and conduct, I further declare that I have thoroughly cited and referenced information and data that are not unique to this work.

Kabiru BALA

28./09/2023

Acknowledgments

Grateful to Allah for guiding my successful PhD journey. Special thanks to Prof. Dr. İlker Etikan, my dedicated supervisor, and my teachers for their support. Appreciation to my PhD jury members and Assoc. Prof. Dr. Özgür Tosun for mentorship. Heartfelt gratitude to my wife, Rashida Kabiru, and family for their unwavering support. Thanks to research colleagues, especially Dr. Sani Isah Abba and Dr. Abdullahi Garba, and friends for their encouragement. Also, appreciation to Taraba State Polytechnic Suntai, TETFUND, and Near East University for the scholarship. My journey wouldn't be complete without acknowledging my parents, especially my Father Bala Abubakar, and my Uncle Gambo Abubakar for their contributions. I'm also thankful for the inspiration from friends like Comrade Yahya Kodabo. Through their emotional and social support, they've been instrumental in my success.

Kabiru BALA

Abstract

Comparative analysis of classic algorithms and methods in exploratory data analysis: a retrospective cohort study of antiretroviral patients, visiting Federal Teaching Hospital Gombe State, Nigeria

Kabiru Bala

Prof. Dr. İlker Etikan

Doktora Tezi, Biyoistatistik Anabilim Dalı

Eylül, 2023, (89) sayfa

The suppression of viral load achievement brought global success for the ongoing antiretroviral therapy (ART) treatment because of an undetected viral load record in patients with HIV. Lack of cooperation from society outside the treatment group is the major setback in sustaining and improving the achievement of ART treatment. This study combined exploratory methods and class algorithm techniques of multivariate analysis by using their coordinate space to project the success of ART-treated patients at Federal Teaching Hospital Gombe State, Nigeria. The exploratory methods employed correspondence analysis (CA), multiple correspondence analysis (MCA), and principal component analysis (PCA) while the class algorithm used multidimensional scaling (MDS) of metric and non-metric MDS, which proxscal dissimilarity and similarity including alscal dissimilarity were used in running the analysis. The proxscal MDS was measured using Torgerson stress while the alscal MDS was measured using Kruskal's stress. Similarly, their performance was evaluated using dispersion accounted for (DAF), Stress and squared correlation (RQS), the variance accounted for (inertia), initial eigenvalue (Total variance explained), and accounted for (proportion of inertia). The exploratory method of MCA accounted for a 0.782 coefficient of determination to offer the best results in predicting similarity compared to the 0.641 results in PCA. In a separate analysis, the CA was revealed as the true constructor of both

MCA and PCA. However, the exploratory class algorithm of proxscal similarity gave a high coefficient of determination of 0.95134 with less stress of 0.16537 proven to be the best for predicting the distance in ART follow-up.

***Key Words*:** Proxscal and alscal proximity, Burt matrix, correlation matrix, antiretroviral therapy, dual analysis.

Ozet

Keşif amaçlı veri analizinde klasik algoritmaların ve yöntemlerin karşılaştırmalı analizi: Nijerya Gombe Eyaleti Federal Eğitim Hastanesi'ni ziyaret eden antiretroviral hastalar üzerinde retrospektif bir kohort çalışması

Kabiru Bala

Prof. Dr. İlker Etikan

Viral yük başarısının baskılanması, HIV'li hastalarda tespit edilemeyen viral yük kaydı nedeniyle devam eden antiretroviral tedavi (ART) tedavisine küresel başarı getirdi. Tedavi grubu dışındaki toplumla işbirliği eksikliği, ART tedavisinin başarısının sürdürülmesinde ve iyileştirilmesinde en büyük engeldir. Bu çalışma, Nijerya Gombe Eyaleti Federal Eğitim Hastanesi'nde ART ile tedavi edilen hastaların başarısını projelendirmek için koordinat uzayını kullanarak çok değişkenli analizin keşif yöntemlerini ve sınıf algoritma tekniklerini birleştirdi. Keşif yöntemlerinde uygunluk analizi (CA), çoklu uygunluk analizi (MCA) ve temel bileşen analizi (PCA) kullanılırken, sınıf algoritması metrik ve metrik olmayan MDS'nin çok boyutlu ölçeklendirmesini (MDS) kullandı; bu, alscal farklılığı da dahil olmak üzere yakın ölçek farklılığı ve benzerliği sağlar Analizin yürütülmesinde kullanıldı. Proxscal MDS, Torgerson stresi kullanılarak ölçülürken, alscal MDS, Kruskal stresi kullanılarak ölçüldü. Benzer şekilde performansları, açıklanan dağılım (DAF), Stres ve kare korelasyonu (RQS), açıklanan varyans (atalet), başlangıç özdeğeri (açıklanan toplam varyans) ve açıklanan (atalet oranı) kullanılarak değerlendirildi. MCA'nın keşif yöntemi, PCA'daki 0,641 sonuçlara kıyasla benzerliği tahmin etmede en iyi sonuçları sunmak için 0,782'lik bir belirleme katsayısını hesaba kattı. Ayrı bir analizde CA'nın hem MCA hem de PCA'nın gerçek kurucusu olduğu ortaya çıktı. Bununla birlikte, yakın ölçek benzerliğine ilişkin keşif sınıfı algoritması, 0,95134'lük yüksek bir belirleme katsayısı vermiş ve 0,16537'lik daha az

stresin, YÜT takibinde mesafeyi tahmin etmede en iyi olduğu kanıtlanmıştır.

***Anahtar Kelimeler*:** Proxscal ve alscal yakınlığı, Burt matrisi, korelasyon matrisi, antiretroviral tedavi, ikili analiz**.**

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AIDS	Acquired immunodeficiency syndrome
ALSCAL	Alternating least square scaling
ART	Antiretroviral therapy
ARV	Antiretroviral
CA	Correspondence analysis
CMDS	Classical multidimensional scaling
DAF	Dispersion accounted for
DC	Coefficient of determination/R-square
DHHS	Department of Health and Human Services
DNA	Deoxyribonucleic acid
ED	Eigenvalue decomposition
EDA	Exploratory data analysis
FDA	Food and drug administration
GFA	General factor analysis
GSVD	Generalized singular value decomposition
HAART	Highly active antiretroviral therapy
HIV	Human immunodeficiency virus
HOMALS	Homogeneity analysis
INDSCAL	Individual difference scaling
INSTIs	Integrase strand transfer inhibitors
KMO	Kaiser meyer-olkin
MCA	Multiple correspondence analysis
MDS	Multidimensional scaling
MMDS	Metric multidimensional scaling
MSA	Measure of sampling adequacy
MTCT	Mother-to-child transmission
NMDS	Non-metric multidimensional scaling
NNRTIs	Non-nucleoside reverse transcriptase inhibitors

NRTIs	Nucleoside/Nucleotide reverse transcriptase inhibitor
PCA	Principle component analysis
PCM	Principal component methods
PCCA	Principal component correlation analysis
PCoA	Principal coordinate analysis
PIs	Protease inhibitors
SAS	Statistical software suite
SDV	Singular value decomposition,
SIV	Simian Immunodeficiency Virus
SIVagm	Simian immunodeficiency virus of African green monkeys
SIVcpz	Simian immunodeficiency virus of chimpanzee
SIVgor	Simian immunodeficiency virus of gorilla
SIVsm	Simian immunodeficiency virus of sooty mangabey
SIVsyk	Simian immunodeficiency virus of syke'smonkeys
SPSS	Statistical package for social sciences
SSA	Sub-Sahara Africa
UNAIDS	Joint United nation programme on HIV/AIDS
WMDS	Weighted multidimensional scaling

CHAPTER I

Introduction

HIV is a class of retrovirus or lentivirus established from the Human Immunodeficiency Virus (HIV) family, the gradual destruction of the immune system by the lentivirus causes significant havoc to the immune system, this, calls for urgent commencement of early treatment to avoid triggering to Immunodeficiency syndrome (AIDS) as ratified by the clinical expert. (Gougeon, 2003; Kim et al., 2014; Desimmie et al., 2014) both agreed that the virus of HIV dysfunctions the immune system rapidly by lowering the CD4 level as a result of immunological weakness and other infections that will advance to death. (Abram et al., 2014; Binka et al., 2012; Nyamweya et al., 2013) disclosed that HIV belongs to a species of retrovirus of the lentivirus type tracing its origination back then in the 20th century from Sub-Saharan Africa with transmitting mode from monkeys. (Campbell-Yesufu & Gandhi, 2011) classified HIV into two (HIV-1 and HIV-2) stages with epidemiological and biological aspects different across West Africa. Thus, HIV-2 species infect only minorities, while HIV-1 species become the global distributor. The current antiretroviral therapy (ART) medications are better drugs for fighting HIV-1 and HIV-2 than ever before, which has been the only medication found to be effective in the fight against HIV/AIDS. (CDC, 2019) endorsed antiretroviral therapy (ART) as the only acceptable treatment for HIV both locally and globally. (Esbjörnsson et al., 2019; Campbell-Yesufu & Gandhi, 2011) confirm that only the current available antibody test can specifically discriminate between the two distinct viruses (HIV-1 & HIV-2). HIV-1 found to be the most terrible virus on the planet accounting for over 95% of all deadly diseases while HIV-2 is the dominant type tied to West Africa and other nations. HIV-2 progresses slowly causing fewer infections, and is not as dangerous as HIV-1. HIV-2 differs genetically by 55% from HIV-1 according to the evaluation. HIV-1 and HIV-2 are nearly identical in action, when these protease enzymes

are not properly regulated, they can cause AIDS and death. Despite the negative side effects, people live with HIV. (Hemkens et al., 2015; Wynberg et al., 2018; Raffi et al., 2014) said currently**,** ART has improved the quality of patients' lives. WHO has reaffirmed the use of Lamivudine 3TC, Tenofovir TDF, and Efavirenz EFV (EFV/TDF/3TC) as the first HIV treatment regimen. In the same vein, the introduction of integrase inhibitors, fusion inhibitors, gp120-CD4 attachment inhibitors, and CCR5 antagonist therapy, combined with other treatments lowers HIV/AIDS transmission rate and death. Normally, vpx proteins in HIV-2 stand replaced by the vpu proteins in HIV-1. (Bangsberg et al., 1997; D. R. Bangsberg et al., 2004) testified that certain regimens have different levels of adherence and resistance to antiretroviral treatment, and some of the resistance shown in patients is due to negligence in patients without following tight medication regimens.

HIV/AIDS global edge trendy

(UNAIDS, 2021) estimate showed there were 37.7 million positive persons worldwide as of June 2021 battling with HIV. (UNAIDS, 2021; Frank et al., 2019) confirm the declining rate of about 50% progress in people with new HIV cases, in contrast to about 64% drop in related fatalities of HIV considering the previous record years of 1999 and 2006. According to (UNAIDS, 2021) 84% of HIV careers recognized their status toward the 2020 windup, and 87% of those with status known, commenced antiretroviral therapy (ART) treatment, based on this, 90% of those taking medications showed their viral load was suppressed. (UNAIDS, 2021) reported that there were 10.2 million people who contracted HIV and are yet to commence their ART medication, new diagnoses hit 1.5 million, and then 680,000 deaths related to HIV were ascertained. According to the UNAID, 67% of people in Africa, found with the virus were specifically from Sub-Saharan Africa (SSA), newly infected showed about 58%, and death as a result of AIDS hit 68%. The SSA record shows mothers are becoming the victims of HIV infection,

while children are not receiving appropriate attention; this is a severe goals setback in terms of eradicating HIV/AIDS cases.

The interface of HIV/AIDS in Africa reached

(UNAIDS, 2021) confirm that the severe rate of high HIV affecting SSA conceded two-thirds of the total numbers of HIV in the world, showcasing their channels of transmission via heterosexual contact in adults, through mother-to-child transmission (MTCT), and therapeutic practices. (UNAIDS, 2021) confirm that the severe rate of high HIV infection in Sub-Saharan Africa reached two-thirds of the global total number of HIV with the means of transmission via heterosexual contact in adults, through mother-to-child transmission (MTCT), and therapeutic practices. Based on the (UNAIDS) evaluation of adults and children in Eastern and Southern Africa contracted with HIV estimated at 20.6 million, newly infected were 670.000 and deaths connected to AIDS were about 310.00. Conversely, the adults and children in Western, Central Africa who contracted HIV accounted for about 4.7 million, newly infected, and deaths related to AIDS gave 200.00 and 150.000. Finally, the adults and children in Middle and Eastern North Africa living with HIV recorded 230.000, newly infected gave 16.000 and dead linked to AIDS gave 79.000 AIDS.

Based on the preceding, patients should adopt complete compliance treatment to avoid limiting their medications, which could lead to an increase in drug resistance and viral spreading. Consequently, any patient involved in censorship or postponement of ART medicine will not only develop AIDS but also deny themselves the ability to exist.

Purpose of the study

This research is targeted to examine the ART success development for the HIV/AIDS-treated patients at Gombe State Federal Teaching Hospital.

The aims and objectives

Is to compare exploratory data analysis (EDA) techniques performance to see the contribution and non-contribution alliance in-patients on ART based on geometrical or visual structure.

Significance of the study

The impact of this research will lead to the recommendation of dimensional design methodologies for analyzing ART data. The findings would aid Teaching Hospital Gombe State, as well as Nigeria as a whole to provide a lasting solution for the ART victims. This will increase the confidence and courage of those on treatment, as well as encourage and motivate those who do not yet know their status to go for an HIV test so that they can be placed on ART if they are positive. Non-believers will be convinced that HIV exists and that ART can cure the virus to some extent.

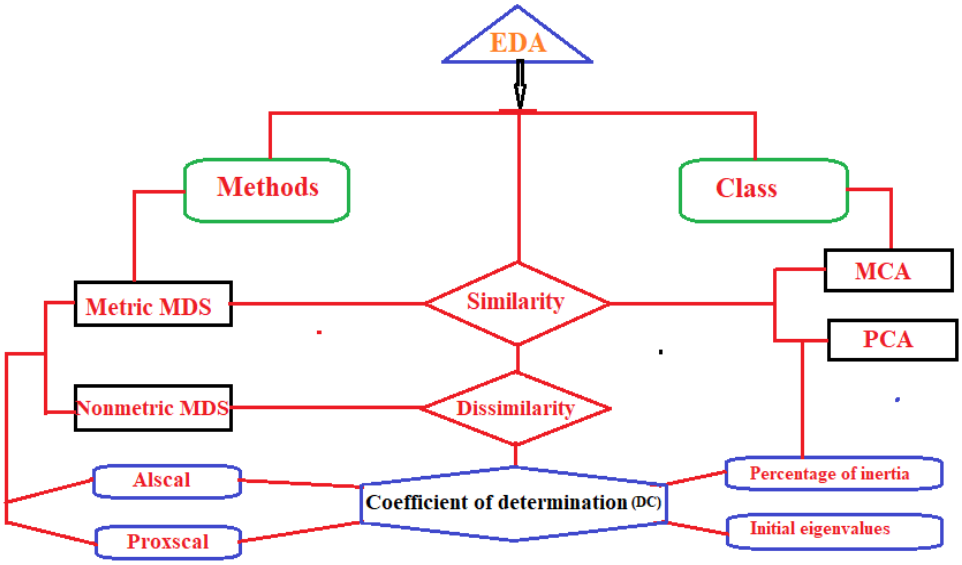
CHAPTER II

Review of Related Literature

Introduction

This section contained the accompanying literature review supporting this current write-up; it was built from sources like journals, textbooks, research theses, dictionaries, and different sources from the internet. The literature used will serve as references for this research. All the kinds of literature used in this research work were connected to classic algorithms and methods of exploratory data analysis comparison, a cohort retrospective study of ART patients visiting teaching hospitals in Gombe State, Nigeria.

Figure 1 EDA framework



Historical background of the pandemic

(Takehisa et al., 2009; Hayflick, 1992) testify that from 1910 through 1930, researchers studied that western wildlife gorillas caused the Simian Immunodeficiency Virus (SIV). However, the zoonotic viral transmission episode between the Pan Troglodyte's chimpanzee and homo sapiens human-caused HIV emerged. Regarding this, the powerful simian immunodeficiency virus (SIVgor) of gorillas affects lowland gorillas in

western central Africa. This is allied to SIVcpz of chimpanzee human immunodeficiency viruses connected to HIV-1. According to (Campbell-Yesufu & Gandhi, 2011) HIV-1 is hazardous compared to HIV-2, because it emerges as the virus leading to AIDS, while HIV-2 is a common type affecting West Africa.

Agent of transmission (SIV) according to types

Different types of SIV according to African region include African green monkeys (SIVagm), mandrill (SIVmnd), Syke'smonkey (SIVsyk), Sooty mangabey (SIVsm), Chimpanzee (SIVcpz), and Asian macaques. Conversely, HIV-1 is the strain of SIV common with chimps, and HIV-2 is the strain of SIV common with sooty mangabey monkeys. However, the genetic variant refers to strain. Is culture within a biological species, which is portrayed as non-natural for specific genetic isolation. Humans and chimpanzee species are primate lentivirus groups common to HIV-1/Sivcpz while sooty mangabeys, macaques, and humans are primate lentivirus groups common to HIV-2/Sivsm/mac and African green monkeys are primate lentivirus groups common to HIV-2/agm.

Figure 2 Monkey to human SIV transmission

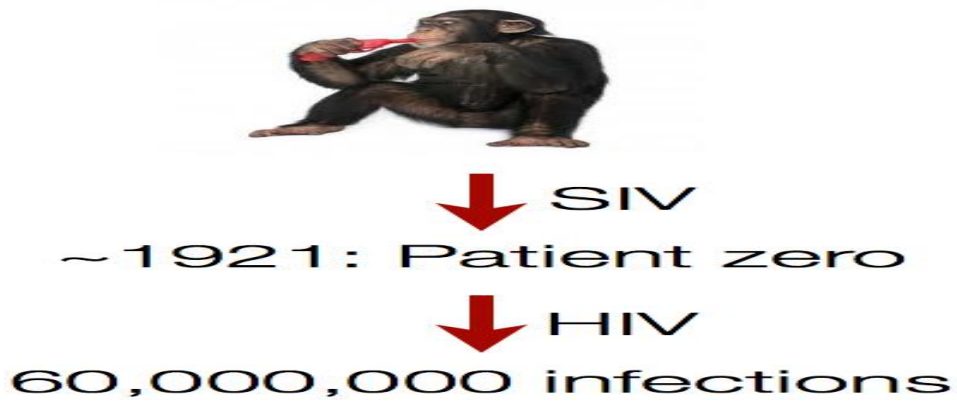


Figure 3. Monkey species

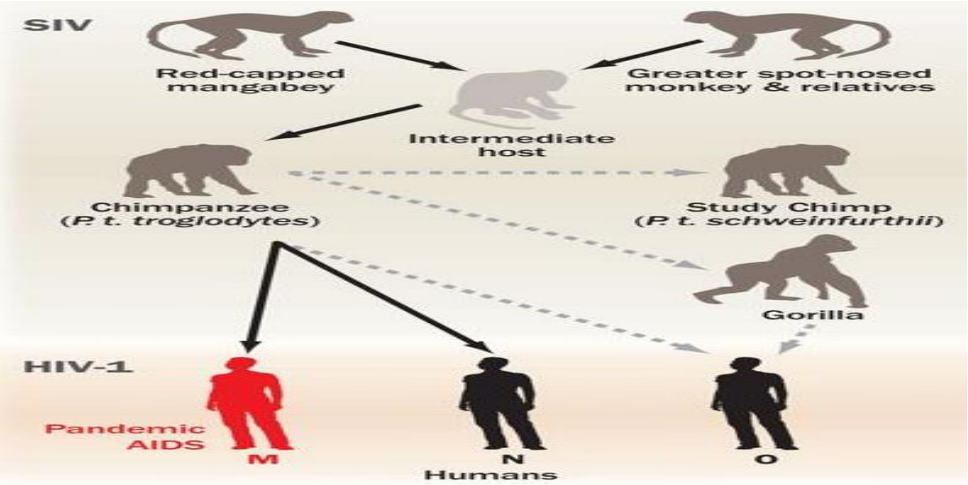
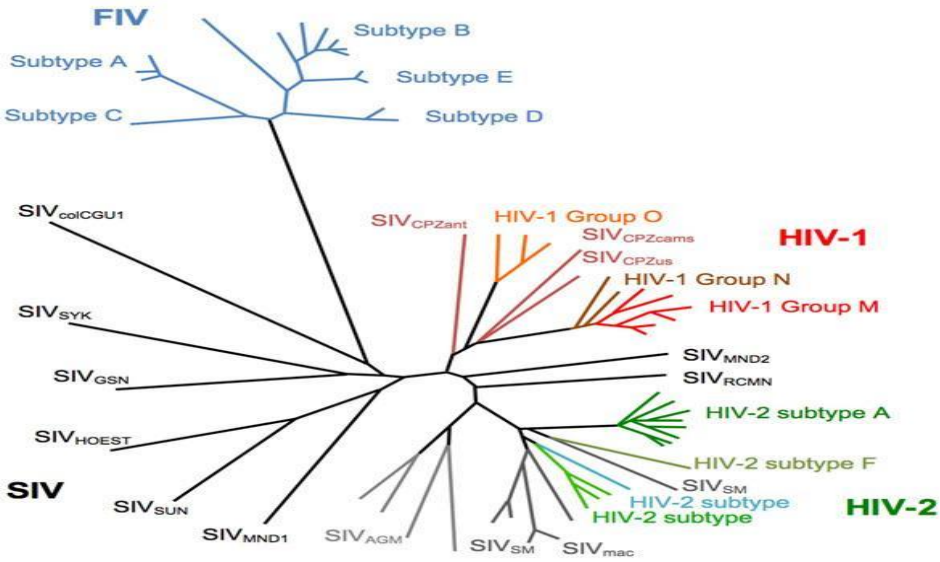


Figure 4: HIV-1 and HIV-2 formation



Prevention of HIV/AIDS care and management

According to (Vlahov & Junge, 1998) HIV and Syphilis blood screening before transfusion, including voluntary counseling and testing are the crucial actions to be taken as HIV control measures. Regarding this, preventative approaches such as intravenous drug and exchange use of

needle programs were implemented. (Alonso & De Irala, 2004) confirmed that after establishing the key channels of HIV infection, preventive measures were devised around the world, including raising public awareness of HIV/AIDS through public media campaigns, and declaring any person with HIV-positive status in public, particularly celebrities. (Frank et al., 2019; Roomaney et al., 2022) demonstrated that the current progress with HIV utilizing antiretroviral (ARV) for a longer period of treatment dramatically reduced HIV morbidity, death, and transmission incidents. (Lu et al., 2018) attest that from 2006 till date, HAART regimens endorsed the approval of the first single fixed-dose combination as the first drugs comprising of Efavirenz (EFV), Tenofovir (TDF), Emtricitabine (FTC) and the second-generation NNRTIs such as Etravirine (ETV) respectively, as well as the introduction of new drugs such as Integrase Strand Transfer Inhibitors (INSTI)-based regimens. (Jesmin et al., 2013; UNAIDS) confirm that abstinence, properly using a condom, and being faithful to one partner were important measures initiated to curtail any kind of disease from spreading further. Despite this, mother transmission prevention of HIV to child through Nevirapine, AZT, or combination therapy is necessary.

(Waning et al., 2009) described the HIV current treatment standard evolving from one-drug therapy to twofold-nucleoside therapy, to 3-drug therapy, among also, 2 nucleoside analogues blending with a non-nucleoside reverse transcriptase inhibitor known as protease inhibitor. Based on this, WHO recommends the following first-line combination of ARV regimens for adult and adolescent ART.

HIV/AIDS care and treatment

A pharmacologic group containing nucleoside reverse transcriptase inhibitors (NNRTI) uses nevirapine-NVP, delaviridine-DLV, and efavirenz-EFV as first-preventive drugs, while rilpivirine-RPV, etravirine-ETR are used as second-generation drugs. The DHHS endorsed

EFV/TDF/FTC, EFV/ABC/3TC, and RPV/TDF/FTC as guidelines. Regarding qualities, the single pill regimens can easily be administered, it has durability and suitable virologic potency. However, HAART drugs deliver lower genetic barriers in terms of advancement to resistance. The PI groups consists of amprenavir-AMP, indinavir-IND, lopinavir-LPV, ritonavir-RTV, and fosamprenavir, respectively. Nelfinavir comprising ritonavir-boost atazanavir-ATV/r, ritonavir boost tipranavir-TPV/r, and ritonavir-boost darunavir-DRV/r serve as the non-boosting sample Based on this, DHHS recommended ATV/r along with TDF/FTC/DRV/r, TDF/FTC/ATV/r along with, ABC/3TC LPV/r, plus TDF/FTC/LPV/r along with ABC/3TC. The drugs possess the qualities of good virologic potency and durability in treating HIV-positive patients. Administering with other related medicines involving inhibition of the cytochrome P (CYP) 450 enzyme and PIs can link to metabolic deformities. These combinations evidenced significant drug-drug interactions. However, the Integrase Strand Transfer Inhibitors (INSTI) group comprises raltegravir-RAL, elvitegravir-EVG, and dolutegravir-DTG, which the guidelines by DHHS recommended RAL-TDF/FTC, DTG-ABC/3TC and DTG-TDF/FTC. The quality of INSTI-based regimens is endorsed according to HAART. The Nucleoside/Nucleotide reverse transcriptase inhibitors (NRTI) involved lamivudine, zidovudine, emtricitabine, abacavir, tenofovir disoproxil fumarate, didanosine, dideoxyinosine, and stavudine. The entry or fusion inhibitors comprise the class of maraviroc enfuvirtide.

The sequential and breaking-through development of ART

Zidovudine (ZDV) or azidothymidine (AZT) become nucleoside analogue reverse transcriptase inhibitors (NRTIs) used for the advancement of viral load handling. Single NRTI agents and dual NRTIs contributed to this success (Anderson et al., 2004). Serving as the first breakthrough in 1987 in the history of HIV medication under the brand name Retrovir. The USA Food and Drug Administration (FDA) approved NRTIs when about 32,000 people were affected, and as a result, many lives were lost. After the trial,

the challenges faced included severe anemia, skin rashes, allergic reactions, liver problems, muscle disease, and blood disorders. Subsequently, highly active antiretroviral therapy (HAART) in 1996 was the initial regimen to use first-line drugs to control viral load and reduce HIV/AIDS death. These drugs are formed from two backbones of NRTIs and one base of PI or NNRTI or NSTI drugs targeting to lower the effect side and suppress the viral load (Bowen et al., 2020; Ayele et al., 2018). At present, the antiretroviral therapy (ART) begun in 2007 became the current breakthrough of HIV taking with or without food. The uniqueness of the ART has reduced the viral load and can control all the cases related to HIV effectively. The mixture of two NRTIs and an integrase inhibitor constitutes the current ART regimen (Lanman et al., 2022). The exceptionality of the ART drugs fits into four different drug classes: NRTIs, NNRTIs, PIs, and integrase inhibitors (IIs), which can be taken on an empty stomach.

Figure 5 Combination of Single NRTI agents and dual NRTIs (Zidovudine or azidothymidine)

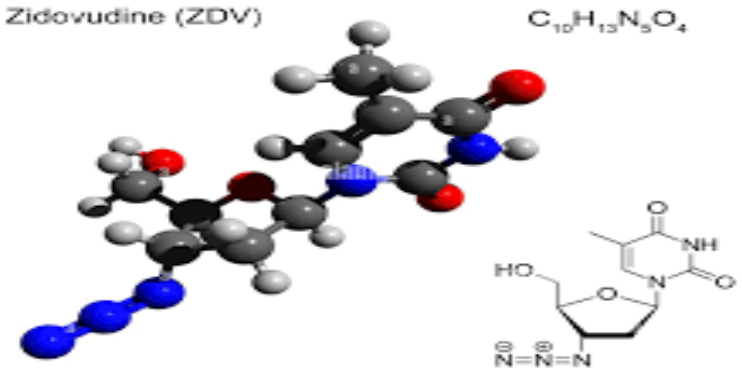
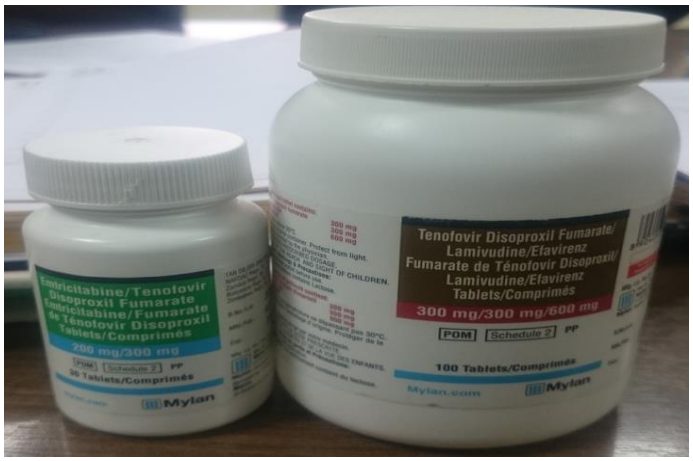


Figure 6 combination of NRTIs and one base of PI or NNRTI or NSTI drugs (HAART)

Figure 7 combination of four different classes; NRTIs, NNRTIs, PIs, and integrase inhibitors (ART)



The life cycle of HIV/AIDS centered on the below six categories of ART drugs

Nucleoside reverse transcriptase inhibitors (NRTIs): serve as enzyme blockers named reverse transcriptase. Here, the drug blends to deny chances for multiple copies of the virus by building blocks as an interface. It is sometimes referred to as nucleoside analogues (Nucleotide reverse transcriptase inhibitors). Examples of these drugs are Abacavir-ABC (Ziagen); Emtricitabine-FTC (Emtriva); Tenofovir alafenamide-TAF (Vemlidy); Lamivudine-3TC (Epivir); Tenofovir disoproxil fumarate-TDF (Viread); and Zidovudine-ZDV (Retrovil).

Non-nucleoside reverse transcriptase inhibitors (NNRTIs): It give a gap from active sites of RT. Also, they oblige as the non-nukes used to fix distracted protein to prevent the replicas of the virus copy. NNRTIs

constitute; Doravirine-DOR (Pifeltro); Efavirenz-EFV (Sustiva); Etravirine-ETR (Intelence); Nevirapine-NVP (Viramune); and Rilpivirine-RPV (Edurant). Integrase inhibitors (Integrase strand transfer inhibitors (INSTIs): This is an enzyme blocker also called integrase, which goes well with raltegravir drug.

Protease inhibitors (PIs): An enzyme blockade referring to as protease that flows with the combination of the following; atazanavir-ATV (Reyataz); darunavir-DRV (Prezista); Lopinavir + ritonavir-LPV/r (Kaletra); and ritonavir-RTV (Norvir). Integrase strand transfer inhibitors (INSTIs) called Integrase Inhibitors (INIs): This is a major player that blocks the protein to stop the virus from making copies into the healthy DNA. The drugs combination includes; Bictegravir-BIC, amalgamated with other drugs such as Biktarvy; Cabotegravir and rilpivirine (Cabenuva), Dolutegravir-DTG (Tivicay), Elvitegravir-EVG (Vitekta), and Raltegravir-RAL (Isentress).

Entry inhibitors (ENIs) forming from enfuvirtide and maraviroc and Pharmacokinetic enhancers serve as the booster drugs.

Dual analysis Correspondence analysis (CA) and Techniques

(Fellenberg et al., 2001) defined CA as a multivariate analysis that handles computational statistics to examine between correlating variables by means of exploratory strategies. That is, it is good at finding relationships between two simultaneous variables, and can communicate visual ideas in low dimension similar to PCA. According to (Pearl, 1994) CA is a bivariate nonlinear network of multidimensional scaling that handles analysis including quantification theory, optimal scaling, method of reciprocal averages, and so on. However, it is made to examine a group's interactions of categorical data concerning two or more dimensional plots. Is an exploratory analysis that strategies the exploration of categorical data (Hoffman & Franke, 1986). Despite this, the CA in terms of using categorical data, shares approaches with MCA, FA, PCA, MDS, Cluster

Analysis, and nonlinear canonical correlation (NLCC) and both use exploratory methods to explain the variance in the model that can be viewed in a low-dimension. (Fellenberg et al., 2001 & Hoffman & Franke, 1986) described correspondence analysis as the best method for examining data validity and controlling outlier management. (Pearl, 1994) confirmed that the CA uses the matrix and dimensionality of a map, which are the output coordinates of the maps for the sites as row objects and for the types as column objects. Regarding this, it offers different types of plots including the decomposition of the inertia very easily. Hence, it decomposes the total inertia from the vectors to the components in the dimension of first, second and so on. The first dimension is achieved using the component to divide by the total components. The precise results explain the percentage of the inertia, leading to a smaller percentage in the second dimension and so on. We used $\mathrm{t} = \min$ for $\mathrm{r} = 1$ and $\mathrm{c} = 1$ dimensions, each is selected to reduce the percentage of the total inertia. (Kudlats et al., 2014) attested that various researchers frequently used CA to collect and analyze categorical data due to its simplicity in nature. These include ecological and marketing research fields. (Sourial et al., 2010) have proven CA as a graphical tool in a multivariate analysis that uses epidemiological data to explore correlations among categorical variables. Sourial also, described that the CA used a contingency table to handle the graphical representation of relative frequencies to generate the average distance of individual row and column profiles. Hence, the visualized plot aids in revealing the association's pattern among the variables in a low dimensional point between the row ''i and row i′(i ≠ i′)''. Expressively Chi-square matric is define as;

$$d(i, i') = \frac{\sum_j (p_{ij} - p_{i'j})}{p_{+j}} \quad (1)$$

Where; p_{ij} and $p_{i'j}$ in CA function as the relative frequencies for row i and i' in column j, marginal relative frequency (p_{+j}) known as the mass, j is seen as an instance defining Chi-square distance involving two dimensions $\Lambda^2 = \sum p_{i+}\, d_i$. Where the marginal relative frequency (p_{+j}) for row i, and $d_{i=} \frac{\sum_j (p_{ij} - p_{ij})^2}{p_{ij}}$ is used for computing the profiles for the rows and the averages defining the Chi-square distance.

Technicality of CA on Contingency tables

The term "contingency table" was invented by a well-known researcher (Franke et al., 2012) as a tool employed to solve problems related to associations relying on categorical data. Is a measure that completely disagrees with the independence involved in the rows and columns structure of data, in which the total number of dimensions is decided using the rows and columns numbers. The marginal frequencies generated are termed as the profiles of the rows and columns. (E. Beh & Lombardo, 2019) substantiated that the CA is a technique that uses contingency tables to present a graphical representation, on which the same scale of data is expected to be used for the measurement. However, the data matrix is handled using the profiles of row and column, in which the data contained in the row is divided by the row total Likewise, ‘‘the data contained in the columns’’ is divided by the column total. Similarly, an individual row and column generates its weight according to the mass, while rows and column summation generates the total results.

CA regarding exploratory techniques

The history of CA began in 1960 by French Linguist and data analyst Jean-Paul Benzercri and his colleagues, defined it as a geometric method of multivariate descriptive computation of data. The CA is a multivariate exploratory technique that works with two sections of component using categorical data of either nominal or ordinal type, hence some of this data will later be converted into a certain distance, and this distance will be further designed for the perceptual map. In the context of exploratory

analysis, the CA foundation generates the map on the types and the map on the sites correspondingly and then generates a low-dimensional geometric representation to define the map. However, when selecting two equal dimensions, the map on the type contains c points in the plane, and then one point will correspond with each type. In addition, when selecting three types of dimensions, the map on the site point contained r for three-dimensional space.

The perceptual map, biplot, and centroid formation of CA

The perceptual map is the correspondence map or plot, known as the output. This output forms the row and column points for the two-dimensional plot. Perhaps, dimensions in CA help to describe the variation or inertia in a data set, normally, two or three dimensions are considered adequate for the biplot. However, the biplot in CA corresponds to the same origin of a contingency table analysis, which is linked to the Chi-Square test for homogeneity. In today's world, CA is used to manage graphical arrangements of contingency tables, it can be utilized also to establish the relationship between variables with two-way categorical data. The centroid is the weighted mean contained the profiles of the row and column forming the data originality of the axis's arrangement in the correspondence map. Presentation of the family tree using categorical data to reveal patterns of association has achieved headway in statistical application. (Beh & Lombardo, 2019 and E. J. Beh, 2004) said the family three exhibition is based on computational results of reciprocal average, centroid position, optimal scaling and so on.

The singular value in CA

The maximum canonical correlation between any analyses of categorical data surfacing to dimension refers to the singular value, which successfully the singular value attainment is cantered on the square-root of the inertia coefficient.

Inertia term (variance)

Generally, the inertia, or percentage denoted as the variance in CA, and the coefficients of the inertia are the basic roots of building PCA. Each dimension contains one eigenvalue, and each eigenvalue explains the amount of variance in the CA table. However, the eigenvalues (inertia) display the virtual meaning concerning the dimensions, while the sum of the eigenvalues gives the total inertia, hence, the total inertia spreads the points all over the centroid. Hence, the first dimension explains the most variance because of its largest eigenvalue, followed by the second and next respectively. Normally, only results in the first two dimensions revealed the correspondence map, to account for the detailed percentage of inertia by the current model in the original table equal to the sum of the eigenvalues.

Frequency Normalization

Mathematically, the hypothesis for the row's homogeneity can be compared with the independent hypothesis. Especially, when the interest is to compare the abundance matrix in the rows, since individual rows can explain the types of distribution by normalizing the rows and dividing the individual row by the row total. Conversely, comparing the columns is also appropriate to make the treatment symmetrical, but is less common in archeology. Thus, we used the sum of the ‘‘row and column’’ in stabilizing the frequency table and then divided the table items through their respective marginal row and column, hence, is defined as;

$$p_{j|i} = \frac{n_{ij}}{n_{i*}} = \frac{p_{ij}}{p_{i*}} \tag{2}$$

$$p_{i|j} = \frac{n_{ij}}{n_{*j}} = \frac{p_{ij}}{p_{*j}} \tag{3}$$

The independence hypothesis $p_{ij} = p_{i*}p_{*j}$ is written as;

$$p_{j|i} = p_{*j},$$

(4)

$$p_{i|j} = p_{i*},$$

(5)

Consequently, the rows and columns' homogeneity are described, such that the homogeneity of rows can explain the types of probability distribution to have the same positioning altogether while that of column homogeneity can explain the position of probability distribution to have the same types altogether. Defining the CA inertias depends on the measures of homogeneity for the rows and columns of results symbolically defined as;

$$\chi^2_{i*} = \sum_{j=1}^{c} \frac{\left(p_{j|i} - p_{*j}\right)^2}{p_{*j}}$$

(6)

$$\chi^2_{*j} = \sum_{i=1}^{r} \frac{(p_{i|j} - p_{i*})^2}{p_{i*}}$$

(7)

Regarding this, the rows with the highest inertia are going to be different from the average row (vector p_{*j}) of column marginal proportions, while the columns with the highest inertia are going to be different from the average column of p_{i*}. We define the total inertia of the weighted sum of row and column as;

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(p_{ij} - p_{i*}p_{*j})^2}{p_{i*}p_{*j}}$$

(8)

$$= \sum_{i=1}^{r} p_{i*}\chi^2_{i*}$$

(9)

$$= \sum_{j=1}^{c} p_{*j}\chi_{*j}^{2}$$

(10)

Measurement of distance in CA

In CA, the distance connecting the row profiles and the column profiles is termed as Chi-square distance, referring to as the weighted profile procedure. Researchers used the distance between two points of data in a multi-dimensional space to define the CA, which Euclidean distance aids in defining the Euclidean space of distance between the two points.

Physical distance

The distance between row points is measured using χ^2 distance, which has few differences from the physical distance between points in vector space. To measure the physical distance between two vectors $x = [x_1, x_2 \ldots x_n]$ and $y = [y_1, y_2 \ldots y_n]$ is defined by;

$$\text{physical distance} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2}$$

(11)

Chi-square distance

Is defined as a distance that uses the corresponding column marginal proportion to weight each squared term in reverse as;

$$\chi^2 = \text{distance} = \sqrt{\frac{(x_1 - y_1)^2}{c_1} + \frac{(x_2 - y_2)^2}{c_2} + \cdots + \frac{(x_n - y_n)^2}{c_n}}$$

(12)

Distance according to Benzécri distances

Multiple correspondence analysis

The MCA is broadly defined as a statistical procedure for identifying, classifying, summarizing, and visualizing categorical data. However, MCA can arrange data provided by individuals to their similarities and differences. It also aids in the connection of data features from individuals so that visual appearance and relationships may be readily identified.

Similarly, the MCA is based on the application of CA to the Burt Matrix ($\mathrm{B} = \mathrm{G}^{\mathrm{T}}G$) referring to a matrix containing all pairwise correlations between variables. The MCA, on the other hand, runs PCA on the G indicator matrix using precise row and column weights; hence, the weights ensure that the Chi-square is used to interpret the distances between the respective rows and the principal components. Also, it serves as the new variables connecting with the set of variables using the squared correlation ratio (η^2) analysis of variance to measure the relationship. MCA according to (Husson et al., 2019; Johs, 2018) is an exploratory technique that does analysis using generated data from questionnaires, ideally categorical or nominal data. (Enzécri, 1979) described homogeneity analysis (HOMALS) and the MCA was developed to express the relationship between categorical variables in a variety of disciplines, including political science, educational research, social sciences research, marketing data, health psychology, educational research, product and food preferences, and mammalian epidemiology, to name a few. (Methods, n.d. and Greenacre & Blasius) used survey data to analyze responses from different participants to consider a set of data containing n rows with J categories variables,($\mathrm{v}_{\mathrm{j\,j=1\,.\,.\,.j}}$) and $\mathrm{K_j}$ categories. Therefore, the dummy variables of the indicator matrix are given as, $\mathrm{G_{n\times K}}$, $\mathrm{K} = \sum_{\mathrm{j}} \mathrm{K_j}$ with $\mathrm{g_{ijk}} = 1$, such that a person i will select the category k class of variable j and $\mathrm{g_{ijk}} = 0$. Below is an illustration of the three variables.

$$G = [G_1|G_2|G_3] = \left[\begin{array}{ccc|ccc|cc} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{array}\right]$$

When the data sets are smaller, the MCA yields a greater percentage, resulting in powerful links between the data. When the data sets are larger,

the percentage of inertia associated with each axis becomes weak. (Audigier et al., 2017; Hoffman & Franke, 1986; Methods, n.d. and Kocatepe et al., 2020) described multivariate exploratory methods such as PCA (correlation matrix), CA (disjunctive binary table or indicator matrix), and MCA (Burt matrix) as dimensional reduction techniques that share common objectives regarding exploration and visualization, it only differ on the type of the study data. In general, they utilize a contingency table (cross-tabulation), either two way or multi-way table, to explain similarities and relationships between rows and columns. (Methods, n.d.) attested that the MCA can be considered as a powerful tool that can process any complex analysis with higher values and high-dimensional data, it can also be viewed as a means of employing cross-tabulation to summarize data that can fit relationships between variables. However, a common type is to display numerical information graphically in the form of patterns and thus numerically relate the variables. The points in MCA maps serve as the foundation for revealing the relationship between the investigated variables. Various sciences scholars, marketing, health biomedical and so on acknowledge the usefulness of MCA application. According to (Salkind, 2012) the indicator matrix is known by various names, including homogeneity analysis, appropriate (optimal) scoring, dual scaling, and quantification methodologies. The indication matrix allows for 0 and 1 entries, making it unique in that, it is more advanced than CA. However, it can evaluate an array associated with categorical multiple dependent variables using similar PCA features. The achievement of the CA on the indicator matrix allows the factor score for the rows and columns to be computed.

Algorithm and notation of MCA

(Salkind, 2012) described that the building of rows and columns factor scores generated by the indicator matrix is based on the factor scores rescaling of CA to form the MCA. The table of $J \times J$ of the Burt Matrix associated with X is given by $B = X^T X$. The MCA table (Burt matrix)

building on CA is often easier to calculate and can yield the same factors as using X. That is, the eigenvalues derived via the table of the MCA provide satisfactory inertia approximation of factors compared to the eigenvalues of X.

Burt (Indicator) matrix

The Burt matrix is a modified sub-matrix on its diagonal, the Burt matrix is a J by J table obtained as $B = X^T X$ which has an association with X. Building CA on the Burt matrix provides features similar to the analysis of X and rarely simpler to analyze. However, it plays an important role in acquiring eigenvalues and can give reliable inertia approximation by the factors compared to the eigenvalues of X. Despite this, the MCA computation can be accomplished with I observations in which the I X J indicator matrix is symbolized as; X, and the K nominal variables will contain J_k levels, then the sum of the J_k will be equal to J. As a result, the proper functioning of the CA on the indicator matrix is stated to be determined by two sets of factor scores from row and column. To allow the variances to be equal to their corresponding eigenvalue, the two-factor scores should be proportionate. On the other hand, when N signifies the total number, then the subsequent process for the analysis will manipulate the running of the probability matrix $Z = N^{-1}X$. The r stands for the total row vector-matrix Z, then $r = Z_1$ and 1 becomes vector of $1's$, c stands for column totals, and $D_c = \text{diag}\{c\}$, $D_r = \text{diag}\{r\}$.

Symbolically, factor scores are based on the singular value decomposition of $D_r^{-\frac{1}{2}} (Z - rc^T) D_c^{-\frac{1}{2}} = P\Delta Q^T$

(15)

Where; the matrix for singular values is denoted as Δ, the eigenvalue matrix represented by $\lambda = \Delta^2$. Thus, the factor scores for the row and column matrix can be attainable as follows;

$$\mathrm{F} = \mathrm{D_r}^{-\frac{1}{2}} \mathrm{P}\Delta \left(\mathrm{D_r}^{-\frac{1}{2}} \mathrm{UT} \right)$$

(16)

$$\mathrm{G} = \mathrm{D_c}^{-\frac{1}{2}} \mathrm{Q}\Delta \left(\mathrm{D_r}^{-\frac{1}{2}} \mathrm{VT} \right)$$

(17)

The rows and column Chi-Square distance with respective barycenter can be achieved by,

$$\mathrm{d_r} = \mathrm{diag}\{\mathrm{FF^T}\}$$

(18)

$$\mathrm{d_c} = \mathrm{diag}\{\mathrm{GG^T}\}$$

(19)

The squared cosine aids in facilitating the importance of factors from any set of variables, which the squared cosine of row i with factor ℓ and column j with factor ℓ is given by;

$$O_{i,\ell} = \frac{f_{i,\ell}^2}{d_{r,i}^2}$$

(20)

$$O_{j,\ell} = \frac{f_{j,\ell}^2}{d_{c,j}^2}$$

(21)

Where; $d_{r,i}^2$ represent the i^{th}element of d_r and j^{th}element of d_c respectively.

The influence of row i to factor ℓ and columns j to factor ℓ supported the locating of the important variables defined as;

$$t_{i,\ell} = \frac{f_{i,\ell}^2}{\lambda_\ell}$$

(22)

$$t_{j,\ell} = \frac{g_{j,\ell}^2}{\lambda_\ell}$$

(23)

Using the below formula, symbolic elements are projected on one factor, in which their coordinates $f_{sup}\ and\ g_{sup}$ are attained mathematically as;

$$f_{sup} = \left(i_{sup}^T 1\right) i_{sup}^T G\Delta^{-1}$$

(24)

$$g_{sup} = \left(j_{sup}^T 1\right) j_{sup}^T F\Delta^{-1}$$

(25)

Where; the projected (i_{sup}^T and j_{sup}^T) represent the row and column

Burt matrix associated with X (weighted least squares approximation of Burt matrix) for $J \times J$ is achieved by $B = \mathrm{X}^T X$ ($B \approx nr\mathrm{r}^T + nDXD_\beta \mathrm{X}^T D$).

Where; ($n,\ r\ and\ D$), represent the total, the row mass, and the diagonal matrix for the mass respectively. Let $S = n^{-\frac{1}{2}} D^{-\frac{1}{2}} BD^{-\frac{1}{2}}$, SVD of S is $S = UD_s \mathrm{V}^T$.

Categories and distance formation of cloud in MCA

Construction of cloud fitting

The cloud fitting for individual categories is expected to give a high dimensional and the best cloud of point representation, however, within the specified space, the eigenvectors u_k, which are associated with the eigenvalues λ_k, are acquired from the eigenvalue decomposition of $PD_c^{-1}P^T D_r^{-1}$, hence $\frac{1}{Q} ZD^{-1} u_k = \lambda_k u_k$. (Blasius, 2000 and Husson et al., 2019) described the analysis of cloud fitting as an interesting analysis that contributes to a given variable by connecting the square correlation ratio between the variable and the principal component using the relationship as follows; $\eta^2\left(f_k, \chi_q\right) = Q \times CTR_k(q) \times \lambda_k$. The $CTR_k(q)$ denotes the contributions of $CTR_k(j)$ to the entire j categories of variable q. Similarly, analyzing the column of cloud from the indicator matrix is performed based

on the representation of the categories constructed. Geometrically, the studying category deals with investigating the appearance of cloud points in $\mathbb{R}^N$. Thus, the distance revealed between two columns is called the chi-square distance symbolically defined as;

$$d^2_{\chi^2}(\text{column profile j, vs column profile j}') = \sum_{i=1}^{N} \frac{1}{r_i} \left(\frac{p_{ij}}{c_j} - \frac{p_{i'j}}{c_{j'}} \right) \quad (26)$$

The cloud of individuals

There is a need to create distance between individuals before constructing a cloud of individuals so that two individuals should be on the same levels of distance that will be equal to zero (0), two individuals should pick all the categories except the uncommon one, to make it far away and lastly the two individuals should have a common unequal level so that there will be closeness even if they are taken different levels from other variables. Thus, the cloud of individuals is defined as;

$$d^2_{ii'} = \frac{I}{J} \sum_{k=1}^{N} \frac{1}{I_k} (\times_{ik} - \times_{i'k})^2 \quad (27)$$

$$\sum_{k=1}^{K} \frac{1}{\frac{I_k}{(IJ)}} \left(\frac{\frac{\times_{ik}}{(IJ)}}{\frac{1}{I}} - \frac{\frac{\times_{i'k}}{(IJ)}}{\frac{1}{I}} \right)^2 \quad (28)$$

$$d_{\chi^2}(\text{row profile i, vs row profile i}') = \sum_{J=1}^{J} \frac{1}{f_{*j}} \left(\frac{f_{ij}}{f_{i*}} - \frac{f_{i'j}}{f_{i'*}} \right)^2 \quad (29)$$

The cloud of levels

Cloud of levels can be achieved using the distance between levels, which two levels are expected to be closer when many individuals are taking the levels, and the asymmetrical levels should demonstrate distance from each other.

$$d^2_{k,k'} = I\sum_{i=1}^{I}\left(\frac{\times_{ik}}{I_k} - \frac{\times_{ik'}}{I_{k'}}\right)^2$$

(30)

$$= \sum_{i=1}^{I}\frac{1}{\frac{1}{I}}\left(\frac{\frac{\times_{ik}}{(IJ)}}{\frac{I_k}{(IJ)}} - \frac{\frac{\times_{ik'}}{(IJ)}}{\frac{I_{k'}}{(IJ)}}\right)^2$$

(31)

$$d_{\chi^2}(\text{column profile j, column profile j}') = \sum_{i=1}^{J}\frac{1}{f_{i*}}\left(\frac{f_{ij}}{f_{*j}} - \frac{f_{ij'}}{f_{*j'}}\right)^2$$

(32)

On the other hand, two given categories of j and j' is used to define between the 'squared distance' as;

$$d^2_{\chi^2}(j, j') = \sum_{i=1}^{N}\frac{1}{\frac{1}{N}}\left[\frac{\frac{z_{ij}}{(NQ)}}{\frac{N_j}{(NQ)}} - \frac{\frac{z_{ij'}}{(NQ)}}{\frac{N_{j'}}{(NQ)}}\right]^2$$

(33)

$$= N\sum_{i=1}^{N}\left[\frac{z_{ij}}{N_j} - \frac{z_{ij'}}{N_{j'}}\right]^2$$

(34)

(Methods, n.d.) conclude that there is difficulty in justifying the chi-square distance between two categories in the indicator matrix. Hence the squared distance in the category j and the center of gravity (G_K) of the cloud categories is given by;

$$d^2(j, G_K) = N\sum_{i=1}^{N}\left(\frac{z_{ij}}{N_j} - \frac{1}{N}\right)^2 = \frac{N}{N_j} - 1$$

(35)

Considering cloud fitting, it is good to note that, the less frequency recorded from each category, the more distance or farther the center of gravity appears, however, the less frequently a category appears, the higher its inertia. In concert categories with lower frequencies that might heavily

affect the overall analysis of the results need to be avoided. Therefore, the squared distance weighted by $\frac{N_j}{NQ}$ (the weight of the category j) is the inertia of the category j, which is defined by;

$$\text{Inertia}(j) = \frac{N_j}{NQ} \times d^2(j, G_K) = \frac{1}{Q}\left(1 - \frac{N_j}{Q}\right) \quad (36)$$

The inertia in MCA

Inertia is expressed by the sum of inertia of its categories, when computing the categories of inertia, its weight counter offsets the squared distance. Conversely, the higher the number of categories recorded from a variable, the greater its inertia generates. Mathematically defined as;

$$\text{Inertia}(q) = \sum_{j=1}^{J_q} \frac{1}{Q}\left(1 - \frac{N_j}{N}\right) = \frac{J_q - 1}{Q} \quad (37)$$

Normally, the total inertia cloud of categories is similar to the cloud of individuals. This is justified based on the duality between two analyses of rows and columns. However, when all the inertias of the variables are collectively added, the total inertia of the cloud of categories can be obtainable by;

$$\text{Inertia} = \sum_{q=1}^{Q} \frac{J_q - 1}{Q} = \frac{J}{Q} - 1 \quad (38)$$

The term inertia or a percentage of inertia is universally accepted as the variance, in which the eigenvalue is divided by the estimated sum of the eigenvalues. However, the Burt matrix evaluates the inertia using the percentage of inertia by the average inertia of the off-diagonal blocks. Hence, two formulas are used to handle eigenvalues that are smaller than $\frac{1}{\text{K}}$. The eigenvalues are identical to the squared eigenvalues of X, relying on the equivalent of the MCA analysis with the Burt matrix. Therefore, a

better way to estimate this inertia is to extract each eigenvalue by the below-given formula

$$c\lambda_\ell = \begin{cases} \left[\left(\frac{K}{K-1}\right)\left(\lambda_\ell - \frac{1}{K}\right)\right]^2 & if\ \lambda_\ell > \frac{1}{K} \\ 0 & if\ \lambda_\ell \leq \frac{1}{K} \end{cases}$$

(39)

The expression of $\overline{g}$ inertia can be calculated as;

$$\overline{g} = \frac{K}{K-1} \times \left(\sum_\ell \lambda_\ell^2 - \frac{J-K}{K}\right)^2$$

(40)

Based on the above formula inertia percentage is used as the ratio defined by

$\tau_c = \frac{c\lambda}{\overline{g}}$ as a substitute for $\frac{c\lambda}{\sum c\lambda_\ell}$

Modification of Inertias

To improve the percentage of the inertia we need to adjust the inertias using the;

$$\frac{1}{K}\ \lambda_s^{adj} = \left(\frac{Q}{Q-1}\right)^2 \left(\lambda_s - \frac{1}{Q}\right)^2$$

(41)

In that, λ_s denotes the singular value suiting $\lambda_s \leq \frac{1}{Q}$ inequality, computed using the adjusted inertias. More so, it states the percentage of the average for the off-diagonal inertia, then it is computed either by using the total inertia of C or through direct computation of the off-diagonal using the Burt matrix table mathematically defined as follows;

$$\frac{Q}{Q-1}\left(inertia\ (C) - \frac{J-Q}{Q^2}\right)$$

(42)

Where; inertia (C) equal to sum **of** the principal inertias ($\sum_s \lambda_s^2$)

Bootstrap resampling procedure

Bootstrapping is a resampling strategy that uses random sampling with replacement to assign accuracy measurements like confidence intervals, prediction error, bias, variance, and so on to sample estimates. However, utilizing the random sampling techniques, the method tolerates the estimation of numerous sampling distributions. The technique is an estimator, regarding the approximation of sampling to variance using the evaluating properties. (Mayssara, Hassanin, 2019; Alonso-Atienza et al., 2012) and Del Giudice et al., 2018) confirmed that bootstrap resample is focused on creating a new dimensional record from sampling with replacement using the previous record. It is assumed that the operator generating the column demonstration's projection coordinates in the matrix is symbolized as Γ, then re-written as;

$$g_{sup} = \Gamma(X)$$

(43)

Also, it defines the column eigenvalue as ν, denoted by Θ, to become the operator used in creating the data matrix given as;

$$\nu = \Theta(X)$$

(44)

Statistical measurement concerning bootstrap replication is attained by analyzing the already quantity in the resampled population. Then in any given iteration, it resamples respectively to project the column and the eigenvector. Nevertheless, the resampling of data matrix X^* identifies the sampling replications of rows up to I times to give an equal size to the original data matrix.

$$g^*_{sup}(b) = \Gamma(\, X^*(b))$$

(45)

$$\nu^*(b) = \Theta\,(\, X^*(b))$$

(46)

Where; the asterisk denotes the Bootstrap resampling as a statistical elements plug-in principle and resampling of process, which is used in separating theoretical statistical factors.

Methods of Principal_Component_Analysis (PCA)

Considering the PCA techniques in handling dimensionality reduction to allow correlation patterns to be discovered in a group of data such that it can transform a significant group of data without leaving behind any important evidence. The PCA is also helpful in machine learning to reduce high-dimension data, when observed from previous data, it deals with removing inconsistencies, handling redundant data, and taking care of highly correlated features while going by the new data, it tries to lesser the features and then retain most of the information as possible. However, the PCA forms the new variables sets created by the initial set of variables, generally, they are computed in a manner that the newly acquired variables are expected to be highly significant and independent from each other. In addition, the PCA is the new fundamental feature that is obtained after accomplishing dimensionality reduction; this has to do with the procedure of cutting or reducing the variables. More so, the PCA is made to compress and possess most of the functional information that is scattered among the variables. (Chen & Chang, 1995) testifying that the PCA is among the neural network-based, related to algorithms of MCA. Regarding updating the equations of the weights based on previous literature, it does not rely

on the eigenvalue estimates rather it only controls the percentage of estimate done by the eigenvalue.

Mathematically, PCA is expressed based on the following eigenvalue problem as $\frac{1}{N-1}X'X_w = \lambda w$. The PCA in general analyses eigenvectors and refers to them as factor weights that can be used for extracting maximum variance in all variables. However, regarding the relationship finding between variables and factors, the factor weights and factor loadings are similar, just that the launching is on different scales. (Lebart et al., 1995) identified principal component methods (PCM) or General factor analysis (GFA) as a technique that handled geometrical methodology presentation of extracting and visualizing any facts in a set of variables x. Generally, it goes with contingency tables or frequency count and in between the row and columns, it explores the similarity of geometrical points. Similarly, the matrices used by the PCA are; the $(I \times K)$ data matrix of X, then the weighting system from the rows, and also the metric from the column space stored in $(I \times I)$ matrix of D, while the weighting system from the columns with metric from the row space is stored in $(K \times K)$ matrix of M. (Lebart et al., 1995) attested PCA as a principal components method that uses an individual's $\times$ table data of quantitative variables type. Despite this, the $(I \times K)$ matrix X includes K columns while I rows with the common term x_{ik} conveying the value of a variable k for a single I. Going by the PCA data, it is a column-based used for calculating the axes of inertia compared to the centroid of the individual cloud.

Principal component analysis_(PCA)_linking to covariance

PCA is built from complex to simplest without losing too much information, with its optimality details outcomes operating as the second-order statistics, known as covariance, in which the covariance tells how they are related while variance simply tells that the mean is zero in equivalent. The covariance in PCA describes how co-dependent two variables are, when it gives negative covariance, it indicates indirectly

proportional, and when it gives positive covariance it means directly proportional.

Rows and columns clouds formation

Standardization is the process of transforming the shape of a cloud by pairing all the angles of its variability from the original variables to give them the same magnitude, however, column standardization is acknowledged when the variables are maintained in different entities. The PCA is said to be standardized if the variables are standardized and centered while it is unstandardized if the variables are only centered. However, the PCA can be operated as an X matrix of data which can be either standardized or centered while with M metric and D weighting system in the space of row or column is seen as PCA (X,M,D).

Column and centroid inertia

The total inertia gives equal to the sum of the squared singular values of the data table, while the inertia of a column is identified as the sum of the squared factors of the column mathematically defined as;

$$\gamma_j^2 = \sum_i^I x_{ij}^2 \quad (47)$$

$$r_j^2 = \sum_i^I x_{ij}^2 \quad (48)$$

Where; γ_j^2 signified I called the total inertia.

The center of gravity

(Abdi & Williams, 2010) attested centroid or barycenter as the center of gravity for the rows denoting as g, which is the mean vector for each column in X. Once, the X is centered, then the center of gravity will be equivalent to $I \times J$ row vector (0^T). The i^{th} observation for the Euclidean distance g will be the same as

$$d_{jg}^2 = \sum_i^J (x_{ij} - g_i)^2$$

(49)

When the data are centered, the above equation is reduced to

$$d_{jg}^2 = \sum_i^J x_{i\,j}^2$$

(50)

Which the sum of all d_{jg}^2 is equal to I τ ℜ ℑ, standing as the inertia for data.

Least mean squared error (LMSE)

(004635259288F49Ffd000000.Pdf, n.d. and Yang, 1995) proved that the PCA is obtainable by adjusting the objective function centered on the approach of gradient descent Mathematically the least mean squared error (LMSE) can be elaborated using modified MSE function as;

$$E(W) = \sum_{t_1=1}^{t} \mu^{t-t_1} \parallel x_{t1} - WW^T x_{t1} \parallel^2$$

(51)

Where; $0 < \mu \leq 1$ signifies the neglecting factor that can handle the nonstationary progressions, then t means the instantaneous current time.

Kaiser Meyer Olkin (KMO) criterion

(Kaiser, 1974) disclosed KMO known to be the measure of sampling adequacy (MSA), used to indicate if a series of group variables may explain the correlations between the variables. The Bartlette test of sphericity is employed to test the null hypothesis by ensuring the correlation matrix is a diagonal type when altogether the non-diagonal elements are zero. (Russell, 2002; Zwick & Velicer, 1986) testified that the PCA uses the KMO and Bertler test to investigate the sampling adequacy of a model. The Kaiser criterion is a criterion acknowledged by specifying the number of factors, advocating more factors than required. However, it extracts factors with an eigenvalue greater than (> 1), to serve as the latent root criterion or Kaiser Criterion, which uses the number of factors as a determinant factor.

Principal axis factoring (principal factor analysis)

(Matsunaga & Masaki, 2010) certified that the objective of the PCA is to reproduce the main structure of data as much as possible as required handling a few factors, in the same vein the factor analysis uses correlations to describe the factor variables. However, PCA and factor analysis sometimes are designated as principal axis factoring or principal factor analysis used for ascertaining the arrangement of variable sets.

Rotation

Rotation serves as the solution provider that can accurately represent the data, and give a more reliable solution than the original solution. It used several numbers of component formations to simplify the interpretation and to retain the required component after the rotation. The rotation is performed based on two types; oblique and orthogonal rotation. Along the line, the oblique rotation is utilized when the new axes refuse to comply with the orthogonal, while the orthogonal rotation is employed to generate the new axis to become orthogonal to everyone. Ideally, the performance of rotation should be executed based on subspace, expecting fewer inertia explanations than the original components in the new axes. The execution of the rotation should be the same before and after explaining the inertia part based on the total subspace, only the portioning of the inertia should be changed. These subspaces form the basis for the rotation, serving as the space for the components to be retained. More so, the term loadings insinuate elements of the Q matrix during the procedure of rotation, which defines the PCA model as;

$$X = tp^T + E = \hat{X} + E \tag{52}$$

The PCA falls under the category of generalized distributions from the exponential family. Usually, the generalized linear model operates on the functions of Bregman distance and also it facilitates based on the mix techniques of dimensionality reduction using several distributions of data

characteristics. Transformation of correlation to an orthogonal means is the foremost aim of PCA. More so, the PCA gives more explanation with metric data and can fix a distance like the correspondence analysis (CA). On the other hand, the non-linear PCA reduces high components in sequence to provide sufficient representation. The nonlinear PCA results are independent compared to the corresponding linear result; in addition, the learning algorithm networks for the non-linear PCA can be organized into symmetric and hierarchical forms. The loading score in PCA ascertains the variables with the main effect and it uses fancy math terms such as Eigen decomposition to coordinate the graph.

Application of singular value decomposition in the PCA

The matrix forms the basic part of the application of singular value decomposition (SDV), that is the SDV is applied with the support of the matrix. The exploratory analysis technique in PCA is achieved either by eigenvalue decomposition (ED) of a sample containing a correlation matrix or by SVD. Thus, ED and SVD serve as a tool for achieving primal analysis, which serves as an alternative to geometric PCA building on common similarity. However, considering a set of points $\{x_j\}_{j=1}^{N}$ in $\mathbb{R}^D$ seeking to find a subspace ($S \subset \mathbb{R}^D$) dimension (d) that will give the best fits points ($x_j \in S$) so that each point can be represented as;

$$x_j = \mu + Uy_j, \quad j = 1, 2, . \quad . \quad . \quad , N \tag{53}$$

Where; $\mu \in S$ denotes the point in the subspace, U = D × d matrix, in which the columns form the basis for the subspace, and $y_j \in \mathbb{R}^d$ represents the vector of new coordinates of x_j in the subspace. Fundamentally, the SVD is a device used for extracting the eigenvector, subspace tracking, and total least squares problem. In addition, the SVD is the generalization of the eigendecomposition, which decomposes a rectangular matrix taken along

with one diagonal matrix, and two orthogonal matrices, symbolically expressed as;

$$A = P\Delta Q^{T} = \sum_{\ell=1}^{L} \delta_{\ell}\, p_{\ell} q_{\ell}^{T} \tag{54}$$

Where; column A describes the left singular vectors same as the eigenvectors of the matrix AA^{T} ($P^{T}P = 1$), the Q column signifies the right singular vectors of A and is the eigenvectors of matrix $A^{T}A(Q^{T}Q = 1)$ and Δ indicate the diagonal matrix of the singular values, and $\Delta = \Lambda^{\frac{1}{2}}$ link with Λ denoting the eigenvalues of the matrix AA^{T}. Also, L refers to the rank of X, then $\delta_{\ell}, p_{\ell,}$ and q_{ℓ} denote as $\ell^{-\text{th}}$ singular value and left and right singular vectors of X. Thus, X is reformed as the sum of L rank matrices ($\delta_{\ell}\, p_{\ell} q_{\ell}^{T}$) requirements. Despite this, we have generalized singular value decomposition (GSVD) to decompose a rectangular matrix to account for the constraints for the rows and the columns matrix. However, there is an assigning of weighted generalized least square estimate by the GSVD for a particular matrix which can decrease the rank matrix for the $I \times J$ in matrix A. Conversely, we used SVD to generalize two exact square matrices that are non-negative based $I \times I$ and $J \times J$ sizes. These two matrices articulate the constraints for the rows and columns of A presentation. However, when M becomes the $I \times I$ matrix conveying the constraints for row A and W and $J \times J$ constraints matrix for columns A, which matrix A can be decomposed as;

$$A = \widetilde{P}\,\widetilde{\Delta}\,\widetilde{Q}^{T} \text{ with } \widetilde{P}^{T}M\widetilde{P} = \widetilde{Q}^{T}W\widetilde{Q} = 1 \tag{55}$$

Generally, the generalized singular vectors are disclosed as orthogonal under the constraints of M and W control. The accomplishment of its decomposition is based on standard value decomposition, which can be established by expressing a matrix of $\widetilde{A}$ as

$$\widetilde{A} = M^{\frac{1}{2}} A W^{\frac{1}{2}} \Leftrightarrow A = M^{-\frac{1}{2}} \widetilde{A} W^{-\frac{1}{2}}$$

(56)

We then compute the SVD of $\widetilde{A}$ as;

$$\widetilde{A} = P \Delta Q^T \text{ with } P^T P = Q^T Q = 1$$

(57)

Therefore, the matrices of the generalized eigenvectors can be seen as;

$$\widetilde{P} = M^{-\frac{1}{2}} P \text{ and } \widetilde{Q} = W^{-\frac{1}{2}} Q$$

(58)

Then, the singular values for the diagonal matrix are equivalent to the matrix of singular values of $\widetilde{A}$; $\widetilde{\Delta} = \Delta$. To be verified with $A = \widetilde{P} \widetilde{\Delta} \widetilde{Q}^T$, and substituted by

$$A = M^{-\frac{1}{2}} \widetilde{A} W^{-\frac{1}{2}} = M^{-\frac{1}{2}} P \Delta Q^T W^{-\frac{1}{2}} = \widetilde{P} \Delta \widetilde{Q}^T$$

(59)

Thus, the condition supports it to give; $\widetilde{P}^T M \widetilde{P} = P^T M^{-\frac{1}{2}} M M^{-\frac{1}{2}} P = P^T P = 1$

$$\widetilde{Q}^T W \widetilde{Q} = Q^T W^{-\frac{1}{2}} W W^{-\frac{1}{2}} Q = Q^T Q = 1.$$

(60)

Eigenvectors and the eigenvalues

The eigenvectors present direction while the eigenvalues are the factors of straightening or arrangement. The arrangement of the results for the eigenvectors and eigenvalues in descending order, will make the eigenvector generate the highest eigenvalues and stand to be the most significant in developing the first principal component.

Clarification of PCA based on eigenvector decomposition

The source of the PCA is centered on the algebraic solution of eigenvector decomposition, which is achieved using some orthonormal matrix of $P\ in\ Y = PX$ such that $C_Y \equiv \frac{1}{n} YY^T$ conveying the diagonal matrix. The rows of P are the principal components of X, then C_Y is the unfamiliar variable that can be rewritten as;

$$C_Y = \frac{1}{n} YY^T \tag{61}$$

$$= \frac{1}{n} (PX)(PX)^T \tag{62}$$

$$= \frac{1}{n} PXX^T P^T \tag{63}$$

$$= P\left(\frac{1}{n} (XX^T)\right) P^T \tag{64}$$

Hence, the covariance matrix of X is identified as; $C_Y = PC_XP^T$. Based on the above we recognize that any symmetric matrix A is diagonal by an orthogonal matrix of its eigenvectors.

Multidimensional Scaling (MDS_)

(Xu et al., 2004) certified MDS as a method that uses nominal or ordinal data and, at the same time, it depends on multivariate normality to model nonlinear relationships among variables, hence it can offer a substitute to methods like factor analysis and smallest space analysis. (Series, 2010) confirmed that the MDS is a descriptive technique with an almost incomplete absence of statistical inference notation; it has an attempt to introduce statistical models that will link the estimate and testing procedures, but are all abortive. For instance, if I notation is initiated then dissimilarities become δ_{ij}, and distances are d_{ij} (X). The i and j are the

objects of interest. Then n × p matrix X is the configuration, and the coordinates of the objects in $\mathcal{R}^p$ also , there are data weights w_{ij} which reflects the accuracy of dissimilarity δ_{ij}. (Ghodsi, 2006 and Cox & Cox, 2000) attested MDS as a method of mapping original high m-dimensional data into d-lower dimensional data. It has the capability of addressing the obstacle regarding the construction of configuration n between the n points of k × k matrix D, titled as the distance of affinity matrix. That' $d_{ii} = 0$, and $d_{ij} > 0$, $i \neq j$ MDS finds n data points of $y_1 \cdot \cdot \cdot y_n$ from the distance matrix D into d dimension space, such that if $\hat{d}_{ij}$ is the Euclidean distance between y_i and y_j then $\hat{D}$ will be similar to D, defining the MDS as;

$$\underbrace{min}_{Y} \sum_{i=1}^{k} \sum_{i=1}^{k} \left(d_{ij}^{X} - d_{j}^{Y}\right)^2$$

(65)

Where; $d_{ij}^{X} = \| x_i - x_j \|^2$ and $d_{ij}^{Y} = \| y_i - y_j \|^2$ are the distance matrix D^X that convert kernel matrix of inner product $X^T X$ assigned by;

$$X^T X = -\frac{1}{2} H D^X H$$

(66)

Where; $H = 1 - \frac{1}{t} e e^T$ and e stand for the column vector of $1's$, given by the below equation

$$\underbrace{min}_{Y} \sum_{i=1}^{k} \sum_{i=1}^{k} \left(x_i^T x_j - y_i^T y_j\right)^2$$

(67)

The solution is $Y = \wedge^{\frac{1}{2}} V^T$ where V is the eigenvectors of $X^T X$ and $\wedge$ is the d eigenvalues of $X^T X$.

Stress in MDS

(Meyer et al., 1992) laid a means of assessing the MDS model known to be the squared correlation index (R^2), it functions as a proportion of variance for the input data. Normally when (R^2) greater or equal to ($R^2 \geq 0.60$) is considered acceptable. (Kruskal, 1964) verified stress as the most general determiner factor used in resolving the best fit of a model, and stands to be the most powerful diagnostic tool used for separating between the proximities of input and output distances in any dimensional map, which is defined as;

$$Stress = S = \sqrt{\frac{\sum_{ij}(\delta_{ij} - d_{ij})^2}{\sum_{ij} d_{ij}^2}}$$

(68)

Where; d_{ij} denotes proximities between items i and j, and d_{ij} signifies the spatial distance. (Kruskal, 1964) confirmed that no specific rule was stated for stress tolerance, just the rule described it to be excellent when it is ≤ 0.1 while not tolerable when it is ≥ 0.15, however, the stress function lies within zero and one range, with the best model to present a smaller stress function. (Xu et al., 2004) verified that arriving at non-zero stress is an indication that few or almost entire distances are subject to distortions on the map. Hence, distortions result in the spreading of overall relationships or concentration on some points. Because of this, for any increase in dimensionality, the input data and the model will give closer fitting. Therefore, increasing the number of dimensions tends to decrease the stress or remain in the same position.

Loss function

The loss function is limited using a least-square loss function such as;

$$\delta(X) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(\delta_{ij} - d_{ij}(X))^2$$

(69)

Is said to be minimized over X, then w_{ij} defines the weights, used for choosing the reflect variability and error measurement.

Forms of the loss function (stress and strain)

Stress is the best MDS numerical algorithm used for solving a target function. The stress contains both the observed dissimilarities and the fitted dissimilarities. However, the SMACOF adds a transformation step to the observed dissimilarities to achieve a small stress value. The distances between entities in the lower dimensions are not preserved exactly, MDS aids in finding the best-fitting configuration. This distortion in distances between the lower dimension and higher-order space is called stress. In distance scaling, procedures concerning loss function include residual sum of squares, ordinarily used to explain standardized stress values that are unit-free, mathematical given as;

$$Stress_D\,(\times_1,\ \ldots\ ,\times_N) = \left(\frac{\sum_{i\,j}(D_{ij} - ||\,x_i - x_j\,||)^2}{\sum_{i\,j} D_{ij}^2}\right)^{\frac{1}{2}} \quad (70)$$

The general forms of loss function refer to strain in classical MDS, while it refers to stress in distance MDS. This can be executed using iterative minimization of loss functions including the application of eigendecomposition.

Embedding

(Scheit et al., 2009) testified that reducing the dimensional metric gap as a result of embedding increases the ability and competence of interpretation and also can be used to relate the responses of geographical features.

Geometrical Model

The geometric models are based on disparities and similarities. The disparities are categorized into two Kruskal's disparities and Guttman's rank image disparities. The symbol δ is called Kruskal's disparities known as d-hat, it is centered on a weak monotone transformation, which allows

tied points to be integrated using the primary method while Guttman's rank image disparities are denoted by δ called d-start which centered on a strong monotone transformation that used a secondary method to fix the ties, and it does not permit fitting unbalanced data with equal disproportions.

Local convergence

According to Hessian f represents $D^2 f = D_{11} - D_{12} D_{22}^{-1} D_{21}$ this permits the local convergence rate for augmentation algorithms to give the largest eigenvalue as $M = \mathfrak{T} - D_{11}^{-1} D^2 f$.

Majorization

Minimizing $f(x)$ on X, is the aim of majorization algorithms to solve a problem, despite this, if for instance, we have a second function $g(x, y)$ on $X \otimes X$ such that;

$$f(x) \leq g(x, y) \quad \forall x, y \in X \tag{71}$$

$$f(x) = g(x, x) \quad \forall x, \in X \tag{72}$$

Thus, the above function can be rewritten to give meaning as;

$$f(x) = \underbrace{min}_{y \in X} \; g(x, y) \tag{73}$$

$$x = argmin_{y \in X} \; g(x, y) \tag{74}$$

In a real sense majorization algorithms exhibit a narrow class compared to augmentation algorithms as a result of $X = Y$ which is the condition for the argmin, solves the sub-problems. However, we also have; $D^2 f = D_{11} + D_{12}$, and $M = -D_{11}^{-1} D_{12}$ global convergence following the sandwich inequality as;

$$f\left(x^{(k+1)}\right) \leq g\left(x^{(k+1)}, x^{(k)}\right) \leq g\left(x^{(k)}, x^{(k)}\right) = f\left(x^{(k)}\right)$$

(75)

Where the first inequality described majorization, and secondly the next expressed majorization function g that is minimized in each phase.

Multidimensional Scaling and Types

The classical and least-squares scaling are two main classes of MDS, which the classical scaling is called Torgerson's metric or interval MDS while, the metric model includes, classical metric scaling known as Torgerson scaling (Torgerson-Gower scaling). (S. Zhang, 2014) attested that any type of function can be used for the mapping. It could be either continuous, parametric, or monotone type. Regarding proximities in MDS, there are two types of metrics (MMDS and NMDS). Thus, the metric MDS proximity used $P_{ij} = \hat{d}_{ij}$ as distance, symbolically the cost function is given as;

$$X = \sum_{i \neq j}\left(\hat{d}_{ij} - d_{ij}\right)^2$$

(76)

A single dimension in MDS is referred to as the latent growth class known (the latent growth profile). Consequently, we approximate dimensions based on the model distance then the scale of the group values indicates the growth of the curve.

Types of Multidimensional Scaling

We have various types of MDS that are based on different loss functions. This can be achieved using Metric scaling versus nonmetric scaling and Kruskal-Shepard distance scaling versus classical Torgerson-Gower inner product scaling.

1. Metric versus nonmetric scaling actual values of the dissimilarities are used in matric scaling, while only ranks are used in non-metric scaling (Jiawen, 2012 & Kruskal, 1964). However, calculating the optimal

monotone dissimilarities transformation $f(D_{ij})$ and configuration is the major target of Nonmetric MDS.

2. Kruskal-Shepard distance scaling versus classical Torgerson-Gower inner product scaling assured that dissimilarities can be measured through ($||\ x_i - x_j\ ||$) distance, while the classical scaling will then transform the dissimilarities (D_{ij}) into a form that will logically fit through the inner products of x_i, x_j. Also transforming dissimilarity (D_{ij}) data to inner-product (B_{ij}) data will satisfy the following; $D_{ij}^2 = B_{ii} - 2B_{ij} + B_{jj}$, which will imitate the characteristics of $||\ x_i - x_j\ ||$ and $(x_i\ ,\ x_j)$.

Link between dissimilarity and Euclidean distance

Several MDS techniques can be projected in achieving adequate coordinate representation of dissimilarities into m-dimensional space. Ideally, the dissimilarities are expected to be mapped and match the Euclidean distances. For that reason, if dissimilarity (P_{ij}) between points ‘‘i and j’’mapped to Euclidean distance (d_{ij}) based on some loss minimum information, then X defines the configuration points of $n \times m$, while m describes the dimensional space and the number of points signifies to be n. Normally, there is a link connecting the dissimilarities with the Euclidean distance through the $f: P_{ij} \rightarrow d_{ij}(X)$ function where; d_{ij} infers be distance connecting X coordinate that is yet to be known. Assorted models of MDS believe in mapping dissimilarities distance building on the $f(P_{ij}) = d_{ij}(X)$ equation. However, dissimilarities P_{ij} can be transformed into disparities by using the f function, that is, $\hat{d}_{ij} = f(P_{ij})$. Various functions define disparities from the dissimilarities when the linear function of $\hat{d}_{ij} = bP_{ij}$ is exhibited referring to the MDS ratio, which gives identity function as; $\hat{d}_{ij} = P_{ij}$. Also, the MDS is defined as $\hat{d}_{ij} = g + hP_{ij}$ similar to interval MDS when g and h parameters are chosen to ensure the relation possibly holds. Similarly, a distance d in the below equation is said to be Euclidean distance, for instance, the distance (dissimilarity) matrix $D = (d_{ij})$,

therefore, MDS is required to find, $\times_1, \ldots, \times_n \in \mathbb{R}^\rho$ so that $d_{ij} \approx \| \times_i - \times_j \|_2$ will be closed as possible, normally for some larger $\rho \; or \; p$, there exists a configuration, $\times_1, \ldots, \times_n$, with an accurate distance equivalent to $d_{ij} \equiv \| \times_i - \times_j \|_2$. On the other hand, there are times when there will be distance (dissimilarity), without configuration in any $\rho \; or \; p$, with perfect matching $d_{ij} \neq \| \times_i - \times_j \|_2$, for some i, j, this distance is described as non-Euclidean distance. Contrary the distance, similarity and dissimilarity (proximity) are a function that gives a distance between two objects referred to as metric, which for several pairs of objects in any space can be defined when the following conditions are made;

$$d(x, y) \geq 0 \tag{77}$$

$$d(x, y) = 0 \text{ if and only if } x = y \tag{78}$$

$$d(x, y) = d(y, x) \tag{79}$$

$$d(x, z) \leq d(x, y) + d(y, x) \tag{80}$$

In similar development, Davison 1983 considered the scale values exhibited by the model of

spatial distance as Euclidean distance connecting two consecutive points, when the succeeding

are fulfilled.

$$d(a, b) \geq 0 \tag{81}$$

$$d(a, a) = 0 \tag{82}$$

$$d(a,\ b) = d\ (b\ ,a\ )$$

(83)

In that, d described the point between a and b functioning as the distance. Such that, when a value presents a sign without modification, is an indication that the scale values between the pairing points are equivalent. Contrary, when a value shows a sign reading difference scale values between pairing points, it is a sign showing the study behavior has changed in the study.

Isomap method

(Functionalizable et al., 2016) confirmed Isomap as a technique of data reduction aimed at structuring mapping dimensions from higher to lower level, equally is the extension of MDS handling practical geodesic distances as an alternative to the configured Euclidean distances between each pair of points. Considering the Isomap is the forerunner of geodesic distances used for computing between each set of points and then construct a graph that will show the relationship between point i and j when the geodesic distance (g_{ij}) becomes smaller than the threshold ($\epsilon,\ g_{ij} < \epsilon$) so that the point value will be equivalent to g_{ij}. The expression for the Isomap loss function is defined as;

$$\xi_i = \sum_{i \neq j} \left(g_{ij} - d_{ij}\right)^2$$

(84)

(Functionalizable et al., 2016) Replaced Euclidean distance with an approximation of the geodesic distance based on its non-linear nature. Below defines the geodesic distance (D) involving data sets with distance d(u, v) and neighborhood (k)

$$\overline{D}(a,b) = \underbrace{min}_{p}\ \sum_i d(\ p_i,\ p_{i+1})$$

(85)

Where; p denotes, the length of the sequence $l \geq 2$ alongside $P_1 = a$, $P_l = b$, $P_i \in D \forall i \in (2, \ldots, l-1)$ and (P_i, P_{i+1}) remain the k nearest neighbors and the length l free in minimization.

Double centering

Multi-Dimensional Scaling (MDS) is a concept of affinity K or difference that is used for computing between each pair of training patterns. (Cox & Cox, 2000) acknowledge that the metric MDS converts normalization of distance to dot products by applying double centering given as;

$$\bar{M}_{ij} = -\frac{1}{2}\left(M_{ij} - \frac{1}{n}S_i + \frac{1}{n}S_j \ \frac{1}{n^2}\sum_k S_k\right) \quad (86)$$

The e_{ik} defines the embedding of structure x_i given by $e_{ij} = \sqrt{\lambda_k}\,\upsilon_{ki}$ (87)

Class of MDS

(Fitzgerald & Hubert, 1987) categorized MDs into four different categories such as the Metric MDS model, Non-metric MDS Model, Individual Differences Model, and Maximum Likelihood MDS model.

Metric MDS models

(Fitzgerald & Hubert, 1987) said that the metric MDS model was established by Torgerson in 1952 and is familiar with Torgerson MDS algorithms. This algorithm assembles a multidimensional map of points based on finding a scale of relative distances between the points. It also transforms comparative distances to ratio distances, to prove the dimensionality initiating the ratio distances.

Non-metric MDS Model

NMDS centered on rank order proximities than using real values. It is called ordinal NMDS because of its ordinal nature. However, the coordinate of points with exposed distance that are nearly possible in ranking order is the main target of NMDS. In this case, there must be the

same ranking order from the distance obtained in the final configuration and the original data. Fitzgerald, 1987 confirms that the Non-metric MDS Model was proved by (Jiawen, 2012 and Kruskal, 1964). The non-metric MDS entails the use of rank order and it considers similar distances as the distance calculated from the already itemized MDS model. Thus, the model does not follow assumptions like the metric MDS model. Kruskal and Coombs also contributed to the development of non-metric MDS. Furthermore, different stages of nonmetric MDS were explained by Kruskal, he initiated the least square fit function while running analysis in the MDS and it reduced and regulated residuals among the transformed rank order data obtainable by a model with distance created on multidimensional space. Kruskal introduced two optimization procedures that will control two pairs of data exhibiting equal distance between them. Along the line, attached data is treated using the primary method while separate data is managed using the secondary method. (Press, 1950) attested that the Kruskal algorithm is used for analyzing data matrix that is not complete, and it achieves MDS distance results without Euclidean distance space equivalent to the city-block distance.

Individuals' differences models

(Fitzgerald & Hubert, 1987) described two options involved in analyzing individual differences in MDS models, the first is the utilization of a single matrix data and averaging it, across all individuals while the second is analyzing each matrix of data believing that the control of independent variables will affect individuals. Hence, the model is further categorized into Individual difference scaling (INDSCAL) and alternating least square scaling (ALSCAL). Also, the weirdness index is applied in weights MDS or INDSCAL, which is used in revealing each unusual subject's weights that are similar to the weights represented by that subject being analyzed. However, the weirdness index has a zero to one range which allows the subject's weights to become absolute with the average subject's weights especially when there zero score. In case where the extreme subject's score

is revealed, then the scaling problem is poorly fitting the subject. (Carroll & Chang, 1970) proved that several matrix results are derived and assumed to be different from each other in systematically nonlinear approaches by using weighted MDS (WMDS) results because the WMDS permits the individual differences scaling (INDSCAL) model to interpret reasonable individual differences. (Carroll & Chang, 1970; N, 1964 and Tucker et al., 1970) confirmed INDSCAL as a well-identified model called the Weighted Euclidean Model. (Takane et al., 1977) rendered MDS user-friendly to researchers by introducing the ALSCAL algorithm features into SAS and SPSS. In a similar development, Individual differences analysis is executed using INDSCAL which in SPSS is denoted as PROXSCAL or ALSCAL while it refers to proc MDS in SAS.(Carroll & Chang, 1970) upgrade the MDS modification to use a one-dimensional weighting model identified as INDSCAL. The weighting of the model yields subjective metrics that mathematically define the INDSCAL as;

$$d_{ijk}(\mathrm{X}) - \sqrt{\sum_{a=1}^{m} w_{ak}\ (x_{ia} - x_{ja})^2\ ,\ \ w_{ak}\ > 0} \tag{88}$$

Maximum likelihood models MDS

Fitzgerald proved that the Maximum Likelihood MDS, is a model capable of using tools of inferential instead of descriptive. MULTISCAL MDS is the most prominent algorithm used as maximum likelihood established by Ramsay, 1977. PROSCAL was taken on by MacKay & Zinnes, 1982. Hence, the MDS inferential character is built based on the maximum likelihood notion admitting to prove the dimensionality significance of analyses.

Euclidean distance

(Torgensen, 1952) define Euclidean distance as the sum of the inner products of $\overline{\mathrm{X}}_{\mathrm{i}}.\overline{\mathrm{X}}_{\mathrm{j}}$, that employed simple matrix operation of double centering to generate Euclidean distance towards the inner product matrix. It also entails the discovery of eigenvalue factorization to the inner-product

matrix succeeding with SVD. However, the connection between inner and outer product matrices based on SVD, made the PCA and classical scaling yield similar results, on this note the classical scaling is sometimes signified as principal coordinate analysis.

Classical multidimensional scaling (CMDS)

(Guttman, 1968) described two barriers that can be used in resolving classical multidimensional scaling (CMDS). The first is identifying a function by transforming dissimilarity into distances, and the second is restricting the configuration of the object so that it can be fixed using Euclidean geometry. The major aim of developing cMDS is to set representatives of some sets into minimal possible intervals using Euclidean distance to link the set of data points. Classical scaling is the oldest version of MDS, initiated by (Torgensen, 1952). The MDS was adopted earlier according to (Kruskal, 1964) and developed as the leading approach. Kruskal identified MDS as a jargon that can minimize the loss function entitled stress, used to measure lack of fit for the dissimilarities (D_{ij}) including the distance fittings ($|| x_i - x_j ||$). Generally, residual sum of squares stress is given as;

$$Stress_D (\times_1, \ldots, \times_N) = \left(\sum_{i \neq j=1\ldots N} (D_{ij} - || x_i - x_j ||^2\right)^{\frac{1}{2}} \quad (89)$$

Where; the outer square root represents a greater distribution for smaller values. Then in any dissimilarity matrix ($D = D_{ij}$) of the MDS, Stress is reduced in all the configurations $(\times_1, \ldots, \times_N)^T$, while N × k-dimensional hypervectors stand for unfamiliar parameters, which can minimize the gradient descent of the $Stress_D$ directly like the function on $\mathbb{R}^{Nk}$.

$$\left(D_{ij} - || x_i - x_j ||\right)^2 + \left(D_{ji} - || x_j - x_i ||\right)^2 = 2 . \left(\left(\frac{D_{ij} + D_{ji}}{2}\right) - || x_i - x_j ||\right)^2 + C \quad (90)$$

Where; C symbolizes illustration that cannot depend on, $|| x_i - x_j ||$. Then the dissimilarities are assumed to be symmetrized.

Classical multidimensional scaling CMDS can be executed based on the following steps;

1. To calculate the squared proximity matrix $P = [P^2]$.

2. To make the center of the proximity double, which is, $B = \frac{1}{2} J P J$, by applying the centering operator. $J = 1 - n^{-1}11'$, where; n expresses as the total number of items. It also subtracts the means in every column and row of the fixed matrix containing the elements and then adds it to the grand mean. This is called double centering

3. To extract m eigenvectors $e_1 . . . e_m$, together with the corresponding eigenvalues $\lambda_1 . . . \lambda_m$.

4. To coordinate the n objects in m-dimensional space by deriving from $X = E_m \Lambda_m^{\frac{1}{2}}$, where; E_m denote m eigenvectors while Λ_m denote the m eigenvalues.

cMDS is used for developing n projections to ×, from a high point of dimension towards d- dimensional space and then coordinating the projections to enable the Euclidean distances linking between pairs of d_{ij}, which is viewed like dissimilarities between the high dimensional space. Similarly, the CMDS creates n projections, and × stand for the ‘‘high dimensionality points in a d- dimensional linear space’’ at the same time organizes the projections so that the Euclidean distances between pairs of d_{ij}, resemble the dissimilarities between the high dimensional spots mathematically minimization of the distance follows;

$$\chi^2 = \sum_{i \neq j} (p_{ij} - d_{ij})^2 \quad (91)$$

Where; p_{ij} conveys the distance between X_i and X_j point, then d_{ij} connotes the distance between the projection in X_i, x_i, and the estimate prediction X_j, x_j. The concern of cMDS is to find the location of points in the X matrix based on the acquired eigenvalue decomposition extracted by the double-centered matrix ($B = XX'$). However, the ($B = XX'$) is assembled based on proximity matrix (P) and then multiplied with the centering matrix $J = 1 - n^{-1}11'$. (G. Zhang & Cheng, 2010) confirm that the Non-metric MDS and cMDS can handle computation ‘‘concerning coordinates of the objects in m-dimensional space’’, this allow the proximities to give balance to the inter-point distances. In NMDS, in between the proximities and the distances, there should be a less constrained relationship. Hence, it signifies a function f that increases monotonically, such that $\mathrm{d_{ij}} = \mathrm{f}(\delta_{\mathrm{ij}})$, creates some distances ($\mathrm{d_{ij}}$) from a given proximities δ_{ij}. The function f operates in a form as; $\delta_{\mathrm{ij}} < \delta_{\mathrm{rs}}$ for $\mathrm{f}(\delta_{\mathrm{ij}}) < \mathrm{f}(\delta_{\mathrm{rs}})$. The data inputted to NMDS is a matrix of dissimilarities.

$$D = |\delta_{ij}| \tag{92}$$

where; d_{ij}, denotes the dissimilarity matrix (i and j).

Singular value decomposition (SVD)

(Brock et al., 2008; Candès & Recht, 2009 and Troyanskaya et al., 2001) assured of fixing any missing values contained in the data using SVD. The SVD is an approach based on the following three types of information;

New data matrix approach*:* In every SVD there will be a new data matrix $\mathrm{X_{new}}$, that will sustain the data points to proceed to a new orthonormal root involving a minimum number of components, supported by the distance connecting the illustrations.

Inertia parameters: In every SVD we have, $c_i = \frac{s_i}{\sum_i s_i}$ with $\sum_i c_i = 1$, signifying the SD and relative support to the point of cloud on every single principal component.

Matrix V: The SVD used matrix V to find contributions in each specific principal component, which can be done by observing the first eigenvalues and eigenvectors contained in the inner and outer results.

(Approach & Computation, 1954) described decomposition as a factorization of a matrix into simpler factors, which the main principle of the decompositional method for the matrix calculation is not just to use the matrix algorithm to clarify problems but to create computational outlooks that can solve problems coming from different angles. Correspondingly, the advantage of using SVD is to bring the dimensionality reduction techniques of dual analysis, Burtz matrix, classical scaling multidimensional scaling, and principal component correlation analysis all together, which their results can be accomplished using SVD after accurate data normalization. The SVD considered both the inner-product and outer-product (XX^t and X^tX) of any X data matrix to give similar eigenvalues as, λ_i with $\lambda_i = s_i^2$. If $X = USV^t$ then;

$$XX^t = USV^t(VS^tU^t) = USS^tU^t \quad (93)$$

$$X^tX = (VS^tU^t)USV^t = VS^tSV^t \quad (94)$$

SVD and vectors

(Cuadras & Fortiana, 1995) described that both the vectors have equivalent statistical distribution, especially at zero distance, then possibly it reduced the distance to Euclidean distance once the normalization of the data is accomplished as shown below;

$$\tilde{X}_{ik} = \frac{x_{ik}\sqrt{W}}{(\sqrt{\sum_I x_{Ik}})\,(\sum_I x_{ik})}$$

(95)

Performing either PCA or cMDS on a new data matrix using SVD helps in finding the minimal surrounding space in any data and then preserves the χ^2 distance information. However, the SVD has the added advantage of having a close link with dimensionality reduction like the PCA, cMDS, and principal component correlation analysis (PCCA). Conversely, for any given data matrix X alongside n rows and p columns and x_{ij} with value in row i and column j, connoting as $\bar{X}_i$, been the p components vector corresponding to row i of the matrix and $\bar{X}_j$, serves as n components vector corresponding to column j matrix. Therefore, sets of variables or vectors $\overline{\mathrm{X}}_{\mathrm{i}}$ implies as the set of events while the vector sets connote the set of variables. Based on this, the following notation used for extracting vectors from X was developed by (Schmidt and Stewart, 1992), testifying rectangular matrix as techniques employed to decompose a singular value as

$$\mathrm{X} = \mathrm{USV}^{\mathrm{t}}$$

(96)

In that, U symbolizes the left singular vectors while V indicates singular vectors. The two of them are square orthogonal matrices, while S denotes a rectangular matrix dealing with the singular values ($\mathrm{s_i}$), expected to be positive($\mathrm{s_{ii}} = \mathrm{s_i}$ and $\mathrm{s_{ij}} = 0$). Subsequently, U, S and V will then be reordered to generate $\mathrm{s}_1 > \mathrm{s}_2 > \cdot\ \cdot\ \cdot > \mathrm{s_r}$. Then, r, symbolized the rank of S. Conclusively, before performing SDV, X needs to be centered so that the averages in every column will be zero. Therefore, $\mathrm{rank\ (X)} = \mathrm{rank\ (S)} \leq \min(\mathrm{n}-1, \mathrm{p})$ if X is n. p.

Confirmatory analysis as opposed to exploratory MDS

The fundamental issue of confirmatory MDS is to relate the space matrix observed with the estimated model's space matrix. The contests of

confirmatory MDS can be theoretically described as a penalty function labelled as pseudo-data matrix, used for solving an existing confirmatory MDS program. According to (MacKay & Zinnes, 1982) confirmatory MDS analysis can be done by comparing two profiles and then checking for growth arrangement equivalence among postulated ones, whereas the second kind can be done with a maximum likelihood MDS analysis application like Proscal. The operationalization of the confirmatory analysis was made toward comparing profiles that are above two sets using the model's estimated correlation for the created data. Theoretically, the two sets to be configured are supposed to correspond to show evidence of backing up the hypothesis. Borg et al. (2013) said that confirmatory MDS could be classified as either regular or weak confirmatory MDS. It employed the hypothesized configuration outline as the initial starting value to investigate the latent configuration in the weak confirmatory MDS. In addition, is more challenging to utilize Confirmatory MDS than using exploratory MDS because there is a need to construct graphics as well as translate them into precise computational language required by the user. Instead of allowing the computer to determine its initial formation, tactically, the weak confirmatory MDS handled it as a user-defined external instituting formation constructed using hypothetical notions. As a result, Confirmatory MDS, also known as regular confirmatory MDS, is used to compare certain elements of configurations expected based on past discoveries to observed the data.

Link between PCA and FA, MDS and PCoA

Both the PCA and FA methods were established to describe correlation forms in any set of given variables. As a result, PCA and FA are related in terms of reaching solutions; in return, a high set of correlated variables discovery as well as implying of structures to the factor is required. The PCA uses a similarity matrix in plotting the graph, not the original data. It gives more emphasis on dimensions by working toward minimizing the explained variance. Moreover, the creation of the plots by PCA is based on

the correlation between samples while the creation of plots by MDS is centered on the distance between samples. Nevertheless, the PCA is a simple form of MDS, both give similar results in terms of visualizing relationships, they account for a percentage of variation for each axis and, it does not necessitate direct distance preservation, but rather it minimizes variance in a set of orthogonal projections. Also, the PCA is just a method of exploratory analysis and regarding mapping, the PCA is a particular case of MDS and FA. Despite this, there is a resemblance in MDS, Principal coordinate analysis (PCoA) and the core PCA. However, converting correlations of 2-D plots in PCA is difficult rather it can only convert distances amid the sample to generate 2-D plots. That means the PCA can only yield a 2-D plot by converting distance correlations between samples while MDS of PCoA creates a 2-D plot by converting distance between samples.

CHAPTER III

Methodology

Introduction

This chapter connects all the material and methods of how the research was conducted. This includes research design, study location, sample study participants, research tools (materials), method of data analysis, and interactive session (interview). The methodological aspect applied in this study will determine the level of success in patients on ART.

Research Design

The study is designed to predict the distance covered by the performance of ART-treated patients at Federal Teaching Hospital Gombe State, Nigeria. Exploratory data analysis techniques (EDA) of CA, MCA, PCA and MDS were employed to achieve this success. Application of this technique will help to predict two distances (dissimilarity and similarity) to understand how well the current improved ART drug is prolonging a patient's life while on HIV treatment.

Location of the study

Federal Teaching Hospital located at Gombe State is the area of interest for this study. The unit of ART in the hospital received HIV patients from the entire North-Eastern Geopolitical Zone. The zone comprises six (6) states

that involve Gombe, Bauchi, Yobe, Borno, Adamawa and Taraba. However, apart from the patients visiting the North East Zone, a few patients also visit the hospital from various states across the country. The ART unit of the hospital was chosen for this study because is one of the best hospitals in the country that stores patient information records in a database. Nigeria is a country in Africa made up of 36 states including Abuja as the Federal Capital. The country is carved up based on six (6) various zones comprising the North-Central, North-West, Northeast, South-South, South-East, and South-West in that order. Below is the map of Nigeria capturing the states and zones across Nigeria (see figures 8 and 9).

Figure 8: Map of North-eastern Nigeria indicating the affected zone

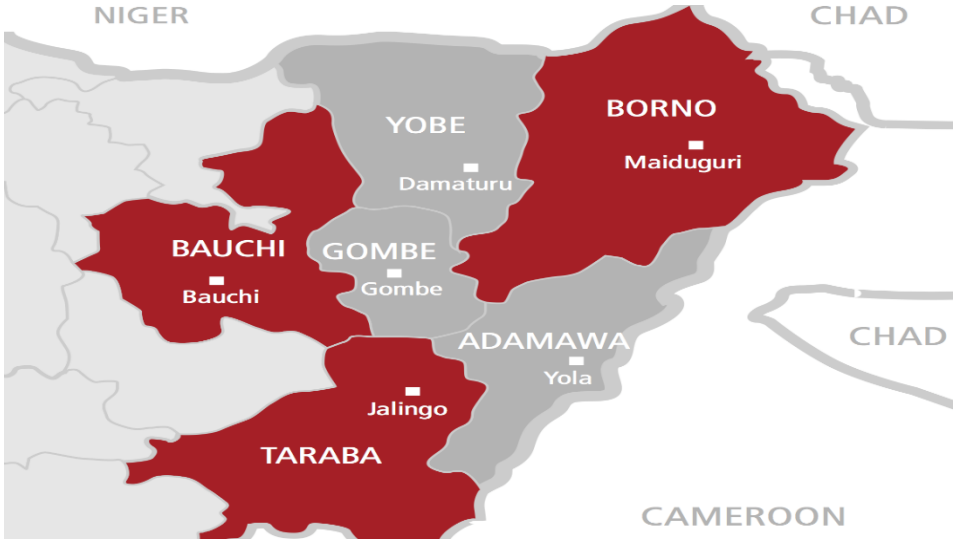
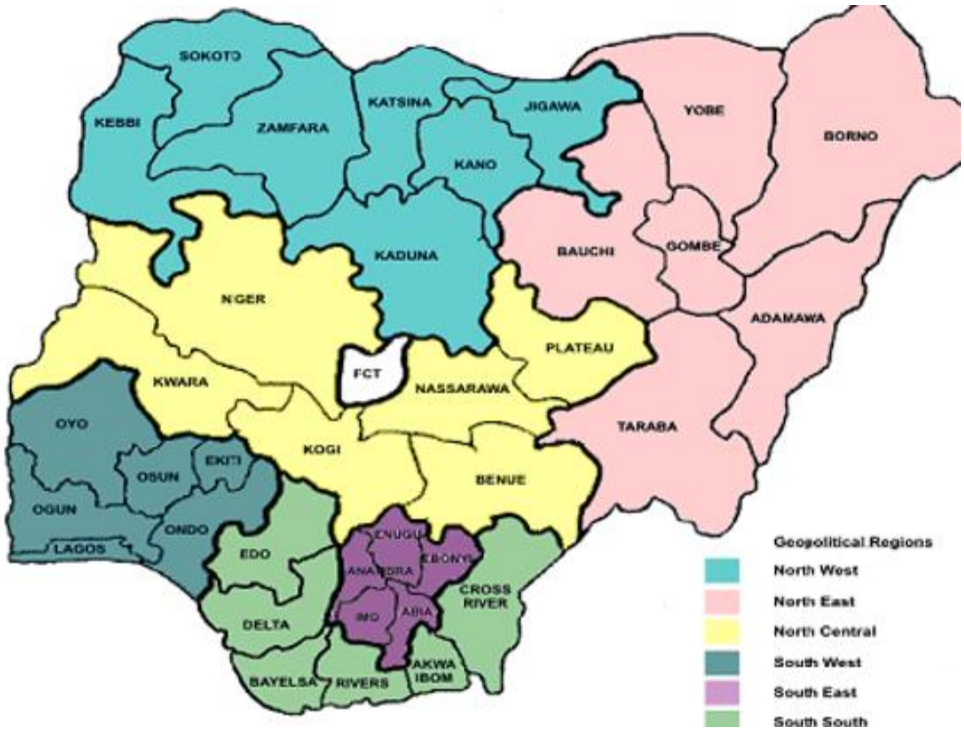


Figure 9: Map of Nigeria based on geopolitical zones



Research tools (materials) for the study

A retrospective cohort study of HIV-treated patients placed on ART triggered the formation of this research finding. The data was extracted from the ART unit database at Federal Teaching Hospital Gombe state, Nigeria. Dimensionality techniques of EDA were applied as a tool to predict similarity distance using CA, MCA, and PCA while proxscal MDS predicted both dissimilarity and similarity distance.

Participants for the study

The study population used only 1500 HIV patients on ART at Gombe State Federal Teaching Hospital. North Eastern Nigeria, is one of the six (6) geopolitical zones, which the study covers.

Method of data analysis

The distance covered by the treated patient's performances was analyzed using IBM SPSS Statistics version 25. In addition, statistical software like Microsoft Word 16, and Excel 16 complimented the outcome of this study findings.

Consultation during clinic session

During the clinic session, the patients, nurses, and medical doctors participated in an interactive session using convenient sampling. The information and data generated during the interactive session showed drugs have effects on individuals visiting the hospital concerning their ART medication. Based on this, about 85% of the treated patients appear to be healthier and better without any signs showing they are HIV positive.

CHAPTER IV

Findings and Discussion

Introduction

This section covered the exploratory data analysis techniques of MCA, PCA, MDS and CA results, employed to compare the success of ART for the ongoing patient receiving HIV treatment in Federal Teaching Hospital Gombe State. The obtainable results were presented in tables and plots.

Exploratory method of MCA result and discussion

Table 1. MCA variance accounted for summary performance

Dimension	Cronbach's Alpha	Variance accounted for (Inertia)	Eigenvalue
Dim 1	0.638	0.408	2.042
Dim 2	0.581	0.374	1.869
Total		0.782	3.911

The results in table 1 describe the predicted performance of the Burt matrix. The Cronbach's Alpha gave (0.638 and 0.581) in dimensions I and II. The inertia variance for the MCA performance gave a cumulative total of 0.782 in both dimensions, with dimension I (0.408) accounting for a high variance and dimension II (0.374) recording a low variance. Dimension I and II gave eigenvalues (2.042 and 1.869) to account for a total eigenvalue of 3.911.

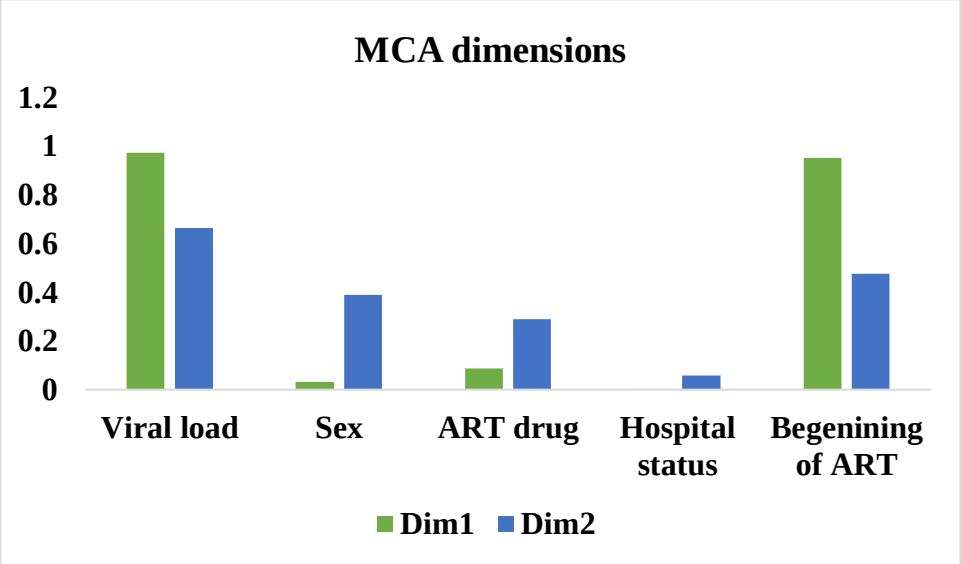
Table 2. Discrimination measures for MCA

	Dimension I	Dimension II
Viral load	0.972	0.661
Gender	0.031	0.388
ART drugs	0.088	0.288
Hospital status	0.001	0.057
Beginning of ART	0.950	0.476

Each patient variable contributing to the follow-up was positioned according to its forms in the coordinate space as seen in table 2. Hence the viral load (0.972) showed a high distance in dimension I, followed by the beginning of ART (0.950), ART drugs (0.088), gender (0.031), and hospital

status (0.001) in that order. However, viral load (0.661) showed a high distance in dimension II, followed by the beginning of ART drug (0.476), gender (0.388), ART drug (0.288), and hospital status (0.057) respectively. Figure 10 is a bar graph representing the dimensions for the coordinate space.

Figure 10 Bar graph describing the MCA coordinate space



Exploratory method of PCA result

Measuring the strength of association between variables *a and b* is the clear target of the KMO. Also, it is employed to determine whether the sampling adequacy for the sample responses will be accepted or not. The KMO result in this study, measured 0.503 of the data, which is above 0.5, indicating the result of the factor analysis has satisfactorily allowed further analysis to proceed. Kaiser (1974) endorses 0.5 as the minimum accepted value used by the KMO, such that a good value stands between the ranges of 0.7- 0.8 while an excellent value stands from 0.9 and above. Bartlett's Test of Sphericity accepted this study's results, implying that a correlation matrix is substantially different from an identity matrix. (Russell, 2002 and Zwick & Velicer, 1986) attested that

the PCA uses KMO and Bertler test to investigate the sampling adequacy of the model. It specifies for number of factors advocating for more factors than necessary. However, it can extract all the factors that revealed eigenvalue (> 1) and serves as the Kaiser criterion used in establishing the required number of factors.

Table 3 PCA total variance summary results

Component	Total	% of variance	Cumulative %
1	1.179	23.578	23.578
2	1.023	20.457	44.035
3	1.003	20.066	64.100
4	0.965	19.295	83.396
5	0.830	16.604	100.00

The initial eigenvalue for the total variance explained in table 3 gave a cumulative percentage of 64.100 with a value greater than 1 to retain three components from five components. Figure 11 describes the cumulative percentage performance and scree plot for the retained and un-retained components.

Figure 11 Initial eigenvalues

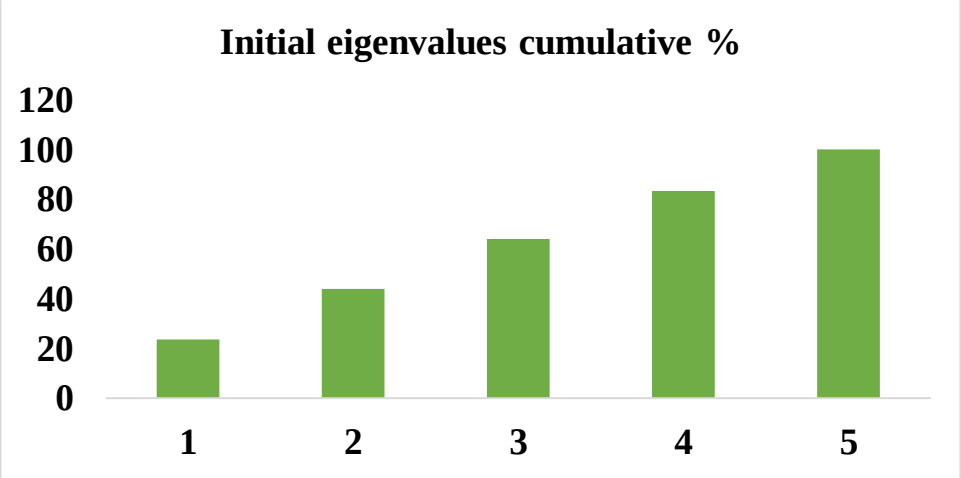


Figure 12 Scree plot for PCA

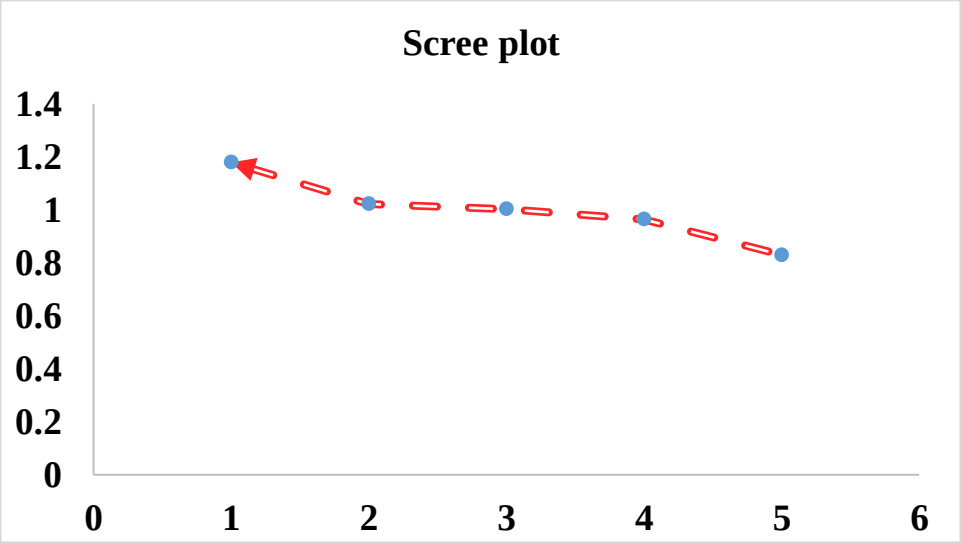
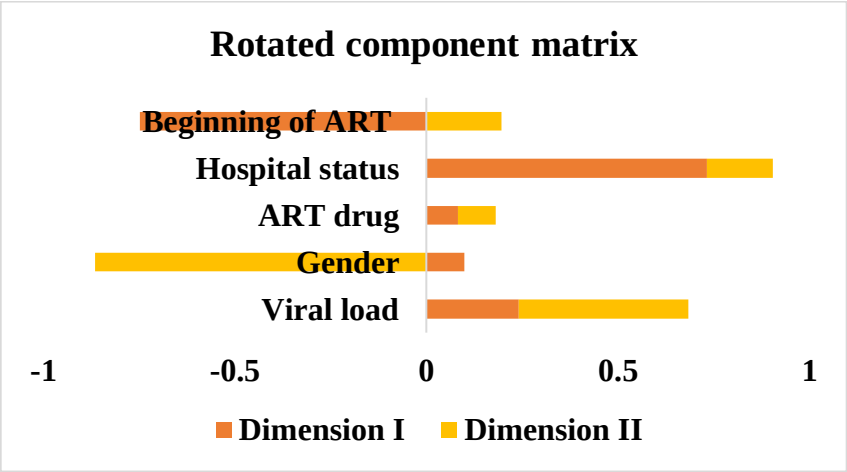


Table 4 PCA rotated component matrix

Rotated component matrix		
Variables	Component I	Component II
Viral load	0.241	0.443
Gender	0.100	-0.865
ART drugs	0.082	0.099
Hospital status	0.733	0.172
Beginning of ART	-0.748	0.197

In the formation of the new component in table 4, the rotated component matrix displayed a higher negative loading at the beginning of ART (-0.748) year, followed by a positive loading in hospital status (0.733), while in component II only gender (-0.865) had a higher negative loading. In this regard, gender contributed to higher loading, followed by the beginning of ART and hospital status as seen in figure 13.

Figure 13 Stacked bar chart for rotated component matrix



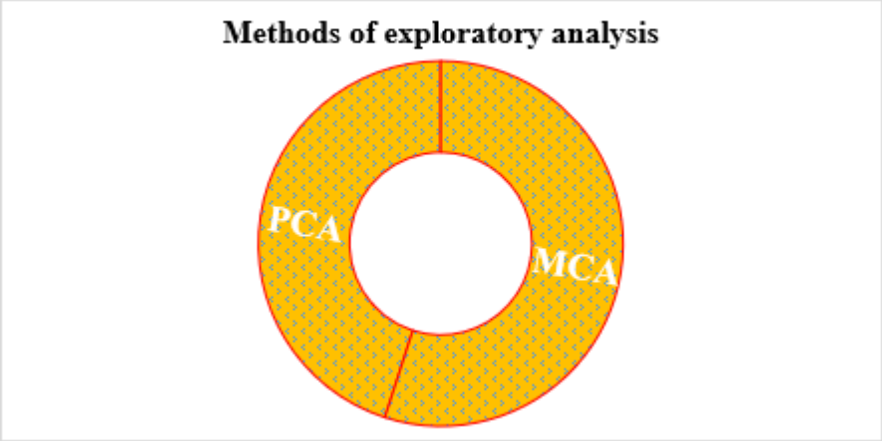
Exploratory methods performance (MCA and PCA)

Table 5 Similarity distance performance of EDA (MCA and PCA)

Models	R-square
MCA	0.782
PCA	0.641

The results in table 5 showed that the performance of exploratory methods in PCA is very low compared to MCA. Figure 14 is the pie chart describing the performance.

Figure 14. Pie chart representing the methods of exploratory analysis performance (MCA and PCA)



Class algorithm of exploratory techniques results

The MDS algorithm techniques in this section employed two respective rates (proxscal and alscal) to measure the dissimilarity and similarity distance performances, just that the alscal measured only dissimilarity. The proxscal stress was measured using Torgerson stress while the Alscal stress was measured using Kruskal stress. The proxscal is the most suitable tool in MDS for computation than the alscal because it fulfills the aim of the MDS in terms of measuring individual similarity and dissimilarity (metric and non-metric).

Table 6 Proxscal dissimilarity and similarity evaluation index using Torgerson stress

Stress and fit measures	Dissimilarity	Similarity
Stress Normalized raw	0.48114	0.04866
Stress I	0.69364	0.22059
Stress II	1.08275	0.86405
S-Stress	0.64167	0.16537

Torgerson stress in table 6 was used in generating the normalized raw stress, which presented a dissimilarity and similarity of 0.48114 and 0.04866 while the S-stress recorded a dissimilarity of 0.64167 and similarity of 0.16537. The proxscal stresses recorded a higher distance in dissimilarity compared to proxscal similarity. The closer the original distance is to the predicted distance, is the indication that the value presented a lower stress, this showed that the proximity (predicted distance and the estimated distance) are similar otherwise are distance. Figure 15 and 16 describe the Torgerson stresses for the dissimilarity and similarity.

Figure 15 Proxscal dissimilarity histogram plot using Torgerson stresses

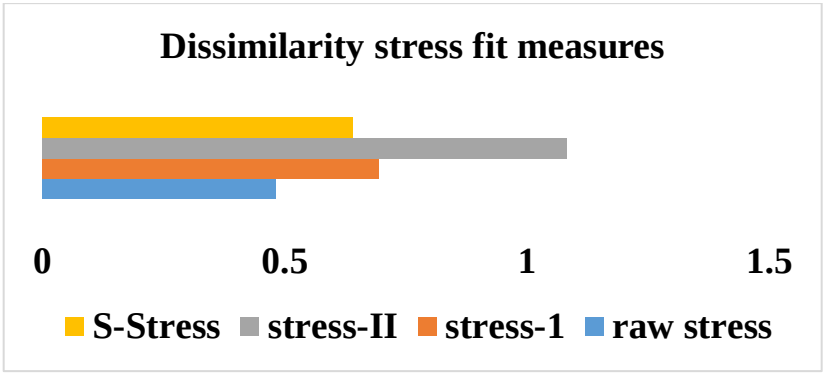


Figure 16 Proxscal similarity histogram plot using Torgerson stresses

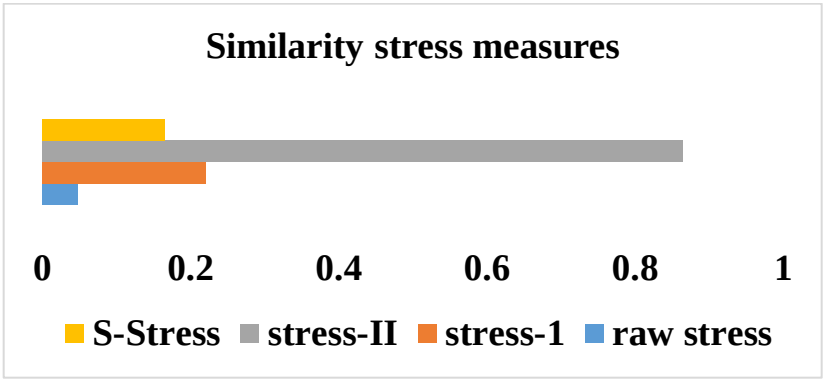


Table 7 Proxscal coordinate common space for dissimilarity and similarity

	Dissimilarity		Similarity	
	Dimension I	Dimension II	Dimension I	Dimension II
Viral load	0.939	0.003	-0.013	0.209
Gender	-0.305	-0.005	-0.548	0.448
ART drugs	-0.212	0.195	0.391	-0.546
Hospital status	-0.233	-0.005	0.600	0.385
Beginning of ART	-0.190	-0.188	-0.455	-0.496

The proxscal coordinate space was presented to describe the direct bearing of each patient's variable on follow-up. The coordinate revealed how the variables contribute to the follow-up. Variables that are closer imply similarity and a strong relationship while those that are not closer indicate distance and no strong relationship. Based on table 7, the dissimilarity

results in dimension I revealed a high positive distance in viral load (0.939) while the remaining variables in the dimension present a negative distance. Dimension II presented a positive distance in ART drug (0.195) and viral load (0.003) while the remaining variables presented a negative distance. Subsequently, the similarity coordinates in dimension I revealed that hospital status (0.600) presented a high positive distance, followed by ART drug (0.391), while the remaining presented a negative distance. And in dimension II, a high positive distance was revealed in gender (0.448), followed by hospital status (0.385), and viral load (0.209) while a negative high and low distance was presented in ART drugs (-0.546) and Beginning of ART (-0.496) respectively. Figure 17, captured the dimensions for the Proxscal coordinate space (dissimilarity and similarity).

Figure 17 Bar-chart representing the coordinate dimensions for proxscal dissimilarity

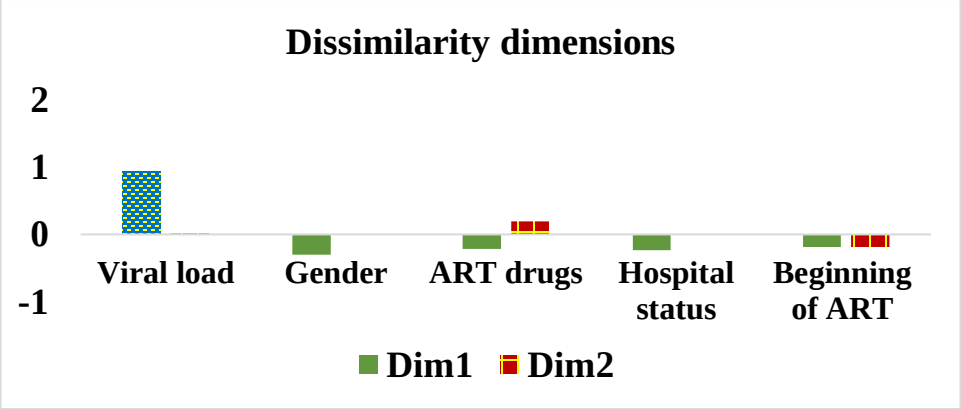


Figure 18 Bar-chart representing the coordinate dimensions for proxscal similarity

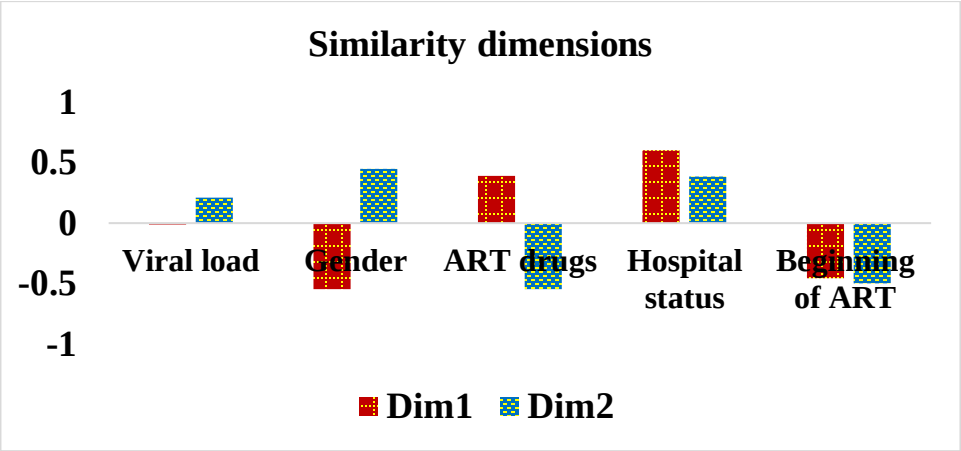


Table 8 Alscal stimulus coordinates space

Stimulus name	Dimension 1	Dimension 2
Viral load	2.7261	0.1071
ART drug	-0.7224	-0.7306
Hospital status	-0.6070	0.1166
ART starting	-0.7376	0.1753
Hospital status	-0.6591	0.3316

The Alscal stimulus coordinate space in table 8 shows dimension I presented a high positive distance in viral load (2.7261) while the remaining presented a negative distance. Dimension II showed that only the ART drug (-0.7306) presented with a high negative distance while the remaining variables showed a positive distance (see, figure 19).

Figure 19 Alscal stimulus coordinates space

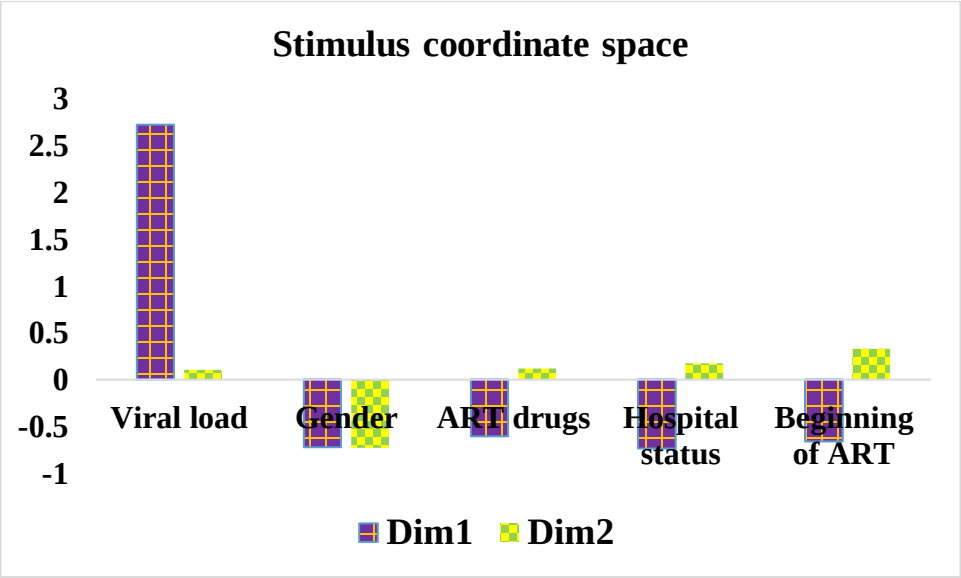


Table 9 Class algorithm evaluation index for stress and R-square

Models	Stress	R-square
MDS proxscal dissimilarity	0.64167	0.51886 (DAF)
MDS proxscal similarity	0.16537	0.95134 (DAF)
MDS alscal dissimilarity	0.71743	0.45727 (RQS)

The distance (stress) and coefficient of determination (R-square) described the MDS exploratory class algorithm performances in table 9, in which the proxscal similarity performed better in terms of DAF (0.95134), followed by proxscal dissimilarity DAF (0.51886) and alscal dissimilarity RQS (0.45727). On the other hand, higher stress was revealed in alscal dissimilarity (0.71743) followed by proxscal dissimilarity (0.64167) and proxscal dissimilarity (0.16537) respectively. The algorithm with lower stress indicates that the variables have a strong association and are closer, but the variables that are farther apart do not have a strong relationship (see, figure 20 and 21).

Figure 20 Histogram plot presenting exploratory class algorithm results

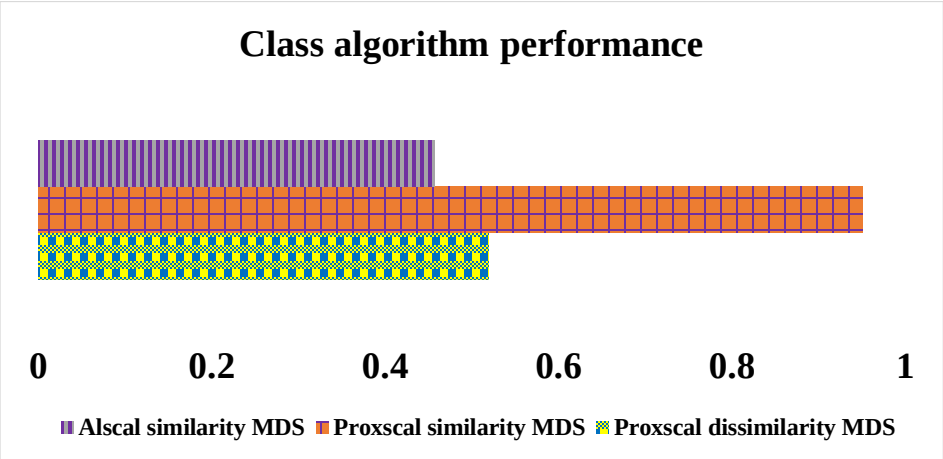


Figure 21 Radar plot describing exploratory class algorithm results

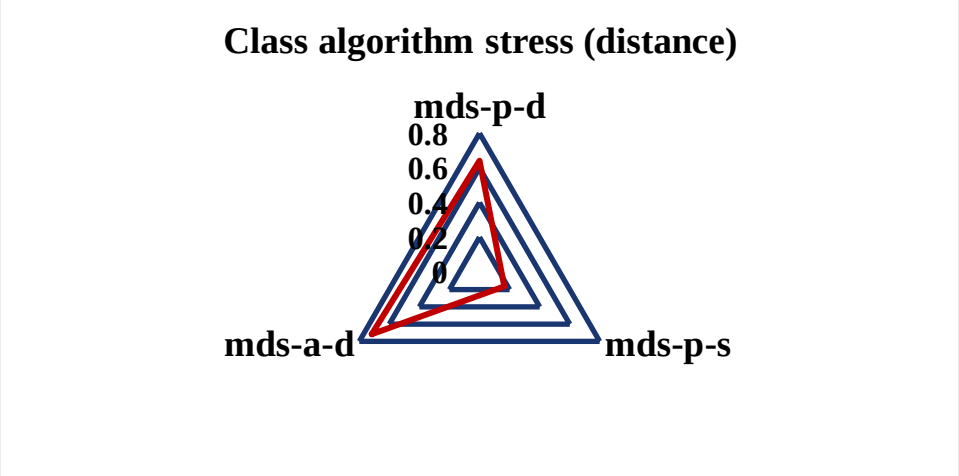


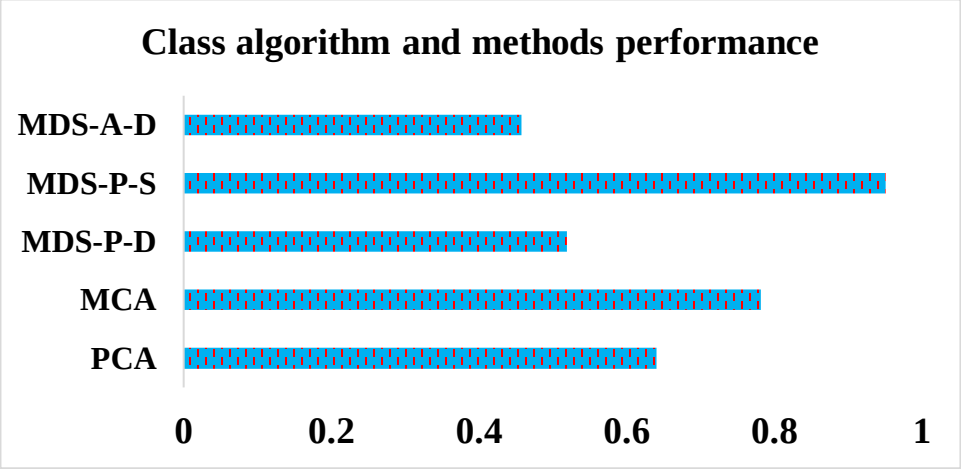
Table 10 General evaluation of EDA performances (methods and class algorithm)

Models	R-square
MCA (inertia)	0.782
PCA (inertia eigenvalues)	0.641
MDS proxscal dissimilarity (DAF)	0.51886
MDS proxscal similarity (DAF)	0.95134
MDS alscal similarity (RQS)	0.45727

In this aspect, we bring together only the coefficient of determination for all the EDA methods and class algorithms to explain their dissimilarity and similarity performances. Based on table 10, the results were presented in

the following order; MCA (0.782), PCA (0.641), Proxscal dissimilarity (0.51886), proxscal similarity (0.95134), and alscal dissimilarity (0.45727) respectively. The performances indicate that the MCA was more effective in presenting a higher value to become the best in predicting similarity than PCA. The class algorithm of MDS showed proxscal similarity was more effective in predicting dissimilarity, considering the high coefficient of determination DAF (0.95134), followed by proxscal dissimilarity DAF (0.51886) and alscal dissimilarity RQS (0.45727). Figure 22 is a bar graph presenting the overall performance.

Figure 22 Graphical presentation performance of class and methods of exploratory data analysis



Building performance and plots of CA on MCA, PCA and MDS

In this section, a separate dual analysis using only two variables' drugs and hospital status was used to confirm if the EDA techniques of MCA, PCA, and MDS can be built on CA techniques. Surely, indeed CA is validated as the major source of MCA, PCA and MDS. Also, in all the desired findings the PCA graph only produced a scree plot rather than the customary 2-D plot seen in the CA, MCA and MDS (Figure 24). The study proved that with only two variables you can use CA to predict similarity distance, while with more than two variables MCA, PCA, FA and MDS can be applied. Herve Abdi and Lynne J. Williams 2010 described PCA and MCA as the

upgrade of CA, in which the CA used qualitative variables while MCA used multiple or heterogeneous sets of variables.

Table 11 CA proportion of inertia summary results

Dimension	S. Value	Inertia	Chi-Square	Sig	Proportion of Inertia	
					Accounted for	Cum.
1	0.124	0.015			0.836	0.836
2	0.055	0.003			0.164	1.000
Total		0.018	18.431	0.782	1.000	1.000

Based on the CA cumulative percentage of 1.000 revealed by the proportion of inertia in table 11, dimension I presented a high distance of 0.836 while dimension II recorded a lower distance of 0.164.

Table 12 MCA variance accounted for summary results

	Variance accounted for		
Dimension	Total (Eigenvalues)	Inertia	% of variation
1	1.124	0.562	56.207
2	1.055	0.527	52.731
Total	2.179	1.089	

Table 12 presented eigenvalues greater than 1 in both dimensions to account for 1.089 variance with dimension I recording a slightest difference of 0.035 than dimension II to become the higher dimension in the MCA table.

Table 13 PCA total variance summary results

Component	Total	% of variance	Cumulative %
1	1.017	50.872	50.872
2	0.983	49.128	100.000

Only dimension I extracted a component value greater than 1 in table 13. Then, the high initial eigenvalues were revealed in dimension I (50.872) and the lower initial eigenvalue was recorded in dimension II (49.129)

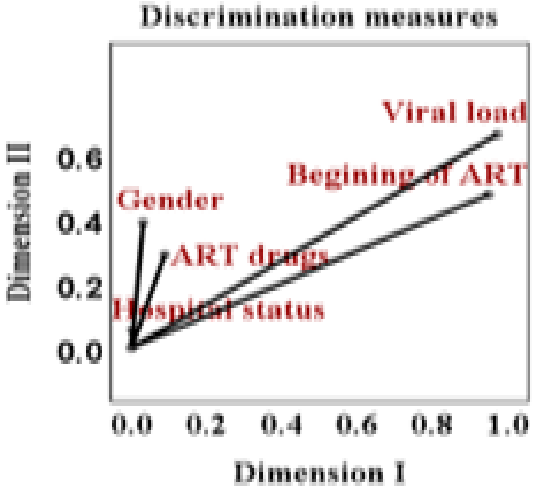
bringing the total cumulative to 1.000. Based on the extraction of one component the PCA therefore cannot produce a 2-D plot but rather to give a scree plot to indicate the trend of the two components.

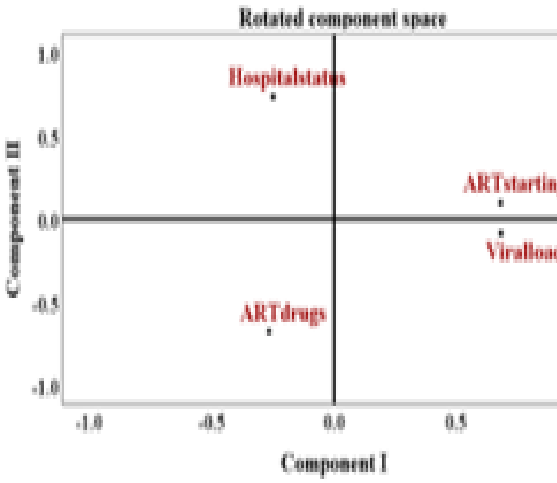
Table 14 MDS summary measures

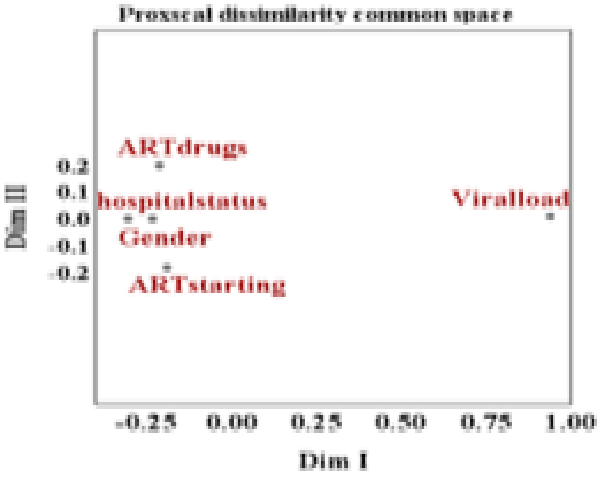
Stress-Badness of fit)	0.0669495
Coefficient of determination-Goodness of fit	0.9465347
Correlation-Kendall's Tau-b	0.7622878

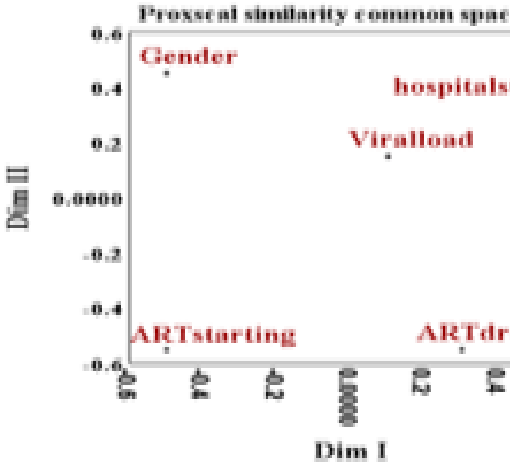
In table 14, the MDS goodness of fit accounted for about 0.9465347 percentage of variation, while the badness of fit for the Normalized stress gave 0.669495 distance and the Kendalls Tau-b Correlation coefficient showed 0.7622878.

Figure 23 Comparison of the EDA techniques visual plots for the methods and class algorithm (MCA, PCA and MDS).


Discrimination measures
Viral load
Begining of ART
Gender
ART drugs
Hospital status
Dimension II
Dimension I
0.0
0.2
0.4
0.6
0.8
1.0


Rotated component space
Hospitalstatus
ARTstartin
Viralloa
ARTdrugs
Component II
Component I
1.0
0.5
0.0
-0.5
-1.0


Proxscal dissimilarity common space
ARTdrugs
hospitalstatus
Viralload
Gender
ARTstarting
Dim II
Dim I
0.2
0.1
0.0
-0.1
-0.2
-0.25
0.00
0.25
0.50
0.75
1.00


Proxscal similarity common space
Gender
hospitals
Viralload
ARTstarting
ARTdr
Dim II
Dim I
0.6
0.4
0.2
0.0000
-0.2
-0.4
-0.6

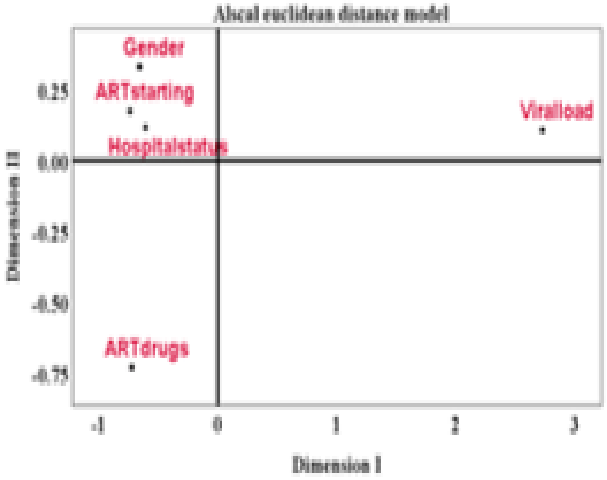
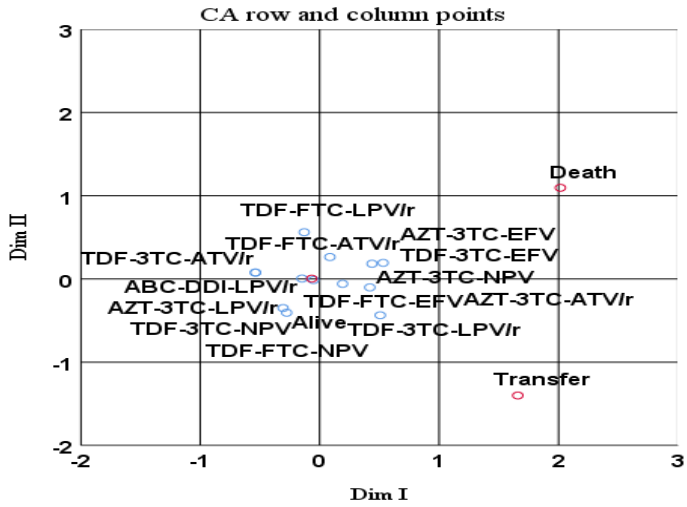
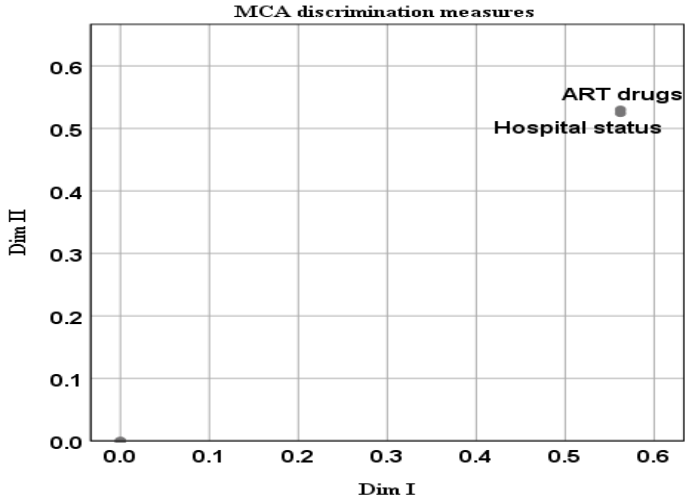
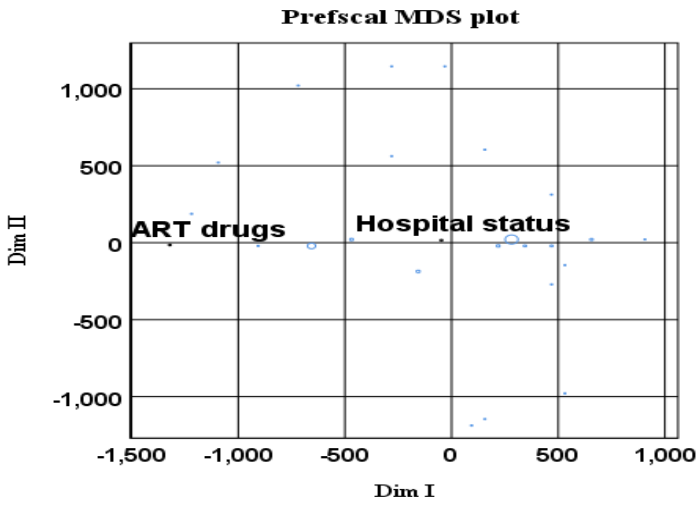
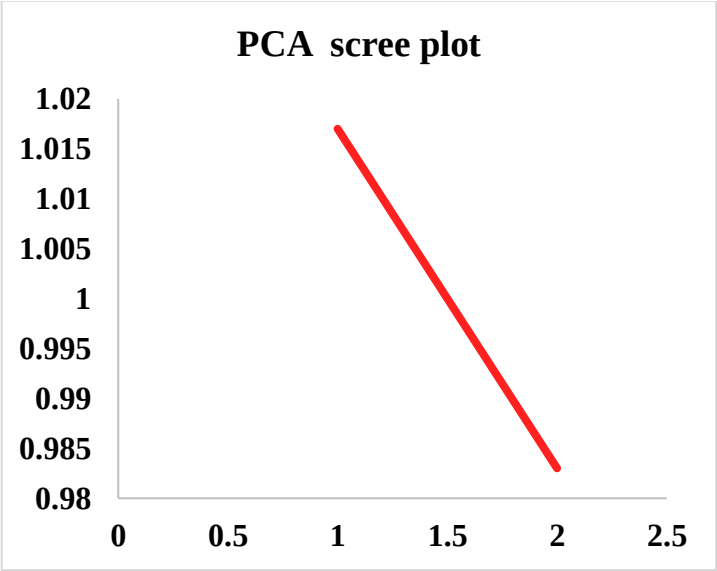

Abcal euclidean distance model
Gender
ARTstarting
Hospitalstatus
Viralload
ARTdrugs
Dimension II
Dimension I
0.25
0.00
-0.25
-0.50
-0.75
-1
0
1
2
3

Figure 24 Comparison of CA, MCA, MDS and PCA dual plot






Prefscal MDS plot
1,000
500
0
-500
-1,000
Dim II
ART drugs
Hospital status
-1,500
-1,000
-500
0
500
1,000
Dim I


PCA scree plot
1.02
1.015
1.01
1.005
1
0.995
0.99
0.985
0.98
0
0.5
1
1.5
2
2.5

CHAPTER V

Discussion of Results

Discussion of the methods and class algorithm of the EDA

The results for the exploratory method of MCA are presented in table 1, and 2. The total eigenvalue in both dimension I and II accounted for 0.782 variances while Cronbach's Alpha gave 0.638 and 0.581 in both dimensions. The contribution of patient toward follow-up in dimension I showed a higher distance in viral load (0.972), followed by starting of the ART year (0.950), ART drugs (0.088), gender (0.031), and the hospital status (0.001) while in dimension II viral load (0.661) showed a high distance, followed by the beginning of ART drug (0.476), gender (0.388), ART drug (0.288), and hospital status (0.057) respectively. The exploratory method of PCA in table 3 retained three (1.179, 1.023 and 1.003) components from five components with eigenvalues greater than 1 based on the 64.100 cumulative percentage of the total variance presented. However, the new component in table 4, displayed a higher negative loading at the beginning of the ART year (-0.748) by the rotated component matrix, followed by a positive loading in hospital

status (0.733). Also, gender contributed to a higher loading, followed by the beginning of the ART year and hospital status in dimension I. Based on these results, the performance of exploratory methods in PCA (0.641) is very low compared to MCA (0.782). Subsequently, the classic algorithm of EDA techniques results in table 6 confirmed that proxscal dissimilarity (0.64167) gave higher stress than proxscal similarity (0.16537). Their coordinate space (dimensions) indicates how the variables contribute to the follow-up, showing dissimilarity results in dimension I gave a high positive distance in viral load (0.939), whereas the remaining variables in dimension II presented a negative distance. A positive distance in ART drug (0.195) and viral load (0.003) was presented in dimension II while the remaining variables present a negative distance. The hospital status in dimension I from the similarity coordinate revealed that hospital status (0.600) presented a high positive distance, followed by ART drug (0.391), while the remaining gave a negative distance. In dimension II, a high positive distance was revealed in gender (0.448), followed by hospital status (0.385), and viral load (0.209) while a negative high and low distance was presented in ART drugs (-0.546) and Beginning of ART (-0.496) respectively. Dimension I and II in table 8 showed that the alscal stimulus coordinate space presented a high positive distance in viral load (2.7261) then, the remaining presented a negative distance. Patients on ART drug in dimension II recorded a high negative distance of (-0.7306) while the remaining variables showed a positive distance.

Regarding the distance (stress), the MDS proxscal similarity (0.16537) gave lower stress followed by proxscal dissimilarity (0.64167), and alscal dissimilarity (0.71743) respectively. This confirmed that proxscal similarity stress is the most suitable tool for computing dissimilarity distance in MDS than the proxscal dissimilarity and alscal dissimilarity. Because it fulfilled the aim of the MDS in terms of measuring individual similarity and dissimilarity (metric and non-metric). Considering the distance (stress) and coefficient of determination (R-square) revealed by

the performance of the EDA techniques, the exploratory class algorithm showed proxscal similarity in table 9 and performed better in terms of DAF with less stress (0.95134 and 0.16537) compared to the remaining algorithm. However, the coefficient of determination results in table 10 combined both the methods and class of exploratory data analysis, indicating that the MCA (0.782) reveals higher performance than the PCA (0.641). In addition, the class algorithm of MDS showed proxscal similarity gave a high coefficient of determination (DAF = 0.95134), followed by proxscal dissimilarity (DAF=0.51886) and alscal dissimilarity (RQS=0.45727). According to the techniques of exploratory methods performance, MCA is the best at predicting similarity, while proxscal similarity MDS is the best at predicting dissimilarity on the side of the class algorithm.

Performances and plots confirming the reality of MCA, PCA and MDS building from CA

To validate the origin of MCA, PCA and MDS building from CA two variables ART drug and hospital status were utilized for this purpose. In table 11, the CA percentage of inertia accounted for about 0.836 variation in the data. The inertia for the MCA variance accounted for 0.527 percent in table 12. The PCA results in table 13 cannot produce component plots due to the extraction of only one component but the Initial Eigenvalues for the PCA gave 0.50872 percentage of variation. Despite, the communalities result for the PCA given above 0.5, the p-value was insignificant (0.582) and the correlation was very weak (0.291). However, the KMO and Bartletts test of Sphericity gave (0.500 and 0.304). Finally, in table 14 the MDS goodness of fit accounted for about 0.9465347 percentage of variation while the Badness of fit for the Normalized stress gave 0.669495 distance and the coefficient of the Kendalls Tau-b Correlation was 0.7622878. Regarding the similarity scale, CA proved to be the best, followed by MCA, and then PCA while in terms of dissimilarity, the MDS outperformed all the EDA techniques.

CHAPTER VI

Conclusion and Recommendation

Conclusion

In this study, a comparative analysis of classic algorithms and methods of exploratory data analysis was employed to predict the distance of performances around the patients receiving ART medication at Federal Teaching Hospital Gombe State, Nigeria. Moreover, in a separate analysis, two variables (ART drug and hospital status) were employed to authenticate the origin of MCA, PCA, and MDS. Conclusively, the exploratory methods of CA, MCA and PCA predicted the similarity performances, while the MDS class algorithm for proxscal and alscal predicted the dissimilarity and similarity performance for the patients on follow-up. To measure this reality, different patient factors such as age, gender, marital status, hospital status, ART drugs, and so on were used. The efficiency performance for the models was rated based on several evaluation indexes (i.e. DAF, RQS, the proportion of inertia, inertia, and initial eigenvalue). The distance for the proxscal was determined using Torgerson stress and the distance for the alscal was measured using Kruskal's stress.

The results for the exploratory methods analysis performance in MCA accounted for a high percentage of 0.782 variances compared to PCA 0.641, but in the separate dual analysis, CA accounted for 0.836 higher variances in dimension I compared to MCA and PCA. Nevertheless, in the overall cumulative, MCA superseded again. However, the MDS exploratory class algorithm performance for the proxscal similarity revealed a high coefficient of 0.95134 variations with low stress of 0.16537 to supersede the separate MDS dual analysis that gave 0.9465347 goodness of fit with lower normalized stress of 0.669495. Based on the overall model performances for the ongoing patients receiving ART medication, it has

proven that MCA was the best in predicting shorter distances compared to PCA but in the separate analysis, CA superseded MCA, and PCA. While the MDS proxscal similarity stands the best to predict dissimilarity measures.

Recommendation

Therefore, it is recommended that ART plays a vital role in HIV-treated patients receiving medication in Federal Teaching Hospital Gombe State, Nigeria. Viral load contributed to this development. Furthermore, different epidemiological data apart from HIV may employ exploratory data analysis for further research. Also, in the future, different epidemiological data may employ exploratory methods of analysis such as factor analysis and canonical correlations to compare with the MDS class algorithm.

References

004635259288F49Ffd000000.Pdf. (n.d.).

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459. https://doi.org/10.1002/wics.101

Abram, M. E., Ferris, A. L., Das, K., Quinones, O., Shao, W., Tuske, S., Alvord, W. G., Arnold, E., & Hughes, S. H. (2014). Mutations in HIV-1 Reverse Transcriptase Affect the Errors Made in a Single Cycle of Viral Replication. *Journal of Virology*, *88*(13), 7589–7601. https://doi.org/10.1128/jvi.00302-14

Alonso-Atienza, F., Rojo-Álvarez, J. L., Rosado-Muñoz, A., Vinagre, J. J., García-Alberola, A., & Camps-Valls, G. (2012). Feature selection using support vector machines and bootstrap methods for ventricular fibrillation detection. *Expert Systems with Applications*, *39*(2), 1956–1967. https://doi.org/10.1016/j.eswa.2011.08.051

Alonso, A., & De Irala, J. (2004). Strategies in HIV prevention: The A-B-C approach [1] (multiple letters). *Lancet*, *364*(9439), 1033. https://doi.org/10.1016/S0140-6736(04)17050-5

Anderson, P. L., Kakuda, T. N., & Lichtenstein, K. A. (2004). *The Cellular Pharmacology of Nucleoside- and Nucleotide-Analogue Reverse-Transcriptase Inhibitors and Its Relationship to Clinical Toxicities*. *80262*.

Approach, D., & Computation, M. (1954). the Approach To Mktrix. *Computing in Science & Engineering*, 50–59.

Audigier, V., Husson, F., & Josse, J. (2017). MIMCA: multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and Computing*, *27*(2), 501–518. https://doi.org/10.1007/s11222-016-9635-4

Ayele, G., Tessema, B., Amsalu, A., Ferede, G., & Yismaw, G. (2018). *Prevalence and associated factors of treatment failure among HIV / AIDS patients on HAART attending University of Gondar Referral Hospital Northwest Ethiopia*. 1–13.

Bangsberg, D. R., Moss, A. R., & Deeks, S. G. (2004). Paradoxes of adherence and drug resistance to HIV antiretroviral therapy. *Journal of Antimicrobial Chemotherapy*, *53*(5), 696–699. https://doi.org/10.1093/jac/dkh162

Bangsberg, D., Tulsky, J. P., Hecht, F. M., & Moss, A. R. (1997). Protease inhibitors in the homeless. *Journal of the American Medical*

Association, *278*(1), 63–65. https://doi.org/10.1001/jama.278.1.63

Beh, E. J. (2004). Simple correspondence analysis: A bibliographic review. *International Statistical Review*, *72*(2), 257–284. https://doi.org/10.1111/j.1751-5823.2004.tb00236.x

Beh, E., & Lombardo, R. (2019). A Geneaology of Correspondence Analysis: Part 2 - The Variants. *Electronic Journal of Applied Statistical Analysis*, *12*(2), 552–603. https://doi.org/10.1285/i20705948v12n2p552

Binka, M., Ooms, M., Steward, M., & Simon, V. (2012). The Activity Spectrum of Vif from Multiple HIV-1 Subtypes against APOBEC3G, APOBEC3F, and APOBEC3H. *Journal of Virology*, *86*(1), 49–59. https://doi.org/10.1128/jvi.06082-11

Blasius, J. (2000). Geometric Data Analysis. *BMS Bulletin of Sociological Methodology/ Bulletin de Methodologie Sociologique*, *68*(1), 54–55. https://doi.org/10.1177/075910630006800123

Bowen, A., Sweeney, E. E., & Fernandes, R. (2020). *Nanoparticle-Based Immunoengineered Approaches for Combating HIV*. *11*(April), 1–9. https://doi.org/10.3389/fimmu.2020.00789

Brock, G. N., Shaffer, J. R., Blakesley, R. E., Lotz, M. J., & Tseng, G. C. (2008). Which missing value imputation method to use in expression profiles: A comparative study and two selection schemes. *BMC Bioinformatics*, *9*(February). https://doi.org/10.1186/1471-2105-9-12

Campbell-Yesufu, O. T., & Gandhi, R. T. (2011). Update on human immunodeficiency virus (HIV)-2 infection. *Clinical Infectious Diseases*, *52*(6), 780–787. https://doi.org/10.1093/cid/ciq248

Candès, E. J., & Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, *9*(6), 717–772. https://doi.org/10.1007/s10208-009-9045-5

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*, *35*(3), 283–319. https://doi.org/10.1007/BF02310791

CDC. (2019). *Guideline for use Antiretroviral Agentes in adults adolescents with HIV*. 1–378.

Chen, L. H., & Chang, S. (1995). An Adaptive Learning Algorithm for Principal Component Analysis. *IEEE Transactions on Neural Networks*, *6*(5), 1255–1263. https://doi.org/10.1109/72.410369

Cox, T. F., & Cox, M. A. A. (2000). A general weighted two-way dissimilarity coefficient. In *Journal of Classification* (Vol. 17, Issue 1, pp. 101–121). https://doi.org/10.1007/s003570000006

Cuadras, C. M., & Fortiana, J. (1995). Metric Scaling Graphical Representation of Categorical Data. *{{}Penn{}} {{}State{}} {{}University{}}*, *October*, 2. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.56.958

Del Giudice, V., Salvo, F., & De Paola, P. (2018). Resampling techniques for real estate appraisals: Testing the bootstrap approach. *Sustainability (Switzerland)*, *10*(9), 1–16. https://doi.org/10.3390/su10093085

Desimmie, B. A., Delviks-Frankenberrry, K. A., Burdick, R. C., Qi, D., Izumi, T., & Pathak, V. K. (2014). Multiple APOBEC3 restriction factors for HIV-1 and one vif to rule them all. *Journal of Molecular Biology*, *426*(6), 1220–1245. https://doi.org/10.1016/j.jmb.2013.10.033

Enzécri, J. P. B. (1979). *J. p. b*.

Esbjörnsson, J., Månsson, F., Kvist, A., da Silva, Z. J., Andersson, S., Fenyö, E. M., Isberg, P. E., Biague, A. J., Lindman, J., Palm, A. A., Rowland-Jones, S. L., Jansson, M., Medstrand, P., Norrgren, H., N'Buna, B., Biague, A. J., Biai, A., Camara, C., Karlson, S., … Wilhelmson, S. (2019). Long-term follow-up of HIV-2-related AIDS and mortality in Guinea-Bissau: a prospective open cohort study. *The Lancet HIV*, *6*(1), e25–e31. https://doi.org/10.1016/S2352-3018(18)30254-6

Fellenberg, K., Hauser, N. C., Brors, B., Neutzner, A., Hoheisel, J. D., & Vingron, M. (2001). Correspondence analysis applied to microarray data. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(19), 10781–10786. https://doi.org/10.1073/pnas.181597298

Fitzgerald, L. F., & Hubert, L. J. (1987). Multidimensional Scaling: Some Possibilities for Counseling Psychology. *Journal of Counseling Psychology*, *34*(4), 469–480. https://doi.org/10.1037/0022-0167.34.4.469

Frank, T. D., Carter, A., Jahagirdar, D., Biehl, M. H., Douwes-Schultz, D., Larson, S. L., Arora, M., Dwyer-Lindgren, L., Steuben, K. M., Abbastabar, H., Abu-Raddad, L. J., Abyu, D. M., Adabi, M., Adebayo, O. M., Adekanmbi, V., Adetokunboh, O. O., Ahmadi, A., Ahmadi, K., Ahmadian, E., … Murray, C. J. L. (2019). Global, regional, and national incidence, prevalence, and mortality of HIV, 1980-2017, and forecasts to 2030, for 195 countries and territories: A systematic analysis for the Global Burden of Diseases, Injuries, and Risk Factors Study 2017. *The Lancet HIV*, *6*(12), e831–e859.

https://doi.org/10.1016/S2352-3018(19)30196-1

Franke, T. M., Ho, T., Christie, C. A., Franke, T. M., & Ho, T. (2012). *of Evaluation The Chi-Square Test :* https://doi.org/10.1177/1098214011426594

Functionalizable, A. C., Material, N., Tma, C., Chui, S. S., Lo, S. M., Charmant, J. P. H., Orpen, A. G., Williams, I. D., Ho, C. T. M. A., Chui, S. S., Lo, S. M., Charmant, J. P. H., Orpen, A. G., & Williams, I. D. (2016). *Published by : American Association for the Advancement of Science Linked references are available on JSTOR for this article : Nanoporous MateriaL*. *283*(5405), 1148–1150.

Ghodsi, A. (2006). Dimensionality Reduction A Short Tutorial. *Science*, 25. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.88.3592&rep=rep1&type=pdf%5Cnhttp://stats.stackexchange.com/questions/7111/pca-for-images-arrays-with-high-dimensionality

Gougeon, M. L. (2003). Apoptosis as an HIV strategy to escape immune attack. *Nature Reviews Immunology*, *3*(5), 392–404. https://doi.org/10.1038/nri1087

Guttman, L. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika*, *33*(4), 469–506. https://doi.org/10.1007/BF02290164

Hayflick, L. (1992). Origin of HIV-1. *The Lancet*, *340*(8817), 484–485. https://doi.org/10.1016/0140-6736(92)91806-J

Hemkens, L. G., Ewald, H., Santini-Oliveira, M., Bühler, J. E., Vuichard, D., Schandelmaier, S., Stöckle, M., Briel, M., & Bucher, H. C. (2015). Comparative effectiveness of tenofovir in treatment-näive HIV-infected patients: Systematic review and meta-analysis. *HIV Clinical Trials*, *16*(5), 178–189. https://doi.org/10.1179/1945577115Y.0000000004

Hoffman, D. L., & Franke, G. R. (1986). Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research. *Journal of Marketing Research*, *23*(3), 213. https://doi.org/10.2307/3151480

Husson, F., Josse, J., Saporta, G., Leeuw, J. De, & Saporta, G. (2019). *Jan de Leeuw and the French School of Data Analysis*. *73*(6). https://doi.org/10.18637/jss.v073.i06

Jesmin, S. S., Chaudhuri, S., & Abdullah, S. (2013). Educating Women for HIV Prevention: Does Exposure to Mass Media Make Them More Knowledgeable? *Health Care for Women International*, *34*(3–4), 303–331. https://doi.org/10.1080/07399332.2012.736571.

Jiawen, C. (2012). Development tendency of the embedded system software. *Lecture Notes in Electrical Engineering*, *137 LNEE*(2), 131–135. https://doi.org/10.1007/978-3-642-26007-0_18

Johs, H. (2018). Multiple correspondence analysis. *Multiple Correspondence Analysis For The Social Sciences*, 31–55. https://doi.org/10.4324/9781315516257-3

Joint United Nations Programme on HIV/AIDS (UNAIDS), 2022. (n.d.). *IN IN IN IN IN IN IN IN IN IN U NAID S G l oba l A I D S U p d a t e 20 22 DANGER DANGER DANGER DANGER DANGER DANGER DANGER DANGER DANGER*.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31–36. https://doi.org/10.1007/BF02291575

Kim, E. Y., Lorenzo-Redondo, R., Little, S. J., Chung, Y. S., Phalora, P. K., Maljkovic Berry, I., Archer, J., Penugonda, S., Fischer, W., Richman, D. D., Bhattacharya, T., Malim, M. H., & Wolinsky, S. M. (2014). Human APOBEC3 Induced Mutation of Human Immunodeficiency Virus Type-1 Contributes to Adaptation and Evolution in Natural Infection. *PLoS Pathogens*, *10*(7). https://doi.org/10.1371/journal.ppat.1004281

Kocatepe, D., Alkan, B. B., Keskin, İ., & Kaya, Y. (2020). Consumer perceptions of food safety of fried mussel: multiple correspondence analysis. *Food and Health*, *6*(1), 9–19. https://doi.org/10.3153/fh20002

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*(2), 115–129. https://doi.org/10.1007/BF02289694

Kudlats, J., Money, A., & Hair, J. F. (2014). Correspondence analysis: A promising technique to interpret qualitative data in family business research. *Journal of Family Business Strategy*, *5*(1), 30–40. https://doi.org/10.1016/j.jfbs.2014.01.005

Lanman, T., Letendre, S., Bang, A., Ellis, R., Jolla, L., & Jolla, L. (2022). *HHS Public Access*. *16*(1), 130–143. https://doi.org/10.1007/s11481-019-09886-7.CNS

Lebart, L., Morineau, A., & Piron, M. (1995). Statistique exploratoire multidimensionnelle. *Statistique Exploratoire Multidimensionnelle*, 439.

Lu, D., Wu, H., Yarla, N. S., Xu, B., Ding, J., & Lu, T. (2018). *HAART in HIV/AIDS Treatments: Future Trends*. 15–22. https://doi.org/10.2174/1871526517666170505122800

MacKay, D., & Zinnes, J. (1982). PROSCAL: a program for probabilistic scaling. *Unpublished Manuscript, School of Business, …*, *January 2000*, 1–148. http://proscal.com/pb06.pdf

Matsunaga, M., & Masaki, M. (2010). How to Factor-Analysis Your Data: Do's, Don'ts, and How-to's. *International Journal of Psychological Research*, *3*(1), 97–110.

Mayssara, Hassanin, dkk. (2019). 済無No Title No Title No Title. In *Paper Knowledge. Toward a Media History of Documents*.

Methods, R. (n.d.). *Multiple Correspondence Analysis and Related Methods*. Micheal Greenacre and Jorg Blasius.

Meyer, J. M., Heath, A. C., Eaves, L. J., & Chakravarti, A. (1992). Using multidimensional scaling on data from pairs of relatives to explore the dimensionality of categorical multifactorial traits. *Genetic Epidemiology*, *9*(2), 87–107. https://doi.org/10.1002/gepi.1370090203

N, B. (1964). *The problem immediately raised is how observations may be combined, as they must be when the errors in data are large if the same space is not common to all. In the Tucker and Messick method, a matrix of observations with rows representing stimulus*. *34*.

Nyamweya, S., Hegedus, A., Jaye, A., Rowland-Jones, S., Flanagan, K. L., & Macallan, D. C. (2013). Comparing HIV-1 and HIV-2 infection: Lessons for viral immunopathogenesis. *Reviews in Medical Virology*, *23*(4), 221–240. https://doi.org/10.1002/rmv.1739

Pearl, J. (1994). Department of Statistics Papers. *UCLA Department of Statistics Papers*, *August*, 36.

Press, I. (1950). *Dimensions of Similarity Author (s): Fred Attneave Source : The American Journal of Psychology, Oct ., 1950, Vol . 63, No. 4 (Oct ., 1950), pp . Published by : University of Illinois Press Stable URL : https://www.jstor.org/stable/1418869 REFERENC*. *63*(4), 516–556.

Raffi, F., Pozniak, A. L., & Wainberg, M. A. (2014). Has the time come to abandon efavirenz for first-line antiretroviral therapy? *Journal of Antimicrobial Chemotherapy*, *69*(7), 1742–1747. https://doi.org/10.1093/jac/dku058

Ramsay, J. O. (1977). Maximum likelihood estimation in multidimensional scaling. *Psychometrika*, *42*(2), 241–266. https://doi.org/10.1007/BF02294052

Roomaney, R. A., van Wyk, B., & Wyk, V. P. Van. (2022). Aging with HIV:

Increased Risk of HIV Comorbidities in Older Adults. *International Journal of Environmental Research and Public Health*, *19*(4). https://doi.org/10.3390/ijerph19042359

Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in Personality and Social Psychology Bulletin. *Personality and Social Psychology Bulletin*, *28*(12), 1629–1646. https://doi.org/10.1177/014616702237645

Salkind, N. (2012). Correspondence Analysis. *Encyclopedia of Research Design*, 1–20. https://doi.org/10.4135/9781412961288.n83

Scheit, C., Park, K., & Caers, J. (2009). *Defining A Random Function From A Given Set Of Model Realizations*. 1–11.

Series, L. N. (2010). Graph Layout Techniques and Multidimensional Data Analysis Author (s): Jan De Leeuw and George Michailidis Source : Lecture Notes-Monograph Series, Vol . 35, Game Theory, Optimal Stopping, Probability and Statistics (2000), pp . 219-248 Published. *Statistics*, *35*(2000), 219–248.

Sourial, N., Wolfson, C., Zhu, B., Quail, J., Fletcher, J., Karunananthan, S., Bandeen-Roche, K., Béland, F., & Bergman, H. (2010). Correspondence analysis is a useful tool to uncover the relationships among categorical variables. *Journal of Clinical Epidemiology*, *63*(6), 638–646. https://doi.org/10.1016/j.jclinepi.2009.08.008

Takane, Y., Young, F. W., & de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, *42*(1), 7–67. https://doi.org/10.1007/BF02293745

Takehisa, J., Kraus, M. H., Ayouba, A., Bailes, E., Van Heuverswyn, F., Decker, J. M., Li, Y., Rudicell, R. S., Learn, G. H., Neel, C., Ngole, E. M., Shaw, G. M., Peeters, M., Sharp, P. M., & Hahn, B. H. (2009). Origin and Biology of Simian Immunodeficiency Virus in Wild-Living Western Gorillas. *Journal of Virology*, *83*(4), 1635–1648. https://doi.org/10.1128/jvi.02311-08

Torgensen, W. S. (1952). Multidimensional Scaling: I. Theory and Method. In *Psychometrika* (Vol. 17, Issue 4, pp. 401–419).

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520

Tucker, L. R., Tucker, L. R., & Inv, P. (1970). *AND THREE-MODE FACTOR*

ANALYSIS Protect on Techniques for Investigation of Structure of Individual Differences in Psychological Phenomena.

United Nations Programme on HIV/aids. UNAIDS. (2021). *UNAIDS data 2021*. 4–38.

Vlahov, D., & Junge, B. (1998). The role of needle exchange programs in HIV prevention. *Public Health Reports*, *113*(SUPPL. 1), 75–80.

Waning, B., Kaplan, W., King, A. C., Lawrence, D. A., Leufkens, H. G., & Fox, M. P. (2009). Global strategies to reduce the price of antiretroviral medicines: Evidence from transactional databases. *Bulletin of the World Health Organization*, *87*(7), 520–528. https://doi.org/10.2471/BLT.08.058925

Wynberg, E., Williams, E., Tudor-Williams, G., Lyall, H., & Foster, C. (2018). Discontinuation of Efavirenz in Paediatric Patients: Why do Children Switch? *Clinical Drug Investigation*, *38*(3), 231–238. https://doi.org/10.1007/s40261-017-0605-1

Xu, B., Kajimoto, H., Konyo, M., Saga, S., Hatzfeld, C., Kühner, M., Söllner, S., Khanh, T. Q., Kupnik, M., Sperling, G., Gescheider, G. A., Ma, Q., Zhang, L., Jia, Q., Lü, X. L., Wu, C., Bylinskii, Z., Judd, T., Oliva, A., … Osman, E. M. (2004). Perceptual scaling of the gloss of a one-dimensional series of painted black samples. *Textile Chemist & Colorist*, *29*(3), 1–18. http://europepmc.org/abstract/AGR/IND606170213%0Ahttps://linkinghub.elsevier.com/retrieve/pii/B978012812875601001X%0Ahttp://dx.doi.org/10.1038/nn.3221%0Ahttps://doi.org/10.1016/j.cobeha.2019.06.001%0Ahttp://www.ncbi.nlm.nih.gov/pubmed/21683947%0Ahttp://dx

Yang, B. (1995). Projection Approximation Subspace Tracking. *IEEE Transactions on Signal Processing*, *43*(1), 95–107. https://doi.org/10.1109/78.365290

Zhang, G., & Cheng, L. L. (2010). Semi-supervised classification with metric learning. *Proceedings - 2010 2nd WRI Global Congress on Intelligent Systems, GCIS 2010*, *3*, 123–126. https://doi.org/10.1109/GCIS.2010.223

Zhang, S. (2014). *Multidimensional scaling and model-based clustering analyses for the clade assignments of the HPAI H5N1 viruses*. *January 2007*.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of Five Rules for Determining the Number of Components to Retain. *Psychological Bulletin*, *99*(3), 432–442. https://doi.org/10.1037/0033-

2909.99.3.432.

Appendix X

Turnitin Similarity Report

FINAL THESIS

ÖZGÜNLÜK RAPORU

%20	%18	%16	%11
BENZERLİK ENDEKSİ	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCİ ÖDEVLERİ

BİRİNCİL KAYNAKLAR

1	docs.neu.edu.tr İnternet Kaynağı	%1
2	academic.oup.com İnternet Kaynağı	<%1
3	Submitted to Corvinus University of Budapest Öğrenci Ödevi	<%1
4	jscholarship.library.jhu.edu İnternet Kaynağı	<%1
5	ouci.dntb.gov.ua İnternet Kaynağı	<%1
6	Submitted to Yakın Doğu Üniversitesi Öğrenci Ödevi	<%1
7	www2.mdpi.com İnternet Kaynağı	<%1
8	dokumen.pub İnternet Kaynağı	<%1
9	tel.archives-ouvertes.fr İnternet Kaynağı	<%1

Appendix A

Resume

Personal information

Kabiru Bala

Principal Lecturer

||ND| HND||PGDE||MSC||PhD|| Researcher || Research Data analyst

Taraba State, Nigeria || Mobile: +234-8068955842 ||
kabirubala7@gmail.com || kbstat@yahoo.com

Education data

2003	ND – Statistics Fedpodam
2006	HND – Statistics Fedpodam
2010	PGDE – Education Usman Danfodio University Sokoto
2018	MSC – Biostatistics NEU Cyprus
2023	PhD – Biostatistics NEU Cyprus

Research Interest

Prediction of classical modeling techniques
Prediction of Bayesian modeling techniques
Prediction of Exploratory data analysis (EDA) techniques

Academic/Teaching Experience

2007-present	Mathematics and Statistics Department Taraba State Polytechnic Suntai, Jalingo Campus. Teaching Statistics, Business Statistics, Mathematics, and Student Project Supervision

Working Experience

21-10-1998	First appointment
2007	Instructor

2010	Senior Instructor
2013	Principal Instructor
2018	Assistant Chief Instructor
2023	Chief Instructor
2024	Principal Lecturer
2010-2015	HOD Mathematics and Statistics
2023-present	HOD Mathematics and Statistics

Polytechnic Committee

2011	Member electoral committee ASUP Taraba State Polytechnic
2013	Member of the electoral committee into Dean School of Basic and Applied Sciences Taraba State Polytechnic Jalingo.
2011/2012	Secretary exam malpractice committee second-semester examination academic session.
2014/2015	Member exam malpractice committee second-semester examination academic session.
2023	Member of the electoral committee into Dean School of Basic and Applied Sciences Taraba State Polytechnic Jalingo.

Applicable skills

Computational skills

- Statistical data analysis: Ability to work with SPSS, Microsoft Excel, and Minitab
- AI modeling: Ability to work with Microsoft Excel and Matlab
- EDA modeling: Ability to work with SPSS and Microsoft Excel
- Tutoring and mentoring, students on data analysis

Sport career and National camp attendance

1996	Nuga prelims Tafawa Balewa University Bauchi (ATBU)
1991-2000	Taraba State Sport Council Badminton Player Team captain
2000-2006	Yobe State Sport Council Badminton Player team Captain

2000-2006	Medalist Federal Polytechnics Games Nipoga Team Captain Fedpodam
2003	Germany Olympic Camp preparation Illorin Phase I and Zaria Phase II
2004	South African Junior Camp preparation Abuja

Professional membership

2010	Member Nigerian Statistical Association (NSA)
2009	Member Mathematics Association (MAN)

Conferences/ Training

2010	Nigerian Statistical Association Pre-Conference workshop Imo State Oweri.
2012	Nigerian Statistical Association Pre-Conference workshop Benue State Makurdi.
2012	Science Teachers Association of Nigeria (Stan Maths Panel)
2013	Nigerian Statistical Association Pre-Conference workshop Akwa Ibom State Uyo.
2013	Advance Digital Appreciation Programme for Tertiary Institutions (ADAPTI).

Research and Publications

Etikan et al., 2017 General Bearing of Students with Sustainable Satisfaction in Higher Institution of Learning. American Journal of Biostatistics 2. DOI: 10.3844/amjbsp.2017.

İlkerEtikan and Kabiru Bala 2017 Combination of Probability Random Sampling Method with Non Probability Random Sampling Method (Sampling Versus Sampling Methods).

Follow-up. DOI: 10.33552/ABBA.2019.02.000528.

İlkerEtikan and Kabiru Bala 2017 Sampling and Sampling Method.

İlkerEtikan and Kabiru Bala 2017 Developing Questionnaire Based on Selection and Designing.

İlkerEtikan and Kabiru Bala 2017 Influence of Residential Setting on Student Outcome.

İlkerEtikan and Kabiru Bala 2017 Review of Prostatic Tumor Using Kaplan Meier and Cox Regression. BBIJ-06-00180 survival analysis.pdf
Etikan et al., 2019 Application of Multivariate Statistical Methods of Patient Surviving ART
İlker Etikan, Ogunjesa Babatope, Kabiru Bala and Savaş İlgi 2019 Child Mortality: A Comparative Study of Some Developing Countries in the World. Child mortality A comparative study of some developing countries.pdf.
Etikan et al., 2019 HIV/AIDS prevalence cases among patients undergoing follow-up in federal teaching hospital Gombe state Nigeria application of multivariate analysis. doi: 10.15761/GDT.1000168.
Etikan et al., 2019 Pro Motives behind Research Survey Reflecting on Some Procedures of Data Assortment. ttp://enlivenarchive.org/submit-manuscript.php.
Thesis/Dissertations: Etikan et al., 2019 Application of Multivariate Statistical Methods of Patient Surviving ART Follow-u. DOI: 10.33552/ABBA.2019.02.000528.
Conferences/Seminar
Bala et al., 2023 Artificial-Intelligence-Based Models Coupled with Correspondence Analysis Visualization on ART—Cases from Gombe State, Nigeria: A Comparative Study. Life 2023, 13, 715. https://doi.org/10.3390/life13030715.

References

1. Professor Ilker Etikan: HOD Biostatistics Department Near East University (NUE) Cyprus
Relationship - Advisor, HOD, MSC and PhD supervisor.

2. Dr. Sani Galadima: Chief Lecturer Department of Mathematics and Statistics Federal Polytechnique Damaturu Yobe State.
Relationship- Lecturer, HOD and Supervisor.

3. Dr. Ayuba Abarshi: Rector Taraba State Polytechnique Suntai, Taraba State Nigeria.
Relationship – Rector and Boss.