International Journal of Advanced and Applied Sciences, 6(12) 2019, Pages: 99-104

Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

A study of principal components analysis for mixed data

Zakiah I. Kalantan *, Nada A. Alqahtani

Department of Statistics, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 18 July 2019
Received in revised form
8 October 2019
Accepted 10 October 2019

Keywords:
Dimension reduction
Mixed data
Principal component analysis
R package

ABSTRACT

Analyzing data requires statistical tools to interpret the data information, which helps to improve the process. This is the interpretation of the qualitative and quantitative status of mixed data. The objective of this paper was to study the implementation of principal component analysis on mixed data and explain how to handle this type of databases and to make it possible to extract statistical information over a population under study. The effectiveness of principal component analysis on mixed data was studied using data sets available in the R package and simulated data.

1. Introduction

Most data applications involve dealing with large data sets, which contain several measures (variables) that can be either numerical or categorical. Thus, increasingly, scientific researchers such as businesses and members of medical fields require powerful visual and analytical tools to visualize and analyze data.

Processing large data becomes more and more difficult as the number of dimensions' increases. Dimension reduction is a collection of statistical methods used to analyze mixtures of big data. This is done in two different ways: By selecting the most significant features from all features, which is used to make model building (this technique is called feature selection) or by transforming the high-dimensional data into low-dimensional and saving the most important information. This procedure saves the data information that must be processed, while still accurately and completely describing the original data set (this technique is called *feature* extraction). Principal component analysis (PCA) is one of the commonly used dimension reduction methods, and it is known as a feature extraction method that is used for mixed data. It was invented in 1901 by Pearson (1901).

The central idea is to find a new coordinate system in which input data can be expressed but at the same time information loss can be minimized. The idea of PCA is to reduce the dimension of original data by computing a few numbers of orthogonal linear combinations with minimal loss of information, which means assigning the principal components (PCs) of the original variables with the largest variance. PCA is used for many applications, for example, image compression, bioinformatics, data mining, psychology, and pattern recognition, among others (Kalantan et al., 2017; Kalantan, 2019).

Practically, principal component analysis (PCA) handles numerical variables, while multiple correspondence analysis (MCA) handles categorical variables. PCA on mixed data is one of the several proposed methods to handle large data. This method can be seen as a mixture of PCA and MCA. It was proposed by De Leeuw and van Rijckevorsel (1980). This paper illustrates this method with details and discusses the effectiveness using the method implementation on a real dataset.

The paper is organized as follows. Section 2 presents a brief review of PCA. MCA is discussed in Section 3. Section 4 demonstrates how PCA is obtained for mixed data. Finally, the interpretation of a case study and associated graphics is discussed in Section 5.

2. Principal component analysis

From an algebraic standpoint, principal components are linear combinations of p random variables, $X_1, X_2, \dots, X_p$. We shall look at the derivation of population principle components when the covariance matrix $\sum$ is known. Suppose we have a mean zero normal random vector $\acute{X} = [X_1, X_2, \dots, X_p]$ that has a covariance matrix $\sum$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Let us now consider the following linear combinations (Johnson and Wichern, 2002):

* Corresponding Author.
Email Address: zkalanten@kau.edu.sa (Z. I. Kalantan)
https://doi.org/10.21833/ijaas.2019.12.012
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-7040-5623

$$Y_1 = a'_1X = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p,$$
$$Y_2 = a'_2X = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p,$$
$$\vdots$$
$$Y_p = a'_pX = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp} \quad (1)$$

from the definition of covariance, we have that:

$$Var(Y_i) = \acute{a}_i\Sigma a_i \qquad i = 1,2,\dots,p, \quad (2)$$
$$Cov(Y_i, Y_k) = \acute{a}_i\Sigma a_k = 0 \qquad i,k = 1,2,\dots,p. \quad (3)$$

Principal components are uncorrelated linear combinations whose variances are as large as possible. Therefore, the first principal component is the linear combination with the maximum variance or $Var(Y_1) = \acute{a}_1\Sigma a_1$ that has the largest variance. Since one can increase a_1 by any constant, we impose the restriction that maximizing $Var(\acute{a}_1X)$ is subject to $\acute{a}_1a_1 = 1$. Thus, the principal components are such that:

1st principal component= linear combination $\acute{a}_1X$ that maximizes $Var(\acute{a}_1X)$ subject to $\acute{a}_1a_1 = 1$

2nd principal component= linear combination $\acute{a}_2X$ that maximizes $Var(\acute{a}_2X)$ subject to $\acute{a}_2a_2 = 1$ and $Cov(\acute{a}_1X, \acute{a}_2X) = 0$

3th principal component= linear combination $\acute{a}_iX$ that maximizes $Var(\acute{a}_iX)$ subject to $\acute{a}_ia_i = 1$ and $Cov(\acute{a}_iX, \acute{a}_kX) = 0$ for $k < i$.

3. Multiple correspondence analysis (MCA)

Multiple correspondence analysis is a statistical technique. It is an extension of simple correspondence analysis (CA) which allows one to study the association and visualize a data table between two or more qualitative variables. It can be seen as an analogue of principal components analysis (PCA) when the variables to be analyzed are categorical variables instead of quantitative variables (Abdi and Valentin, 2007).

There are K categorical variables, and each categorical variable has J_k levels where $J = \sum_j J_k$. There are I observations. Let X be an indicator matrix with $I \times J$ dimensions. MCA is performed by applying CA on the indicator matrix. Then, the two sets of factor scores are obtained for the rows and the columns. These factor scores are standardized where their variance equals their corresponding eigenvalue.

Firstly, we compute the probability matrix $Z = N^{-1}X$, where N is the whole number. Let $D_c = diag\{c\}, D_r = diag\{r\}$, where the vector of the row totals and the columns totals of Z is denoted by r and c, respectively. We obtain the factor scores by applying the following SVD:

$$D_r^{-\frac{1}{2}}(Z - rc^T)D_c^{-\frac{1}{2}} = P\Delta Q^T \quad (4)$$

where Δ is the diagonal matrix of the singular values and $\Lambda = \Delta^2$ is the matrix of the eigenvalues.

Then, we obtain the rows factor scores which are denoted by F and the columns factor scores which are denoted by G as follows (Abdi and Valentin, 2007):

$$F = D_r^{-\frac{1}{2}}P\Delta \quad (5)$$

and

$$G = D_c^{-\frac{1}{2}}Q\Delta \quad (6)$$

4. Principal component analysis for mixed data

In this paper, we implemented the PCA on mixed data following the approach proposed by Chavent et al. (2014). The dataset to be analyzed by PCA mix consists of n observations described by p_1 numerical variables and p_2 categorical variables. Let X_1 be an $n \times p_1$ matrix which represents the numerical variables and X_2 be an $n \times p_2$ matrix that represents the categorical variables. Let d denote the total number of all variables. An indicator matrix G with $n \times d$ dimensions contains binary coding from each level of categorical variables. A numerical matrix $Y = (Y_1|Y_2)$ is constructed with dimension $n \times (p_1 + d)$ where Y_1 is the standardized matrix constructed by centered and normalized columns of X_1, and Y_2 denotes the centered indicator matrix X_2.

Now, let N be the diagonal matrix of the weights of the rows of Y, where $\frac{1}{n}$ represents the weights of n rows, then $N = \frac{1}{n}I_n$. Suppose $D = diag(1, \dots, \frac{n}{n_1}, \frac{n}{n_s})$ is the diagonal matrix of the weights of the columns of Y and $s = 1, \dots, n$ represents the number of observations appearing at the sth level. Then, the eigenvalue of Y is obtained using the generalized singular value decomposition (GSVD) as:

$$Y = U\Lambda V^T \quad (7)$$

where $\Lambda = diag(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_r})$ is the $r \times r$ diagonal matrix, such that $\lambda_1, \lambda_2, \dots, \lambda_r$ are the eigenvalues of Y and r denotes the rank of Y. U is a matrix with $n \times r$ dimensions, where the first r eigenvectors of ZDZ^tN such that $U^TNU = I_r$. V is the $p \times r$ matrix of the first r eigenvectors of Z^tNZD such that $V^tDV = I_r$. Therefore, the principal component of PCA mix can be computed as:

$$Y^{mix} = YDV \quad (8)$$

with the dimensions of $n \times r$. The scores of rows computed as $R = U\Lambda$ represent the principal component scores. The scores of columns $C = DV\Lambda$ and the standard PCA will be $C = V\Lambda$.

5. Experimental results

In this section, we discuss the effectiveness of PCA on mixed data that contain both numerical and categorical data. This is illustrated with a simulation case and real data available in R packages.

5.1. Simulation case

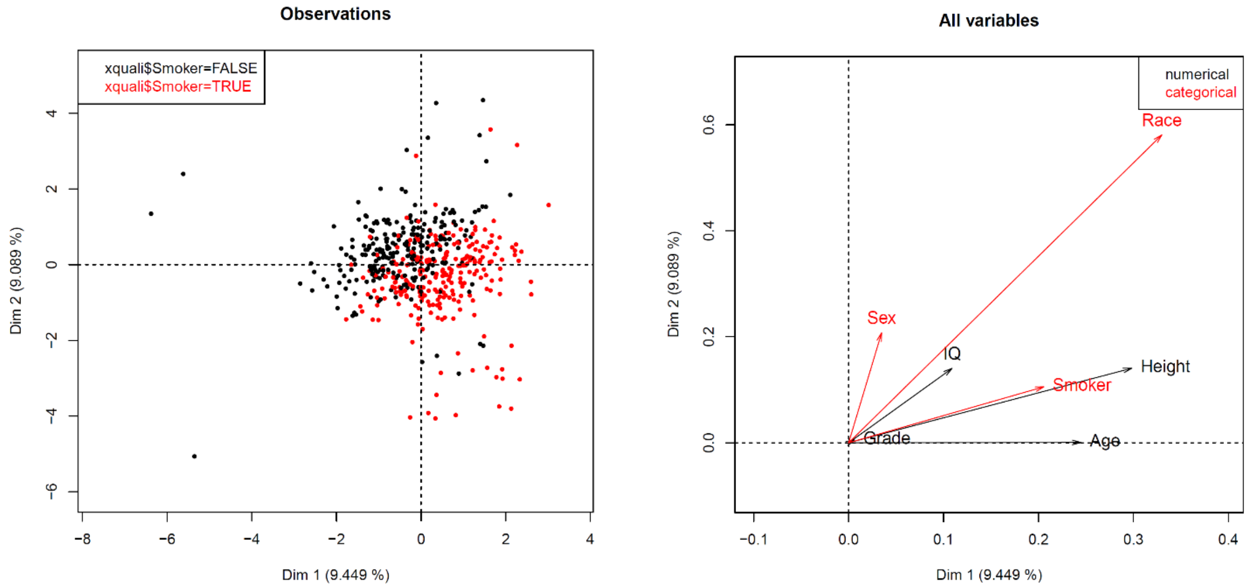
A generalized sample of size 500 consists of seven variables. The first four are quantitative variables: Age, IQ, grade, and height, while the variables race, sex, and smoker are considered as qualitative variables; the data are available in the 'Wakefield' package (Rinker, 2018). As a pre-processing step, we split the data into two data matrices: A 500 × 4 numerical data matrix named data A, and data B, representing the categorical variables as a matrix of 500 × 3. We established the analysis with the implementation for PCA, and the results are summarized in Table 1, which shows that 80.84% of the total variance is explained via 10 PCA components.

Fig. 1a displays the graphical output of the results of the factor coordinates, absolute contribution, and the squared cosinus for all variables. Table 2 presents the contributions of all variables; the contribution squared correlation for each quantitative variable and the contribution correlation ratio of qualitative variables are shown in Fig. 1b in a graphical output.

Table 1: The results of the simulation case

	Eigen Value	Proportion of Variance	Cumulative Proportion
Comp 1	1.2284	0.0945	0.0945
Comp 2	1.1815	0.0909	0.1854
Comp 3	1.1296	0.0869	0.2723
Comp 4	1.0666	0.0820	0.3543
Comp 5	1.0423	0.0802	0.4345
Comp 6	1.0257	0.0789	0.5134
Comp 7	1.0000	0.0769	0.5903
Comp 8	0.9834	0.0756	0.6660
Comp 9	0.9547	0.0734	0.7394
Comp 10	0.8976	0.0690	0.8084
Comp 11	0.8844	0.0680	0.8765
Comp 12	0.8109	0.0624	0.9389
Comp 13	0.7949	0.0611	1

More graphical outputs are presented in Fig. 2a and Fig. 2b. Fig. 2a shows the factor coordinates, absolute contribution, and the squared cosinus of the qualitative variables. The results for the quantitative variables are presented in Fig. 2b.



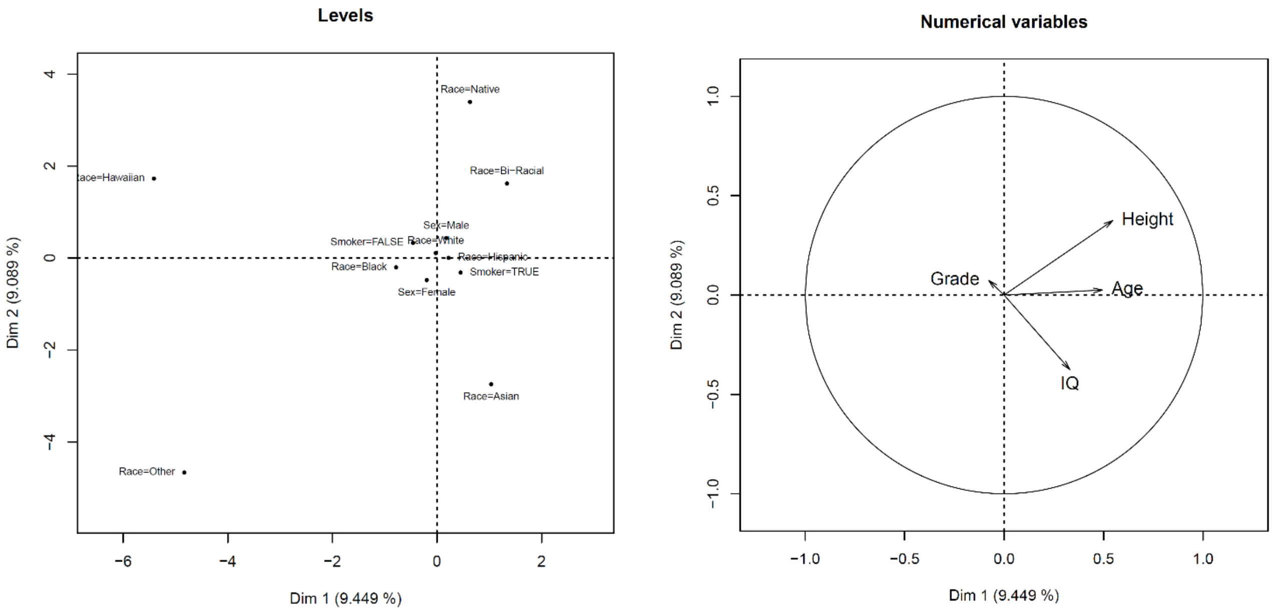
a b

Fig. 1: Simulation case; (a) results for the individuals; (b) results of squared loadings

5.2. Application case

We implemented the PCA mix method on an R dataset from the "ElemStatLearn" package and named it "SAheart". It is a sample of males in a heart-disease high-risk region of the Western Cape, South Africa. The dataset consists of 462 observations on the following 10 variables, two of which are qualitative variables and the rest are quantitative variables, as shown in Table 3.

As a pre-processing step, we split the data into two data matrices: A 462 × 8 numerical data matrix named data A, and data B, representing the categorical variables as a matrix of 462 × 2. We established the analysis with the implementation for PCA, and the results are summarized in Table 4, which shows that 81.23% of the total variance is explained via 6 PCA components.



a b

Fig. 2: Simulation case; (a) results for the levels of the qualitative variables; (b) results of for the quantitative variables

Table 2: The levels of contributions for all variables

	dim 1	dim 2	dim 3	dim 4	dim 5
Age	0.49461335	0.0253062	0.32267444	0.321041741	0.0156075
IQ	0.33015349	0.3747271	0.23761784	0.476332544	0.2191435
Grade	0.07749287	0.0737638	0.69664527	0.200292231	0.1108619
Height	0.54613751	0.3750766	0.02244047	0.041334596	0.2269596
Race	0.57441641	0.7621728	0.64928206	0.674168311	0.8426543
Sex	0.18645019	0.4557888	0.19505485	0.001166874	0.4677109
Smoker	0.45358812	0.3251022	0.15354866	0.490235036	0.0375936
	dim 6	dim 7	dim 8	dim 9	dim 10
Age	0.34787313	7.38404×10^{-28}	0.240260707	0.3127593	0.12191561
IQ	0.19436481	3.33998×10^{-24}	0.260624551	0.2363748	0.07086126
Grade	0.10124499	2.17422×10^{-14}	0.037497596	0.1016380	0.56607122
Height	0.20012710	2.93860×10^{-14}	0.008892727	0.1882476	0.28788321
Race	0.88367977	1.00000000	0.911678682	0.7032087	0.67709763
Sex	0.16800716	6.56282×10^{-14}	0.034110589	0.4740927	0.05923535
Smoker	0.08636048	8.04556×10^{-15}	0.154885059	0.1896927	0.11133049
	dim 11	dim 12	dim 13		
Age	0.02182873	0.5018176549	0.05972314		
IQ	0.08659104	0.3594158770	0.34047717		
Grade	0.18099061	0.0008429403	0.27504554		
Height	0.38419922	0.2871377599	0.34471365		
Race	0.63658797	0.4260427922	0.59264882		
Sex	0.38996161	0.2905133947	0.08099889		
Smoker	0.37248689	0.2856115243	0.35085572		

Table 3: The variables' description

Variables Types	Variable Name	Description
quantitative variables	Sbp	systolic blood pressure
	Tobacco	cumulative tobacco (kg)
	Ldl	low density lipoprotein cholesterol
	adiposity	a numeric vector
	typea	type-A behavior
	obesity	a numeric vector
	alcohol	current alcohol consumption
	Age	age at onset
qualitative variables	famhist	family history of heart disease, a factor with levels Absent and Present
	Chd	response, coronary heart disease, a factor with levels 0 and 1

Fig. 3a displays the graphical output of the results of the factor coordinates, absolute contribution, and the squared cosinus for all variables. Table 5 presents the contributions of all variables; the contribution squared correlation for each quantitative variable and the contribution correlation ratio of qualitative variables are shown in Fig. 3b in a graphical output.

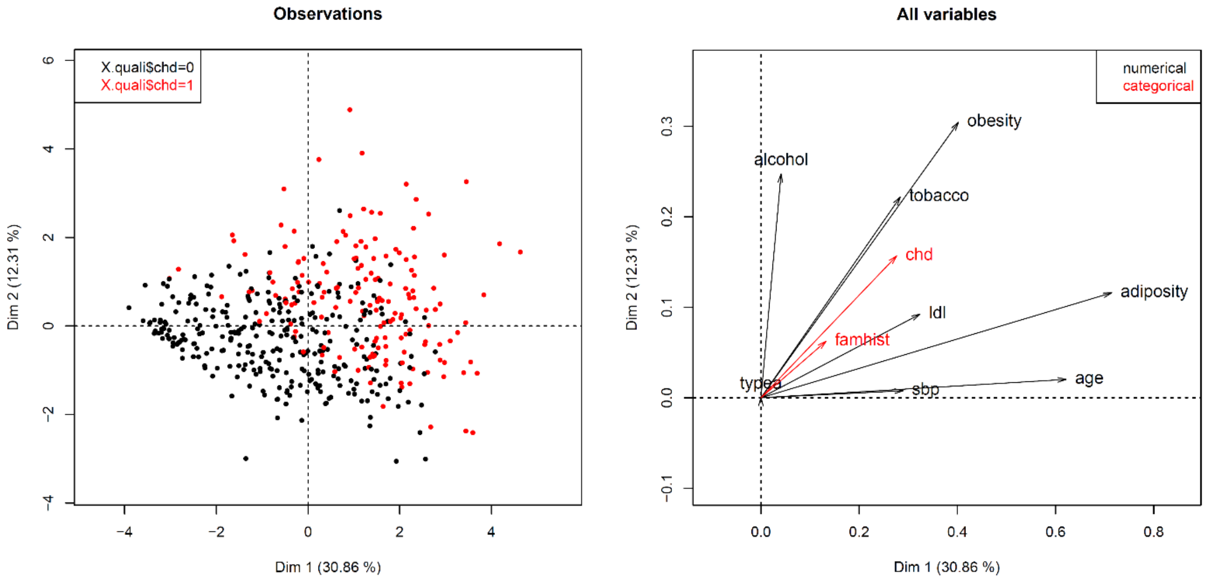
More graphical outputs are presented in Fig. 4a and Fig. 4b. Fig. 4a shows the factor coordinates, absolute contribution, and the squared cosinus of the qualitative variables. The results for the quantitative variables are presented in Fig. 4b.

6. Conclusion

The PCA is a powerful technique for mixed data to interpret the variables status for different data types. The objective of this process is to reduce the number of dimensions by selecting the components that describe 80% of the variance of the data. It was found that through this method, we can analyze a mixture of numerical and categorical variables and extract relevant information without having to deal with each type separately.

Table 4: The results of the application case

	Eigen Value	Proportion of Variance	Cumulative Proportion
Comp 1	3.0865	0.3086	0.3086
Comp 2	1.2307	0.1231	0.4317
Comp 3	1.1462	0.1146	0.5463
Comp 4	1.0199	0.1020	0.6483
Comp 5	0.8725	0.0872	0.7356
Comp 6	0.7676	0.0768	0.8123
Comp 7	0.6728	0.0673	0.8796
Comp 8	0.5740	0.0574	0.9370
Comp 9	0.4551	0.0455	0.9825
Comp 10	0.1748	0.0175	1

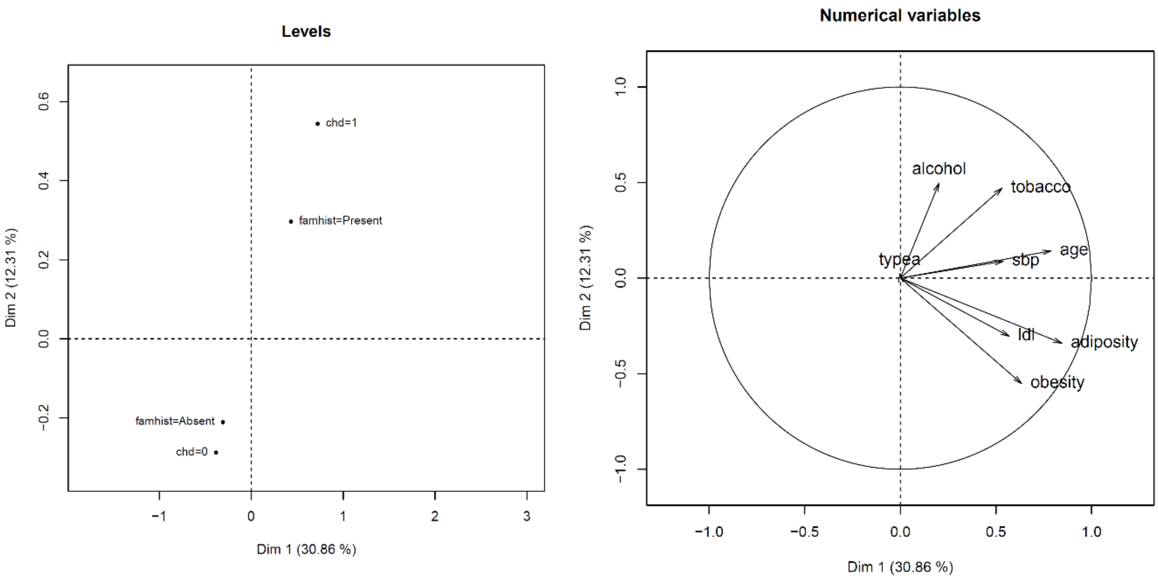


a b

Fig. 3: Application case; (a) results for the individuals; (b) results of squared loadings

Table 5: The levels of contributions for all variables

	dim 1	dim 2	dim 3	dim 4	dim 5
Sbp	2.89672×10^{-1}	0.0078017394	0.1063237008	1.04341×10^{-2}	3.859448×10^{-7}
tobacco	2.83864×10^{-1}	0.2219591833	0.0278278381	9.66279×10^{-5}	1.55977×10^{-1}
Ldl	3.24117×10^{-1}	0.0924692414	0.0708663614	2.74967×10^{-2}	1.38961×10^{-2}
adiposity	7.14502×10^{-1}	0.1160730548	0.0100080553	1.01866×10^{-2}	3.16736×10^{-3}
typea	6.91093×10^{-6}	0.0005861629	0.5351210938	2.76608×10^{-1}	8.69287×10^{-2}
obesity	4.02313×10^{-1}	0.3045250723	0.0002482346	9.61208×10^{-2}	1.34622×10^{-2}
alcohol	4.10580×10^{-2}	0.2475114945	0.0326211866	4.28246×10^{-1}	9.67767×10^{-2}
Age	6.21652×10^{-1}	0.0203379960	0.0281102874	3.33931×10^{-2}	5.15464×10^{-3}
famhist	1.32839×10^{-1}	0.0626074482	0.1937124803	6.253280e-02	4.62151×10^{-1}
chd	2.76437×10^{-1}	0.1567957227	0.1413710331	7.477092e-02	3.49385×10^{-2}
	dim 6	**dim 7**	**dim 8**	**dim 9**	**dim 10**
Sbp	0.5146484168	0.02641214	0.024686861	0.0199847300	3.52167×10^{-5}
tobacco	0.0895634567	0.06291979	0.074281783	0.0831068971	4.01776×10^{-4}
Ldl	0.0487664030	0.34379058	0.077501528	0.0001964012	8.99367×10^{-4}
adiposity	0.0071836630	0.01489831	0.012032233	0.0116347445	1.00313×10^{-1}
typea	0.0411248198	0.01990486	0.021200711	0.0182282302	2.90213×10^{-4}
obesity	0.0059762331	0.03844182	0.031564633	0.0629166662	4.44307×10^{-2}
alcohol	0.0503818195	0.08833238	0.009722416	0.0051579590	1.91700×10^{-4}
Age	0.0000495666	0.02726837	0.001947765	0.2339743159	2.81111×10^{-2}
famhist	0.0007046157	0.03110983	0.046201574	0.0079933407	1.46904×10^{-4}
chd	0.0092358722	0.01969713	0.274829644	0.0119234272	1.37883×10^{-7}



a b

Fig. 4: Application case; (a) results for the levels of the qualitative variables; (b) results of for the quantitative variables

Acknowledgment

This paper is a component of a Master thesis undertaken by the second author under the supervision of the first author. The authors would like to thank the reviewers for their helpful comments.

Compliance with ethical standards

Conflict of interest

The authors declare that they have no conflict of interest.

References

Abdi H and Valentin D (2007). Multiple correspondence analysis. Encyclopedia of Measurement and Statistics, 95: 116-128.

Chavent M, Kuentz-Simonet V, Labenne A, and Saracco J (2014). Multivariate analysis of mixed data: The R package PCAmixdata. Available online at: https://bit.ly/2KSd9cG

De Leeuw J and Van Rijckevorsel J (1980). HOMALS and PRINCALS-Some generalizations of principal components analysis. Data Analysis and Informatics, 2: 231-242.

Johnson RA and Wichern DW (2002). Applied multivariate statistical analysis. Volume 5, Prentice Hall, Upper Saddle River, USA.

Kalantan Z, Adham S, and Bahashwan A (2017). Studying of gamma principal components: Analysis of molding noise. International Journal of Advanced Scientific and Technical Research, 7(4): 62-67.

Kalantan ZI (2019). Implementing correlation dimension: K-Means clustering via correlation dimension. In the Third International Conference on Computing, Mathematics and Statistics, Springer, Singapore: 359-366. https://doi.org/10.1007/978-981-13-7279-7_44

Pearson K (1901). LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11): 559-572. https://doi.org/10.1080/14786440109462720

Rinker TW (2018). Wakefield: Generate random data. Version 0.3.3., Buffalo, USA. Available online at: https://bit.ly/34lLqZB