Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2023;9(4): 500-515

Original Article

Socio-Economic Status of Individuals in Tehran University of Medical Sciences Employees` Cohort Study Using PCA, MCA and FAMD Methods

Faezeh Ramezanzadeh Tabriz, Saharnaz Nedjat, Kamal Azam*, Mehdi Yaseri*

Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran.

ARTICLE INFO

Received 13.06.2023
Revised 27.06.2023
Accepted 08.07.2023
Published 15.12.2023

Key words:
Socio-economic status;
Principal component analysis;
Multiple correspondence analysis;
Factor analysis of mixed data;
Cohort study

ABSTRACT

Introduction: Determining socio-economic status (SES) can greatly help decision makers in the field of social health. Because SES can play an important role in accessing medical services or welfare amenities. We aimed to determine the SES using principal component analysis (PCA), multiple correspondence analysis (MCA), and factor analysis of mixed data (FAMD) methods.

Methods: In this cross-sectional study (2023), 4448 employees aged 19 to 75 years were included to the study from Tehran University of Medical Sciences employees` cohort (TEC). Demographic variables and socio-economic factors were considered. Considering the weaknesses of PCA and MCA methods, we calculated the SES score using PCA, MCA and FAMD methods, and the percentile of people was determined. These weaknesses include normality assumption and considering only linear relationship for PCA, inability to interpret the relationships between variables and considering each level of classification variables as a new variable for MCA.

Results: We studied 4448 people (39.3% men) with mean age of 42.3 and a standard deviation of 8.7.The correlation between the percentiles obtained through PCA, MCA and FAMD methods was very high, and the highest correlation was related to the percentiles obtained through PCA and FAMD methods with a value of 0.994. The intraclass correlation coefficient value was 0.996. Also, this value was 0.996 and 0.994 in the random samples of 250 and 100 individuals from the original data, respectively.

Conclusion: All of the three methods worked similarly on determining the SES and calculating the percentile of people. PCA and FAMD methods had better agreement than others. Therefore, in studies that have both quantitative and qualitative variables, the choice of analysis method depends on the opinion of the researcher.

Introduction

Socio-economic status (SES) is defined as access to physical, human and social capitals, and it is known as one of the important and influential pillars on health and welfare, which plays an important role in access to medical services and welfare amenities.[1,2] For example, it has been shown that people with end-stage cancer diagnosis were in low SES.[3]

Various factors affect the SES, including education level, job status, income level,

*.Corresponding Author: kazam@tums.ac.ir; & m.yaseri@gmail.com.

Copyright © 2023 Tehran University of Medical Sciences. Published by Tehran University of Medical Sciences.
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International license (https://creativecommons.org/licenses/by-nc/4.0/). Noncommercial uses of the work are permitted, provided the original work is properly cited.

gender, race[4] and place of residence.[5]

In recent years in Iran, the SES has affected the inequality in the health system. Among the factors that have affected the SES, high liquidity and inflation, high income inequality, and economic structure can be mentioned.[6] Therefore, providing a tool to determine the SES that works better than common methods can be very helpful for decision makers in the field of social health.

Studying SES makes it easy to monitor SES patterns in health indicators. It is possible to compare the health of people who are in a low SES with the health of others. Also, health differences are identified in middle class subgroups compared to the wealthy.[7]

Although many methods have been proposed to determine the SES, these methods are often based on indirect measurement. Recently, the principal component analysis (PCA) method has been used to construct an index of SES in relation to various health outcomes. In some studies, to determine the SES of each person, the property has been used as an indicator of the economic status, and then by a statistical method such as PCA, the hidden SES variable related to it is evaluated. But it should be paid attention to the fact that in this method all variables under investigation must be continuous. This method is not appropriate for analyzing data in categorical form because the categories cannot be assigned a meaningful quantitative scale. To address this issue, it is recommended to transform qualitative categorical variables into binary variables. If their assets are examined, the multiple correspondence analysis (MCA) is used instead of the PCA method, which is applied for categorical variables, but some SES questionnaires include questions that measure other variables (e.g., a combination of quantitative, ordinal, or categorical data), which can't be considered by these methods. Thus, the factor analysis of mixed data (FAMD) method is suitable for considering quantitative and categorical variables to determine the SES. Compared to PCA and MCA methods, the FAMD is able to identify patterns in the data and help to investigate the relationships between variables. In fact, the algorithm of FAMD method consists of the combination of PCA and MCA methods and it can work with categorical and continuous data. We chose three methods of PCA, MCA and FAMD to determine the Individuals` SES due to their ability to analyze multivariable data, to help investigate the relationships between variables by simultaneously analyzing several variables and discovering patterns in the data. Also, these methods have the ability to reduce the data dimension, which makes data analysis more simple and more understandable. In addition, the use of these methods to evaluate SES, due to their high interpretability, makes better and more accurate decisions about community status.[8-10]

We aimed to determine the SES of individuals in Tehran University of Medical Sciences (TUMS) employees` cohort study using PCA, MCA and FAMD methods, which help introducing the best method to determine the SES, and can be used for policy making in healthcare system.

Methods

Study design, setting, and participants

This was a cross-sectional study that used data from the Tehran University of Medical Sciences employees' cohort (TEC) study. The study

included 4448 people who had an employment relationship with Tehran University of Medical Sciences (TUMS) or its affiliated centers. Data collection was conducted from January 2018 until March 2021.[8] It should be noted that the study protocol was approved by the Research Ethics Committee (REC) of TUMS (REC approval ID: IR.TUMS.VCR.REC.1398.246).

Study variables and measurements

In the TEC study, data were collected through questionnaires, including demographic characteristics and social and economic status. Demographic characteristics were age, sex, father's education level, mother's education level, individual`s education level, marital status and number of children.

In the socio-economic status questionnaire, the SES was evaluated through variables include SES in childhood, current SES, the rate of fluctuation of SES during life , number of family members, rooms, the infrastructure level of the residence, housing ownership status, ownership of agricultural land, garden or villa, ownership of shop or office or commercial property, ownership of residential property (except current residence), having dish washing machine, microwave, personal computer or laptop, washing machine, color television, video players or home theaters, car, internet access at home, the total price of household cars, going to concert, cinema, theater, restaurant, traveling by plane, family internet cost per month, the number of books read in the previous year, the number of non-pilgrimage foreign trips in the last ten years, the number of domestic tourist and pilgrimage trips in the last ten years, desire to migrate to another city, desire to migrate to another country.

In the PCA method, all variables were considered continuous and in the MCA method, all variables were considered categorical. But in the FAMD method, all variables were entered into the analysis without transformation. That is, both types of continuous and categorical variables were present in the analysis.

Statistical methods

Due to the heterogeneity regarding age groups and SES level in the studied population, PCA, MCA and FAMD methods were used to determine the SES of the population.

Therefore, in this study, the SES score of people was calculated using three methods: PCA, MCA and FAMD. But because the raw scores of the tests may not provide much information on their own, for a better interpretation of the data, it was determined what percentile each person is placed in using these methods. Therefore it became clear that where each person's score ranks relative to others who took the same test. Because in order to compare the ranking of people's SES under different models, it is necessary to determine the percentile of people based on their scores. Percentiles give a clearer picture of how well someone performed compared to the rest of the group.

Also, in this study, the performance of these three methods in calculating the SES score of people has been compared. To check whether the agreement of these three methods is affected by the high sample size or not, all the methods were re-run in two random samples of the original data with sizes of 100 and 250. Considering that PCA, MCA and FAMD methods are some kind of factor

analysis methods, the minimum sample size required to implement these methods is 3 to 20 times the number of variables and in the range of 100 to more than 1000 reported[11] Therefore, the minimum sample size required for the implementation of these methods, i.e. 100, was selected, and the sample size of 250 was selected to further check the reliability of the results and pay attention to the mentioned range.

(Comprehensive descriptions of statistical methods are given in the supplementary methods file).

Statistical analysis was performed using R version 4.2.3, the FactoMineR and Factoextra packages. Also, we used SPSS software version 26.

Results

In this study, 4448 participants (39.3% men and 60.7% women) with mean (SD) age of 42.3 (8.7) were included to the study analysis. The majority of participant (72.3%) aged between 30 and 50 years, were married (79.1%), and had high school diploma (40.1%) and had two children or more (39.5%). The most of participants` parents (father and mother) had low level of education (45.7% and 46.6%, respectively) (Table 1).

Table 2 shows that the correlation between the percentiles obtained through PCA, MCA and

Table 1. Demographic characteristics according to gender

Study variables		Women (n= 2699) 60.7 %	Men (n= 1749) 39.3 %	Total
Age	< 30 yr	246 (9.1)	130 (7.4)	376 (8. 5)
	30 to 40 yr	1027 (38.1)	606 (34.6)	1633 (36.7)
	40 to 50 yr	974 (36.1)	609 (34.8)	1583 (35.6)
	> 50 yr	452 (16.7)	404 (23.2)	856 (19.2)
Father's education level	illiterate	347 (12.9)	503 (28.8)	850 (19.1)
	low literacy	1232 (45.6)	802(45.8)	2034 (45.7)
	High school diploma	817 (30.3)	342(19.6)	1159 (26.1)
	Academic degree	303 (11.2)	102 (5.8)	405 (9.1)
Mother's education level	illiterate	541 (20.0)	676 (38.7)	1217 (27.4)
	low literacy	1325 (49.1)	747 (42.7)	2072 (46.6)
	High school diploma	725 (26.9)	282 (16.1)	1007 (22.6)
	Academic degree	108 (4.0)	44 (2.5)	152 (3.4)
Individual`s education level	illiterate	146 (5.4)	217 (12.4)	363 (8.2)
	low literacy	574 (21.3)	634 (36.2)	1208 (27.2)
	High school diploma	1234 (45.7)	551 (31.6)	1785 (40.1)
	Academic degree	745 (27.6)	347 (19.8)	1092 (24.5)
Marital status	Single	635 (23.5)	151 (8.6)	786 (17.7)
	married	1933 (71.6)	1588 (90.8)	3521 (79.1)
	Divorced or widowed	131 (4.9)	10 (0.6)	141 (3.2)
Number of children	without children	427(15.8)	279(16.1)	706 (15.9)
	one child	705 (26.1)	493 (28.2)	1198 (26.9)
	Two children or more	932 (34.5)	826 (47.2)	1758 (39.5)

In the variable number of children, 786 people are single.

FAMD methods was very high. The highest correlation was related to the percentiles obtained through PCA and FAMD methods with a value of 0.994.

Table 2. Correlation coefficients of SES percentiles

Methods	PCA	MCA	FAMD
PCA	1	0.982	0.994
MCA	0.982	1	0.990
FAMD	0.994	0.990	1

According to Table 3, the ICC value in the Average Measures line is very high. The number 0.996 shows that the percentile reliability obtained through these 3 methods, when their average is examined, is completely acceptable. According to the p-value, the null hypothesis, that is, the ICC is equal to zero, is rejected. Therefore, ICC is significantly greater than zero.

Table 3. Intraclass Correlation Coefficient

	Correlation coefficient 95% (CI)	F statistic	P value
Single Measures	0.989 (0.988, 0.989)	264.941	< 0.001
Average Measures	0.996 (0.996, 0.996)	264.941	< 0.001

In Figures one to three, Bland-Altman plots are shown. According to these diagrams, the limits of agreement for PCA and MCA methods are between -10.8 and 10.8, for PCA and FAMD methods between -6.4 and 6.4, and for MCA and FAMD methods between -8.10 and 8.10. Also, in all the plots, the average line and the loess regression coincided.

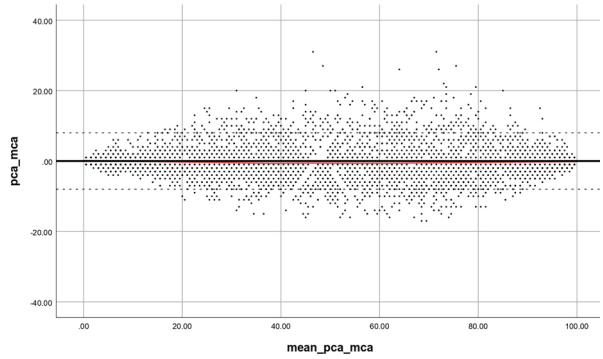

Figure 1. Difference between PCA and MCA methods

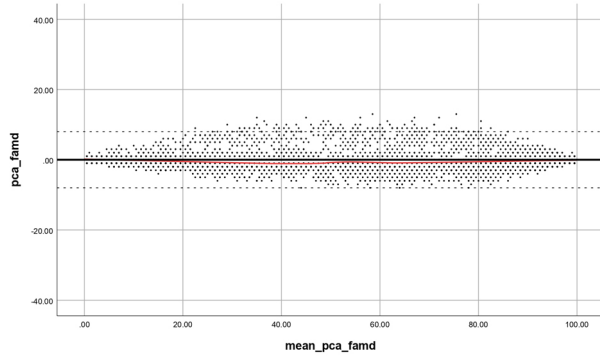

Figure 2. Difference between PCA and FAMD methods

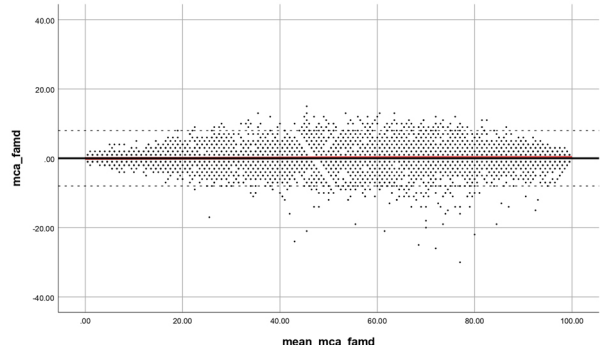

Figure 3. Difference between MCA and FAMD methods

Figures 4 to 6 show the association between the percentiles calculated through the methods was drawn against each other, which indicated a strong linear relationship. Also, the regression line corresponds to the median.

Additionally, the ICC value of 0.996 was obtained in a random sample of 250, and the

ICC value of 0.994 was obtained in the random sample of 100.

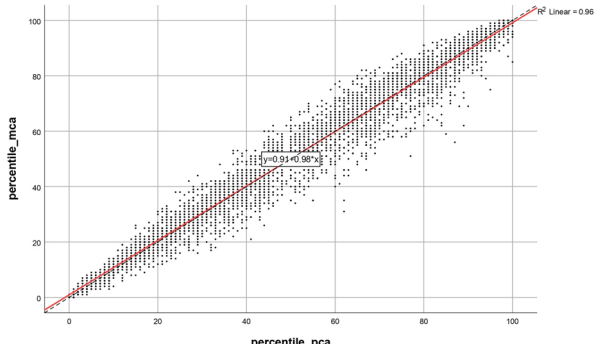

Figure 4. The relationship between percentiles calculated through PCA and MCA methods

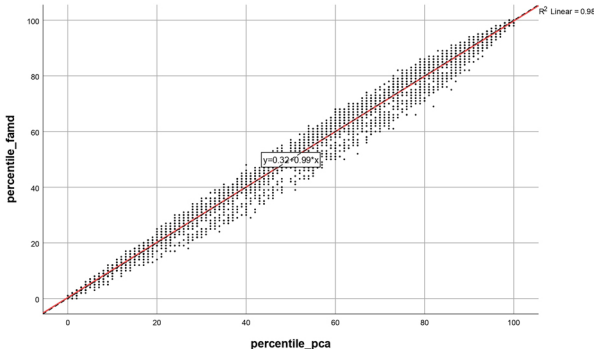

Figure 5. The relationship between percentiles calculated through PCA and FAMD methods

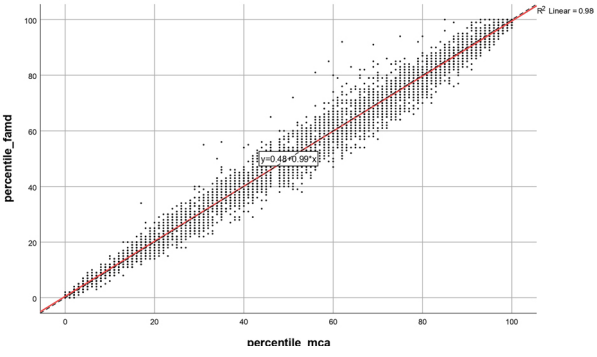

Figure 6. The relationship between percentiles calculated through MCA and FAMD methods

Discussion

Based on this study which was conducted on 4448 people and according to the values of correlation coefficient, intraclass correlation coefficient and Bland-Altman plots, it was concluded that PCA, MCA and FAMD methods in determining the SES of people and their percentage calculation has a high agreement and they work the same. Also, the highest level of agreement was related to the agreement between PCA and FAMD methods. According to the Bland-Altman plots, these two methods had the lowest amount of difference, which showed a maximum of 6 percent difference. The highest amount of difference was related to PCA and MCA methods, which showed a maximum of 11 percent difference. The results obtained from 250 and 100 samples, which were randomly selected from the original data, confirmed the agreement of the mentioned methods and the stability of the results against changes in the sample size was proved. Therefore, the obtained results were not affected by the high amount of data.

Examining the indicators of the SES is very important, because these indicators can help in understanding the SES of the society. By examining these indicators, it is possible to identify the problems and deficiencies in the socio-economic system and consider their improvement in order to help the economic growth of the society, because economic growth has been one of the effective methods to increase welfare.[12] Also, socio-economic factors are among the factors affecting health and food consumption, so that people who are in low SES groups receive fewer calories.[13] On the other hand, studies show that the existence of financial problems and the low level of

education can cause the spread of poverty in society.[14] Variables such as economic growth, unemployment level and inflation have had a significant impact on well-being, which shows that the country's economic policy makers should provide a suitable platform for increasing the level of economic growth in Iran by identifying the SES and the variables affecting growth.[15] Therefore, the examination of these indicators helps to design economic and social development programs more accurately and the society moves towards sustainable development. As a result, checking the SES indicators is very important and should be given enough attention.

Some SES questionnaires have both quantitative and qualitative variables. In some studies have been conducted for determining the SES, only the property of each person, which is in the form of binary variables, has been introduced as an indicator of the SES, and the PCA method has been used to create this index. In a cross-sectional study with the aim of measuring the SES and comparing it in eight countries, the PCA method was used to create an index of the SES. In this study, a subset of dual factors including assets, housing and facilities were selected using Cronbach's alpha coefficient and this method was performed on the correlation matrix of the selected factors. The first main component was defined as a criterion for defining the SES of each household.[16] In the current research, the first principal component was considered as an indicator of SES, and there were asset variables and housing characteristics in the set of examined variables. But the MCA method is used to analyze the categorical variables, so if questions have two or more levels, this method can be used. In a study conducted in 2022, the MCA method was used to classify SES into three classes: low, middle and high. It is mentioned in this study that the MCA method has become a more popular method for determining the SES, because it creates greater weight for assets compared to the PCA method. In this study, a set of variables including household property (personal car, pets), housing characteristics (type of housing, rental status, availability of electricity) and sources of drinking water were included to define the SES index of households. High scores were interpreted as higher SES and low scores were interpreted as low SES.[17] But based on the results of this study, PCA and MCA methods work similarly in determining the SES and are not superior to each other. If the SES questionnaire has both quantitative and qualitative variables, the use of PCA and MCA methods may be pose problems, and it also requires spending more time for data analysis. Considering that all variables must be categorical in the PCA method, qualitative variables that have more than two levels must be converted into binary variables. Also, considering that the variables must be qualitative in the MCA method, all the quantitative variables must be converted into categorical variables and then into binary variables before the analysis. Therefore, the FAMD method, which does not require converting data (quantitative to qualitative or vice versa), was considered. In a study to measure SES, an index reflecting education level, occupation and income was introduced and the FAMD method was used to analyze this index, which has quantitative and qualitative variables, with education and occupation as classification variables and income as a quantitative variable. The first dimension of FAMD, which explains 30% of the total

variance, was introduced as normalized scores. A higher score indicated a higher contribution of SES for income and each category of education and occupation.[18] In another study, this method was used to calculate the SES score, due to the combination of quantitative and classification variables, and these variables were used to create the score. Finally, the first dimension, which explained 19.7% of the variance, was retained in the final model.[19] In this research, the variable of education level was used, but income was not included due to potential participant bias or reluctance to disclose this information. Using the FAMD method can be useful in determining the SES. In this method, the desired data matrix is first prepared by applying pre-processing methods like standardization. Then, the main influencing factors in the data are extracted. In the next step, using the FAMD method, these factors are used to determine the SES. By using this method, people can be placed in different groups based on various variables such as income, education level, occupation, size of residential house, etc. and their SES can be determined. This method is widely used as a powerful analytical tool in many research and scientific fields. The main purpose of this article was to compare the PCA, MCA and FAMD methods in determining the SES in order to check the similarity or difference between these methods in determining the SES. We did not find a study that compared these three methods together and checked their agreement. In a study that aims to compare PCA and MCA methods using biomechanical signals and the example of steering wheel rotation, it is stated that the PCA method shows linear relationships and MCA is a more suitable option for showing more complex relationships. However, the implementation of the MCA method is more complicated than the PCA method and requires more calculations, as a result, more time is spent on the analysis. Also, the MCA method requires more space to display the results in the form of charts, but this method can be considered important because of its ability to reduce the dimensions and display of questionable data. In this study, the differences between the two methods were observed in the way that the variance of the principal components was higher in the PCA method than in the MCA method. In addition, in the MCA method, the variance of the main dimensions is not a suitable criterion for the displayed information.[20] It is suggested to investigate the effect of the number and type of variables on the performance of these methods in future studies. Because examining the number of variables can provide important information about data coverage and the information that can be extracted. By examining the number of variables, it is possible to choose the best method for data analysis and ensure data completeness. It is also possible to design a questionnaire specifically to determine the SES of people, which includes both quantitative and qualitative variables in the same way.

The implementation of these methods has limitations. In this research, there were 37 primary variables that included a wide range of variables such as age, gender, total price of household cars, education level, etc., but three classification variables including marital status, number of children, and housing ownership status have more than two levels, which were first converted into dummy variables and then the PCA method was implemented. So 44 variables were entered for analysis. This problem can be considered as a limitation in

the implementation of this method. Before implementing the MCA method, quantitative variables that included age, price of cars, etc. were defined in a categorical manner. One of the problems of this method is that each level of classification variable is entered as a new variable in the analysis. As a result, additional information is included in the study. In this research, there were 37 primary variables, and according to the above point, 110 variables were included in the analysis during the implementation of the method, which shows that a lot of additional information was entered. However, to implement the FAMD method, variables do not need to be converted (quantitative to qualitative or vice versa), so due to the existence of 37 primary variables, 37 variables were included in the analysis, which can be the advantage of this method over PCA and MCA methods. Converting variables can be prone to errors and time-consuming, it also prevents the entry of additional information into the study.

Conclusion

The compatibility of PCA, MCA and FAMD methods was accepted according to ICC values and Bland-Altman diagrams. Therefore, in case of large sample size and a various types of variables, the performance of the three methods to determine the SES are not significantly different, but the PCA and FAMD methods has more agreement. We saw the same results in the small sample size, as well. Therefore, all three methods agree on the number of different samples, and the choice of analysis method depends on the researchers. But, due to the advantages of FAMD method, it is suggested to use FAMD method for analysis in studies that have both types of quantitative and categorical variables

Acknowledgment

This article relates to the thesis of "Application of Factor Analysis of Mixed Data to determine the SES and its relationship with the quality of life of employees in the cohort data of staff health of Tehran University of Medical Sciences" for a Master`s degree in Biostatistics (REC approval ID: IR.TUMS.SPH.REC.1402.001) at the School of Public Health of Tehran University of Medical Sciences (TUMS). We thank the TEC cohort study of the Tehran University of Medical Sciences for providing the data and permission to use it to write this article.

Conflict of Interest

The authors declare that they have no conflict of interests.

References

1. Kagamimori S, Gaina A, Nasermoaddeli A. Socioeconomic status and health in the Japanese population. Social science & medicine (1982). 2009;68(12):2152-60.

2. Marmot M. Inequalities in health. The New England journal of medicine. 2001;345(2):134-6.

3. Clegg LX, Reichman ME, Miller BA, Hankey BF, Singh GK, Lin YD, et al. Impact of socioeconomic status on cancer incidence and stage at diagnosis: selected findings from the surveillance, epidemiology, and end results:

National Longitudinal Mortality Study. Cancer causes & control : CCC. 2009;20(4):417-35.

4. Health inequities and their causes 22 February 2018 [Available from: https://www.who.int/news-room/facts-in-pictures/detail/health-inequities-and-their-causes.

5. Maryam N, Mohammad Javad G, Amir T, Zahra G, Narges G. Socioeconomic status of patients with pemphigus vulgaris. Journal of Biostatistics and Epidemiology. 2017;3(1).

6. Zandian H, Olyaeemanesh A, Takian A, Hosseini M. Contribution of Targeted Subsidies Law to the Equity in Healthcare Financing in Iran: Exploring the Challenges of Policy Process. Electronic physician. 2016;8(2):1892-903.

7. Braveman PA, Cubbin C, Egerter S, Williams DR, Pamuk E. Socioeconomic disparities in health in the United States: what the patterns tell us. American journal of public health. 2010;100 Suppl 1(Suppl 1):S186-96.

8. Nouraei Motlagh S, Asadi Piri Z, Asadi H, Bajoulvand R, Rezaei S. Socioeconomic status and self-rated health in Iran: findings from a general population study. Cost effectiveness and resource allocation : C/E. 2022;20(1):30.

9. Traissac P, Martin-Prevel Y. Alternatives to principal components analysis to derive asset-based indices to measure socio-economic position in low- and middle-income countries: the case for multiple correspondence analysis. International journal of epidemiology. 2012;41(4):1207-8; author reply 9-10.

10. Vyas S, Kumaranayake L. Constructing socio-economic status indices: how to use principal components analysis. Health Policy Plan. 2006;21(6):459-68.

11. Mundfrom D, Shaw D, Ke T. Mundfrom DJ, Shaw DG, Lu Ke TMinimum sample size recommendations for conducting factor analyses. Int J Testing 5:159-68. International Journal of Testing. 2005;5:159-68.

12. Abida Z, Sghaier IMJZIRoE, Business. Economic growth and income inequality: Empirical evidence from North African countries. 2012;15(2):29-44.

13. Seyyed Reza S, Hassan E-Z, Arezoo R. Socioeconomic Status and Changes in Iranian Household Food Basket Using National Household Budget and Expenditure Survey Data, 1991-2017. Iranian Journal of Public Health. 2022;51(4).

14. Fleisher B, Li H, Zhao MQ. Human capital, economic growth, and regional inequality in China. Journal of Development Economics. 2010;92(2):215-31.

15. Mohammad Nabi Shahiki Tash SM, Khadijeh Dinarzehi. Examining the Relationship between Economic Growth and Coefficient of Social Welfare under the Bayesian Approach in Iran. SID. 2014.

16. Psaki SR, Seidman JC, Miller M, Gottlieb M, Bhutta ZA, Ahmed T, et al. Measuring socioeconomic status in multicountry studies: results from the eight-country MAL-ED study. Popul Health Metr. 2014;12(1):8.

17. Were V, Foley L, Turner-Moss E, Mogo E, Wadende P, Musuva R, et al. Comparison of household socioeconomic status classification methods and effects on risk estimation: lessons from a natural experimental study, Kisumu, Western Kenya. International Journal for Equity in Health. 2022;21(1):47.

18. Wang Y, Wang Y, Xu H, Zhao Y, Marshall JD. Ambient Air Pollution and Socioeconomic Status in China. Environmental health perspectives. 2022;130(6):67001.

19. Pooseesod K, Parker DM, Meemon N, Lawpoolsri S, Singhasivanon P, Sattabongkot J, et al. Ownership and utilization of bed nets and reasons for use or non-use of bed nets among community members at risk of malaria along the Thai-Myanmar border. Malaria journal. 2021;20(1):305.

20. Loslever P, Schiro J, Gabrielli F, Pudlo P. Comparing multiple correspondence and principal component analyses with biomechanical signals. Example with turning the steering wheel. Computer methods in biomechanics and biomedical engineering. 2017;20(10):1038-47.

Supplementary Material

The PCA method is one of the oldest and most widely used methods that reduces the dimensions of the dataset and tries to preserve the information, and expresses the importance of variables affecting a phenomenon. In this method, the initial datasets are transformed into new variables without correlation, so that the created components are a weighted linear combination of the initial variables and maximize the variance. By using the matrix of eigenvalues, the main components are made as a linear combination of primary variables and are independent of each other and are used instead of primary variables in data analysis. Each principal component can be specified by the following sequence:

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + \ldots a_{ip}X_p$$

where Z_i is the desired component, a_{ij} are the coefficients of the primary variables and X_i is the primary variable. The coefficients related to the primary variables are obtained by solving the following equation:

$$|R-\lambda I|=0$$

where I is the unit matrix, R is the correlation matrix between the primary variables and λ is the eigenvalues. From these eigenvalues, eigenvectors are obtained.[1-3]

Also, many variables are included in this method, which may have different measurement units. Because this method depends on the measurement units, which means that the variable that has more variance has more influence on the final results, the standardization method is used to solve this problem.

In recent years, the PCA method has been widely used to construct the socio-economic status variable in studies related to this issue. The socio-economic status index for person *i* is a linear combination of the primary variables as follows by the PCA method based on the first principal component:[4]

$$Z_i = a_1\left(\frac{x_1-\overline{x_1}}{s_1}\right)+a_2\left(\frac{x_2-\overline{x_2}}{s_2}\right)+\ldots+a_k\left(\frac{x_k-\overline{x_k}}{s_k}\right)$$

Where $\overline{x_k}$ and s_k are the mean and standard deviation of the primary variable x_k and a_k indicate the weight for each variable x_k in the first principal component.

The weaknesses of this method include the necessity of normalizing variables, considering only linear relationships between variables, and some information may be lost when reducing data dimensions, which can lead to data distortion, for example, some patterns which can be seen in higher dimensions. Also, if the correct number of principal components that explain sufficient variance in the data set is not selected, it can lead to information loss.

This method is used to analyze quantitative data, therefore considering that there were

qualitative questions in the questionnaires of this study before calculating the socio-economic status score of the people, these questions were first Binary mode was converted and then the analysis was done. Because it is not possible to implement the PCA method on classification variables, because its values are not numerical and it is not possible to compare the variance. Therefore, converting categorical variables into a sequence of binary variables with values of 0 and 1 is one of the ways to perform PCA in a dataset with categorical variables.

In the MCA method, it is possible to study the relationship between two or more qualitative variables and complex data sets are simplified by reducing the number of variables in order to discover the patterns in the data. Each qualitative variable has several levels and each of these levels is coded as a binary variable. The data table consists of binary columns of which only one column has the value 1 for each qualitative variable. The main goal in this method is to study the variability of people from a multidimensional perspective.

The first step in the implementation of this method is to calculate the probability matrix in the form of $Z=N^{-1}X$, where N represents the total of the primary data table and X is the $I\times J$ indicator matrix (if a data set includes I observe and k nominal variables, each nominal variable has J_k levels and the sum of J_k is equal to J):

r represents the vector of the sum of rows z and c the vector of the sum of columns z, and we have $D_r=diag\{r\}$ and $D_c=diag\{c\}$ where diag represents the diagonal matrix.

The principal components are obtained from the following singular value decomposition:

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

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

One of the limitations of this method is the inability to interpret the relationships between variables. Because in this method, data normalization is not done, as a result, we are not allowed to compare the relationships between variables, and the raw data must be checked. For this reason, the risk of data misinterpretation increases. Another weakness of this method is that each level of classification variable is entered as a new variable in the analysis. As a result, additional information is included in the study. In this method, the emphasis is on non-linear relationships. Also, some information may be lost when reducing the dimensions of the data.

Considering that there were quantitative questions in the questionnaires of this study, before calculating the socio-economic status score of the people, these variables were first converted into classification mode and each category of these variables was binary coding and then analysis was done.

Unlike the PCA and MCA methods, the FAMD method is used to analyze a data set that has both quantitative and qualitative variables at the same time, and there is no need to convert qualitative variables into quantitative or vice versa. Also, in this method, it is possible to analyze the similarity between people by considering different types of variables.

We have a data set containing I person with weight p_i so that $\sum_i p_i=1$. To simplify the explanation, it has been assumed that all people have the same weight. People are described by the following variables:

1. Quantitative variables K_1 are standardized variables that, due to the presence of two types

of variables in the study, standardization of variables is required.

2. Categorical variables *Q*, category variable q with K_q and sum of all categories are displayed as $\sum_q K_q = K_2$. Also, p_{kq} shows the proportion of people with category k_q.

Therefore, $K=K_1+K_2$ shows the total sum of quantitative and qualitative variables.

In this method, the balance between the influences of both types of variables in these relationships should be ensured. The effect of the variable is measured by its participation in the variance of the points. If $G_s(k)$ is the coordinate of column k on the rank axis of s, then a quantitative variable:

$$G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_i p_i x_{ik} F_s(i) = r(k, F_s)$$

where $F_s(i)$ represents the coordinates of individual i on the axis of rank s.

Also, qualitative variable q with category k_q and weight p_{kq}:

$$G_s(k_q) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{p_{kq}} \sum_i p_i y_{ik_q} F_s(i) = \frac{1}{\sqrt{\lambda_s}} F_s(k_q)$$

where $F_s(k_q)$ is the coordinate of the center of gravity of people with category k_q along the rank axis s.

The following relationship defines the person's position according to the categories he has and this relationship is the foundation of the interpretation of the FAMD method:

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{k \in K_1} x_{ik} G_s(k) + \frac{1}{\sqrt{\lambda_s}} \sum_{k_q \in K_2} p_{kq} \left(\frac{y_{ikq}}{p_{kq}} - 1 \right) G_s(k_q)$$

The first part of the relationship above belongs to the PCA method, which indicates that a person is placed alongside variables for which they have values higher than the mean and opposite variables for which they have values lower than the mean. The second part of this relationship belongs to the MCA method.

Therefore, considering that the principle of FAMD method is based on balancing the influence of continuous variables and classification, variables are weighted in a way that each variable of both types participates equally in the construction of the main components. The approach used in PCA and MCA methods to deal with missing values involves ignoring them, but in FAMD method, it is possible to predict missing values using inter-individual similarities and relationships between variables simultaneously, i.e., between continuous variables, classification variables, and different types of variables.[5, 6]

Another advantage of FAMD method is that it does not encode continuous variables based on the levels of classification variables, in other words, each variable is studied without modification according to its own type.[7] Therefore, there is no need to convert variables (quantitative to qualitative or vice versa) to implement this method. However, this method also has limited weaknesses. For example, the requirement of normalization of quantitative variables and the fact that this method is not suitable for small data sets because it requires a large number of observations and variables to be effective are among the weaknesses of this method. Considering the above points and weaknesses of PCA and MCA methods, using FAMD method in determining the socio-economic status of individuals may perform better than PCA and MCA methods.

Considering that PCA, MCA and FAMD methods are some kind of factor analysis methods, the minimum sample size required

to implement this method is 3 to 20 times the number of variables and in the range of 100 to more than 1000.[8] Therefore, the minimum sample size required for the implementation of these methods, i.e. 100, was selected, and the sample size of 250 was selected to further check the reliability of the results and pay attention to the mentioned range. In these two random samples the agreement of the three methods were analyzed together to check whether the large amount of data affected the results or not.

Intraclass correlation coefficient (ICC)

ICC was used to examine the agreement of percentiles obtained through three methods. ICC is used when we want to measure the degree of agreement or correlation between two or more measurement scales that include a group of people or objects, and it is usually recommended to evaluate the reliability of measurement scales. Also, ICC is stable against various statistical assumptions such as normality and variance.[9] One of the reasons for the popularity of this method is its easy interpretation.10, 11 In the ICC table we can see single measures and average measures. Single measures shows the reliability of the ratings for single rater. On the other hand the average measures shows the reliability of different raters averaged together.

Bland-Altman plot

Bland-Altman plots are used to analyze the level of agreement between the two methods. In this plot, the difference between two methods is plotted against their average. The horizontal line and the X axis in this graph are always the average, and the Y axis shows the difference between the two methods. It also shows the upper and lower limits of differences (limits of agreement). In these graphs, the average is plotted on the X-axis, which shows the zero value, and loess regression was used to check that the amount of agreement changes significantly during the scores of individuals. It is important to use loess regression to check the agreement of the methods. Due to the complexity of the data and the need for different settings, loess regression has a high adjustability that allows the researcher to choose the best settings to check the agreement of the methods. By using a local weight function, loess regression can respond more accurately to noisy data, which is very important in checking the agreement of the methods. Considering the complexity of the relationship between variables in checking the agreement of the methods, the ability to use loess regression in non-linear data can help the researcher to choose the best method for checking the agreement. Also, due to the complexity of the relationship between variables, loess regression can help the researcher in detecting non-linear parameters. In general, loess regression does not depend on specific assumptions about data distribution, so it can work well in cases where specific assumptions cannot be defined.[12] By using loess regression, the changes in each part of the data can be described point by point.

References

1. Tabachnick BG, Fidell LS, Ullman JB. Using multivariate statistics: pearson Boston, MA; 2013.

2. Ouyang Y. Evaluation of river water

quality monitoring stations by principal component analysis. Water research. 2005;39(12):2621-35. doi:10.1016/j.watres.2005.04.024.

3. NOURI RE, Ashrafi K, Azhdarpour AA. Comparison of ANN and PCA based multivariate linear regression applied to predict the daily average concentration of CO: A case study of Tehran.

4. Ranjbaran M, Soori H, Etemad K, Khodadost MJJoJUoMS. Relationship between socioeconomic status and health status and application of principal component analysis. 2015;1(1):9-19.

5. Audigier V, Husson F, Josse J. A principal component method to impute missing values for mixed data. Advances in Data Analysis and Classification. 2016;10(1):5-26. doi.org/10.1007/s11634-014-0195-1.

6. Wold HJMa. Estimation of principal components and related models by iterative least squares. 1966:391-420.

7. Bertrand F, Maumy M, Fussler L, Kobes N, Savary S, Grosman JJCSIB, Industry, et al. Using factor analyses to explore data generated by the National Grapevine Wood Diseases survey. 2007;1(2):183-202.

8. Mundfrom DJ, Shaw DG, Ke TLJIjot. Minimum sample size recommendations for conducting factor analyses. 2005;5(2):159-68.

9. Bobak CA, Barr PJ, O'Malley AJ. Estimation of an inter-rater intra-class correlation coefficient that overcomes common assumption violations in the assessment of health measurement scales. BMC Medical Research Methodology. 2018;18(1):93. doi.org/10.1186/s12874-018-0550-6.

10. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation. 2010;19(4):539-49. doi:10.1007/s11136-010-9606-8.

11. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. COSMIN checklist manual. 2012.

12. Jacoby WG. Loess:: a nonparametric, graphical tool for depicting relationships between variables. Electoral Studies. 2000;19(4):577-613.