Faculty of Science - University of Benghazi

Libyan Journal of Science & Technology

journal home page: www.sc.uob.edu.ly/pages/page/77

Using Multiple Factor Analysis to Study Students' GPA: Case Study Faculty of Science, University of Benghazi.

Osama H. Othman [a], Rami S. Gebril [b], Adel M. AlSharkasi [b,*], Ashraf A. Younis [b]

[a]*Department of Statistics, Faculty of Arts and Science, University of Ajdabiya, Ajdabiya, Libya.*

[b]*Department of Statistics, Faculty of Science, University of Benghazi, Benghazi, Libya.*

Highlights

- **The Multiple Factor Analysis is proved to be a powerful tool for variable selection, especially when applied through multiple sequential stages in large datasets.**
- **Both the Multiple Factor Analysis and the Block Principal Component Analysis showed very similar good results using the Students' GPA data.**
- **The amount of variation laying within a large number of variables of a dataset can be summarized in a much smaller number of variables without losing most of the information hidden in that dataset.**

ARTICLE INFO

Article history:

Received 02 May 2020
Revised 02 June 2021
Accepted 03 June 2021

Keywords:

Multiple Factorial Analysis, Principle Component Analysis, Data Reduction, Variable Selection.

**Address of correspondence:*

E-mail address: adel.alsharkasi@uob.edu.ly

A. M. AlSharkasi

ABSTRACT

Data or dimensionality reduction is very common and critical in statistics because, in the model building phase in any statistical research or study, the researchers are always annoyed by the number of variables to be included or excluded in a model. In this sense, Multiple Factor Analysis (MFA) can be employed to select variables and reduce data. In other words, it can be used to know the variables that have the largest amount of variation in the data understudying. The aim of this paper is to study the students learning behavior to find out the variables that have the largest amount of variation by applying the Multiple Factor Analysis (MFA) to the students' GPAs in different departments in the Faculty of Science-University of Benghazi and to find out whether the variables (courses), in the database, have a considerable variation that worth studying and investigating. The variables considered in this study are the different courses in each department and the criteria used to measure these courses. The data collected in this study also contains information related to the course subjects and grading systems in these departments. The results of this study show that fifteen variables (Courses) absorb most of the variation in the data sets. The finally selected fifteen variables vary in the amount of the distributed variation. The results also show that the variables used in this study correspond to some of the variables and data reduction used in Othman and Gebril, 2014, using Block Principle Component Analysis (BPCA). It is hoped that the findings of this study will be a useful contribution to our faculty to improve teaching methodology and provide these different departments with modern educational facilities.

1. Introduction

Data or dimensionality reduction is a very common, desired, and critical topic in statistics because, in the model building phase in any statistical research or study, researchers are always annoyed by the number of variables to be included or excluded in such a model. However, data reduction techniques can be applied to reduce data set to a smaller size and to maintain most of the original data. Moreover, mining on the reduced data set should be more efficient to produce the same or approximately the same results.

Multiple Factor Analysis (MFA) is a recent technique that originated from the work of the French statisticians Escofier and Pagès (Escofier and Pagès, 1994). MFA has been applied extensively in many areas and disciplines such as psychology, food quality, and tests of suiting materials. In addition, it has been applied with other multivariate techniques in many other applications. Pagès, 2004 applied MFA to sensory data to show the method clearly with its main properties and an application to sensory data. His result concluded that MFA has two main features: The balancing of the sets of variables and Output septic of the partition of the variables in different sets. Mainly, 1) The superimposed representations of individuals and categories and 2) The groups' representation. Abdi *et al.*, (2013) showed clearly the origin and the use of MFA with a comprehensive empirical example for computations and partition.

Multiple Factor Analysis can be used to select variables and reduce data. In other words, it can be used to know the variables that have the largest amount of variation in the underlying data set. Thus, we decide, in this study, to use the students' GPAs that are available in the Faculty of Science Database. The reason for this decision is to see the problems that would face the administration of this faculty and to investigate whether the courses (variables) in this database have any considerable variation that worth studying.

In our previous work (Othman and Gebril, 2014), the Block Principle Component Analysis technique was applied to the same data set and the results show that the technique can be relied on to reduce a large number of variables to a smaller number that reflects a largest extracted percentage of variation.

ISSN: Online 2663-1407, Print 2663-1393; National Library of Libya, Legal number: 390/2018

2. Methodology

Multiple Factor Analysis, or Multiple Factorial Analysis, can be considered as a generalization of Principle Component Analysis (PCA) that has the advantage of simplicity and easiness in practice. The main idea of MFA is to give the most important variables among many data sets that have different variables measured on the same observations or vice versa (Pagès, 2014).

MFA consists of five different sequence steps. These steps are: i) Collecting on the same observations of the dataset (the ***K*** departments in our data), where $X = \left[\left|X_{[1]}\right| \left|X_{[2]}\right| \ldots . \left|X_{[K]}\right|\right]$, ii) Computing generalized PCA on each of the ***K*** department individually $X_{[i]} = \mathrm{U}_{[i]}\Gamma_{[\mathrm{i}]}\mathrm{V}_{[\mathrm{i}]}^{\mathrm{T}}$ with $\mathrm{U}_{[\mathrm{i}]}^{\mathrm{T}}\mathrm{U}_{[\mathrm{i}]} = \mathrm{V}_{[\mathrm{i}]}^{\mathrm{T}}\mathrm{V}_{[\mathrm{i}]} = \mathrm{L}$, iii), Normalizing each department by dividing its first singular value (Eigenvalue) λ, to make the departments comparable, iv) Merging the ***K*** normalized departments into one department and v) Computing a generalized PCA on the final aggregated data set $\mathrm{X} = \mathrm{P}\Delta\mathrm{Q}^{\mathrm{T}}$ with $\mathrm{P}^{\mathrm{T}}\mathrm{MP} = \mathrm{Q}^{\mathrm{T}}\mathrm{AQ} = \mathrm{L}$ (for more details see (Abdi *et al.*, 2013)).

In these steps, PCA was applied twice in steps (ii) and (v). In these two steps, the first two principal components (PC's) were utilized in graphical representation with the selected variables. Presenting and displaying the chosen variables graphically through Biplot is of paramount importance in PCA to have a deep insight into the hidden and vague relationships between the selected variables and respective PCs (Jolliffe, 2002). The MFA has its objectives, but in this paper, it is utilized in variable selection and data reduction tasks to achieve a new objective.

3. Data and Results

The data used in this study is the students' GPAs collected from different departments in the Faculty of Science at the University of Benghazi, as shown in Table 1.

4. Data description

Table 1 shows the construction of the data matrix where its columns contain the courses (variables) and the rows contain the semesters (cases). Our empirical analysis is performed on forty-one semesters from the Fall semester (to be referred to by S) of the academic year (1990/1991) to the Spring semester of the academic year (2009/2010). As for the number of variables, just 251 out of 616 courses were chosen as they were the most studied courses in the whole duration of the study. The courses, which were neglected, were not studied as much during the period of the study. The process of variable coding was easy and explained clearly, because what wanted is a way by which it can manage and distinguish between different variables**;** every variable has three indices as follows:

- The First index refers to the serial number of the variable and the count is from 1 to 251.
- The Second index, which consists of one letter, refers to the Department by the first letter as following: M: Mathematics. C: Chemistry. S: Statistics. G: Geology. P: Physics. B: Botany. Z: Zoology.
- The third index refers to the semester in which that course must be taken or studied.

Table1

Description of the used data matrix

	Courses						
Semesters	X_{1g1}	X_{2g1}	X_{3g2}	X_{4m3}	X_{6m1}	…………..	X_{251c6}
1990-1991 Fall	1.27	2.05	1.23	1.79	1.68	. . . . .	1.27
1990-1991 Spring	1.42	1.85	1.36	2.45	0.67	. . . . .	1.42
1991-1992 Fall	2.19	1.72	2.85	2.21	2.43	. . . . .	2.19
1991-1992 Spring	1.97	1.75	2.32	2.14	2.07	. . . . .	1.97
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
2009-2010 Spring	1.65	1.64	2.49	0.95	1.62	. . . . .	1.65

5. Empirical Results

In this subsection, our results with interpretations are presented and summarized in graphs and tables. Table 2 shows the number of retained PC'S and selected variables after performing PCA on each department before and after normalizing the departments. It can be noted that the number of retained PC'S from these departments are equal; the number of selected variables is equal in some departments such as Mathematics, Zoology, and Physics. Table 3 presents the selected variables from the departments before and after normalizing.

Table 2

Some selected PCs and variables before and after normalization.

Department		Chemistry	Mathematics	Statistics	Botany	Zoology	Physics	Geology	Total
selected PCs	Before	8	9	7	11	9	10	9	63
	After	8	9	7	11	9	10	9	63
selected variables	Before	10	9	8	8	10	9	8	62
	After	7	9	7	9	10	9	9	60

Table 3

The selected variables from each department before and after normalization.

Chemistry		Mathematics		Statistics		Botany		Zoology		Physics		Geology	
Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
X88	X88	X5	X5	X34	X34	X178	X169	X209	X209	X56	X56	X142	X142
X99	X99	X7	X7	X36	X36	X181	X178	X211	X211	X57	X57	X143	X143
X101	X101	X11	X12	X42	X47	X184	X181	X215	X215	X62	X62	X148	X148
X105	X105	X12	X13	X47	X50	X186	X184	X220	X220	X65	X67	X150	X150
X109	X118	X13	X21	X50	X51	X191	X186	X223	X223	X67	X70	X159	X159
X111	X129	X21	X22	X51	X52	X193	X191	X227	X227	X70	X72	X161	X161
X118	X136	X22	X25	X52	X55	X196	X193	X229	X229	X72	X74	X162	X162
X129		X25	X28	X55		X241	X196	X232	X232	X74	X75	X164	X164
X136		X28	X30				X241	X234	X234	X75	X77		X167
X239								X235	X235				

In addition, the PCA was used after normalizing each department to select variables. As well as an extra step in the context of variable selection has been added to make the research more meaningful and to extract more characteristics from the underlying database. The extra step involves dealing with the selected variables before normalization, which are 62 selected variables distributed on various departments differently. After normalizing the 62 selected variables first time and applying PCA, these variables have been merged into one data set that was normalized and analyzed by PCA to select the variables that have the largest amount of variation to be compared with the finally selected variables in the fifth and final step.

Table 4 shows the departments and the selected variables (courses) from each department with their codes and names, in order to explain the nature of each variable after normalizing the 62 variables one more time and applying PCA.

The selected variables, which are given in Table 4, are presented in Biplot to detect some patterns and explore some hidden features that help in the comparison with the finally selected variables in step (v).

Table 4. Names of selected variables after the second normalization.

Department	Codes of the variables	Courses
Chemistry	X101c8	Graduation Project
	X105c3	Inorganic Chemistry 1
	X136c6	Analytic Chemistry 5
Statistics	X34s6	Probability Theory
	X50s4	Applied Statistics
	X52s4	Sampling Techniques 1
Botany	X184b4	Algology
	X196b3	Research Methodology of Botany
	X241b8	Graduation Project
Physics	X57p7	Selected Topics
	X65p8	Solid State Physics
Geology	X143g7	Geophysics Project
	X164g1	Natural Geology

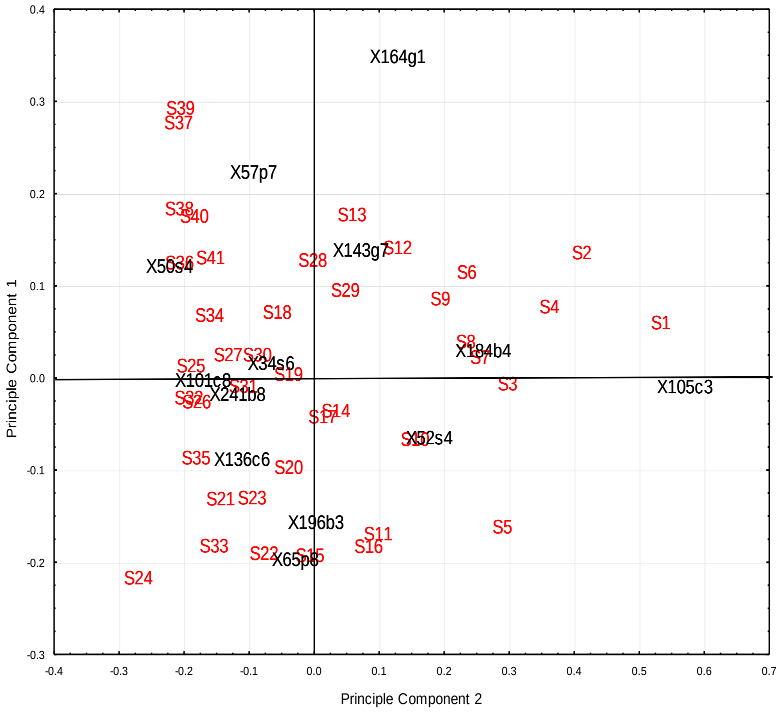

Fig. 1. Biplot of the selected variables after the second normalization.

The pictorial depiction of the Biplot, presented in Fig. 3, can display some relationships that can help in understanding the general pattern of those variables if compared with the fifteen finally selected variables. Some similar behaviors of the students in some courses and semesters can be seen in the second quarter between **X57p7**, **X50s4**, **S40**, **S41**, **S38,** and many other semesters. The odd behavior of students can be seen in many semesters and courses such as **X164g1**, **S5,** and **S24**.

6. General PCA

The step (v) suggested by Escofier and Pagès (1994) shows the application of generalized PCA on the last merged group which contains the selected normalized variables from the eight departments. Table 5 exhibits the selected variables resulted from applying the generalized PCA as it was done in step (iv) to introduce the names of the courses of these variables with their codes.

Table 5

Names of selected variables after applying general PCA.

Department	Codes of the variables	Courses
Mathematics	X30m3	Foundations of Mathematics
Statistics	X50s4	Applied Statistics
Botany	X196b3	Research Methodology of Botany
	X241b8	Graduation Project
Zoology	X235z3	Lower Invertebrates
Physics	X57p7	Selected Topics
	X62p6	Physics Laboratory 5
	X70p4	Non Relative Mechanics
	X72p3	Waves and Vibrations
	X74p8	Nuclear Physics 2
	X75p6	Nuclear Physics 1
	X77p1	General Physics 1
Geology	X143g7	Geophysics Project
	X162g8	Geological Field Work
	X164g1	Natural Geology

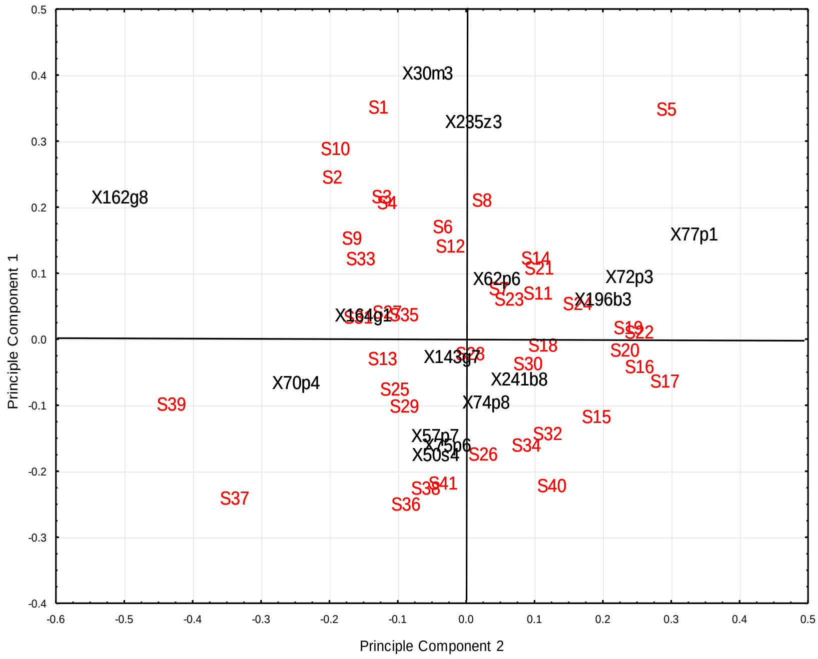

Fig 2. Biplot of the selected variables after applying general PCA.

The Biplot, which is given in Fig. 2, has great importance for the fifteen finally selected variables. From Fig. 2, it can be seen that many mutual relationships were detected by Biplot, in the sense that the relationship between variable (courses) and cases (semesters) can be extracted graphically and the relationships between variables and the first two PCs and cases with the first two PCs also between variables and cases.

By inspection at the first quarter in Fig. 2, it can be seen that the variables $\mathbf{X_{77p1}}$, $\mathbf{X_{72p3}}$, $\mathbf{X_{196b3}}$, and $\mathbf{X_{62p6}}$ are near to each other and near to the semesters $\mathbf{S_{24}}$, $\mathbf{S_{14}}$, $\mathbf{S_{11}}$, $\mathbf{S_{23}}$, and some other semesters. This proximity between those variables (courses) and cases (semesters) indicates that the behavior of students in those courses was very similar to average in those semesters. Those previous courses must have a positive relationship with the first principle (PC) component because they are in the first quarter. Some patterns can be explored as outliers, such as the cases $\mathbf{S_5}$, $\mathbf{S_{37}}$, $\mathbf{S_{39}}$ and $\mathbf{X_{162g8}}$; their behavior can be interpreted as odd, because there are no other semesters or courses near to them or similar to them. This means that the average results of students in those courses and semesters are different from any other course and semester.

7. Inertia Comparison

The inertia of the first principle components is made before and after the normalization process for all departments to check the amount of variation extracted and explained by the first PCs for each department and also to check whether the normalization process made any change to the amount of explained variation.

Inertia values can be obtained by displaying the Biplot graph to get the proportions of explained variation for the first principle component which is presenting in the horizontal axis. From Table 6, it can be noticed that the values of inertia are almost the same in many departments; Physics and Geology inertia's values are approximately the same, whereas some differences can be noticed in Chemistry and Mathematics. As for the other departments, a semi-consistency or similarity is obvious with regards to their inertia values.

Table6. Inertia comparison of first principle component before and after normalization for all departments.

	Departments						
	Chemistry	Mathematics	Statistics	Botany	Zoology	Physics	Geology
Before Normalization	54.72%	33.46%	48.12%	29.86%	35.13%	24.33%	31.40%
After Normalization	57.69%	29.88%	50.45%	31.06%	35.13%	24.76%	29.80%

8. Previous Study

The results of the present work agree with some results of our previous work (Othman and Gebril, 2014). The technique used was Block PCA, which was suggested originally by Zhang, *et al.* (2002). The last selected variables were twelve out of 616 variables representing the largest possible amount of variation among the other variables. The variables, their departments, and the names of courses found in the previous study by using the Block PCA technique are presented in Table 7.

The reason for presenting the previous study here is that, both studies are in the context of variable selection and data reduction for the same data. In this respect, some results of the previous study are similar to the results of this study when two different techniques were applied.

Table 7. Names of selected variables from previous study.

Department	**Codes of variables**	**Name of variables**
Physics	X57p7	Selected Topics
	X56p6	Quantum Mechanics 1
	X75p6	Nuclear Physics 1
Geology	X159g4	Exploration Geophysics
	X162g8	Geological Field Work
Botany	X186b4	Bacteriology
	X191b8	Molecular Biology
Mathematics	X5m8	Graduation Project
	X29m6	Mechanics 1
	X28m7	Mechanics 2
	X7m5	Independent Study
Chemistry	X105c3	Inorganic Chemistry 1

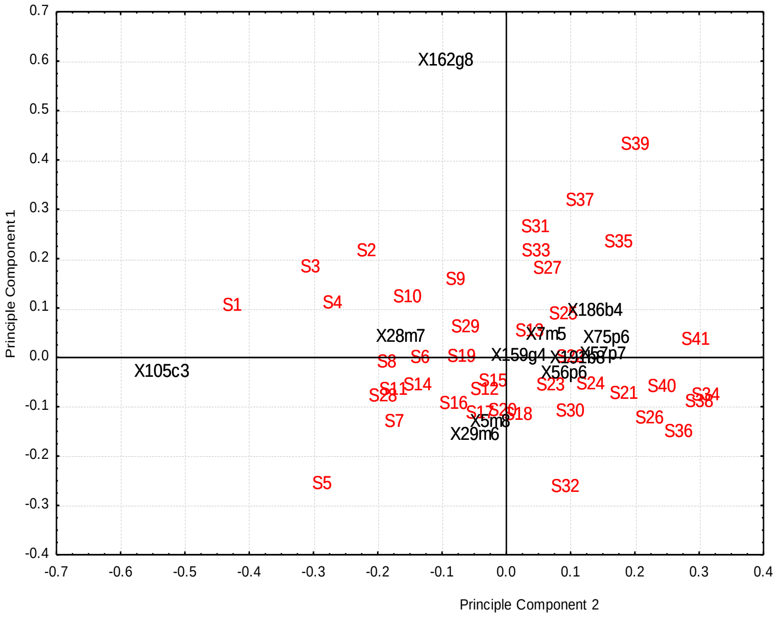

Fig 3. Biplot of the selected variables from the previous study.

From Fig. 3, it can be seen that there are some strange behaviors for the selected variables and selected cases in the previous study, as for the variables $\mathbf{X_{105c3}}$ at the top of the graph in the second quarter and $\mathbf{X_{162g8}}$ at left in the third quarter are outliers. The cases $\mathbf{S_5}$, $\mathbf{S_{32}}$ at the bottoms of the third and fourth quarter, and $\mathbf{S_{39}}$ at the top of the first quarter are also outliers. These outliers' variables and cases have some different and odd behavior compared to the others.

For instance, the variable $\mathbf{X_{105c3}}$ has a large negative correlation with all other variables because it stands alone opposing the other variables. This means that this course has marks in all semesters which are different from the other variables. As for the point (semester) $\mathbf{S_{39}}$, it has very low average scores with the variables in the second and the third quadrants except the only variable in the top left of the second quadrant which is $\mathbf{X_{162g8}}$. Case $\mathbf{S_5}$ has high average scores with the variables that are close to it, which are $\mathbf{X_{105c3}}$ and $\mathbf{X_{29m6}}$, but it has low average scores with the other courses that are far from it. In general, the interpretation of these outliers' courses and semesters may be due to the performance of the students which was different from other courses and semesters.

9. Conclusion

After applying the MFA technique with all its steps as well as the illustrated extra step and the subsection of the previous study, many comparisons were made in the context of variable selection and data reduction. By looking at the departments that have the selected variables, they were retained on three stages in this study including the previous study. Some resemblance between the selected variables in the three stages can be seen. The variables $\mathbf{X_{50s4}}$, $\mathbf{X_{196b3}}$, $\mathbf{X_{241b8}}$, $\mathbf{X_{57p7}}$, $\mathbf{X_{143g7}}$, and $\mathbf{X_{164g1}}$ are mutual between the second normalization process and the application of the general PCA on the concentrated department.

The similarities between the application of the general PCA on the concentrated department and previous studies by Block PCA can be seen in $\mathbf{X_{57p7}}$, $\mathbf{X_{75p6}}$, and $\mathbf{X_{162g8}}$. In general, the variable $\mathbf{X_{57p7}}$ is mutual between the three steps of variable selection. That the similarity between the three stages, for some selected variables, means that the three methods for variables selection are joint and agreed about the behavior of these mutual variables. MFA has the advantage of dealing with such large and extended databases easily like other techniques of variable selection and data reduction, and that was employed to reveal the variables that have the largest amount of variation among other variables to reveal the reasons for that sizable variation in future studies. Four variables out of 251 variables in common between Block PCA and MFA methods show that the two methods are approximately in the same direction and even though they have different criteria and steps in the application, the objectives and the tools are similar and that can be seen in the PCA technique which was used in both of them.

Acknowledgment

We acknowledge and appreciate the assistants obtained from Dr. Ahmed Abunowara in editing the final draft of this paper.

References

Abdi, H., Williams, L. J., and Valentin, D. (2013) 'Multiple factor analysis: principal component analysis for multiple and multiblock data sets', *WIREs Comp Stat*, 5, pp. 149–179. doi: 10.1002/wics.1246.

Escofier, B. and Pagès, J. (1994) 'Multiple Factor Analysis (MFULT package)', *Computational Statistics and Data Analysis,* **18**, pp. 121-140.

Jolliffe, I. T. *Principle Component Analysis*, Springer, New York, 2002, pp. 112-119.

Othman, O. H., and Gebril, R. S. (2014) 'Detecting most influencing courses on students grades using block PCA', *AIP Conference Proceedings*, 1635, 831; https://doi.org/10.1063/1.4903679

Pagès, J. (2014) Multiple Factor Analysis by Example Using R. New York: Chapman and Hall/CRC, https://doi.org/10.1201/b17700

Pagès, J. (2004) "Multiple Factor Analysis: Main Features and Application to Sensory Data', *Revista Colombiana de Estadi´stica*, 27(1), pp. 1-26.

Zhang, A. L. Y., Edmund, G., and Clarke, R. (2002) 'Statistics in Medicine', *Letters*, 21, pp. 3465-3474.