Principal Component Analysis Technique for Finding the Best Applicant for a Job: Case Study at Cihan University-Erbil

Abbood M. Jameel[1], Qusay H. Al-Salami[2]

[1]Department of Accounting, College of Administrative and Financial Sciences, Cihan University-Erbil,
Kurdistan Region, Iraq
[2]Department of Business Administration, College of Administrative and Financial Sciences, Cihan University-Erbil,
Kurdistan Region, Iraq

Abstract—**This paper focuses on the use of principal component analysis (PCA) technique in choosing the best applicant for a job in Cihan University-Erbil. Cihan University has a panel of judges (University staff) to help in choosing the applicants for a job by evaluating or rating each one on different scale of preference and different type of characteristics. This process usually creates complicated multivariate data structure, which consists of 25 applicants for a job rated by a panel of judges on 17 characteristics (25 rows, applicants, and 17 columns, characteristics). PCA plays a crucial role in conducting impactful research as it offers a potent technique for analyzing multivariate data. Researchers can utilize this method to extract valuable information that aids decision-makers in problem-solving. To ensure the appropriateness of data for PCA, certain testing procedures are necessary. In this study, two tests, namely, the Kaiser-Meyer-Olkin measure of sampling adequacy and Bartlett's test of sphericity, were performed, and their significance is vital. The findings indicate that the data employed in this research are suitable for PCA. Scoring and ranking procedures as extra tools were used to see that applicant No. (1) Is the first accepted for a job, applicant No. (17) Is the second, applicant No. (12) Is the third, and so on.**

Keywords—**Allocating scores and ranks, Eigen values and Eigen vectors, Matrices, Multivariate analysis, Principal component analysis.**

I. Introduction

Cihan University is a new private university in Erbil-Kurdistan. Cihan University was established in 2007 to help students to get B.Sc. degree in different fields. This university was started with three departments but now more than 25 departments. The university has a panel of judges (University staff) to help in choosing the applicants for a job by evaluating or rating each one on different scale of preference and different type of characteristics.

Assessing the employability of future students is crucial in education, and collaboration between the Institute of Higher Learning and career centers plays a key role in creating a successful plan. This project uses machine learning to predict employability based on signals from undergraduates, following the PRISMA criteria. The study shows that higher education is a reliable predictor of undergraduates' employability. The findings help to develop a roadmap for simplified use of predictive analytics (Khaiser et al. 2023).

This process usually creates complicated multivariate data structure. In the most cases of complicated data structure with different types of scales for choosing the best applicant for a job, we usually use PCA as the best tool to analyze this type of data structure. Merely conducting data analysis and comparing two variables is inadequate for supplying the essential information required by a panel of judges to arrive at a decision. Conversely, multivariate analyses extract crucial information from a multitude of variables and furnish supplementary details that can enhance the decision-making procedure. Although these analytical techniques can be challenging to compute, contemporary information systems can assist researchers in acquiring the most valuable information (Smith and Sasaki, 1979).

The proper utilization and interpretation of multivariate methods are of utmost importance. This study seeks to assist judges in understanding and applying principal component analysis (PCA), which is a widely used and influential technique for analyzing multivariate data. PCA

Cihan University-Erbil journal of Humanities and Social Sciences (CUEJHSS)
Volume VII No. 1 (2023) 5 pages
DOI: 10.24086/cuejhss.vol7n1y2023.pp 121-125
Received: 04 May 2023; Accepted: 10 June 2023; Regular research paper: Published: 30 June 2023
*Corresponding author's e-mail: Abood.Jameel@cihanuniversity.edu.iq

effectively reduces the complexity of extensive datasets by converting numerous variables into a smaller set that retains essential information. Although accuracy may be slightly compromised during this process, the objective is to strike a balance between accuracy and simplicity. Ultimately, PCA enables the creation of concise and flexible representations of large datasets that offer valuable insights.

A. Research Objective

The main objective of this paper is to find the most preferable applicant for a job in Cihan University-Erbil using PCA, allocating scores and ranks.

B. Research Importance

The procedure of selecting or choosing the best applicant for a job depends most of times on subjective opinions. Using PCA technique gives the judges scientific procedure to help them in choosing the best applicant for a job.

II. Literature Review

PCA is a data processing technique used to extract a limited set of composite variables, known as principal components, from a larger set of measured variables. These principal components aim to capture and explain a specific phenomenon (Hastie et al., 2009; Constantin, 2014).

PCA is widely acknowledged as a valuable technique for reducing dimensions and compressing data. It generates orthogonal factors that account for a significant portion of the variation observed in the variables meeting specific criteria (Hastie et al., 2009). However, determining the number of factors to retain in the analysis is a decision the left to the researcher’s discretion (Lefter et al., 2006).

Various extraction rules and methods exist for determining the number of factors to retain. Among these, Kaiser’s criteria are widely recognized and suggest retaining only those factors with eigenvalues >1. Another approach involves using the scree plot or the cumulative percentage of extracted variance (Williams et al., 2010).

Qi and Luo (2014) suggested using PCA in Hilbert space to capture variations and interconnections between variables. They also introduced sparse PCA with generalized elastic network constraints to identify key features in high-dimensional data. Their methods involve efficient algorithms for optimization and a proposed guideline for selecting tuning parameters.

In Ada’s (2020) study on the impact of internal and external academic control on prospective teachers’ success, academic performance was assessed using GPA and proficiency measures. The study included 180 pre-service teachers. Findings revealed a negative correlation between emphasizing external academic control and both GPA and proficiency. Conversely, emphasizing internal academic control showed a positive correlation with efficiency. T-test results indicated that groups focusing on external academic control achieved higher GPAs and proficiency, while the group emphasizing internal academic control displayed higher proficiency levels (Al-Salami et al., 2022a).

When considering plot creation, usually two or three main components are enough. However, for modeling objectives, it is vital to precisely ascertain the suitable quantity of significant components (Wold et al., 1987; Patil, 2021; Al-Salami et al., 2022b).

In a study conducted by Wang (2023), the intelligent slope stability prediction (PCA) method was employed. By analyzing visual exploratory data from 77 *in situ* conditions, along with utilizing the Kernel PCA method, a total of seven slope stability prediction models were developed and assessed for their reliability through random cross-validation. The outcomes of this study introduce a fresh approach for predicting slope stability within the field of geotechnical engineering.

A. Data Collection

We gathered 25 forms from a panel of judges who interviewed 25 job applicants at Cihan University-Erbil. The judges rated the applicants on various scales for each of the 17 described characteristics listed on the form (Table I).

Therefore, researchers acquired a dataset containing information on 25 job applicants and 17 variables (characteristics). The measurements are organized in a table or matrix format with 25 rows and 17 columns, as illustrated in Table II.

Understanding the available information within the dataset, consisting of 25 rows (representing applicants) and 17 columns (representing characteristics), can be challenging due to its complexity.

B. Data Analysis

To examine a data set, similar to the one found in Table II, we employ the PCA technique. Before commencing our data analysis, it is crucial to perform two testing procedures to ascertain whether the data are appropriate for this method. To accomplish this, we utilize the following two tests.

1. Kaiser-Meyer-Olkin (KMO) index or measure of sampling adequacy (Williams et al. 2010). The KMO index is a numerical value that falls between 0 and 1. If this index is equal to or >0.50, it indicates that the sample is appropriate for performing PCA
2. For Bartlett’s test of sphericity to be considered significant ($P < 0.05$) according to Constantin (2014), the analysis is conducted using the JAMOVI 2.2.5 package from the Jamovi project (2021). The obtained results are as follows:
   a. KMO = 0.561, it is >0.50; hence, the sample is considered suitable for PCA
   b. Furthermore, the Bartlett’s test of sphericity shows a significance level of ($P < 0.002$) with degrees of freedom (df) equal to 136, indicating a highly significant result. The findings presented in Table III indicate that the data utilized in our example are suitable for conducting PCA.

Following these two tests, we can employ PCA to determine the appropriate number of principal components to retain in the model. Initially, the number of components matches the number of variables (17 variables) included

TABLE I
The Variable's Description

V.	Description	V.	Description
X_1	General Appearance, Health, and Build, 1–3	X_{10}	Other Languages, 0–5
X_2	Dress; Formal, Informal, and Clean, 1–3	X_{11}	Personality: Weak, Hesitate, Strong, 2–6
X_3	Voice: Too loud, Too Low, Pitchy, 1–3	X_{12}	Leadership: Ability to control, 0–12
X_4	Lecturing Experience: One mark for each year, max. 6.	X_{13}	Communication, 0–10
X_5	Job experience: One mark for each year, max. 7	X_{14}	Presentation: Stability, lecture method, lecture subject, 0–10
X_6	Academic Rank: A. Lect., Lect., A. Prof., Prof., 6, 8, 10, 12	X_{15}	Time punctuality: On time: 2, ≥15 min: 0, <15 min: 1
X_7	Researches: One mark for each research, max. 5	X_{16}	Arabic Languages: 0–10
X_8	English Languages, 0–10	X_{17}	Department Recommendation: 0–4
X_9	Computer Skills, 0–2		

TABLE II
Interview Assessment from Judgment Panel of Judges - Cihan University-Erbil

No.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	Total
1	2	3	4	6	5	8	2	8	5	2	4	4	8	7	2	10	3	83
2	2	3	3	5	4	6	2	6	5	2	4	4	8	6	2	10	2	74
3	2	3	4	6	6	8	2	8	5	2	4	5	7	8	2	8	2	82
4	3	2	4	5	5	6	3	8	5	3	5	4	8	8	2	10	3	84
5	2	2	4	4	5	6	2	6	6	2	4	5	7	7	2	8	3	75
6	3	3	3	5	5	6	3	6	5	2	5	5	8	7	2	8	3	79
7	2	2	3	4	5	6	2	6	5	2	4	4	6	6	2	8	2	69
8	3	2	2	4	4	6	2	6	5	2	4	4	6	6	2	8	3	69
9	3	3	4	4	5	8	2	6	5	3	5	5	8	6	2	10	3	82
10	2	2	3	3	4	6	2	6	6	2	4	4	6	7	2	8	2	69
11	2	3	3	4	3	6	3	6	4	3	4	4	6	6	2	8	3	70
12	3	3	4	5	5	8	3	6	5	3	5	5	7	8	2	10	3	85
13	3	2	3	4	4	6	2	6	5	2	4	4	6	7	1	8	2	69
14	3	3	4	4	4	6	3	6	5	3	4	5	6	6	1	10	2	75
15	2	2	3	4	4	7	2	6	4	2	4	4	6	6	2	8	2	68
16	3	2	2	4	4	6	2	6	5	3	4	5	6	6	1	8	2	69
17	3	3	4	5	6	7	3	6	5	3	5	5	8	7	2	10	3	79
18	2	3	4	4	3	6	2	4	4	2	4	4	6	7	2	8	3	68
19	3	2	4	5	4	5	3	4	5	3	4	4	5	6	2	8	2	69
20	3	2	4	4	4	6	3	5	5	3	4	5	5	6	2	10	3	78
21	2	2	3	4	5	5	3	5	6	3	5	4	5	6	2	8	2	70
22	3	3	4	5	6	6	3	6	5	3	4	5	7	7	2	10	3	82
23	2	3	4	4	5	6	3	6	5	3	5	5	7	8	2	10	3	81
24	3	3	4	5	6	5	3	6	6	3	4	5	8	8	2	8	3	82
25	2	3	3	4	6	6	3	5	5	3	4	4	7	7	2	8	2	74

Data set created by authors

TABLE III
KMO and Bartlett's Test

Bartlett's test of sphericity			KMO measure of sampling adequacy (MSA)
Chi-square	df	*P*-value	Overall=0.561
188	136	<0.002	

in the model. Each component possesses an eigenvalue, which indicates the amount of variance explained by that particular component. The eigenvalues and eigenvectors of the correlation matrix were derived using JAMOVI package. The first principal component accounted for 28.077% of the total variance; the second a further 19.413%; the third a further 10.004%; the fourth 9.153%; the fifth 6.056% making 72.7% of the total variance "explained" by five uncorrelated combinations of the original variables (Table IV).

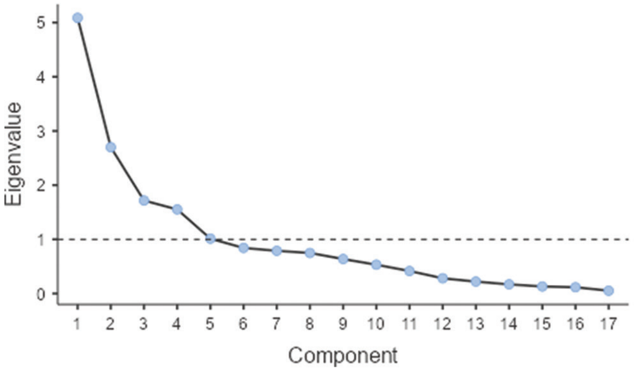

Fig. 1: Illustrating the scree plot for the original variables.

The Kaiser's criterion, also referred to as the eigenvalue-one criterion, is a widely used approach for selecting principal components. According to this criterion, only variables with eigenvalues >1 are retained in the new model. As a result, the 6th variable, which has an eigenvalue of 0.8200 (Table IV), will be excluded from the model.

Typically, the first components have the highest eigenvalues. In this case, we observe that only the first five components have eigenvalues >1. When considering the cumulative percentage of variance explained by these five components, it amounts to only 72.7% (Table IV).

To identify the main components of the data, it is necessary to compute eigenvectors and eigenvalues, which are principles derived from linear algebra. Before exploring these principles, it is crucial to comprehend that principal components are fresh variables generated through the linear blending or amalgamation of the original variables.

These mixtures are formulated in such a way that guarantees the absence of correlation among the newly created variables, known as principal components. In addition, a substantial amount of information contained within the original variables is condensed or compacted into the initial components. Consequently, despite beginning with data in 17 dimensions, the objective of PCA is to maximize the information captured in the first component, then proceed to capture the maximum remaining information in the second component, and so on. This iterative process continues until we attain a representation resembling the scree plot as illustrated in Fig. 1.

If we are concerned with a method of selection of a few indicator variables from a larger set, then PCA has a practical

TABLE IV
The Amount of Variance Accounted for (Initial Eigenvalues)

Component	Initial Eigenvalues			Extraction sums of squared loadings		
	Total	% of Variance	Cumulative%	Total	% of Variance	Cumulative %
1	4.7731	28.077	28.1	4.7731	28.077	28.1
2	3.3002	19.413	47.5	3.3002	19.413	47.5
3	1.7007	10.004	57.5	1.7007	10.004	57.5
4	1.5560	9.153	66.6	1.5560	9.153	66.6
5	1.0295	6.056	72.7	1.0295	6.056	72.7
6	0.8200	4.823	77.5			
7	0.7747	4.557	82.1			
8	0.6435	3.785	85.9			
9	0.5833	3.431	89.3			
10	0.5284	3.108	92.4			
11	0.3625	2.132	94.5			
12	0.2870	1.688	96.2			
13	0.2214	1.302	97.5			
14	0.1850	1.088	98.6			
15	0.1378	0.811	99.4			
16	0.0564	0.332	99.8			
17	0.0405	0.238	100.0			

TABLE V
Eigenvectors for the First Five Components (Component Loadings)

Variable	Component				
	1	2	3	4	5
X_1	0	0	0	0	0
X_2	0.827	0	0	0	0
X_3	0.668	0	0.302	0	0
X_4	0.727	0	0	0	0
X_5	0.446	0	0	0	0.751
X_6	0.395	0.699	0	0	0
X_7	0.332	−0.728	0	0	0
X_8	0	0.741	0	0	0
X_9	0	0	0	0	0.855
X_{10}	0	−0.623	0	0.529	0
X_{11}	0	0	0.763	0	0
X_{12}	0.343	0	0	0.609	0
X_{13}	0.540	0.454	0.371	0	0.315
X_{14}	0.499	0	0	0	0.499
X_{15}	0	0	0.550	−0.678	0
X_{16}	0.431	0	0.470	0.443	0
X_{17}	0	0	0.774	0	0

“varimax” rotation was used

TABLE VI
Score and Rank for each Applicant

Applicant No.	Score	Rank	Applicant No.	Score	Rank	Applicant No.	Score	Rank
1	29.064	1	6	26.692	10	13	22.937	19
17	29.063	2	2	25.934	11	7	22.884	20
12	28.971	3	25	25.528	12	15	22.833	21
3	28.95	4	14	25.47	13	21	22.281	22
22	28.128	5	5	24.934	14	10	22.21	23
24	27.91	6	20	24.103	15	16	22.113	24
4	27.551	7	18	23.986	16	8	21.77	25
23	27.454	8	19	23.23	17			
9	27.454	9	11	23.151	18			

advantage, for example, in the dataset (25 × 17) matrix, the first five components have high positive loading as shown in Table V. PCA offers the necessary weights to obtain a new variable that effectively captures the variation present in the entire dataset to a certain extent. This newly derived variable, along with its corresponding defining weights, is referred to as the first principal component. The calculation of these new variables, representing the principal components, involves a linear combination of the initial variables. Specifically, the JAMOVI system directly computes these variables as PCA1, PCA2, PCA3, PCA4, and PCA5. Based on the analysis, it is recommended to retain five principal components as the primary set of components in the model (Table V).

Now, we are not interested only in the interpretation of the components in this research; but we wish also to consider the component of scores which can be produced by post-multiplying the original data matrix (25 × 17) by the matrix of eigenvectors (17 × 5) using (Mathcad 15 M050) software, due to the large size of the matrices. This process produces and allocates score to each applicant, and then researchers find the component of ranks by ranking the score component as described in Table VI below (Jameel, 2019).

The result of the previous procedure analysis reveals that the applicant No.1 has the greatest score (29.064), and ranked No.1 and the applicant No. 17 has scores (29.063) and ranked No.2, and so on (Table VI).

III. Conclusions

PCA is a valuable approach when dealing with numerous variables that measure the same underlying concept. In situations where complex data structures arise, such as evaluating 25 applicants using 17 variables with different scale types during the assessment process, the issue of multicollinearity emerges. In such cases, alternative analysis methods like regression models are not suitable. The crucial steps in conducting PCA involve assessing the suitability of the data for this method and selecting the most appropriate factors that capture the total variance among the original variables. In this context, the interpretation of the resulting factors is of utmost importance, especially when it comes to selecting the best applicant for a job, which is a critical concern for Cihan University. The obtained principal components can be utilized for further analysis. The derivation of these components involves calculating new variables, which serve as representations of the principal components, by combining the initial variables in a linear manner. In the JAMOVI system, these variables are directly computed as standardized values. New variables can also obtain by post-multiplying the original data matrix (25 × 17) by the matrix of eigenvectors (17 × 5). This process produces and allocates score to each applicant, and then we find the component of ranks by ranking the score component as described in Table V. The process of allocating scores and ranks of the previous procedure analysis reveals that the applicant No.1 has the greatest score (29.064) ranked No.1, the applicant No. 17 has score (29.063) ranked No.2 and applicant No.12 has score (28.971) ranked No.3. As illustrated in Table VI. Finally, they conclude that this PCA is a very good and power full scientific technique can be used to help in analyzing complicated data structure to give advice in selecting the best applicant for a job.

References

Ada, Ş. (2020). Competence of low-high academic control focus and its place in academic success. *International Journal of Psychology and Educational Studies*, 7(2), 1-9.

Al-Salami, Q.H., & Abdalla, S.N. (2022a). The impact of academic satisfaction as a mediator on international conferences. *Cihan University-Erbil Journal of Humanities and Social Sciences*, 6(1), 19-26.

Al-Salami, Q.H., Saleh, R.K., & Al-Bazi, A.F. (2022b). An efficient inventory model-based ga for food deterioration products in the tourism industry. *Pesquisa Operacional*, 42. Doi: 10.1590/0101-7438.2022.042.00257447.

Constantin, C. (2014). Principal component analysis-a powerful tool in computing marketing information. *Bulletin of the Transilvania University of Brasov. Economic Sciences. Series V*, 7(2), 25.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* 2nd ed. Berlin: Springer.

Jameel, A.M. (2019). Proposed statistical model for scoring and ranking sport tournaments. *Cihan University-Erbil Journal of Humanities and Social Sciences*, 3(1), 15-19.

Jamovi Project. (2021). Jamovi, (*Version 2.2*). [*Computer Software*]. Sydney, Australia. Available from: https://www.jamovi.org

Khaiser, F.K., Saad, A., & Mason, C. (2023). Systematic Review of Qualitative and Quantitative Studies on Perceived Employability of Graduates. In: *2023 17th International Conference on Ubiquitous Information Management and Communication* (*IMCOM*). United States: IEEE. pp1-8.

Lefter, C., Brătucu, G., Bălăşescu, M., Chiţu, I., Răuţă, C., & Tecău, A. (2006). *Marketing*. vol. 2. Braşov: Universităţii Transilvania.

Patil, I. (2021). Visualizations with statistical details: The "ggstatsplot" approach. *Journal of Open Source Software*, 6(61), 3167.

Qi, X., & Luo, R. (2014). Sparse principal component analysis in Hilbert space. *Scandinavian Journal of Statistics*, 42, 270-289.

Smith, K., & Sasaki, M.S. (1979). Decreasing multicollinearity: A method for models with multiplicative functions. *Sociological Methods and Research,* 8(1), 35-56.

Wang, G., Zhao, B., Wu, B., Zhang, C., & Liu, W. (2023). Intelligent prediction of slope stability based on visual exploratory data analysis of 77 *in situ* cases. *International Journal of Mining Science and Technology*, 33(1), 47-59.

Williams, B., Brown, T., & Onsman, A. (2010). Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine*, 8(3). Available from: http://ro.ecu.edu. au/jephc/vol8/iss3/1

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3), 37-52.