International Journal of Civil Engineering and Infrastructure

Vol. 3, No. 2, Oktober 2023

MULTIVARIATE ANALYSIS ON THE STUDY OF PORT FACILITIES DEVELOPMENT: PCA ANALYSIS FOR LOW CORRELATION DATASET

Rachmad Irwanto[1]; Budi Satiawan[2], Irnanda Satya Soerjatmodjo[3]; Andika Setiawan[4]
[1, 2, 3, 4] Civil Engineering Study Program, Muhammadiyah University of Jakarta
Cempaka Putih, Jakarta, Indonesia
Correspondence email: rachmad.irwanto@umj.ac.id

ABSTRACT

Efforts to develop existing infrastructure facilities are highly regarded in order to keep up with capacity demand and upgrade changes in data trends. To obtain the best users' interest in the facilities that aligns management development plan and schedule, a questionnaire is commonly conducted. Datasets acquired from questionnaire featuring satisfactory level such as Likert scale tends to be ordinal. Ordinality using standard Pearson correlation lean towards weak relationship. Traditional PCA, relying on Pearson correlation, may struggle to capture the nuanced relationships within such ordinal data, leading to a loss of valuable information. Through a comparative analysis of PCA results using both covariance matrix and conventional Pearson correlation, this paper demonstrates the efficacy of the proposed methodology in uncovering latent patterns and relationships within the questionnaire responses.

Keywords: *questionnaire, Likert scale, PCA, ordinal dataset, Pearson correlation, low correlation, variance, covariance matrix.*

1. PRELIMINARY

This paper intends to present a method used in reviewing a data set. The data set is an output from questionnaires distributed among respondents asking their level of satisfactory over an existing ferry port terminal. The purpose of the questionnaire is to assess facilities within the port terminal for further development. The intended development has been including capacity increase. To align with budget limit and economic feasibility, not all facilities are to be expanded. Some facilities are limited to have new interior while the others would be expanded in capacity. To ensure the development plans accommodate commercial mission of the management and the users' comforts, the management decided to carry out the questionnaire. One way the purpose of the questionnaire is carried out is by observing facilities with least satisfactory responds from the users (Ferry passengers). The satisfactory data that was acquired, set to be in a range for each question.

The objective of the analysis, statistically is to assess the component with major influence with Principal Component Analysis (PCA). Due to its non-linearity

and low-correlation features, the dataset turned out to be one of PCA pitfall in pulling conclusion.

Principal Component Analysis (PCA) stands as a powerful method of multivariate analysis, renowned for its ability to extract meaningful patterns and dimensions from complex datasets. Its application spans across fields, from image processing to social sciences, enabling researchers to uncover hidden relationships, reduce dimensionality, and pave the way for more concise data representation. However, in the pursuit of these goals, the efficacy of PCA can be challenged when working with low correlation data. While PCA thrives in scenarios where variables exhibit substantial pairwise correlations, it encounters limitations when confronted with datasets where correlations are notably weakened.

Traditional PCA techniques, which rely on maximizing variance and exploiting correlations between variables, may yield results that appear less insightful and less indicative of underlying structures in the absence of strong relationships between variables. In response, this study embarks on a focused exploration of conducting PCA on datasets characterized by low correlations among variables. By dissecting the intricacies of low correlation data and its implications, we uncover strategies to adapt and harness PCA's potential even in situations where traditional correlations are diminished.

This paper delves into the challenges of PCA when applied to Likert-scale questionnaire data. Likert-scale responses, tends to be ordinal. It often exhibits low correlation, which is the common pitfall of traditional PCA. Recognizing these limitations, this study proposes an approach by harnessing covariance matrix to enhance variance exploitation in Likert-scale data analysis.

The research underscores the inherent complexities of Likert-scale data, where respondents provide subjective ratings that may not adhere strictly to numerical order. Traditional PCA, relying on Pearson correlation, may struggle to capture the nuanced relationships within such ordinal data, leading to a loss of valuable information. To address that, the authors attempted to apply covariance matrix within PCA to optimize the use of variance.

2. LITERATURE REVIEW

As Jolliffe *et al*. explained in his paper [1], PCA is basically a technique in multivariate statistics. Addressing above issues with multiple variables obtained from respondents' responses, the common first step is usually dimensionality reduction of a dataset. In this case, one thing to keep in mind while reducing data dimensionality is retaining as much of the original variation as possible. A method of standardized data incorporated in PCA's early step processes data dimensionality reduction [2] while retaining original variation. The PCA transforms the data into a new set of variables, known as principal components. Basically, we could use PCA analysis to determine the most important variables among multiple variables assumed to be linearly related with the unknown underlying factors. In their theoretical words, principal component sets are linear combinations of the original variables and are ordered in terms of their nature to explain the variation in the data.

PCA can technically be used for a variety of purposes, including data compression, feature extraction, and data visualization. A paper by Karamizadeh *et al*. [3] through its references mentions that PCA is particularly useful when dealing with datasets that have a large number of variables, as it allows for a more manageable representation of the data. Karamizadeh [3] also concluded PCA's key advantages such as its low noise sensitivity, the decreased requirements

for capacity and memory, and increased efficiency given the processes taking place in a smaller dimension.

PCA's pitfall is when the dataset is small and has well distributed classes. In that case PCA becomes less relevant [3]. An interesting paper by Jeong et al. in 2009 [4] explains that PCA is also commonly considered as a black box due to its difficult interpretations. Shlens in 2014 [5] has made a good tutorial PCA reference. In 2019 another paper by Björklund [6] has underlined what to notice during PCA analysis. This has something to do with how we look at our principal components. Another reference that discusses PCA method is by Abdi *et al*. [7] in 2010. A paper by Guerra et. al [8] in 2021 provides good guidance for PCA that processes non-ordinary data. As our data from questionnaire may not provide good linearity, the correlation from standard Pearson is weak. Questionnaire data that features Likert scale is usually ordinal. Therefore, the use of other correlation factors such as Spearman's rank or Kendall's tau [9], [10] is suggested.

PCA is a good data summary when the interesting data set patterns increase the variance of projections onto orthogonal components. Yet, PCA also has limitations that must be considered when interpreting the output: the underlying structure of the data must be linear, patterns that are highly correlated may be unresolved because all PCs are uncorrelated, and the goal is to maximize variance and not necessarily to find clusters. As implied by Lever *et al*. [11], conclusions made with PCA must take these limitations into account. As with all statistical methods, PCA can be misused. The scaling of variables can cause different PCA results, and it is very important that the scaling is not adjusted to match prior knowledge of the data. If different scaling sets are tried, they should be described. PCA is a tool for identifying the main axes of variance within a data set and allows for easy data exploration to understand the key variables in the data and spot outliers. Properly applied, it is one of the most powerful tools in the data analysis tool kit.

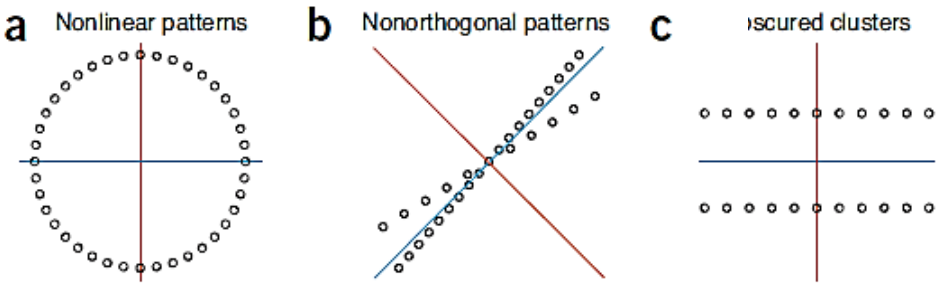

Figure 1 – Several data sets where PCA has limited use [11].

The first step as mentioned earlier is to standardize the data. This step is to have a mean of zero and a standard deviation of one to all variables. By subtracting the mean from all the data set, basically we shift the center of data to zero. The purpose is to ensure that all variables are on the same scale and have equal importance in the analysis. Following step is then to determine the principal components calculated using eigenvectors and eigenvalues of the covariance matrix from the standardized data.

From what the author learns, this dataset transformation by determining eigen vectors and eigenvalues of a matrix is the core of PCA. Through covariance matrix, the goals are to minimize data redundancy while maximizing variance. The covariance matrix that we all know, will be

the measure of how much the dimensions vary from the mean.

The first principal component explains the most variation in the data, with subsequent components explaining decreasing amounts of variation. The number of principal components to retain depends on the amount of variation one wishes to preserve and the purpose of the analysis. Typically, one retains enough principal components to explain a substantial portion of the total variation, while discarding the rest.

3. OUR CASE

Figure 2 describes the situation with our data. Though the data points look only very few, there are actually about two hundred data points in the plot. Many of them are repeating, so they overlap to each other and still not visible even with 3D plot (figure 2 – top left). An ideal cumulative variance should be above 90% for comfort interpretation. The plot shows about 80%. A comparison for an ideal PCA analysis is displayed in figure 3.

Among implications of low cumulative variance is weak correlation as referred in [12] and [13]. Also noticeable from figure 2 (bottom left), the principal components were plotted in blue and red lines from the first two columns of the entire dataset. Though the lines are almost not visible, red line is slightly longer than the blue lines saying that principal component (PC1) 1 is more dominant than PC2. However, as the lines are too short compared the data plot, we can tell the whole PCA analysis in this case is not ideal.

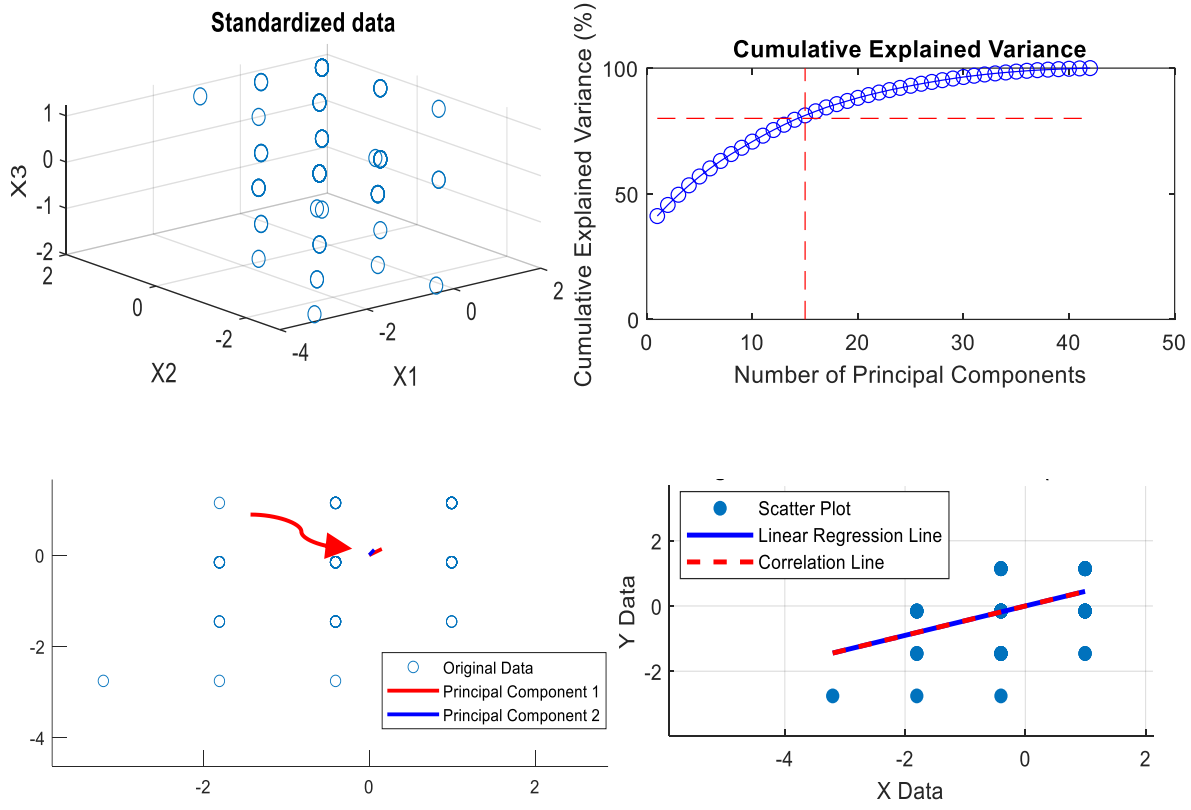

Figure 2 - PCA output from our case. PC1 vs. PC2 (arrowed).

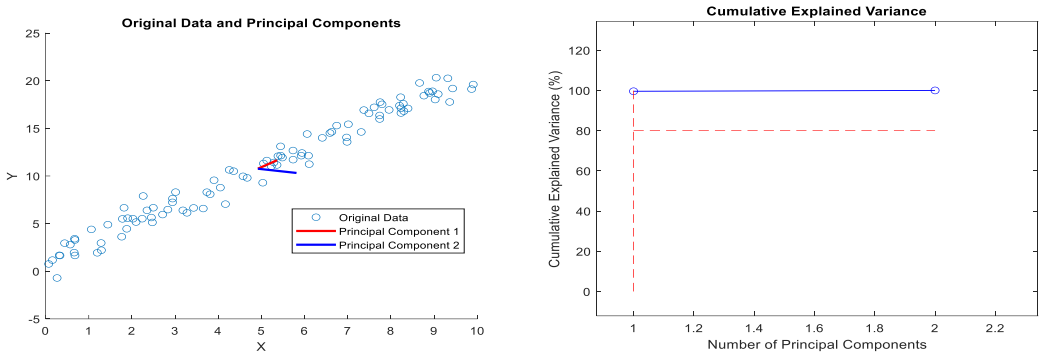

Figure 3- Ideal data set for ordinary PCA

Intriguingly, one context in which the challenge of low correlation data is often encountered is the analysis of questionnaire outputs. Consider a scenario where respondents evaluate various aspects of a service, facility, or experience using Likert scale responses. Such data is inherently characterized by the limited range of response options, which can result in responses exhibiting low pairwise correlations across variables. For instance, a user satisfaction questionnaire for port facilities may involve respondents assigning scores to different criteria, such as cleanliness, accessibility, and staff behavior. While these aspects are all important, they might not necessarily correlate strongly due to individual variations in perceptions and preferences. In the presence of low correlation data, PCA's interpretability is notably affected for several reasons:

Diminished Information Capture: Traditional PCA relies on capturing variance and correlations to create principal components. In the absence of strong correlations, the captured variance might be dominated by noise, leading to principal components that do not meaningfully represent the underlying data structure.

Lack of Clear Dimensionality: Principal components, derived from correlated variables, often align with underlying dimensions in the data. However, in the case of low correlations, the relationships between variables are weak, making it challenging to identify distinct dimensions that can be easily interpreted.

Unstable Loadings: In low correlation scenarios, the loadings of variables onto principal components become unstable and susceptible to minor variations in the data. This instability can make it difficult to confidently interpret which variables contribute most to a given component.

Further we illustrate the challenges and complexities of applying PCA on low correlation data, delve into the fundamental numerical formulas that underpin PCA. First thing, the covariance matrix of the variables, we denoted as (Cov), quantifies the relationships between pairs of variables. For highly correlated variables, (Cov) has significant off-diagonal elements, indicating strong pairwise associations. In contrast, low correlations lead to a nearly diagonal covariance matrix with small off-diagonal elements.

Carrying out PCA involves finding the eigenvectors and eigenvalues of the covariance matrix (Eigen Decomposition). Eigenvectors (v) represent directions in the original variable space, and eigenvalues (λ) quantify the variance captured by each eigenvector. Eigenvalues indicate the proportion of total variance explained by each principal component. In low correlation data, eigenvalues may show less distinct separation, making it harder to determine the significant components.

Principal components (PCs) are linear combinations of the original variables that capture maximal variance. The first PC (PC1) corresponds to the direction of maximum variance, with subsequent PCs orthogonal to previous ones. For low correlation data, PCs might not represent clear underlying dimensions, leading to less informative component interpretation.

$$PC_j = \sum_{i=1}^{P} \phi_{ij} X_i$$

$$\phi_{ij} = \text{loading of variable } X_i \text{ on } PC_j$$

In the following sections, we will delve deeper into the implications of low correlation data on PCA, discussing strategies for mitigating challenges and enhancing interpretability. Through illustrative examples and practical insights, we aim to provide a comprehensive understanding of applying PCA to scenarios where correlations among variables are limited.

Correlation vs. Covariance in the Context of PCA:

One fundamental aspect that significantly influences the application and interpretation of Principal Component Analysis (PCA) is the choice between working with correlation matrices or covariance matrices. Both correlation and covariance matrices serve as the foundation of PCA, yet they embody different perspectives that can impact the results and insights gained from the analysis. Understanding the nuances of using correlation versus covariance matrices is pivotal for effectively conducting PCA, particularly in scenarios involving low correlation data [14].

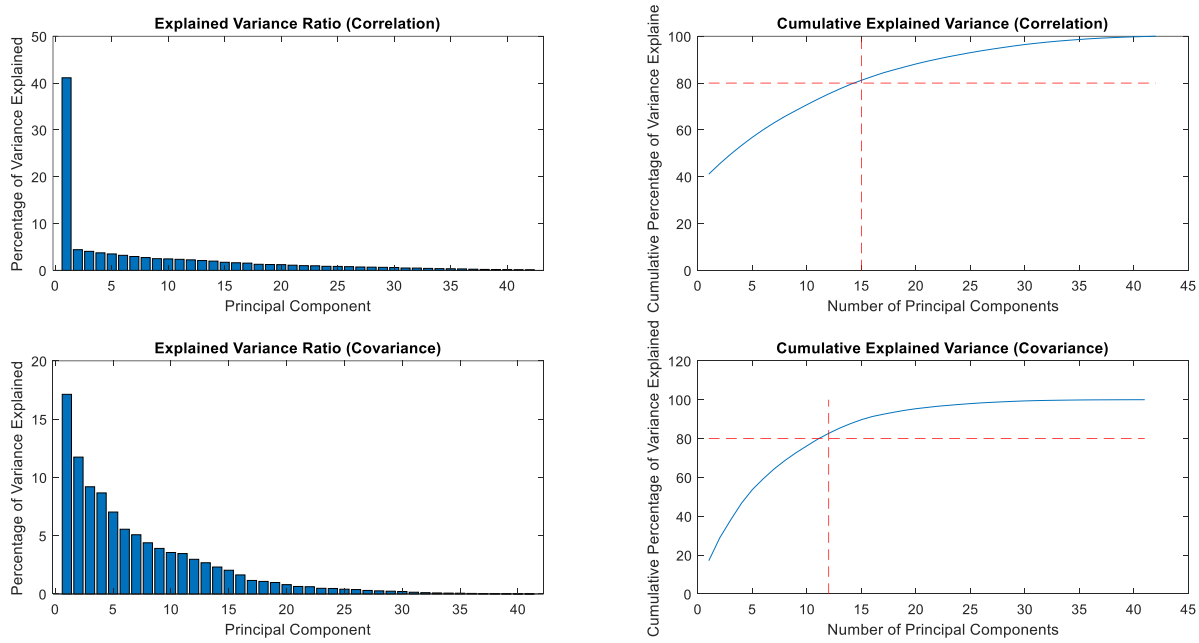

Figure 4 – PCA with correlation vs. PCA with covariance

Covariance Matrix:

The covariance matrix, denoted as Σ, quantifies the strength and direction of linear relationships between pairs of variables in a dataset. Each element in the matrix represents the covariance between two variables, reflecting how their values change together. Mathematically, the covariance between two variables X_i and X_j is given by:

$$Cov(X_i, X_j) = \frac{1}{N-1} \sum_{k=1}^{N} (X_i^{(k)} - X_i)(X_j^{(k)} - X_j)$$

In the context of low correlation data, the covariance matrix can exhibit a pronounced diagonal structure with small off-diagonal elements. This occurs when variables exhibit limited linear relationships. As a result, PCA based on the covariance matrix may emphasize variance primarily along the axes of high variance variables, potentially masking valuable insights and overlooking dimensions that contribute to data variation but are not strongly correlated.

Correlation Matrix:

In contrast, the correlation matrix, often denoted as R, standardizes the covariance matrix by dividing each element by the product of the standard deviations of the two variables. The resulting values, known as correlation coefficients, range between -1 and 1, representing the strength and direction of linear relationships while accommodating for differing scales. The correlation between variables X_i and X_j is calculated as:

$$\rho\left(X_i, X_j\right) = \frac{Cov\left(X_i, X_j\right)}{\sigma_i \ \sigma_j}$$

Applying PCA to a correlation matrix is advantageous in scenarios where low correlation data is prevalent. By transforming the original variables into standardized versions, the correlation matrix accentuates the relative strength of relationships, facilitating the identification of underlying patterns that might otherwise remain obscured by scale discrepancies. Moreover, correlation-based PCA yields principal components that emphasize dimensions defined by the patterns of highest joint variability, rather than being overly influenced by high-variance variables.

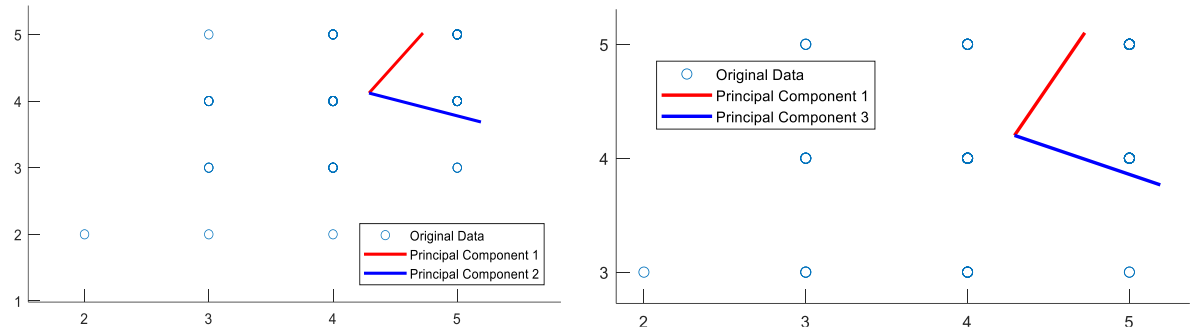

Figure 5 – Principal Components after covariance matrix. PC1 vs PC2 (left); PC1 vs. PC3 (right)

Considerations and Interpretation:

As we can look at figure 4 and figure 5, though it is not significant, the use of covariance helps increasing the use of variance during PCA analysis. When employing PCA on low correlation data, the choice between covariance and correlation matrices influences the outcome, interpretation, and applicability of the analysis. While the covariance matrix preserves the information about both the magnitude and direction of relationships, it may inadvertently magnify variables with greater variances. The correlation matrix, on the other hand, addresses scale disparities, enabling a more balanced representation of variable relationships.

Ultimately, the decision between using covariance or correlation matrices in PCA depends on the research objectives, the nature of the data, and the intended emphasis on different aspects of relationships. When variables exhibit low correlation, utilizing the correlation matrix can provide a clearer view of latent dimensions that contribute to data variation, enhancing the interpretability of PCA results and uncovering insights that

might otherwise remain hidden in the shadows of raw covariance.

As we delve further into the application of PCA on low correlation data, we will explore practical strategies for navigating these choices and leveraging the inherent strengths of both correlation and covariance matrices to extract meaningful dimensions and patterns from the data.

What Next

The authors limit the discussion in this paper until the use of covariance matrix as the comparison of correlation matrix in PCA. The discussion below this sub point is intended to be featured in other paper.

In cases with our questionnaire data featuring Likert scale, that lacks strong correlations, as we notice above, traditional PCA might not yield meaningful results. Since the mission is to maximize variance among data points to capture more interpretable multivariate aspects, it is advisable to consider more techniques like Factor Analysis (FA).

In similar cases to datasets with low correlation, preprocessing data is necessary. This could be carried out with other methods other than FA. In our case, the nature of dataset seems non-linear.

The correlation matrix, as we know, is a normalized version of the covariance matrix. This normalization process, therefore, scales the values to lie between -1 and 1, which makes them easier to interpret. Please be noted, whenever a covariance matrix is identical to correlation matrix, this could imply that all of the variables (columns in the matrix) have a standard deviation of 1. This is due to the dataset has been standardized, which is a common preprocessing step in many data analyses. This may lead to a false perception that the dataset is perfectly positively correlated. Even if that is true, it does not mean the dataset is perfectly correlated, while the actual linear relationships (correlations) are weak.

Preprocessing data by handling missing data points and outliers could be carried out carefully, if necessary, just to ensure the dataset is not discontinued [15]. Specifically for dataset that features Likert scale, it is more likely to be ordinal. The categories are in good order but not consistently different among the data points. This is the cause why the dataset has weak correlation. The use of standard Pearson correlation may not be suitable for that case. It is suggested by papers [9], [10] to use Spearman's rank or Kendall's tau. Both measures are generally adequate and suitable for ordinal data.

4. CONCLUSION

PCA is a useful tool for analyzing dataset to find major influencing component within. However, PCA has limitation in its linearity. PCA exploits variances among data points within the dataset. When a dataset has low variance due to low correlation, this will affect the outcome. The produced principal components will not be significant. This is measured by its use of variance that is lower than 90%. The correlation using standard Pearson may be altered to optimize the use of variance.

Questionnaire output usually features Likert scale. The scale such as customer satisfactory index that ranges the answers from determined scale (for example 1 to 5), has meaningful order. However, among questions in the questionnaire, when plotted in a scatter plot without looking at the questions, statistically only shows repeating numbers and has almost no different compared to the other questions. This ordinal nature of dataset causes weak relationships among the data points and therefore low correlation.

An option to look at the covariance matrix of the dataset to be implemented in PCA turns to be increasing the use of variance.

Through a comparative analysis of PCA results using both covariance matrix and conventional Pearson correlation, this paper demonstrates the efficacy of the

proposed methodology in uncovering latent patterns and relationships within the questionnaire responses. The findings not only contribute to the refinement of PCA applications in ordinal data analysis but also offer practical insights for decision-makers involved in port facility development planning. This approach offers an alternative to extract meaningful information from Likert-scale data characterized by low correlation.

ACKNOWLEDGEMENT

The authors would like to thank to the rector of UMJ for the funding and facilities. This work is financially supported by LPPM - UMJ.

The authors also declare that there is no competing financial, professional or personal interest during the writing and publication of this paper.

REFERENCES

[1] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, Apr. 2016, doi: 10.1098/rsta.2015.0202.

[2] S. Lespinats, B. Colange, and D. Dutykh, *Nonlinear Dimensionality Reduction Techniques*. Cham: Springer International Publishing, 2022. doi: 10.1007/978-3-030-81026-9.

[3] S. Karamizadeh, S. M. Abdullah, A. A. Manaf, M. Zamani, and A. Hooman, "An Overview of Principal Component Analysis," *Journal of Signal and Information Processing*, vol. 04, no. 03, pp. 173–175, 2013, doi: 10.4236/jsip.2013.43B031.

[4] D. H. Jeong, C. Ziemkiewicz, W. Ribarsky, and R. Chang, "Understanding Principal Component Analysis Using a Visual Analytics Tool," 2009. doi: 10.1.1.157.1469.

[5] J. Shlens, "A Tutorial on Principal Component Analysis," Apr. 2014, [Online]. Available: http://arxiv.org/abs/1404.1100

[6] M. Björklund, "Be careful with your principal components," *Evolution (N Y)*, vol. 73, no. 10, pp. 2151–2158, Oct. 2019, doi: 10.1111/evo.13835.

[7] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscip Rev Comput Stat*, vol. 2, no. 4, pp. 433–459, Jul. 2010, doi: 10.1002/wics.101.

[8] R. Guerra-Urzola, K. Van Deun, J. C. Vera, and K. Sijtsma, "A Guide for Sparse PCA: Model Comparison and Applications," *Psychometrika*, vol. 86, no. 4, pp. 893–919, Dec. 2021, doi: 10.1007/s11336-021-09773-2.

[9] M.-T. Puth, M. Neuhäuser, and G. D. Ruxton, "Effective use of Spearman's and Kendall's correlation coefficients for association between two measured traits," *Anim Behav*, vol. 102, pp. 77–84, Apr. 2015, doi: 10.1016/j.anbehav.2015.01.010.

[10] C. Croux and C. Dehon, "Influence functions of the Spearman and Kendall correlation measures," *Stat Methods Appt*, vol. 19, no. 4, pp. 497–515, Nov. 2010, doi: 10.1007/s10260-010-0142-z.

[11] J. Lever, M. Krzywinski, and N. Altman, "Points of Significance: Principal component analysis," *Nature Methods*, vol. 14, no. 7. Nature Publishing Group, pp. 641–

642, Jun. 29, 2017. doi: 10.1038/nmeth.4346.

[12] U. Sunarya *et al.*, "Feature Analysis of Smart Shoe Sensors for Classification of Gait Patterns," *Sensors*, vol. 20, no. 21, p. 6253, Nov. 2020, doi: 10.3390/s20216253.

[13] Z. Wan, Y. Xu, and B. Šavija, "On the Use of Machine Learning Models for Prediction of Compressive Strength of Concrete: Influence of Dimensionality Reduction on the Model Performance," *Materials*, vol. 14, no. 4, p. 713, Feb. 2021, doi: 10.3390/ma14040713.

[14] Y. Z. Ma, "Principal Component Analysis," in *Quantitative Geosciences: Data Analytics, Geostatistics, Reservoir Characterization and Modeling*, Cham: Springer International Publishing, 2019, pp. 103–121. doi: 10.1007/978-3-030-17860-4_5.

[15] A. Ilin and T. Raiko, "Practical Approaches to Principal Component Analysis in the Presence of Missing Values," 2010.