International Journal of Engineering & Technology, 7 (2.29) (2018) 488-490

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Rice Odours' Readings Investigation Using Principal Component Analysis

Nurul Aini Abdul Wahab[1]*, Shamshuritawati Sharif[2]

[1]Department Of Statistics, Decision Sciences and Actuarial Sciences, Universiti Teknologi Mara Cawangan Negeri Sembilan Kampus Seremban
[2]School Of Quantitative Sciences, College Of Arts And Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah
**Corresponding Author E-Mail: Shamshurita@Uum.Edu.My*

Abstract

The use of electronic nose (e-nose) devices plus principal component analysis can help the process of categorizing the 16 different rice into its type. Generally, the physical feature of an e-nose own more than one hole to capture the odour of rice. For example, the portable e-nose so-called Insniff does have 10 holes (or variables). In this situations, we will have a dataset that consist high-dimension dataset where lead to the presence of interdependencies between all variables under study. Therefore, this study is presented to investigate the odour of rice for identifying the most important variables contributing to the rice odour readings. The principal component analysis (PCA) is implemented to determine the component that best represent the all 10 variables in order to eliminate the interdependency problem, and (2) to identify which variable is considered as important and influential to the newly-formed principle component (PC). The results from PCA suggested that the first two principle components is chosen. It is based on three assessments which are Kaiser's criterion larger than 1, cumulative proportion of total variance, and scree plot. These two principle components explained 89% of total variance. Results showed that sensor 1 (0.931) and sensor 2 (0.966) are the two important variables that highly contribute to PC1. On the other hand, for PC2, the highest contribution is from sensor 8 (0.828). This study demonstrate that PCA is effective for investigating rice odour readings.

Keywords: *Data Reduction; e-nose; Principle Component Analysis (PCA)*

1. Introduction

A dataset of odour of rice was collected using an electronic nose (e-nose). The physical feature of an e-nose that own more than one hole to capture the odour, give a high-dimension of set of data. For example, *Insniff* does have 10 holes to capture the odour of rice. Thus, the collected rice odour's readings will eventually give a data set of n-obervations x 10 holes (variables). In this study, a sample of 4767 rice from 16 types of rice that were captured by 10 sensors of an *Insniff*, produce a set of data with a dimension of 4767 x 10. Therefore, this is truly a high-dimensional data set where dimension that was built from a several readings of odours is considered large. A problem in handling high-dimensional dataset because there is always unclear pattern and not easily interpretable (1). The issue of interdependency is exist between variables in such dataset Clarke, Ressom (2) and Fan and Lv (3). All statistical procedure does not only generating the intended results but also stress on the accuracy interpretability and simplicity of the developed model (3). In developing predictive model or classification model, identifying the pattern of data is important to avoid either over fitted model or misclassified model (3, 4)

One of multivariate statistical methods that can help us to identify the pattern in data is Principal Component Analysis (PCA)(5). PCA is a superior tool for analyzing data of high dimension. Once we found the pattern in the data, PCA able to reduce the dimension without loss of information. By using PCA, a recommendable number of interpretable components are proposed to replace the original sets of variables. PCA guarantee that there is no correlation between each newly-formed components and also no original variables were left out from the analysis.

In this paper we present the investigation results on rice odours' using principal component analysis. The analysis was performed to achieve two main objectives which are (1) to determine the best number of component that represent 10 holes in order to eliminate the interdependency problem, and (2) to identify which variable is considered as important and influential to the newly-formed principal component (PC). The outline the remainder of this paper is as follows. In section 2, we reviewed on PCA in details. In section 3, an empirical results based on e-nose data set is discussed. Finally, the conclusion is presented in section 4.

1.1. Review on Principal Component Analysis (PCA)

The Principal component analysis (PCA) can provide on the most meaningful parameters that aid data set interpretation, data reductions, and summarize the statistical correlation among constituents in the water with minimal loss of original information(6, 7). The first principal component loading represents most of the variance in the observed variables, while each subsequence component explains progressively less variance. PCA is used for understanding the characteristics of water quality in many fields and it has provided important information for environment (8, 9)

In the topic of high dimensional dataset or multivariable of dataset, there is always the discussion of PCA as a dimension reduction technique (5, 10, 11). The need of restructuring the original dataset are mass, huge, multidimensional, and sometimes with irrelevant measures existed. Reducing the large dimension to smaller, but still carrying the amount of information as much as the original

variables plus all the noises are taken care off very well is considered as an important phase before the higher statistical analysis is performed (10).

The important of performing dimension reduction process was highlighted by (4) in their study.

The purpose of PCA is to discover a new set of variables, Z_1 , Z_2 , ..., Z_p in a form of a linear combination of all variables which is $Z = a^T X$. The first principal component, Z_1 is the linear combination of the original features which mathematically written as $Z_1 = a_1X_1 + a_2X_2 + \cdots + a_pX_p$ accumulate as the largest as possible of variance of p variables subject to $a_1 + a_2 + \cdots + a_p = 1$. Then, the second principal component, Z_2 is selected to have the second largest possible variance. The remaining principal component is defined similarly.

Statistically, the decision on coefficients values a is subjected to its mathematical formulation. Thus, the obtained principal components are in decreasing order of variance $var(Z_1) \geq var(Z_2) \geq$, ..., $\geq var(Z_p)$. It is equivalent to $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. In practice, only the first k numbers of principal components account for most of the variability of the original data, thus keeping all the p principal components sound impractical. This mean, only the first k principal components will be used in further analysis while the $p - k$ principal components will be ignored. A number of procedures to determine k have been suggested. Among the most common procedures are discussed in section 3.

2. Methodology/Materials

PCA was performed using rice odour gathered by e-nose. E-nose is a non-destructive intelligent electronic sensing instrument, which mimics the human olfactory system to detect, discriminate, and classify odour samples (11). From literature, there are various factors that influence the rice quality of rice such as cultivated location, climatic conditions, genetics, and post-harvest activities (12). The quality characteristic is evaluated based on the physical form such as size, shape, and cleanliness (13).

In this analysis, PCA developed ten series of principle components (**Z**), and that are actually sets of linear combinations of all 10 hole sensors. The extraction process was done based on the standardized or correlation matrix of the sensors. The root characteristics extracted were eigenvalues, λ, and sets of eigenvectors, a_i. The a_i. then were used to write the equation of the linear combination formed as PC.

For the dataset used in this paper, the ten PC are expressed as follows.

$Z_1 = a_{1,1}SENSOR_1 + a_{2,1}SENSOR_2 + \cdots + a_{10,1}SENSOR_{10}$

$Z_2 = a_{1,2}SENSOR_1 + a_{2,2}SENSOR_2 + \cdots + a_{10,2}SENSOR_{10}$

$. Z_{10} = a_{1,10}SENSOR_1 + a_{2,10}SENSOR_2 + \cdots + a_{10,10}SENSOR_{10}$

Solving the expression produced sets of PC's scores organized Z_1, Z_2 , ..., Z_{10} and will then be used by the researcher for future use such as for classification model development.

Later, further assessments were performed to confirm on the final number of principle component, k that best placing the original set 10 sensors. The three common and straight forward tools used for determining the number of component that sufficient to replace the original sensors are Scree plot, Kaiser criterion and cumulative percentage of total variance explained (5). The summary of assessments is shown in Table 1.

Table 1: Assessment tools for determining the number of PC.

Name	Rule
Scree plot	Consider the number of component before the first bend from the plot.
Kaiser Criterion	Look for eigenvalues $\lambda > 1$
Cumulative percentage of total variance explained	Number of components that exceed 70% of variance explained would be satisfy

In assessing the most important and influential variable to a component, correlation values were considered to give that useful information. The highest correlation between variable and component will be indicated as the most important and influential among all other variables (5).

3. Results and findings

In this section, we present the results of statistical analysis. The procedure on how to determine the number of principal component is shown in details, and followed by the discussion.

3.1. Number of component to retain

Based on the scree plot in Figure 1, it is clear to visualize that the first bend occur at the third component. Therefore, it recommended to retain two components. Next, the inspection using Kaiser criterion in Table 2, also recommended that two components to be retained.

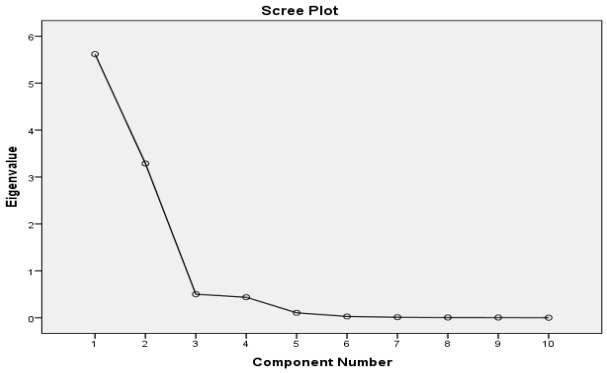

Fig. 1: Scree plot

The eigenvalues produced by Z_1 and Z_2 are 5.618 and 3.286, respectively. Both are greater than the cut off value (equal to 1). Meanwhile, the eigenvalues for the other 8 components are observed to have value less than 1.

Table 2: Eigenvalues and total variance explained

Component	Eigenvalues	Proportion of total variance (%)	Cumulative percentage of total variance explained (%)
Z_1	5.618	56.185	56.185
Z_2	3.286	32.864	89.048
Z_3	.502	5.021	94.069
Z_4	.438	4.378	98.447
Z_5	.105	1.052	99.499
Z_6	.027	.271	99.771
Z_7	.011	.114	99.884
Z_8	.006	.055	99.939
Z_9	.004	.044	99.984
Z_{10}	.002	.016	100.000

The principle component analysis for the rice odour's readings is shown in Table 2. It includes, total initial eigenvalues for each component, proportion of variance and cumulative percentage of total variance explained by each component. The results show 10 components is analyzed, two out of 10 have the eigenvalue greater than 1 where explained 89.048% of the total variances. From the result, Z_1 has a total of 56.2% of total variance explained, and with Z_2 in the system, it adding the total percentage of the variance explained turn into 89.1%. Based on the cut-off point 70%, it can concluded that, two (2) components are enough to be retained in the system.

3.2. Important and the most influential sensor

Further assessment on the two components retained, is to investigate the most important sensor among all sensors. According to the value of correlation (factor loading) in Table 3, Sensor 1 and Sensor 2 have a contribution of more than 90% to Z_1. But then, Sensor 2 (96.6%) is slightly higher than the contribution made by Sensor 1 (93.1%). For Z_2 (in Table 3), it is obvious to conclude that, only one

sensor that contribute to the highest among all sensors, which is Sensor 8 with 82.8%. All other sensors' contribution are below

Table 3: Correlation of each sensors for each component

Component	Z_1	Z_2
$SENSOR_1$	.931	.188
$SENSOR_2$	.966	.133
$SENSOR_3$	.687	.670
$SENSOR_4$	.764	.600
$SENSOR_5$	.514	.628
$SENSOR_6$	-.730	.663
$SENSOR_7$	-.636	.752
$SENSOR_8$	.137	.828
$SENSOR_9$	.903	-.299
$SENSOR_{10}$	-.854	.501

4. Conclusion

In this paper, the researcher highlighted the application of PCA onto one set of high-dimensional data where PCA is used a method to overcoming the issue of interdependency. The presented results showed, how PCA works to cater the problem.

For the dataset used in this study, originally there are 10 sensors that were used to capture the odour of rice sample. The direct uses of all the odours' readings from all 10 sensors are in the development of classification model to classify the rice into its type. However, after PCA was performed, only two (2) components are suggested to be used for the model development. In the process of building up the component, PCA did not left out any original variables because, it is believe that such sensor brings much information about the subject under studied.

In conclusion, PCA as a dimension reduction technique suggest fewer number of component that developed through a linear combination of the original set of variables, instead of using the original set of variables that were found to be having interdependency problem between the original variables in the set.

References

[1] Clark NR, Ma'ayan A. Introduction to statistical methods to analyze large data sets: principal components analysis. Science signaling. 2011;4(190):tr3.

[2] Clarke R, Ressom HW, Wang A, Xian J, Liu MC, Gehan EA, et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat Rev Cancer. 2008;8(1).

[3] Fan J, Lv J. A Selective Overview of Variable Selection in High Dimensional Feature Space. 2010;20(1).

[4] Mwangi B, Tian TS, Soares JC. A review of feature reduction techniques in neuroimaging. Neuroinformatics. 2014;12(2):229-44.

[5] Johnson R, Wichern D. Principle Components. Applied Multivariate Statistical Analysis. United States of America: Pearson; 2014. p. 430.

[6] Mizumukai K, Sato T, Tabeta S, Kitazawa D. Numerical studies on ecological effects of artificial mixing of surface and bottom waters in density stratification in semi-enclosed bay and open sea. Ecological Modelling. 2008;214(2-4):251-70.

[7] Wunderlin DA, Pilar DMD, Valeria AM, Fabiana PS, Cecilia HA, Angeles BMDL. Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study : Suquia River Basin (Cordoba-Argentina). Wat Res. 2001;35(12).

[8] Harada M, Yoshida I. Distribution of Sulfides and Heavy Metals of Bottom Sediments in lake Koyama. Transactions of Rural Planning. Journal of Rural Planning Association. 2001;20:64-6.

[9] Oketola AA, Adekolurejo SM, Osibanjo O. Water Quality Assessment of River Ogun using Multivariate Statistical Techniques. Journal of Enviromental Protection. 2013;4(05):466.

[10] Koch I. Analysis of Multivariate and High-Dimensional Data. United States of America: Cambridge; 2014.

[11] Principle Component Analysis- A Realization of Classification Success in Multi Sensor Data Fusion [Internet]. 2012.

[12] Champagne TE. Rice Aroma and Flavor : A Literature Review. Cereal Chemistry Journal. 2008;85:445.

[13] Bergman CJ, Bhattacharya KR, Ohtsubo K. Rice End-Use Quality Analysis. Rice: Chemistry and Technology. 3rd ed. USA: AACC Intl. PRESS; 2004. p. 415-72.