PRINCIPAL COMPONENT ANALYSIS OF MEASURES ON SOME VARIABLES OF MANGO LEAVES

Abdelfatai Mohammed[1*], Thompson E. Idike[1], Yahaya Gana[1] and Agnes Oyetunji[2]

[1]Department of Mathematics and Statistics, Niger State Polytechnic, Zungeru
[2]Department of Liberal Studies, Niger State Polytechnic, Zungeru
[*]Corresponding author e-mail: abdelfataimohammed@yahoo.com

ABSTRACT

This research work used a multivariate analysis technique called Principal Component Analysis (PCA) to analyze the data collected (mango leaves) on some mango trees on the main campus of Niger State Polytechnic, Zungeru in 2016. The researcher used direct observation method to collect data on two hundred (200) mango leaves which were sampled using systematic random sampling. Ten (10) mango trees were randomly selected out of the mango trees on the main campus. Out of the ten (10) mango trees, five (5) trees were systematically selected and on each of the five (5) trees, 40 leaves were randomly selected which made the total sample size, n = 200. Six components were measured on each of the leaves, which are; the leaf weight, length, width, stock-length, diagonal and vein branches. The statistical analysis using PCA, was conducted to obtain variables' reduction. The statistical results indicated that only two principal components, (P_1 and P_2) were to be retained based on the results of series of techniques applied. It is recommended that other researchers should further their research work on more variables as well as use more observations than we have done.

Keywords: Principal Component Analysis, Sampling Adequacy, Scree Plot, Loadings and Variance Explained

INTRODUCTION

Principal component analysis (PCA) originated in 1901 with research by Karl Pearson into 'common factors' emerging from correlated variables (Robert & Brooks, 2015). The time consuming labyrinthine mathematical computations restricted the use of this complex mathematical process until the arrival of computers in the 1950's. These packages (SAS and SPSS) contain principal component analyses programmed with easily used default settings. Such software has broadened the popularity of the process to neophytes. It is imperative that neophytes be aware of circumstances suitable for employing principal component analysis. To make valid generalizations they must be knowledgeable about techniques, analysis and interpretation of results. Critical evaluation of principal component analysis studies is also an essential skill.

Principal component analysis is a popular multivariate method in classical data analysis; Situations where PCA is most applicable include problems with high dimensional data and problems with highly correlated data, (Damon & Maron, 2014). PCA reduces the dimension of the data, hence reducing the computing power required to solve the problem. In another situation where correlation exists among the variables, PCA creates uncorrelated variables, hence eliminates the problem of collinearity. After the principal components

are computed, analysis and interpretation can then be performed on the principal components instead of the original variables. Due to its wide applicability, theories of PCA are well developed. The principal components may then be used as predictor or criteria variables in subsequent analyses. Principal component analysis is a large-sample procedure. To obtain reliable results, the minimal number of subjects providing usable data for the analysis should be 5-10 times the number of variables being analyzed.

Principal component analysis parsimoniously reduces information to the few common uncorrelated components or higher-level relationships (Everitt & Hothorn, 2011). Principal components represent most of the original data and make the data easier to understand and use. The process may also make prediction of the original variables possible (Robert & Brooks, 2015).

Principal component analysis is sometimes confused with factor analysis, and this is understandable, because there are many important similarities between the two procedures, since both are variable reduction methods that can be used to identify groups of observed variables that tend to hang together empirically and they sometimes even provide very similar results.

Principal components analysis technique has been widely applied to many scientific domains as diverse as engineering, astronomy, biology, remote sensing and economics (Jollife, 2002). The dimensionality of scientific data could be more than thousands, such as digital images and videos, gene expressions and DNA copy numbers, documents, and financial time series. As a result, data analysis on such data sets suffers from the curse of dimensionality. To solve this problem, the dimensionality reduction (subspace learning) algorithms have been proposed to project the original high-dimensional feature space to a low-dimensional space, wherein the important statistical properties are well preserved. Among these methods, the unsupervised dimensionality reduction methods are more useful in the practical applications since labeled data are usually expensive to obtain and we often have no any prior knowledge for new scientific problems (Feiping *et al*., 2014). Due to the recent advances in low-rank matrix, robust principal component analysis was suggested to decompose a given data matrix into a low-rank matrix and others (Wright *et al*., 2009).

Mango accessions were visually evaluated on site using the International Plant Genetic Resources Institute (IPGRI) descriptors for mango. The attributes of interest were recorded on a data sheet and pictures of the same were recorded (Burcu *et al*., 2009). These included measurements on tree height, stem circumference, tree growth habit (angle at which the main branches join the stem) and crown shape. For the leaves, the leaf attitude, color of young and fully mature leaves (CYL and CFL), fragrance strength, leaf blade shape (LBS), leaf blade length (LBL), leaf blade width (LBW), petiole length, leaf apex shape (LAS), leaf base shape, leaf margin type, leaf texture, pelvis thickness, leaf pubescence, angle of secondary veins to midrib and presence of secondary veins were evaluated.

Burcu *et al*. (2009) gave that the tree height was measured with the help of a ladder to a height of 10m. Trees taller than that were labeled as ‘over 10m.’ Stem circumference was measured at 50cm above the ground on mature trees. Leaf blade length was

determined from an average of ten mature leaves per tree, measured from the base to the tip of the leaf blade. The leaf width was determined by measuring the widest part of the leaf blade for ten leaves per tree. The petiole length was also measured as an average of ten leaf petioles per tree, measured from the base of the leaf blade to the stem. Fragrance strength was determined from a fully matured leaf when crushed. All other attributes including colors, habits and shapes were evaluated using IPGRI visual appraisals, with the colors for young leaves being determined on newly sprouted shoots while the colors of fully grown leaves were evaluated on normal-looking fully matured leaves (on an average of ten leaves per tree sample).

Robert and Brooks (2015) used principal component analysis to check the construct validity of the Multidimensional-Multiattributional Causality Scale (MMCS). The scale measures attribution of success or failure to factors such as ability, effort, context and luck. Three hundred and fifty American community college students enrolled in developmental reading classes were surveyed on the achievement MMCS scale on eight, three item subscales. These were rated from 0 disagree to 4 agree. The intercorrelation matrix of the 24 item responses analyzed via principal component analysis. Seven factors with eigenvalues of one or greater were extracted and rotated by the varimax method. These seven factors accounted for 56.7% of the total variance. Findings indicated high loadings with only one item appearing on more than one factor. The factor structure indicated items distinguished between attributions of success or failure to ability, effort, context and luck.

METHODOLOGY

Method of Data Collection

The researcher used direct observation method to collect data on two hundred (200) mango leaves which were sampled using systematic random sampling, in the main campus, Niger State Polytechnic, Zungeru, September, 2016. Ten (10) mango trees were randomly selected out of the mango tress in the main campus. Out of the ten (10) mango trees, five (5) trees were systematically selected and on each of the five (5) trees, 40 leaves were randomly selected which make the total sample size (n) of 200. Six components were measured on each of the leaves, the six components measured were: the leaf weight, length, width, stock-length, diagonal and vein branches. The weights (mg) were measured by weighing balance machine, while the vein branches were counted, however tape rule was used to measure the leaf length, width, stock-length and diagonal

Principal Component Analysis

The aim of principal component analysis (PCA) is to construct out of a set of variables x_{js}(j = 1, 2, 3, 4,…,k), a new set of variables, (p_i) called principal components, which are linear combination of the x_{js} is obtained below as given in (Onyeagu, 2003).

That is;

$$p_1 = a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1k}x_k$$
$$p_2 = a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots + a_{2k}x_k$$
$$p_3 = a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \ldots + a_{3k}x_k$$
$$\vdots \quad \vdots \quad \vdots \quad \cdots \quad \vdots$$
$$p_k = a_{k1}x_1 + a_{k2}x_2 + a_{k3}x_3 + \ldots + a_{xk}x_k$$

It should be noted that the method of principal component analysis can be obtained by using the original values of x, the deviation of mean from each x_i or the standardized value where $z_i = {}^{x_i}/_{\delta}$, this is better because $x_i$takes care of variable that

are recorded in different units. The a_{is} constructed principal component satisfies two conditions:

I. The principal component are uncorrelated.

II. The first principal component (p_1) absorbs the greater proportion of the total variation in the set of all x_{is}

That is, $p_1$accounts for the maximum possible proportion in the variation. Also p_2 accounts for the maximum possible proportion of the remaining balance of the variance, after p_1 has taken maximum proportion and so on. Only principal components whose latent roots (λ_{is}) that are greater than one, that is $\lambda_i > 1$, should be retained. All others that are $\lambda_i < 1$ should be discarded. By implication, we are to stop computing the remaining principal components, at the point when ($\lambda_i < 1$) (Onyeagu, 2003).

The correlation between $x_I \; and \; x_j$ is given by the formular.

$$\rho_{ij} = \frac{cov\,(X_i X_j)}{\sqrt{v(X_i)v(X_j)}}$$

Where $-1 < \rho_i < +1$

Then $X_i X_j = \sum X_i X_j - \frac{\sum X_i \sum X_j}{N}$

And $X_i^2 = \sum X_i^2 - \frac{(\sum X_i)^2}{N}$

and then $A = \begin{bmatrix} \rho_{11} & \cdots & \rho_{16} \\ \vdots & \ldots & \vdots \\ \rho_{61} & \cdots & \rho_{66} \end{bmatrix}$

Proportion

The proportion of variance explained by the k[th] principal component

$$= \frac{Y_k}{Y_1 + Y_2 + Y_3 + \cdots + Y_p}$$

Where Y_k is the k[th]eigenvalue

Cumulative proportion

The cumulative proportion of variance explained by the first k principal component

$$= \frac{Y_1 + Y_2 + Y_3 + \cdots + Y_k}{Y_1 + Y_2 + Y_3 + \cdots + Y_p}$$

Where Y_k is the k[th]eigen value.

Eigenvectors

Eigenvectors comprise of coefficients corresponding to each variable, the weight for each variable used to calculate the principal component scores.

However, before choosing to analyze data using Principal components analysis, some process must be put into consideration, part of the process involves checking to make sure that the data you want to analyze can actually be analyzed using PCA. You need to do this because it is only appropriate to use PCA if your data "passes" four assumptions that are required for PCA to give you a valid result. In practice, checking for these assumptions requires you to use SPSS Statistics to carry out a few more tests, as well as think a little bit more about your data, but it is not a difficult task.

In establishing the reliability of the data, some tests of statistics were conducted, such statistics are Kaiser-Meyer-Olkin, a Measure of Sampling Adequacy (usually called the MSA) and Bartlett test of Sphericity, all are available in SPSS.

RESULTS AND DISCUSSION

Table 1**:** Summary of data collected (measurements) on some variables of selected mango leaves, Niger State Polytechnic main campus, Zungeru, 2016/2017

Variable	Mean	Standard Deviation	Piece of leaves	Minimum	Maximum
Weight (gm)	4.0108	1.21491	200	1.50	**6.95**
Stock length (cm)	5.0815	1.53532	200	2.50	**10.80**
Length (cm)	28.4980	6.86482	200	18.5	**77.5**
Width(cm)	6.6635	3.85600	200	3.0	**10.2**
Diagonal(cm)	27.2675	5.78159	200	16.0	**45.5**
Vein branches (count)	22.5750	3.59709	200	18	**29**

The aim of this study is to conduct a principal component analysis to determine the number of components to retain in the data set. The essence of this work was to know the number of principal components to be extracted (Onyeagu, 2003), using kaiser's criterion as well as to plot the latent roots against the order of extraction of the p's and to use the shape of the resulting curve to judge, how many p's are to be retained in the analysis using Cattel's criterion (Scree plot).

The data were run using SPSS statistical package and the results were given in the tables below:

Table 2: KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.726
Bartlett's Test of Sphericity	Approx. Chi-Square	526.864
	Df	15
	Sig.	.000

This result gave the permission to continue running the analysis based on the KMO result (0.726) of measure of sampling adequacy (SMA) as well as the result of Bartlett's test of Sphericity (0.000).

Table 3: Anti-image Correlation

	Weight	Stock length	Length	Width	Diagonal	Vein branches
Weight	.822[a]	-.376	-.003	-.105	-.336	-.031
Stock length	-.376	.805[a]	-.105	-.079	.045	-.142
Length	-.003	-.105	.682[a]	-.006	-.806	-.035
Width	-.105	-.079	-.006	.772[a]	-.028	.160
Diagonal	-.336	.045	-.806	-.028	.661[a]	-.031
Vein branches	-.031	-.142	-.035	.160	-.031	.785[a]
Measures of Sampling Adequacy(MSA)						

We can see that we have good values for all variables for the Measures of Sampling Adequacy (MSA) except length and diagonal but the overall value is a bit good at 0.755, however Bartlett's Test of Sphericity has an associated P value (sig in the table) < 0.001 as by default SPSS reports p values of less than 0.001 as 0.000. So, from the above results, it was glaring that we can now continue and perform a valid principal component analysis.

Data Analysis using SPSS

Table 4: Descriptive statistics

	Mean	**Std. Deviation**	**Analysis N**	**Coefficient of variation (CV)**
Weight	4.0108	1.21491	200	30.3
Stock length	5.0815	1.53532	200	30.2
Length	28.4980	6.86482	200	24.1
Width	6.6635	3.85600	200	57.9
Diagonal	27.2675	5.78159	200	21.1
Vein branches	22.5750	3.59709	200	15.9

The table above simply shows the means, standard deviations, sample size for each variable and coefficient of variation. It was observed that width has the highest coefficient of variation (57.9), while vein branches have least coefficient of variation (15.9).

Table 5: Correlation Matrix

	Weight	Stock length	Length	Width	Diagonal	Vein branches
Weight	1.000	.551	.617	.220	.676	.191
Stock length	.551	1.000	.428	.172	.425	.223
Length	.617	.428	1.000	.167	.887	.204
Width	.220	.172	.167	1.000	.181	-.096
Diagonal	.676	.425	.887	.181	1.000	.203
Vein branches	.191	.223	.204	-.096	.203	1.000

Table 5 above is a correlation matrix showing the correlation of the variables. It was observed that the correlation coefficient of diagonal and length was very high (r=0.887), followed by the correlation coefficient of weight and diagonal (r=0.676), while length and weight correlation coefficient gave (r=0.617)

Table 6: Communalities

	Initial	**Extraction**
Weight	1.000	.716
Stock length	1.000	.473
Length	1.000	.775
Width	1.000	.656
Diagonal	1.000	.807
Vein branches	1.000	.635

Extraction Method: Principal Component Analysis

Table 6 above is the table of estimated communalities, this is the amount of the variable's variance that is accounted for by the components (since the loadings are correlations between variables and components and the components are orthogonal, a variable's communality represents the R^2 of the variable predicted from the components). For this data the communalities are weight, 72; stock length, 47; length, .78; width, 66; diagonal, .81; vein branches, 64. All have good communalities except stock length.

Table 7: Rotated Component Matrix

	Component	
	1	2
Weight	.845	.048
Stock length	.686	-.043
Leave length	.880	-.013
Width	.301	.752
Diagonal	.898	.001
Vein branches	.326	-.727

Extraction Method: Principal Component Analysis

Rotation Method: Varimax with Kaiser Normalization

Table 7 above shows the factor loadings that resulted from varimax rotation: These two rotated factors are just as good as the initial factors in explaining and reproducing the observed correlation matrix. In the rotated factors, Diagonal, Stock length, Leave length and Weight all have high positive loadings on the first factor (and low loadings on the second). Only width has high loadings on the second factor.

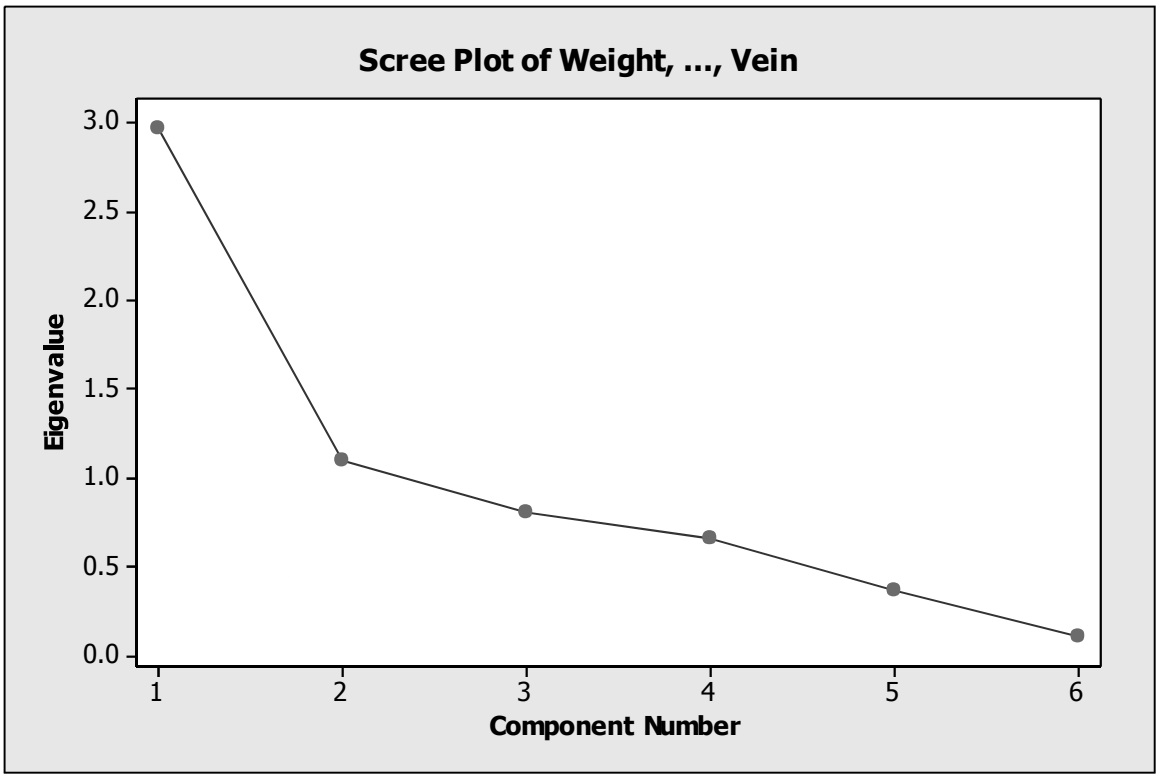

Figure 1: Scree Plot

The Scree plot (Fig. 1) confirmed the results in table 7 that only two principal components (1 and 2) are to be retained while others are neglected.

Table 8: Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.963	49.386	49.386	2.963	49.386	49.386	2.963	49.385	49.385
2	1.099	18.309	67.695	1.099	18.309	67.695	1.099	18.310	67.695
3	.805	13.425	81.120						
4	.661	11.011	92.131						
5	.364	6.068	98.199						
6	.108	1.801	100.000						

Extraction Method: Principal Component Analysis

Table 8 above showed the importance of each of the six principal components. We observed that only the first two components have eigenvalues >1.00. The first principal component has eigen value = 2.96 and accounted for 49.386% of variance, while the second principal component has eigenvalue = 1.09 and accounted for 18.309% of variance. The two amounted to about 67.7% of the total variability in the data. This led us to the conclusion that a two-factor solution will be adequate. This conclusion is supported by the scree plot (which is actually simply displaying the same data visually).

Furthermore, the middle part of the table showed the eigenvalues and percentage of variance explained for just the two factors of the initial solution that are regarded as important. Clearly the first factor of the initial solution was much more important than the second. However, in the right-hand part of the table, the eigenvalues and percentage of variance explained for the two rotated factors were displayed. The variance explained was equal to the sum of squared loadings (SSL).The eigenvalue for the first component can be obtained by adding up the square of all the factor loadings in the rotated component matrix for the first component, that is:

$$(.845^2 + .686^2 + .880^2 + .301^2 + .898^2 + .326^2) = 2.96.$$

Second component:

$$(0.048^2 + 0.043^2 0.013^2 + 0.752^2 + 0.001^2 + 0.727^2) = 1.09$$

Table 9: Reproduced Correlations

		Weight	Stock length	Length	Width	Diagonal	Vein branches
Reproduced Correlation	Weight	.716[a]	.578	.743	.290	.759	.241
	Stock length	.578	.473[a]	.605	.174	.617	.255
	Length	.743	.605	.775[a]	.254	.791	.297
	Width	.290	.174	.254	.656[a]	.271	-.449
	Diagonal	.759	.617	.791	.271	.807[a]	.292
	Vein branches	.241	.255	.297	-.449	.292	.635[a]
Residual[b]	Weight		-.026	-.126	-.070	-.083	-.050
	Stock length	-.026		-.177	-.002	-.192	-.032
	Length	-.126	-.177		-.088	.096	-.093
	Width	-.070	-.002	-.088		-.090	.353
	Diagonal	-.083	-.192	.096	-.090		-.089
	Vein branches	-.050	-.032	-.093	.353	-.089	

Extraction Method: Principal Component Analysis.

a. Reproduced communalities
b. Residuals are computed between observed and reproduced correlations
There are 11 (73.0%) non redundant residuals with absolute values greater than 0.05.

Table 9 above showed the extent to which the original correlation matrix can be reproduced from the two factors. The small residuals showed that there is very little difference between the reproduced correlations and the correlations actually observed between the variables. The two factors solution provided a very accurate summary of the relationships in the data.

CONCLUSION AND RECOMMENDATIONS

From the analysis carried out, it was observed that the data collected were fitted for principal component analysis. Only two components have eigenvalues >1.00. The first component accounted for the largest possible amount of variance (49.386%), the second component formed from the variance remaining after that associated with the first component has been extracted, accounted for the second largest amount of variance (18.309%). The first principal component has eigen value = 2.963, while the second principal component has eigenvalue = 1.099. The first two components explained for about 67.7% of the total variability in the data. However, from the results obtained it was concluded that only the first two components should be retained. This conclusion was supported by the scree plot (which actually simply displayed the same data visually, that is, the first two components should be retained). It is recommended that othe researchers should attempt to measure more variables for further statistical analysis.

REFERENCES

Burcu A., Gabor P, Haonan W., Bullitt E and Maron J. S. (2009). A principal component analysis for trees, *The Annals of Applied Statistics*, **3** (4), 1597-1615.

Damon J and Maron J. S (2014). Backward principal component analysis and principal nested relations, *Journal of Mathematical Imaging and Vision*, **50** (1-2), 107-114.

Feiping N, Jianjun Y and Heng H (2014). Optimal mean robust principal component analysis, *International Conference on Machine Learning*, JMLR: W&CP, **32**, Beijing, China.

Onyeagu S. I (2003). *A first course in multivariate statistical analysis*, MegaConcept, Awka, Nigeria.

Qian Z, Deyu M, Zongben X and Lei Z (2014). Robust Principal Analysis with Complex Noise, *International Conference on Machine Learning*, JMLR: W&CP, **32**, Beijing, China

Robert R and Brooks J. P (2015). Principal Component Analysis and Optimization: A Tutorial, *Journal of Informs Computing Society*, Richmond, Virginia, **14,** 221-225.

Jollife I. T (2002). *Principal Component Analysis* (2nd ed.). Springer-Verlag, New York