International Journal of Multidisciplinary Research and Development
Online ISSN: 2349-4182, Print ISSN: 2349-5979; Impact Factor: RJIF 5.72
Received: 08-11-2019; Accepted: 09-12-2019
www.allsubjectjournal.com
Volume 7; Issue 1; January 2020; Page No. 34-37

Analysis of agriculture data using principal component analysis

Vikas Singh[1], Alka Singh[2]
[1] Department of Statistics, Central University of South Bihar, Gaya, Bihar, India
[2] Department of Economics, Baba Saheb Bhimrao Ambedkar University, Lucknow, Uttar Pradesh, India

Abstract
Principal component analysis is a multivariate statistical method of data analysis which is used to reduce the dimension of the data. The reduction of dimension is achieved by forming new variables which are linear combinations of variables in the data set. These linear combinations are chosen to account for as much of the original variance-covariance/correlation structure in the original variables as possible. In PCA, our main aim to maximize the variance of a linear combination of the variables. This paper analyses Indian agricultural crop data which consists of the eight major crops reported to the country for the period 2016. The crops consist of rice, bajra, jwar, maize, marua, wheat, barley, and oth_care. In this paper, we applied the principal component analysis (PCA) to explain the correlation between the crops. PCA has recommended that only two PC's explain 93% of the total variability of the data set. Data analysis was carried out using R- Software. The Scree and Loading plot shows that correlation exists between crops. The datasets for this study has been taken from Agriculture department, Govt. of India for the year 2016.

Keywords: principal component analysis, scree plot, loading plot, crop data

1. Introduction

Principal component analysis is used as instrument or technique which is most of the time utilized in different sectors and dimensions. Because be trying to find out the relationship between the variables. There are so many areas where we used this technique. Among all the sectors one of the important sector is agriculture. Where wildly used this technique. (Paniskhan et al., 2012). There are many authors indicate that Agriculture is an employment creator and self-motivator sector in India where the good amount of goods and services contributing in GDP (Limbore and Khillare, 2015). On an average approximately 70% of population and 10% of the urbanized population depends on agricultural activities as his/her survival. Basically farmers' livelihood directly related to the agriculture. From the ancient time, India is one of the best crop cultivators and today's time farmers developed the crop growing skills, technologies, education, chemical fertilizers, and different seeds varieties. Therefore, India is major producers or supplier in terms of crop production of respective agricultural food items like oil meals, fresh fruits, fresh vegetables, meat or non-vegetarian items, marine area side products, tea, coffee, and rice this is basic require things in the national and the international markets. India is an eminent producer country of respective agricultural products. India is one of the top producer countries in the world in terms of milk and got second rank for rice and rice wheat production. Quality of wheat and rice is also undecomposed. One of study found that is we are use inorganic fertilizers that time increased grain quality of maize (ALmaz, 2017) [1].

There are many studies shows that if few percent increases in crop production then also increases the growth rate. Basically agriculture is the main source of livelihood for all and especially for the farmers. Our country also depends on agriculture, because this is the direct source of income, gross domestic product and per capita income. There are many other factors also contributing in the production process likewise, the land, capital labor and many other resources. Without these assets farmers cannot grow or produce anything. Many factors are significantly related to the economic growth and some are not. In different studies and proficient authors said that these factors are benchmark of growth era period. One of adroit author has been concentrate to the aspect, which is growth decomposition method in agriculture sector. That is useful for the policy makers and growth target achievers. There are various components is reflect the growth status viz., area, yield and cropping intensity. To provide the infrastructure projects with ultimate goals and target or polices, to improvement the growth in the different sectors such as important in cropping intensity, yield and area. Non parametric measure shows for correlation but according to Ajay (2018) in his study shows the highly significant positive correlation among the measure.

Accordingly Rahman (2011) explained that there are two important causes for a huge expansion in agriculture and non-agriculture sector. This is an outcome of agriculture increments. First insights in agriculture has substantial background to copulation high firm output industries with using less inputs, let in high chemicals, fertilizers, modern technology in the forms of machinery as well as good quality of food and fiber provider. Second insights, promoting the income through agriculture which is mostly exhausted on goods and services produced by locally. Particularly, demonstrate richly income elasticity which create demand and generate the employment. Although, increased agricultural production is significantly interrelated with entire economic growth perspective. This is the indication that to increase employment opportunity and decrease the poverty.

The cultivation of High-Value Crops (HVCs) has caught attention in the agricultural sector in recent years. The

consumption basket of general mass has undergone a change reflecting a decrease in the share of traditional food grains (rice, wheat, pulses, and cereals); and an increase in the share of non-traditional foods and beverages (fruits, dry fruits, vegetables, fish, poultry, milk, etc.). As a result, the market demand for high-value food crops has increased substantially from 2% (1992-93) to 14% (2003). The agricultural policy analyst is also being observed that increased share of HVCs in the gross value of the agriculture output. In this study main focus beyond farm production (Shruti, 2019) [15].

1. Applicability of Principal Component Analysis

According to Rotaru (2012) [11], Principle components analysis (PCA) is a component technique that was preferred by greatest author Pearson in the period of 1901 and which has developed by Hoteling in 1933 (Rotaru, 2012) [11]. Principle component analysis method comprises or converted huge data in a small datasets (Helmy, 2009) [5]. Few specifications of the method which is required some steps viz., associations between the factors or variables (correlations) and at the preferred the less variable for showing the variability. Theses phenomenon or event called factor analysis which is divided in some factors and in some components. This method uses many fields. But one of the most important fields is agriculture. Where use this method reversely.

Whenever talking about the factor analysis in agriculture sector. That time also talking about the very important component that is water uses. Means what is the existence of water in the agriculture sector. This is a major concern in Kenyan public irrigation schemes low productivity and less water uses. (Muema, 2018) [14]. Public irrigation schemes are monitor and evaluate the necessary performance. The author Analyze the performance of three rice growing irrigation schemes in western Kenya using benchmarking evaluation. The performance of the irrigation schemes was evaluated for the period from 2012 to 2016 using eleven performance indicators under agricultural productivity using by principal component analysis and some other categories like water supply, and financial performance also include.

According to author Li Chen (2011) [3], the gap of finance between two well-known areas urban and rural occupier using data of Anhui from 2006 to 2010 shows in Anhui province is continuously progressive (Chen, 2011) [3]. The aim is to reduce the gap in agriculture between urban and rural occupants, three aspects viz., agriculture industrialization, the income of peasants increasing and rural urbanization. PCA is the tool of discussing agricultural industrialization.

2. Principal Component Analysis (PCA)

The methods of PCA consists that summarization and visualization of the data context. Cope with the vulnerable information with the multivariate data set. This is the simplification statistical method used for Principal component analysis. The method creates a new set of variables, called principal components. This is the linear combination of the original variables in the each principal component. Principal components are orthogonal to each other, so there is no exhibiting information. In principal component analysis, we seek to maximize the variance of a linear combination of the variables. The reduction of dimension is achieved by forming new variables which are linear combinations of variables in the original data set. These linear combinations are chosen to account for as much of the original variance-covariance/correlation structure in the original variables as possible. The new variables formed by these linear combinations are uncorrelated which is a desirable property.

2.1 Technology and Concept

The i^{th} principal component (Y_i) is a linear combination of the original of the *p* variables in the data set as given by

$$Y_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p, \qquad i = 1, \ldots, p$$

The coefficients (a_{ij}) are called the loadings for the i^{th} principal component (Y_i). The interpretation of loadings can be an important part of the PCA (Jolliffe, (2002)) [6]. First, we can gain important insight into how the original variables relate to one another from them. Also examining the loadings leads to our understanding of what the scores (values) for the i^{th} principal component (Y_i) are measuring. We hopefully can capture much of the information in the original variables $(X_1, \ldots, X_p)$ with a much smaller number of principal components $(Y_1, \ldots, Y_k)$ where $k << p$.

2.2 Statistical Methodology

As we know that principal component analysis does not necessarily rely on the multivariate normal distribution; however, if this is assumed a PCA (Rencher, (2002)) definitely performs better. The principal components are linear combinations of the columns of our data matrix. Principal components depend only on the variance/covariance matrix (or more generally on the correlation matrix) of the original *p* variables $X_1, X_2, \ldots, X_p$.

Consider the following linear combinations of the data matrix $\boldsymbol{X}$ that has variance/covariance matrix Σ.

$$Y_1 = \mathbf{a_1}'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + \ldots + a_{1p}X_p$$
$$Y_2 = \mathbf{a}_2'\mathbf{X} = a_{21}X_1 + a_{22}X_2 + a_{23}X_3 + \ldots + a_{2p}X_p$$
$$\vdots$$
$$Y_p = \mathbf{a}_p'\mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + a_{p3}X_3 + \ldots + a_{pp}X_p$$

The variance/covariance term for each linear combination is given by

$$Var(Y_1) = \mathbf{a_1}'\Sigma\mathbf{a_1} \quad , \quad Var(Y_2) = \mathbf{a}_2'\Sigma\mathbf{a}_2 \quad ,\cdots, \quad Var(Y_p) = \mathbf{a}_p'\Sigma\mathbf{a}_p$$

and

$$Cov(Y_1, Y_2) = \mathbf{a_1}'\Sigma\mathbf{a}_2 \quad , \quad Cov(Y_1, Y_3) = \mathbf{a_1}'\Sigma\mathbf{a}_3 \quad ,\cdots, \quad Cov(Y_1, Y_p) = \mathbf{a}_p'\Sigma\mathbf{a}_p$$
$$Cov(Y_2, Y_3) = \mathbf{a}_2'\Sigma\mathbf{a}_3 \quad ,\cdots, \quad etc.$$
$$etc.$$

The principal components are defined to be the uncorrelated linear combinations that achieve maximum variances for $Var(Y_1), Var(Y_2), \cdots, Var(Y_p)$. In particular,

First Principal Componen	The linear combination $\mathbf{a_1}'\mathbf{X}$ that maximizes $Var(\mathbf{a_1}'\mathbf{X})$ subject to the constraint of $\mathbf{a_1}'\mathbf{a_1} = 1$

t:

Second Principal Componen t: The linear combination $\mathbf{a}_2'\mathbf{X}$ that maximizes $Var(\mathbf{a}_2'\mathbf{X})$ subject to the constraint of $\mathbf{a}_2'\mathbf{a}_2 = 1$ And $Cov(\mathbf{a}_1'\mathbf{X}, \mathbf{a}_2'\mathbf{X}) = 0$

Third Principal Componen t: The linear combination $\mathbf{a}_3'\mathbf{X}$ that maximizes $Var(\mathbf{a}_3'\mathbf{X})$ subject to the constraint of $\mathbf{a}_3'\mathbf{a}_3 = 1$ AND $Cov(\mathbf{a}_1'\mathbf{X}, \mathbf{a}_3'\mathbf{X}) = 0$ AND $Cov(\mathbf{a}_2'\mathbf{X}, \mathbf{a}_3'\mathbf{X}) = 0$ etc.

Result 1

Let $\hat{\Sigma}$ be the sample variance/covariance matrix associated with $\mathbf{X}$ (the data matrix). Let $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ be the eigenvalues/eigenvectors of Σ where it is assumed $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, then the principal components are given by

$$\begin{aligned} Y_1 &= \mathbf{e}_1'\mathbf{X} \quad \text{and} \quad Var(Y_1) = \mathbf{e}_1'\hat{\Sigma}\mathbf{e}_1 = \lambda_1 \\ Y_2 &= \mathbf{e}_2'\mathbf{X} \quad\quad\quad Var(Y_2) = \mathbf{e}_2'\hat{\Sigma}\mathbf{e}_2 = \lambda_1 \\ &\vdots \quad\quad\quad\quad\quad\quad \vdots \\ Y_p &= \mathbf{e}_p'\mathbf{X} \quad\quad\quad Var(Y_p) = \mathbf{e}_p'\hat{\Sigma}\mathbf{e}_p = \lambda_p \end{aligned}$$

and because the eigenvectors are orthogonal we have the desired covariance results, namely $Cov(e_j'X, e_k'X) = 0$ for all $j \neq k$. Also if we standardize the variables first then $\hat{\Sigma} = R$, the sample correlation matrix.

Result 2

Let Σ be the variance/covariance matrix associated with $\mathbf{X}$ (the data matrix) and let $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ be the eigenvalues/eigenvectors of Σ where it is assumed $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$.

$$\sum_{j=1}^{p} Var(X_j) = \sigma_{11} + \sigma_{22} + \ldots + \sigma_{22} = \lambda_1 + \lambda_2 + \ldots + \lambda_p$$

Result 2 imply that λ_1 can be interpreted as the contribution to the total variance that is due to the 1st principal component (i.e. the linear combination of the original variables with maximal variance), λ_2 can be interpreted as the contribution to the total variance that is due to the 2nd principal component, etc.

3. Data Analysis and Result

Data analysis using PCA has been done through R-software (Soren, 2006).Table 1 shows the correlation matrix between different types of crops in India. We can see from the table that there is no significant correlation between the major crops which means that none of the variables can be used to predict (explain) one another excluding wheat, rice, maize, bajra and etc. Table (2) gives the descriptive statistics of the crop production data.

Table 1: Correlation between different Crops

	Rice	Jwar	Bajra	Maize	Marua	Wheat	Barley	OTH_Cerelas
Rice	1.00	-0.77	-0.54	0.86	-0.62	0.96	-0.94	-0.90
Jwar	-0.77	1.00	0.81	-0.87	0.95	-0.85	0.82	0.93
Bajra	-0.54	0.81	1.00	-0.59	0.80	-0.64	0.66	0.75
Maize	0.86	-0.87	-0.59	1.00	-0.79	0.92	-0.83	-0.87
Marua	-0.62	0.95	0.80	-0.79	1.00	-0.73	0.71	0.87
Wheat	0.96	-0.85	-0.64	0.92	-0.73	1.00	-0.95	-0.93
Barley	-0.94	0.82	0.66	-0.83	0.71	-0.95	1.00	0.95
OTH_Cerelas	-0.94	0.93	0.75	-0.87	0.87	-0.93	0.95	1.00

Table 2: Descriptive Statistics of different crop

	Rice	Jwar	Bajra	Maize	Marua	Wheat	Barley	OTH_Cerelas
Min	29991	5755	7668	3250	1117	9624	620	687
1st Qu.	35864	9915	9744	5119	1816	13625	789	1563
Median	40511	16100	10961	5878	2333	23179	1493	3681
Mean	39270	14050	10749	5921	2124	21073	1830	3332
3rd qu.	42744	17059	11497	6376	2426	25991	2828	4962
Max	45456	18426	14132	9020	2682	32078	3547	6057

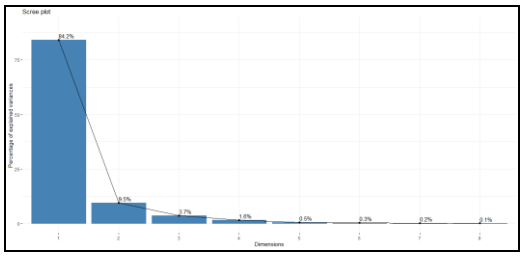

Fig 1: A Scree Plot for Crop production of Variability explained and No. of Components

The scree plot (Cattell, (1966)) [2] shows the rough bar plot of the cumulative proportions of the explained variance and the number of principal components is displayed in figure (1). From the scree plot, only the first two principal components explain maximum variability. The first PCs explain 84% of the total variability of the data while second PCs explain9% of the total variability. Therefore, we might want to stop at the second principal component. 93% of the information (variances) contained in the data are retained by the first two principal components.

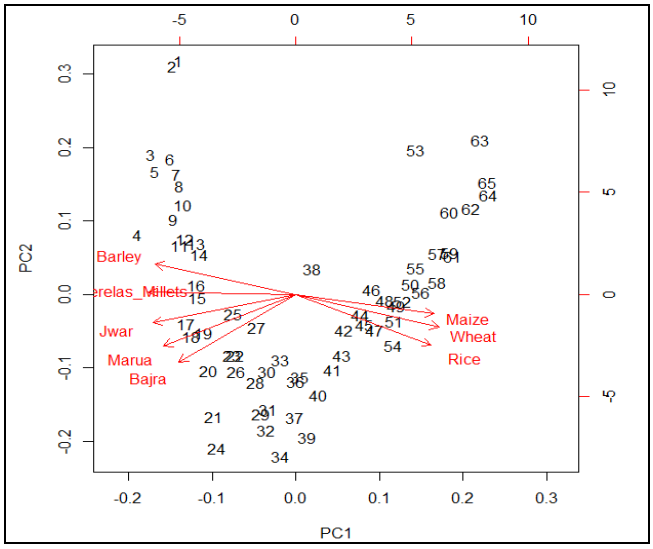

Fig 2: Bi-plot for different crop production

Figure (3) shows the loading plot for crop data. The loading plot shows that rice, wheat, and maize are the major production in India whereas the crop bajra, marua, jwar, etc. are produced in India but very less.

8. Conclusions

In this paper, we have taken the eight different variables (crops) from the agricultural department, Govt. of India and applied the method of Principal Component Analysis to explore the number of PC's to be retained. With the help of the multivariate method of PCA, the dimensionality of the data will reduced i.e. we reduce the eight distinct variables that can affect the Indian crop production to only 2 or 3 variables. The results shows that only two principal component explain up to 93% of the variability of the data set which gives evidence that the Indian State has a low productivity record.

There are low correlations between crops of type jwar, bajra, barley, etc. against rice, wheat, maize, etc. and therefore we cannot be used to explain one another. So this paper concludes that only the first two-component is sufficient to explain the total variability of the production of the crops and we can see from the result final data is reduced and only two variables are sufficient.

So for the conclusion point of view, it is also very important to focus on all crops including jwar, bajra, barley, mustard, arahar, etc. and the government needs to pay attention in this direction also so that the production could be increased in a better and simplified way.

9. References

1. Almaz MG, Halim RA, Yusoff Wahid. Effect of incorporation of Crop Residue and Inorganic Fertilizers on Yield and Grain quality of Maize. Indian Journal of Agricultural Research. 2017; 51;6.
2. Cattell RB. The Scree test for the number of factors. Multivariate Behavioral Research, 1966, 245-276.
3. Chen L. Principal Component Analysis of Anhui Agricultural industrialization.5th Computer and Computing Technologies in Agriculture (CCTA), Oct 2011, Beijing, China, Pp. 430-435.
4. Directorate of Economics and Statistics. Agricultural Situation in India. Department of Agriculture, Cooperation and Farmers Welfare Ministry of Agriculture and Farmers Welfare, Govt. of India, 2019, Vol. LXXVI, No. 6.http://eands.dacnet.nic.in/publications.htm
5. Helmy AK, Taweel GHS. Authentication Scheme Based on Principal Component Analysis for Satellite Image. International Journal of Signal Processing Image Processing and Pattern Recognition. 2009; 2:3.
6. Jolliffe IT. Principal Component Analysis. 2nd edn, Springer-Verlag, New York, 2002.
7. Linderman RH, Peter F, Gold RZ. Introduction to Bivariate and Multivariate Analysis. Scott, Foresman, and Company, 1980.
8. Panishkan K, Swangjang K, Sanmanee N, Sungthong D. Principle Component Analysis for the Characterization in the Application of Some Soil Properties. International journal of Environmental and Ecological Engineering. 2012; 6:5.
9. Rencher AC. Methods of Multivariate Analysis. 2nd edn, John Wiley & Son, New York, 2002.
10. Richard AJ, Dean WW. Applied Multivariate Statistical Analysis. 3rd edn, Prentice-Hall, New Delhi, 2001.
11. Rotaru AS, Pop ID, Vactca A, Cioban A. Usefulness of Principal Components Analysis in Agriculture. Bulletin UASVM Horticulture. 2012; 69:2.
12. Soren H. Example of multivariate analysis in R – Principal component analysis (PCA). httpgene:// tics.agrsci.dk/statistics/courses/Rcourse-Djf2006 / day3/ PCA-notes.
13. and K.K.S. An Analytical Study of Indian Agriculture Crop Production and Export with Reference to Wheat. Review of Research. 2015; 4:6.
14. Muema FM, Home PG, Raude JM. Application of Benchmarking and Principle Component Analysis in measuring Performance of Public Irrigation Schemes in Kenya. Agriculture. 2018; 8:162.
15. Shruti S, Sharma JP, Gills R. Potential and Prospects of Value chain Development for fruits and Vegetable in India. Indian Journal of Agricultural Science. 2019; 89:1.
16. Verma A, Singh j, Kumar V, Kharab AS, Singh GPGx E Interaction Estimation for Forage Yield of Dual Purpose Barley by Nonparametric Measures. Indian Journal of Agricultural Research, 2018; 52:6.