Research Article

Reason-A Technical Magazine
ISSN 2277-1654
Volume - XI ● 2012

REDUCTION OF ENVIRONMENTAL PARAMETERS USING PRINCIPAL COMPONENT AND FACTOR ANALYSIS

Satyendra Nath Mandal

Department of Information Technology, Kalyani Government Engineering College, Kalyani- 741235, Dist. Nadia, West Bengal, India. Email: satyen_kgec@rediffmail.com

Abstract : The output of any physical problem is likely to be dependent on huge number of parameters. But, many of them are not significant and some are highly correlated with other parameters. Same result can be produced by fewer parameters in stead of considering all parameters. In this paper, an algorithm has been proposed to reduce parameters based on principal component analysis and factor analysis. The algorithm has been applied to reduce the environment parameters which are needed for the healthy growth of mustard plant. It has been observed that the growth of mustard plant has not been disturbed if the significant environment parameters have been supplied sufficiently.

Keywords : Physical problem, environmental parameters, principal component analysis, factor analysis, significant parameters, plant growth.

1. Introduction

Factor analysis and principal component analysis have been applied as data reduction or structure detection methods. The term factor analysis has been introduced [1] by Thrustone in 1931. A hands-on how-to approach has been found by Stevens in 1986; more detailed technical descriptions have been provided [1] by Cooley and Lohnes in 1971, Harman in 1976, Kim and Muller in 1978, Lawley and Maxwell in 1971, Lindeman, Merenda and Gold in 1980, Morrison in 1967 and Mulaik in 1972. The interpretations of secondary factors in hierarchical factor analysis, as an alternative to traditional oblique rotational strategies, have been explained in detail by Wherry in 1984. When mining a dataset comprises of numerous variables, it is likely that subsets of variables are highly correlated with each other. Given high correlation between two or more variables, it can be concluded that variables are quiet redundant and they share the same driving principle in defining the outcome of the interest. They can be replaced by a single variable. In order to demonstrate the argument, a basic example may be taken. Measurement of two parametric properties of a planar shape, such as length and width of the shape may be considered that determines certain outcome of interest. As these two properties seem to be positively correlated, they can be replaced with a single new variable, i.e., the area of the shape. The area can still capture most of the information about the shape supplied by its length and width. The use of principal component analysis technique [2] has been well established in many fields such as pharmacology, climatology, numerous aspects of life science, economics, etc.

In this paper, an algorithm has been proposed to reduce parameters based on principal component analysis, and factor analysis. The algorithm has been applied

on environmental parameters which are necessary for healthy growth of mustard plant. The algorithm has been used to select the most significant environmental parameters among necessary parameters. It has been observed that the growth of mustard plant will not be suffered if these significant parameters are supplied sufficiently. This type of work has not been applied in prediction the growth of mustard plant previously. Detail of the methodology used and results obtained are discussed in this paper.

2. Theoretical Illustration of Principal Component and Factor Analysis

2.1. Principal Component Analysis

Principal Component Analysis [3-6] transforms the original set of variables into a smaller set of linear combination that account for most of the variance of the original set. The principal components are extracted so that the first principal component, denoted here by $PC_{(1)}$, accounts for the largest amount of the total variation in the data. $PC_{(1)}$ is the linear combination of the observed variables X_j, j = 1,2,3,....p given by :

$$PC_{(1)} = w_{(1)1}X_1 + w_{(1)2}X_2 + \ldots\ldots + w_{(1)p}X_p$$

where, the weights $w_{(1)1}$, $w_{(1)2}$, ... $w_{(1)p}$ have been chosen to maximize the ratio of the variance of $PC_{(1)}$ to the total variation.

The second principal component, $PC_{(2)}$ is that weighted linear combination of the observed variables which is uncorrelated with the first linear combination and which accounts for the maximum amount of the remaining total variation not already accounted for by $PC_{(1)}$. In general, the m^{th} principal component is that weighted linear combination of the X's given by :

$$PC_{(m)} = w_{(m)1}X_1 + w_{(m)2}X_2 + \ldots\ldots + w_{(m)p}X_p$$

which has the largest variance of all linear combinations that are uncorrelated with all of the previously extracted principal components.

2.2. Factor analysis

Factor analysis attempts to identify underlying variables, or factors, that explains the pattern of correlations within a set of observed variables [7].

It is assumed that X variables are related to a number of functions operating regularly, that is,

$$\begin{aligned} X_1 &= \alpha_{11}F_1 + \alpha_{12}F_2 + \alpha_{13}F_3 + \ldots + \alpha_{1m}F_m \\ X_2 &= \alpha_{21}F_1 + \alpha_{22}F_2 + \alpha_{23}F_3 + \ldots + \alpha_{2m}F_m \\ X_3 &= \alpha_{31}F_1 + \alpha_{32}F_2 + \alpha_{33}F_3 + \ldots + \alpha_{3m}F_m \\ &\ldots \\ X_n &= \alpha_{n1}F_1 + \alpha_{n2}F_2 + \alpha_{n3}F_3 + \ldots + \alpha_{nm}F_m \end{aligned} \quad \textbf{(1)}$$

where X's are variables with known data, α is a constant and F is a function of some unknown variables. For any of the X variables of equation 1, it may be written that

$$X = \alpha_1F_1 + \alpha_2F_2 + \alpha_3F_3 + \ldots + \alpha_mF_m \quad (2)$$

With the F's representing factors and α's representing loading.

3. Data Used in the Paper

A statistical survey has been conducted by a group of agricultural scientists on different mustard plants under the supervision of Prof. Dilip De of Bidhan Chandra Krishi Viswavidyalaya, West Bengal, India. In the present paper, environmental data from this survey have been taken between November to February 2003 as input data which effect the growth of mustard plant as furnished in

Table 1. The objective is to reduce the number of environmental parameters in prediction of growth of mustard plant. In Table1, D_1, D_2 and D_3 are soil moisture contents at different depths in the same place. The proposed algorithm to reduce parameters is furnished in section 4.

Table 1 Data related to environmental parameter

Maximum temp. (°C)	Minimum temp. (°C)	Rain fall (mm/hr)	Max. humidity (%)	Minimum humidity (%)	D1 (mm)	D2 (mm)	D2 (mm)	Sun shine (BTU)
27.4	15.6	28.8	95.75	58.35	16.9	21.96	28.13	7.77
25.9	14.15	19.2	96.68	59.48	14.83	19.63	26.1	8.11
24.4	12.7	9.6	97.6	60.6	13.23	17.7	23.76	8.32
24	12.33	29.68	97.78	60.73	11.86	16.33	21.63	7.44
23.6	11.95	49.75	97.95	60.85	18.86	14.9	19.76	7.18
23.2	11.58	69.83	98.13	60.98	11.03	15.03	19.3	7.62
22.8	11.2	89.9	98.3	61.1	10.76	14.93	19.13	8.48
24.08	12.38	74.73	97.93	59	10.8	14.83	18.96	7.32
25.35	13.55	64.55	97.55	56.9	9.8	13.8	17.96	7.14
26.63	14.73	51.88	97.18	54.8	9.3	13.1	17.03	6.05
27.9	15.9	39.2	96.8	52.7	8.13	12.36	16.03	7.48
28.53	16.38	79.3	96.43	51.83	7.86	11.86	15.46	9.42
29.15	16.85	119.4	96.05	50.95	6.96	10.13	14.33	7.22

Table 2 Correlation matrix

Variable	Correlation coefficient								
	Maximum temp.	Minimum temp.	Rain fall	Max humidity	Minimum humidity	D1	D2	D3	Sun shine
Maximum temp.	1.000	0.999	0.178	-0.917	-0.911	-0.425	-0.303	-0.279	0.039
Minimum temp.	0.999	1.000	0.146	-0.927	-0.897	-0.400	-0.269	-0.246	0.038
Rain fall	0.178	0.146	1.000	0.0094	-0.434	-0.581	-0.719	-0.731	-0.007
Maximum humidity	-0.917	-0.927	0.009	1.000	0.678	0.110	-0.070	-0.098	-0.114
Minimum humidity	-0.911	-0.897	-0.434	0.678	1.000	0.690	0.656	0.643	0.039
D1	-0.425	-0.400	-0.581	0.110	0.690	1.000	0.772	0.779	-0.002
D2	-0.303	-0.269	-0.719	-0.070	0.656	0.772	1.000	0.993	0.164
D3	-0.279	-0.246	-0.731	-0.098	0.643	0.779	0.993	1.000	0.164
Sun shine	0.039	0.038	-0.007	-0.114	0.039	-0.002	0.164	0.164	1.000

4. Proposed Algorithm

Algorithm : Parameter Reduction by PCA and FA
Input : Related parameters for particular problem
Output: Reduce parameters
Method:

Step 1. To collect data and store it in matrix form.

Step 2. To apply principal component analysis

- 2.1. To compute correlation matrix
- 2.2. To compute eigen values, total variances, cumulative eigen vector and percentage of contribution
- 2.3. To select the eigen vectors for which eigen value > 0.5
- 2.4. To select the maximum value from each component in reduced eigen vector
- 2.5. To reduce 10% of maximum value in each component in reduce eigen vector and to select other values if they are within the 10% of maximum value.
- 2.6. To select all variables corresponding to the maximum and 10% of maximum values.

Step 3. To apply factor analysis

- 3.1. To calculate the eigen vector
- 3.2. To calculate factor loading
- 3.3. To find the maximum value of each variable in factor loading
- 3.4. To select the variables corresponding the maximum values in each factor.

Step 4. To select the variables from principal component and factor analysis without repetitions

5. Implementation

The algorithm has been applied on environmental data for healthy growth of mustard plant furnished in Table 1. The algorithm has been applied in two steps. In the first step, principal component analysis has been tested and in the second step, factor analysis has been applied. The detailed steps have been given in sections 5.1 and 5.2.

5.1 Principal Component Analysis

The data from Table 1 has been taken and principal component analysis has been applied on it. The detailed steps as follows :

Step 1.

After the plantation, the environmental parameters that have been collected during the harvest period of growing stage of mustard plant are furnished in Table 1. Using Statistica 7 software package, the correlation matrix of Table 1 is evaluated and furnished in Table 2.

Step 2.

The eigen values, total variances, cumulative eigen vector and percentage of contribution is furnished in Table 3.

Step 3.

When analyzing correlation matrices in Table 2, the sum of the eigen values is equal to the number of variables from which the factors have been computed and the "average expected" eigenvalue is equal to 1.0. The factors have been selected with eigen values greater than 0.5. In this example, only the first three eigen values are greater than 0.5, accounting for approximately 92% of total variation.

The eigen values in Table 4 have been arranged in decreasing order, indicating the importance of the respective factors in explaining the variation of the data. The factor corresponding to the largest eigen value (4.7424) accounts for approximately 52.7% of the total variance. The second factor corresponding to the second eigenvalue (2.5763) accounts for approximately 28.7% of the total variance, and so on.

Step 4.

The eigen vector corresponding to Table 1 has been furnished in Table 5. The number of components has been displayed corresponding to the larger eigen value, i.e components 1, 2 and 3.

Step 5.

As the value of three eigen value has been calculated greater than 0.50, three components from Table 5 have been taken and furnished in Table 6.

Step 6.

To find out the significant variable from Table 6, the following method has been applied. In principal component analysis, one component is linear combination of all variables. To find the particular variable on which the component is mostly dependent on, the following method has been used.

Table 3 The eigen values computed from Table 2

Component	Eigen values	% total variance	Cumulative eigen value	Cumulative %
1	4.7426	52.7	4.7426	52.6958
2	2.5762	28.6	7.3188	81.3197
3	1.0022	11.1	8.3210	92.4556
4	0.4298	4.8	8.7508	97.2307
5	0.2363	2.6	8.9871	99.8566
6	0.0106	0.1	8.9977	99.9744
7	0.0023	0.0	9.	99.9997
8	0.0000	0.0	9.	100.000
9	0.0000	0.0	9.0000	100.000

Table 4 Number of components, eigen values, total and cumulative variance as obtained from PCA

Component	Eigen values	% total variance	Cumulative eigen value	Cumulative %
1	4.7424	52.7	4.7424	52.6937
2	2.5763	28.6	7.3188	81.3195
3	1.0022	11.1	8.321	92.4555

Table 5 Eigen vector

Variable	Eigenvector and number of components are 9								
	Comp1.	Comp2.	Comp3.	Comp4.	Comp5.	Comp6.	Comp7.	Comp8.	Comp9.
Maximum temp.	0.3768	0.3511	-0.0493	0.0048	-0.0609	0.0452	0.1098	0.8422	0.0723
Minimum temp.	0.3673	0.3689	-0.0571	0.0121	-0.0509	-0.0998	-0.0321	-0.3804	0.7532
Rain fall	0.2784	-0.3358	0.1863	-0.8272	0.2507	-0.0929	0.101	0.0340	0.0825
Max humidity	-0.2552	-0.5041	0.0206	0.2470	-0.0945	-0.4320	0.4006	0.2497	0.4483
Minimum humidity	-0.4473	-0.1143	0.0472	-0.1075	0.1377	0.5402	-0.4272	0.2479	0.4673
D1	-0.3627	0.2290	-0.1344	-0.4574	-0.7591	-0.0997	0.0534	0.005	0.0022
D2	-0.3549	0.3732	0.0238	-0.1090	0.4054	-0.6512	-0.3463	0.1165	-0.021
D3	-0.3499	0.3879	0.0168	-0.1046	0.3600	0.2586	0.7149	-0.0871	0.0287
Sun shine	-0.0211	0.1364	0.9685	0.1049	-0.1812	-0.0052	0.0001	-0.0047	0.0003

Table 6 Eigenvector computed from Table1

Variable	Eigenvector and number of components : 3			
	Variable number	Component 1	Component 2	Component 3
Maximum temp.	1	0.3768	0.3511	-0.0493
Minimum temp.	2	0.3673	0.3689	-0.0571
Rain fall	3	0.2784	-0.3358	0.1863
Max humidity	4	-0.2552	-0.5041	0.0206
Minimum humidity	5	-0.4473	-0.1143	0.0472
D1	6	-0.3627	0.229	-0.1344
D2	7	-0.3549	0.3732	0.0238
D3	8	-0.3499	0.38798	0.0168
Sun shine	9	-0.0211	0.1364	0.9685

The first component corresponding to the first eigen value 4.7424 is most correlated with minimum humidity (high negative correlation). So, component 1 is dependent on minimum humidity. The other dependency can be found for those variable which are under 10% of minimum humidity (highest value in component 1) or -0.4026. From component 1 (Table 6), it has been observed that the value of other variables less than 0.4026 (negative correlation).

So, no other variable plays dominant role in component 1. If, more than one variable have been predicted as significant variables, one correlation matrix will be computed and depending on correlation, the significant variable will be calculated.

Thus, in component 2, corresponding eigen value is 2.5763 and it is dominated by the variable, maximum humidity, and after reduction of 10% of this, it becomes 0.4537 indicating no correlation with other variables. Finally, from component 3 sunshine is found to be the most dominant variable.

Step 7.

It has been observed that component 1, component 2 and component 3 have three dependent variables, such as minimum humidity, maximum humidity and sun shine. So, without considering 9 variables, these three give 92% solution of this problem.

5.2 Factor Analysis

Factor analysis has been applied on the data taken from Table 1. The detailed steps are as follows:

Step 1.

The same data furnished in Table 1 has been used in factor analysis and using Statistica 7 software package, factor loadings have been calculated in factor analysis and are furnished in Table 7 and Table 8. The eigen values have been taken which are greater than 0.5. The factors have been taken the same as number of eigen values.

Table 7 Eigen values

Value	Eigen values	% total variance	Cumu. eigen value	Cumu. %
1	4.7426	52.7	4.7426	52.7
2	2.5762	28.6	7.3188	81.3
3	1.0022	11.1	8.321	92.5

Step 2.

In factor analysis, one variable is linear combination of all factors. The factor value which is the greatest of all factors has been marked in the row of all variables. In each factor, it has been found the greatest value from all marks values; the corresponding variable has been taken. Using this method, the variable minimum humidity has been selected in factor 1 from Table 8. From the other two factors, maximum humidity and sun shine have the lowest (-0.8054) and the highest (0.9696) factor loadings, and are selected.

So, three variables maximum humidity, minimum humidity and sun shine have been considered as the significant variables.

6. Results

Applying principal component and factor analysis, it has been observed that out of nine environmental parameters, three of them (maximum humidity, minimum humidity and sun shine) have played significant roles for growing the mustard plant. If these three parameters are available sufficiently, the growth of mustard plant will be healthy and they will produce huge yield.

Table 8 Factor Loadings

Variable	Factor Loadings (Unrotated) (PCAapplication) Extraction : Principal components (Marked loadings are > 0.7000)		
	Factor 1	Factor 2	Factor 3
Max. temp.	**-0.8278**	0.5581	-0.0480
Min. temp.	**-0.8073**	0.5869	-0.0559
Rain Fall	-0.5994	-0.5442	0.1884
Max Humidity	0.5660	**-0.8054**	0.0177
Min Humidity	**0.9765**	-0.177	0.0473
D1	**0.7852**	0.3737	-0.1324
D2	**0.7652**	0.6042	0.0258
D3	**0.754**	0.6276	0.0189
Sun Shine	**0.0432**	0.2132	**0.9696**

7. Conclusion and Future Scope of Work

In this paper, an algorithm has been proposed to reduce the parameters for any physical problem. The algorithm has been applied on environment parameters which have been used for healthy growth of mustard plant. The algorithm has been applied on the environment data and it has been observed that out of nine environmental parameters, three of them (maximum humidity, minimum humidity and sun shine) have to be applied sufficiently for healthy growth of mustard plant. The ratio of three parameters will be applied for getting maximum yields. The growth of mustard plant would be optimized using the artificial neural network (ANN) in future. The algorithm may be applied in different problems which are dependent on many parameters.

Acknowledgement : The author would like to thank the All India Council for Technical Education (File No.: 1-51/RID/ CA/28/2009-10) for funding this research work.

References

[1] Statsoft Electronics Statistics Textbook, How to Reduce Number of Variables and Detect Relationships, Principal Components and Factor Analysis, http://www. Statsoft.com/textbook/ stfacan.html, Date of access 21.10.2012

[2] Suh, C., Rajagopalan, A., Li, X. and Rajan, K., The Application of Principal Component Analysis to Materials Science Data, Data Science Journal, Vol.1, pp.19-26, 2002.

[3] Bernstein, I.H., Some Basic Statistical Concepts. Applied Multivariate Analysis, Chapter 2, pp.2-46.

[4] Bernstein, I.H., Chapter 6: Explorotory Factor Analysis. Applied Multivariate Analysis, pp.157-182.

[5] Shashua, A., Introduction to Machine Learning, in Algebraic Representation of PCA, pp.1-8, 2003.

[6] Anderson, T.W., Principal Components. An Introduction to Multivariate Statistical Analysis, Chapter 11, pp. 451-460.

[7] Rammel, R.J., Understanding Factor Analysis, Journal of Conflict Resolution, Vol.XI, pp.444-480, 1967.