Tarım Makinaları Bilimi Dergisi (Journal of Agricultural Machinery Science)
2011, 7 (3), 305-308

Principal Component Analysis Applied to Agricultural Equipments

Constantin TÂRCOLEA, Adrian Stere PARIS, Paula VOICU
University Politehnica of Bucharest, Faculty of Biotechnical Systems Engineering
Spl. Independentei, 313, 060042, Bucharest, ROMANIA
constantin_tarcolea@yahoo.com

Received (Geliş Tarihi): 08.05.2011 Accepted (Kabul Tarihi): 10.07.2011

Abstract: The present paper aims at the reduction of attributes for a family of agricultural equipments, bran finisher type. Seven attributes are taken into consideration for each equipment. Software XLSTAT processes the experimental data, computes the eigenvalues, draughts a scree plot and a biplot. The first two new properties cover approx. 80% from the total variance in the studied case, while the first three, new attributes describe 100% of the variance, offering even a visual representation, which simplifies significantly the choice of the adequate bran finisher equipment.
Key words: Principal component analysis (PCA), bran finisher, eigenvalues and eigenvectors

INTRODUCTION

The multitude of agricultural and food industry equipments and their characteristics, as well as their different performances, make it difficult to choose the industrial appliances for a given application. PCA was originally proposed by K. Pearson and independently developed by Hotelling (Hotelling, 1936). The goals of PCA are to extract the most important information from the available data, compress the size of the data set by keeping only this important information, simplify the description of the data set, and analyze the structure of the observations and the variables. The choice of tools in the agricultural area is selected from a wide spectrum (Voicu et al., 2008). Usually the choice is not based on all the attributes, but on a combination of properties. In the first stage, for each device, it is taken into consideration, as many attributes as possible. The PCA is the standard technique to reduce multivariate data sets in a subspace of small dimension, regularly three or two dimensions. The purpose of preprocessing is to try to transform the data into the most suitable form for the main purpose of this research.

MATERIALS and METHOD

At the beginning the data are given, as a nxp-matrix, objects/attributes, a table $\mathbf{y}= (y_{ij})$, i=1,2,..,n j=1,2,…,m. Each row of the matrix represents an object (individual) with its attributes, and each column is an attribute (property, variable). The number of attributes gives the dimension of the initial representation space of the objects. Anyway it is considered an m-dimensional coordinate system, each coordinate being an attribute. Instead of actual attributes the PCA uses new factors, but only a few, which are artificial ones.

The problem is mathematically formulated as follows: it is supposed that $\mathbf{y}^t=(y_1, y_2,..., y_m)$ is a random vector with the center of dispersion **m** and the covariance matrix **Σ**. The PCA procedures try to identify new uncorrelated variables $z_1, z_2,..., z_m$, whose variance decreases, when the index increases from 1 to m (Croux et al., 2005). The first PC explains the maximum variance in the data; the second PC explains the maximum variance that has not been accounted by the first PC, and so on. The PCA solves the problem of finding the directions of the greatest variance of the linear combination of the old coordinates. This means that a set of the coefficient vectors $a_1, a_2,..., a_k$ should be found, each new variable being a linear combination of the initial variables. The first principal component:

$$z_1 = a_{11} y_1 + ... + a_{m1} y_m \qquad (1)$$

is chosen, so that

$$Var(z_1) = Var(a_1^{tr}\ y) = a_1^{tr} \sum \cdot a_1 \qquad (2)$$

is maximum, under the restriction:

$$a_1^{tr}\ a_1 - 1 = 0 \qquad (3)$$

A so-called Lagrange function is used to find the conditional extreme of a function, given the relationship:

$$L(a_1;\lambda) = a_1^{tr} \sum \cdot a_1 - \lambda\left(a_1^{tr}\ a_1 - 1\right) \qquad (4)$$

where λ is an undetermined multiplier. The necessary conditions for the extreme are:

$$\begin{cases} 2\sum \cdot a_1 - 2\lambda a_1 = 0 \\ a_1^{tr}\ a_1 - 1 = 0 \end{cases} \qquad (5)$$

The directions of the new coordinate axes, called principal components, or factors, have been chosen, in such a way, that the deformations of the original cloud implied by this representation are minimal (Târcolea et al., 2009). The coordinates of the objects (samples) in the new system are called scores. The corresponding relationships between the original variables and the new principal components are called loadings.

RESULTS and DISCUSSION

The PCA is a standard technique to reduce multivariate data sets in a subspace of small dimension, in this case a tri-, respectively bivariate. The number of noticeable attributes gives the dimension of the initial representation space of the objects.

Instead of the former attributes, the PCA uses new factors, but artificial ones. The dimensional reduction of attributes for a family of equipments is the concern of the present researches.

As a relevant example let's take an application from the agricultural machinery (Ranken et al., 1997), concerning the selection of usual dehusking equipment in the Romanian market (Voicu and Casandroiu, 1995): seven important mechanical and technological properties are presented for a detailed analysis. Let's consider 4 equipments (objects), each of them having 7 attributes (Table 1).

Table 1. Bran finishers characteristics

Characteristics		FTO	FT 30/60	FT 40/80	BRAN BRUSH
Var1	Mean yield capacity, kg/h	687	275	550	650
Var2	Necesary area for the equipment, m^2	0.93	1.064	1.73	1.322
Var3	Installed power, kW	4	2.2	5.5	4.4
Var4	Equipment mass, kg	285	320	650	530
Var5	Dependability coefficient	0.92	0.85	0.88	0.83
Var6	Air flow for aspiration, m^3/min	5	3.5	4.5	5.5
Var7	Specific loading	85	24.5	27.5	175

The ANOVA method was applied as a preliminary test, to verify if the attributes are statistically identical; the null hypothesis is rejected, based on the result of the F test and p-value.

XLSTAT is a Microsoft Excel add-in, the main product of the company Addinsoft (www.xlstat.com). The results calculated by applying this software are presented bellow (Tables 2,3,4 and Figures 1,2,3). The three largest eigenvalues are 3.463, 2.131 and 1.406 (Table 2). This suggests that the corresponding PC's (F1, F2, F3) are enough.

Table 2. The eigenvalues of the model

Characteristics/ Factors	F1	F2	F3
Eigenvalue	3.463	2.131	1.406
Variability (%)	49.471	30.446	20.083
Cumulative (%)	49.471	79.917	100.000

The representation of the data in a limited number of dimensions (3 dimensions in this case) facilitates to a great extent this analysis. The quantitative relationships between the old variables and the new ones (principal components) are represented in the Table 3.

Table 3. The eigenvectors of the model

Characteristics / Factors	F1	F2	F3
Var1	0.439	-0.363	0.190
Var2	0.321	0.548	0.044
Var3	0.490	0.159	0.287
Var4	0.391	0.466	-0.072
Var5	-0.011	-0.243	0.788
Var6	0.460	-0.344	-0.101
Var7	0.315	-0.384	-0.493

Figure 1 has two parts: the rectangles show the the fraction of the total variance of the primary data for each principal component, while the broken line describes the cumulative variance explained by the first three components.

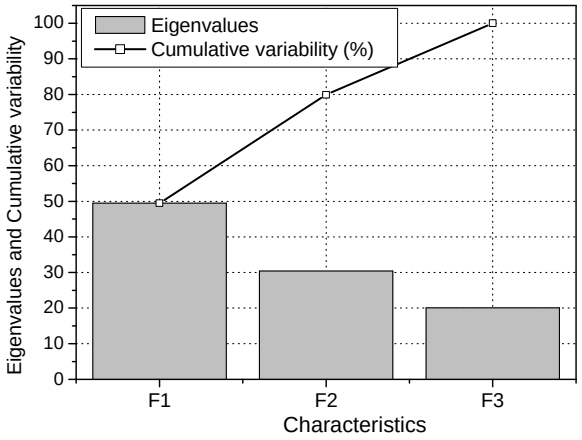

Figure 1. Pareto scree plot

The factor loadings, also called component loadings in PCA, are the correlation coefficients between the variables (rows) and factors (columns). Similarly to Pearson's r coefficient the squared factor loading is the percent of variance in that indicator variable explained by the factor. To get the percent of variance in all the variables accounted for by each factor, add the sum of the squared factor loadings for that factor (column) and divide by the number of variables (Table 4).

Table 4. Factor loadings

Characteristics/ Factors	F1	F2	F3
Var1	0.818	-0.530	0.225
Var2	0.598	0.800	0.053
Var3	0.911	0.232	0.340
Var4	0.728	0.681	-0.085
Var5	-0.020	-0.355	0.935
Var6	0.856	-0.503	-0.120
Var7	0.586	-0.561	-0.585

The correlation circle (herein below having the axes F1 and F3) shows a projection of the initial variables in the factors space. In Figure 2 the variables are far from the center and variable 1,2, 3, 4, 6 are close to each other; they are significantly positively correlated. This can be confirmed either by looking at the correlation matrix or by looking at the correlation circle on axes F1 and F3. The correlation circle is useful in interpreting the meaning of the axes. The Figure 2 shows a projection of the initial variables in the factors space.

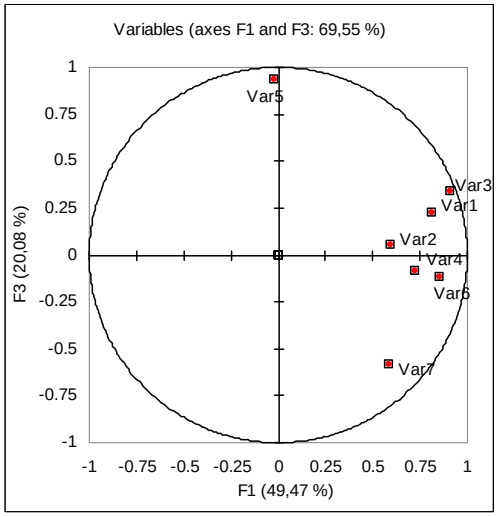

Figure 2. Correlation circle

In Table 5 are given the contribution of the variables (%) (Paris and Târcolea, 2009).

Table 5. Contribution of the variables, (%)

Characteristics/ factors	F1	F2	F3
Var1	19.312	13.160	3.610
Var2	10.331	30.005	0.197
Var3	23.978	2.535	8.226
Var4	15.292	21.734	0.515
Var5	0.011	5.926	62.121
Var6	21.159	11.868	1.021
Var7	9.917	14.772	24.310

To confirm that a variable is well linked with an axis, take a look at the squared cosines table: the greater the squared cosine, the greater is the link with the corresponding axis. The closer the squared cosine of a given variable is to zero, the more careful you have to be when interpreting the results in terms of trends on the corresponding axis. Looking at Table 6 it may be noticed that the pc F1 covers five of the variables, while F2 an F3 explain the last two. The table 7 gives the factors scores, Table 8 - contribution of the observations % and 9 – the squared cosines of the observations.

Table 6. Squared cosines of the variables

Characteristics/ factors	F1	F2	F3
Var1	**0.669**	0.280	0.051
Var2	0.358	**0.639**	0.003
Var3	**0.830**	0.054	0.116
Var4	**0.530**	0.463	0.007
Var5	0.000	0.126	**0.873**
Var6	**0.733**	0.253	0.014
Var7	**0.343**	0.315	0.342

Values in bold correspond for each variable to the factor for which the squared cosine is the largest.

Table 7. Factor scores

Observations	F1	F2	F3
Obs. 1	1.316	1.956	0.996
Obs. 2	1.791	-0.561	-1.646
Obs. 3	1.316	1.956	0.996
Obs. 4	1.791	-0.561	-1.646

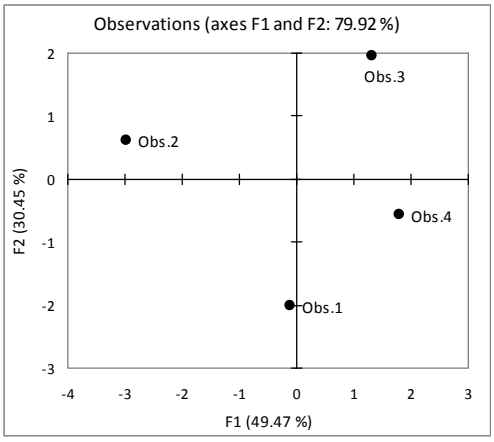

Figure 3. Observations plot (F1 and F2)

Table 8. Contribution of the observations (%)

Observations	F1	F2	F3
Obs. 1	0.110	47.098	27.792
Obs. 2	64.238	4.351	6.410
Obs. 3	12.501	44.861	17.638
Obs. 4	23.150	3.690	48.160

Table 9. Squared cosines of the observations

Observations	F1	F2	F3
Obs. 1	0.003	**0.718**	0.279
Obs. 2	**0.924**	0.039	0.037
Obs. 3	0.264	**0.584**	0.151
Obs. 4	**0.515**	0.051	0.435

Once the results have been obtained, they may be transformed in order to make them easier to interpret, for example by trying to arrange that the coordinates of the variables against the factors are either high (in absolute value), or close to zero.

CONCLUSIONS

Usually in the first stage, for each product, there are taken into consideration as many properties as possible. In the second stage, based on PCA, there were chosen two or three, given by the principal components. An artificial subspace with three (two) dimensions (Târcolea and Paris, 2008) with XLSTAT 2011 software is developed in the present research. The initial attributes for each tool should be expressed with a precision of 80% as function of two artificial axes. The application of this model simplifies the design of the bran finisher equipments. The presented method enables many other possible extensions in the design process.

REFERENCES

Croux C., A. Ruiz-Gazen, 2005. High Breakdowns Estimators for Principal Components The Projection-Pursuit Approach Revisited 2000, Journal of Multivariate Analysis, 95 (1): 206-226.

Hotteling, H., 1936. Relation between two sets of variates. *Biometrica*, 28 (3-4): 321-377.

Paris A. S., C. Târcolea, 2009. Computer aided selection in design processes with multivariate statistics, Proceedings of the International Conference on Manufacturing Systems – ICMaS, 4: 335-338.

Ranken, M. D., R. C. Kill, C. Baker, 1997. Cereals and cereal products. Chap. 5, pp. 175-210. In: *Food industries manual* Ed. 24, D.J. Wallington (ed.), Ed. Blakie Academic & Professional London.

Târcolea C., A. S. Paris, 2008.The Joreskog technique applied for materials design, Proceedings of the 17th International Conference on Manufacturing Systems – ICMaS, 17: 309-312.

Târcolea C., A. S. Paris, A. Demetrescu –Târcolea 2009. Statistical methods applied for materials selection The International Conference DGDS-2008 & MENP-5 Applied Sciences (APPS*)* 11: 145-150.

Voicu Gh., T. Casandroiu 1995. Procese şi utilaje pentru morărit. Chap.1. In :*Utilaje pentru morărit şi panificatie*, U.P.Bucharest

Voicu Gh., T. Căsăndroiu, C. Târcolea, 2008. Testing stochastic models for simulating the seeds separation process on the sieves of a cleaning system, and a comparison with experimental data, „ACS-Agriculturae Conspectus Scientificus", Zagreb 73(2): 95-101.

www.xlstat.com.