ISSN: 2535-1133 (Online)
ISSN: 1312-3882 (Print)

ORIGINAL PAPER

Fundamentals of principle component analysis

Emil Penchev[1] • Sonya Doneva[1]

[1]Dobroudha Agricultural Institute, General Toshevo, 9521, General Toshevo, Bulgaria

Coresponding Author: Emil Penchev; E-mail: emo_ap@mail.bg

Received: June 2021 / Accepted: December 2021 /

Published: December 2021 © Author(s)

Abstract

Penchev, E. & Doneva, S. (2021). Fundamentals of principle component analysis. Field Crops Studies, XIV(2-3-4), 65-72.

In the study are discussed the basic principles of the principal component analysis which have a fundamental role in the researches by factorial experiment. On the base of expanding the variance, is evaluated the significant of the factors in the studied relationships. PCA is applied when the database contains information from only a few variables, but it becomes especially effective when a large number of statistical quantities such as spectral data need to be analyzed. The method makes it possible to discover new variables, called „principal components", which assess the majority of the variables in the database. PCA is applied by the study of the most important characteristics of the quality by winter soft wheat. The main components of the quality of winter soft wheat evaluated by this method are - dough stability, dough development time and sedimentation.

Key words: Principle component analysis (PCA), Common winter wheat, Quality indicators

Introduction

Statistical analysis, as a branch of applied mathematics, is very widely used in various fields of science and public life. The phenomena that are the subject of study are interrelated, therefore their study requires the application of multiple statistical methods (Miranda et al., 2008). Principal component analysis (PCA) is a method based on the multiple covariance of the studied statistical variables and allows to assess their role in the studied relationships. The method has a major role in the study of the factorial experiment as well as the variance experiment, correlation and regression analyzes.

The main task in the selection of winter soft wheat is the creation of varieties combining high productivity and quality. The purpose of the present study is to clarify the algorithm of the method of principal component analysis and its application for evaluation of the main components of quality in winter soft wheat.

Material and Methods

The following quality indicators sedimentation (ml), wet gluten content (%) in 70% flour, bread volume (cm^3), number of farinograph, dough resistance (min), degree of softening (fu), hectoliter, H were evaluated: D, quality of the medium (0-5 points) and vitreous (%) of the varieties Aglika, Enola, Lazarka, Karina, Korona, Kosara, Nicodemus, Dragana, Chiara, Bozhana, Katarzyna, Sladuna, Kalina, Kristina, Pchelina, Marilyn, Goritsa, Iveta, Fani, Jana and Kami for the period 2018 - 2020. The purpose of the analysis is to identify the main components forming the quality.

Principle component analysis (Warmuth et al., 2008) and ANOVA 2 (Anderson and Jeff, 1996) were applied. The statistical package with which the experimental data were analyzed is SPSS 21.0.

Results and Discussion

Nature of the principle component analysis

Principle component analysis (PCA) is a mathematical method for reorganizing information into a database network. It can be applied when the database contains information from only a few variables, but it becomes especially effective when a large number of statistical quantities such as spectral data need to be analyzed. The PCA makes it possible to discover new variables, called “principal components”, which assess the majority of the variables in the database (Kronenberg, 1995). Thus, it is possible to analyze the database with a significantly smaller number of variables than the original ones. For example, many studied objects are characterized by more than 20 random variables, but when applying PCA it turns out that the variables with the first 4-5 in the rank principal components contain the most important information and therefore the study should focus on them (Forkman et al., 2019).

PCA is an analytical procedure for transforming a set of random variables into another set of variables having the following properties:

1. They are a linear combination of the original variables.
2. They are orthogonal, independent of each other.
3. The total variation among them is equal to the total variation of the original variables, therefore the information contained in the observed random variables is not lost during the transformation.

4. The variance associated with each component decreases in the following order - the first component calculates the largest possible proportion of the total variation and the second the largest proportion of the remainder.

The most significant differences between PCA and factor analysis are the following:

1. In the factor analysis p the number of original variables is reduced to m <p non-correlating "factors" having non-correlating residual components; in PCA p correlated variables are transformed into p non-correlated variables, not all of which are statistically significant.

2. Unlike factor analysis, PCA has the potential to rotate the orthogonal axes, which represent the factors in a new inclined position, so that the theoretical postulates contained in the model can be tested.

3. The PCA evaluates the observed variation by highlighting the variables that cause it, examining all original variables. In the factor analysis, however, the share of the correlated variables in the total variation is examined.

Although the application of PCA is not complicated as calculations and is implemented in many statistical packages such as SAS (Litell et al., 1996), SPSS, Statistica and others. The user must know the essence of the algorithm of the method.

Algorithm of principle component analysis

Suppose that $x_1, x_2, \ldots, x_p$ are random variables, on each of which we have made n observations. They form a matrix **X** with dimension **n x p** independent rows and columns. Let us denote the matrix of covariances of **X** by **S**, and the correlation matrix by **R**.

Then the principal components are defined as a linear combination of the original variables $\mathbf{x_i}$ and we can denote them by $\boldsymbol{\lambda_i}$. Then $\boldsymbol{\lambda_i}$ has the form (Warmuth et al., 2008) :

$$\boldsymbol{\lambda_i = a_{i1}\, x_1 + a_{i2}\, x_2 + a_{i3}\, x_3 + \ldots + a_{ip}\, x_{p,}} \qquad (1)$$

where the index i is taken to be values from 1 to n. Thus, the vector **Λ** is defined as a linear combination of the columns of the matrix X and is called an eigenvector. Each eigenvector is related to the variance by means of the eigenvalues of the matrix **{X - λE}**.

The geometric meaning of the method is as follows: data from n variables in the p - dimensional space are represented. PCA is the rotation of the coordinate axes in such a way that the variance of the vertical projections of the points on the first axis is maximal and this is the first component (Gabriel, 1971). The second axis (the second principal component) is chosen orthogonally to the first axis and is calculated as the possible residual variance (Yan and Kang, 2003). The linear combinations $\boldsymbol{\lambda}_i$ are the lengths of the projections along the new axes, and the cosines of the angles of the projections with the axes are the coefficients of the eigenvector $\boldsymbol{\Lambda}$. The variances of the projections are the eigenvalues $\boldsymbol{\lambda}_i$. The following properties of PCA are of particular importance in the interpretation of the data - the variance of each principal component is its eigenvalue $\boldsymbol{\lambda}_i$ and their sum is equal to the total variance of the experimental data. The equality is valid:

$$\Sigma\boldsymbol{\lambda}_i = \Sigma\boldsymbol{\sigma}_i^2 \qquad (2)$$

PCA study of the main components forming the quality of winter soft wheat

The data from the research show that the new varieties Pchelina, Goritsa and Lazarka in most cases exceed the physical properties of the dough standard Aglika, but their fluctuations over the years are higher (Table 1).

Table.1 Analysis of variance (Mean of squares) of the studied indicators

Indicators	MSG	MSC	MSGxC	MSerror
Sedimentation (ml)	1147.3 ***	422.1 *	376.1 *	59.4
Wet cluten content (%)	35.6 *	254.6 ***	31.2 *	9.8
Hectoliter weight (kg)	18.8 **	14.1 *	5.7	2.9
Farinograph number (%)	20.5 *	60.8 ***	7.7	4.2
Vitreous (%)	816.3 *	1933.6 **	941.4 *	204.1
Dough stability (min)	26.2 ***	4.4	6.2	2.2
Degree of softening (fu)	4135.2 *	8534.6 **	1644.6	625.3
Bread volume (cm ^{3})	22136.2**	29359.1 **	16424.2 *	2244.3
H:D	2.1 *	1.6*	1.8 *	0.4
Quality of the environment (scores 0-5)	0.63 *	0.92**	0.27	0.18
Dough development (min)	25.4 **	6.6	7.3	2.8
df	20	2	40	128

* – significant at P=0,05, ** – significant at P=0,01, *** – significant at P=0,001

Kalina and Kiara are characterized by low values of the physical characteristics of the grain. On average for the three years, Pchelina and Lazarka are reliably identified as strong wheat by sedimentation. Significant fluctuations in sedimentation over the years have been registered in Bozhana (32ml), Kiara (26ml), Kosara (22ml) and Aglika (20ml). On average for the period, the varieties Pchelina, Goritsa and Kosara compete in the yield of wet gluten in 70% flour with the other varieties, including the Aglika standard. Katarzyna, Kalina and Kiara, with relatively not very high grain quality, are inferior to the standard estimates for the content of wet gluten in 70% flour. In most of the varieties, the physical properties of the dough determined on the farinograph and alveograph vary greatly over the years. The rheological characteristics of the dough from the farinograph at Pchelina, Goritsa and Aglika are especially variable. The fluctuations in the alveographic assessments of Goritsa and Pchelina are weaker. In this respect, they have an advantage over Aglika, and the other varieties (Kalina, Kosara and Katardzhina) are inferior to them. The analysis of variance sheds light on the differences in the phenotypic expression of the quality indicators depending on the variety, year and the variety x year interaction.

Table 2. Values of the main components

Rotated Component Matrix[a]		
Indicators	Components	
	1	2
Sedimentation, ml	0,420	0,850
Wet gluten content in 70% dough, %	-0,398	-0,855
Dough development, min	0,921	0,195
Dough stability, min	0,870	0,481
Farinograph number,	-0,340	0,659
Degree of softening	-0,993	0,065

In order to refine the relations between the quality indicators in the metameric variability of the bread volume, a principal component analysis was applied, and for this purpose the experimental data were unified with appropriate algebraic transformations (Liao et al., 2010). The results show that the variables entering the first component have a significant maximum positive contribution to the volume of the bread: time for dough development and stability of the dough from the farinograph. They are subjected to the highest metameric variability followed by sedimentation (0.420). It can be assumed that this component is a reflection of the general variability in the volume of bread. Sedimentation, dough stability, quality number from the farinograph and degree of softening forming the second main component have a positive indirect effect on the volume of bread. High negative variability in the volume of bread causes the content of wet gluten in 70% flour, whose contribution is dominant in the second main component (-0.855). The obtained data give grounds for differentiating the varieties into 2 main groups: 1 - Pchelina, Goritsa, Kiara and Aglika; 2 - Kalina, Kosara and Katarzhina.

The data from Table 3 show that with the highest eigenvalues and share in the total covariance of the two-dimensional matrix of quality indicators x volume of bread, the sedimentation (61.7%), the content of wet gluten in 70% flour (25.0%)

and the time stand out. for the development of the dough by the farinograph (13.2%). In total, the three characteristics determine 99.9% of the total variance, which is a reflection of the extreme complexity of the relationships between them and the volume of bread.

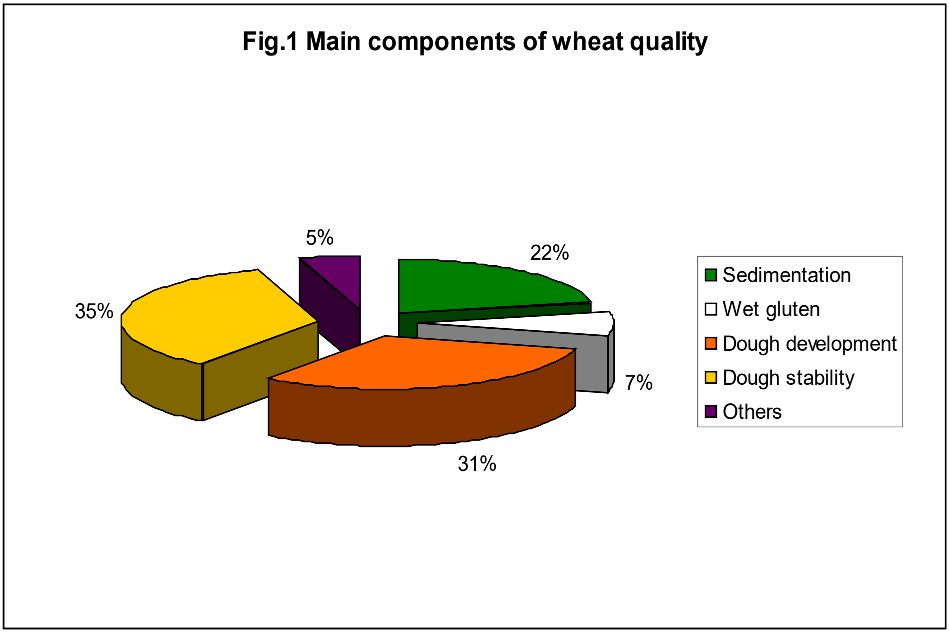

Figure 1. The main components forming the quality of winter soft wheat.

Table 3. Eigenvalues and share of the variance of the studied indicators

	Initial Eigenvalues		
Indicators	Total	% of Variance	Cumulative %
Sedimentation, ml	3,702	61,698	61,698
Wet gluten content in 70% dough, %	1,502	25,031	86,729
Dough development, min	0,794	13,228	99,956
Dough stability, min	0,002	0,030	99,987
Farinograph number	0,000	0,008	99,995
Degree of softening	0,000	0,005	100,000

Conclusions

1. The principal component analysis is a multiple statistical method by which the main indicators forming the studied statistical object can be estimated.

2. Compared to the factor analysis, PCA allows for a detailed study of the components of the studied objects.

3. The main components of the quality of winter soft wheat wer evaluated by the PCA method - dough stability, dough development time and sedimentation.

References

Anderson, C. & Jeff, C. (1996). Dispersion Measures and Analysis for Factorial Directional Data with Replications. *Applied Statistics*, 45(1), 47-61.

Gabriel, R. (1971). The biplot graphic display of matrices with application of principal component analysis. *Biometrika*, 58, 453-467.

Forkman, J., Josse, J., Piepho, H. P. (2019). Hypothesis tests for principal component analysis when variables are standardized. *Journal of Agricultural, Biological and Environmental Statistics*, 24(2): 289–308.

Miranda, A., Le Borgne, Y. & Bontempi, G. (2008). New Routes from Minimal Approximation Error to Principal Components, Volume 27, N.3, Neural Processing Letters, Springer

Kronenberg, M. (1995). Introduction to biplots for G x E tables. Department of Mathematics, Research Report N.51, University of Queensland.

Liao, T., Jombart, S., Devillard, F. & Balloux (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, 11: 86-94.

Litell, C., Miliken, G., Stroup, W. & Wolfinger, R. (1996). SAS systems for mixed models. SAS Institute

Warmuth, M. K. & Kuzmin, D. (2008). Randomized online PCA algorithms with regret bounds that are logarithmic in the dimension. *Journal of Machine Learning Research*, 9: 2287–2320.

Yan, W. & Kang, M. (2003). CGE biplot analysis. A graphical tool for breeders, geneticits and agronomist. CRC Press. Boca Ration. Fl.