Bulletin of the *Transilvania* University of Braşov
Series V: Economic Sciences • Vol. 7 (56) No. 2 - 2014

PRINCIPAL COMPONENT ANALYSIS - A POWERFUL TOOL IN COMPUTING MARKETING INFORMATION

Cristinel CONSTANTIN[1]

Abstract: *This paper is about an instrumental research regarding a powerful multivariate data analysis method which can be used by the researchers in order to obtain valuable information for decision makers that need to solve the marketing problem a company face with. The literature stresses the need to avoid the multicollinearity phenomenon in multivariate analysis and the features of Principal Component Analysis (PCA) in reducing the number of variables that could be correlated with each other to a small number of principal components that are uncorrelated. In this respect, the paper presents step-by-step the process of applying the PCA in marketing research when we use a large number of variables that naturally are collinear.*

Key words: *multivariate analysis, multicollinearity, principal component analysis, marketing research.*

1. Introduction

In the most cases of marketing research the descriptive analysis and the univariate or bivariate inferential analyses are not enough for obtaining that information needed by the decision factors that face with a marketing problem and order such a research. The multivariate analyses extract the main information from a large number of variables and offer additional details that can support the decision process. The computation of such methods is quite complicated but the modern information systems can assist the researchers to obtain the best information. Nevertheless the correct using of the multivariate methods and the results interpretation are very important. In this respect, the present research aims to assist mainly the young researchers in using the Principal Component Analysis (PCA) as one of the most popular multivariate data analysis methods. The theoreticians and practitioners can also benefit from a detailed description of the PCA applying on a certain set of data.

2. Literature review

Principal component analysis (PCA) is a method of data processing consisting in the extraction of a small number of synthetic variables, called principal components, from a large number of variables measured in order to explain a certain phenomenon.

Principal components are a sequence of projections of the data, mutually uncorrelated and ordered in variance,

[1] Faculty of Economic Sciences and Business Administration, *Transilvania* University of Braşov.

which are obtained as linear manifolds approximating a set of N points [1].

Using the Regression model with many variables that are highly correlated each other will not return the best estimators [4]. In such cases when we try to analyse a large set of *p* variables that are usually much correlated and generate the multicollinearity phenomenon, the PCA is recommended. PCA is also known in literature as Factor analysis even if the critics consider the two methods as being different from each other.

Talking about Factor analysis, there are two major classes of research purposes: Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). EFA is heuristic and the investigator has no expectations of the number or nature of the variables. It allows the researcher to explore the main dimensions to generate a theory, or model from a relatively large set of latent constructs. In contrast with EFA, in CFA the researcher uses this approach to test a proposed theory or model [5].

We can see that both Principal Components Analysis and Factor Analysis deal with more variables that usually are correlated in order to reduce the dimension of the analysis to a small number of factors that are not correlated (independent factors). Thus the negative effects of the multicollinearity are avoided. In conclusion the Principal Components Analysis carries information about not only the patterns of variations in individual variables but also the relationships between variables [3].

Principal Components Analysis is considered a useful tool for dimension reduction and compression as the resulted factors are orthogonal and every factor explains a large part of the variation given by the variables that satisfy a certain condition [1]. The principal components that are to be taken into consideration are those factors that can explain the largest part of the information given by the initial variables. In this respect the number of factors which should be retained in the analysis is a decision matter for the researcher [2]. For plotting purposes, two or three principal components are usually sufficient, but for modeling purposes the number of significant components should be properly determined [6]. There are many extraction rules and approaches in the determination of the number of factors that are to be retained. One of the most popular is Kaiser's criteria which state that only those factors with eigenvalue higher than 1 will be retained in the model. Also the Scree test, the cumulative percent of variance extracted and parallel analysis could be used [5].

However the logical judgment of the researcher should be involved in this selection process in order to determine the meaning of every factor retained in the model. For a better interpretation of the results it is recommended to use a rotational method, which maximises high item loadings and minimises low item loadings. The most popular rotation technique is Orthogonal Varimax [5].

3. Research objectives and methodology

The main objective of this paper is to support young researchers in their efforts to use multivariate data analysis methods. In this respect a step-by-step Principal Component Analysis is presented in the following sections of this paper. For exemplification purpose a data base from a survey regarding the social services for students was used.

The sample counts 396 respondents who answered to several questions that used a rating interval scale with five levels equally distanced. As these questions refers to many aspects regarding the accommodation in "Transilvania" University's residence halls, a certain redundancy exists in the measured construct and the variables are imminent correlated each other. Thus the multicollinearity phenomenon is present and the PCA method is recommended. For PCA computation the SPSS system has been used.

4. Applying the PCA

In applying the PCA we have to ensure that the variables used are metric ones (measure with interval or ratio scale). Also the sample size is important even if there is not a general agreement in the literature regarding the number of observations and the ratio between the sample size and the number of variables [5]. Anyway the number of observation have to be bigger than the number of variables included in the analysis with the mention that big samples can lead to more accurate results.

The PCA used in this paper for exemplification purpose takes into consideration 10 items that measure the students' satisfaction regarding various aspects of their accommodation in students' residence hall. For every item a numerical scale with 5 levels has been used (5- very satisfied and 1- very dissatisfied). Starting from the assumption that these variables are collinear, the purpose of PCA is to reduce the number of variables that measure the students' satisfaction to a small number of factors that are not correlated.

For the beginning of the analysis a testing step is necessary in order to determine the suitability of data for such a method. In this respect Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy and Bartlett's Test of Sphericity are computed by SPSS system.

KMO and Bartlett's Test Table 1

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.739
Bartlett's Test of Sphericity	Approx. Chi-Square	783.946
	df	45
	Sig.	.000

The KMO index ranges from 0 to 1and the sample is considered suitable for PCA if this index is equal or higher than 0.50. Also the Bartlett's Test of Sphericity should be significant ($p<0.05$). The results presented in Table 1 reveal that the data used in our example are adequate for PCA.

After these tests we have to take a decision regarding the number of factors (principal components) that should be retained in the model. In the initial solution the number of components is equal to the number of variables included in the model (see Table 2). Every component has an eigenvalue which represents the amount of variance that is accounted for by a given component. Usually the first variables have the greatest eigenvalues.

One of the most commonly used criteria for principal component selection is the Kaiser's criterion known also as eigenvalue-one criterion. According to this one only the variables with the eigenvalue greater than 1 will be retained in model.

Using of eigenvalue-one criterion is not considered the best decision when the actual differences between the eigenvalues of successive variables are quite small. Thus a variable with an eigenvalue of 0.99 will be excluded from the model in spite of its significant contribution to the total variance. For these reason, the proportion of variance accounted for by every factor and the cumulative percentage of variance could be used in the process of factor selection. We can establish to retain in the model all those factors that account for at least 10% or 5% of variance.

Total variance explained Table 2

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.218	32.182	32.182	3.218	32.182	32.182
2	1.456	14.561	46.743	1.456	14.561	46.743
3	.949	9.485	56.228	.949	9.485	56.228
4	.857	8.566	64.795			
5	.817	8.168	72.963			
6	.735	7.352	80.315			
7	.618	6.184	86.499			
8	.581	5.813	92.312			
9	.455	4.554	96.866			
10	.313	3.134	100			

In table 2 we can see that only the first two components have the eigenvalue greater than one but the value of the third component is very closed under one and it explains 9.48% of the total variance. If we look at the cumulative percent of variance explained by the first three factors it counts only 56.23% so that we also can include in the model some of the next variable. Taking into account the cumulative percent of variance explained, according to Hair et al. cited by Williams et al. "in the natural sciences factors should be stopped when at least 95% of the variance is explained. In the humanities, the explained variance is commonly as low as 50-60%". [5].

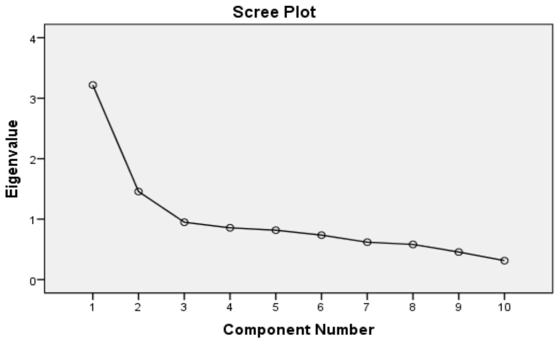

Fig. 1. *The scree plot for the initial variables*

Another method used for factor extraction is the analysis of the Scree plot. This one is a subjective method which requires the researcher judgement. According to this criterion the significant factors are disposed like a cliff, having a big slope while the trivial factors are disposed at the base of the cliff. In the Figure 1 we can appreciate that starting with the fourth factor the slope of the curve is quite small and these factors could be excluded from the model. Nevertheless the method is very subjective because the cut-off point of the curve is not very clear in the above chart.

Whatever method of factor extraction is used it is recommended to analyse the meaning of every principal component according to the variables with significant loadings on the retained factors. In order to apply this meaning interpretation a rotated solution is computed. A rotation is a linear transformation that is performed on the initial factor solution for the purpose of making an easier interpretation. The most common rotation method is Orthogonal Varimax, which is provided by the majority of statistical software.

Rotated component matrix Table 3

	Component		
	1	2	3
Existing reading rooms	**.699**	.102	.111
Silence in residence halls	**.692**	.130	.151
Existing parking	**.662**	.118	-.085
Communication with administration	**.609**	-.223	.357
The guard of residence halls	**.548**	.405	.158
Internet access	.128	**.776**	-.006
Bathroom equipment	.025	**.744**	.337
Room equipment (furniture)	.186	**.423**	.342
Residence halls' cleanliness	.201	.082	**.852**
Bathrooms' cleanliness	.024	.396	**.742**

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

When we apply the rotation method a factor pattern matrix is obtained, which contain the loadings of every variable on the retained factors (see Table 3). In order to make the interpretation of the meaning of every factor the variables that have the greatest loadings on a factor are analysed in terms of their similarity regarding the measured construct. After this interpretation the principal components could be labelled according to their relevant meaning. If the group of variables that determine a factor are meaningless we have to reconsider the number of factors that are included in the model.

In table 3 we can see that the first component is determined by variables related to the safety of students or their cars and to various managerial aspects. The second component is determined by the residence halls' equipment, while the third component refers to cleanliness.

Taking into account the above patterns, we can label the component retained in the model as follows: Component 1 – "Safety and management", Component 2 – "Equipment" and Component 3 – "Cleanliness".

Using the above validation criterion is very important because finally the interpretation of the results should lead to components with a certain meaning for the research purpose.

5. Discussions and conclusions

The Principal Component Analysis can be used when many variables are used to measure the same construct. In such cases the multicollinearity phenomenon appears and using of other analysis methods like regression model is not proper. The most important steps in performing PCA consist in testing the data suitability for this method and in selection of the best factors that describe the total variance produced by the initial variables. In this process the meaning of resulted factors plays a crucial role because the purpose of every marketing research is to support decision makers in their efforts to find solutions for the marketing problem they face with. The resulted factors (principal components) could be used in further analysis regarding to the population description (e.g. the relationship between the factors and population demographics) or as explanatory variables in regression models. The computation of the new variables that represents the principal components could be made as a linear combination of the initial variable. Such variables are directly computed by SPSS system as standardised values. We can also obtain new variables by simply adding the values of every variable that determine a certain component or by computing the mean of these values. For example in order to obtain the new variable "Cleanliness" in the above model we can add the values of the last two variables for every case or we can calculate the mean of these values.

In conclusion the Principal Component Analysis could be considered a powerful tool in computing marketing information.

References

1. Hastie, T., Tibshirani, R., Friedman, J.: *The elements of statistical learning. Data mining, Inference, and prediction.* 2nd edition. Springer, 2009.
2. Lefter, C., Bratucu, G. et al.: *Marketing*, vol. 1. Brasov. "Transilvania" University of Brasov Publishing House, 2006.
3. Qi, X., Luo, R.: *Sparse Principal Component Analysis in Hilbert Space*. In: Scandinavian Journal of Statistics (2014) doi: 10.1111/sjos.12106.
4. Smith, K., Sasaki, M.S.: *Decreasing Multicollinearity: A Method for Models with Multiplicative Functions*. In: Sociological Methods & Research, 8 (1979), no.1, p. 35-56.
5. Williams, B., Brown, T., Onsman, A.: *Exploratory factor analysis: A five-step guide for novices.* In: Australasian Journal of Paramedicine, 8 (2010), no. 3. Retrieved from http://ro.ecu.edu.au/jephc/vol8/iss3/1
6. Wold, S., Esbensen, K., Geladi, P.: *Principal component analysis.* In: Chemometrics and Intelligent Laboratory Systems, 2 (1987), no. 1-3, p. 37-52.