The Principal Components Analysis and Cluster Analysis as Tools for the Estimation of Poverty, an Albanian Case Study

Msc.Evgjeni Xhafaj[1], Phd. Ines Nurja[2]

[1] Department of Mathematics, Faculty of Information Technology, University Alexandër Moisiu (Durrës, Albania),

[2] University of New York Tirana (Tirana, Albania)

Abstract: ***The measurement and analysis of poverty have traditionally relied on reported income or consumption expenditure as the preferred indicators of poverty and living standards. Income is generally the measure of choice in developed countries but a number of methods have been used to assess poverty levels and trends which rely not on consumption or income data but rather on non-monetary dimensions of living conditions. The purpose of this study is to make an estimation of the poverty level by using a multivariate statistical technique called Principal Components Analysis (PCA). The purpose of this technique is the reduction of the variables in a data set into a smaller number of 'dimensions'. The data used for the analysis in this paper come from Living Standards Measurements Surveys (LSMS) in 2012. The principal components analysis was used to create an asset index which gave the social economic status (SES) of each household. The cluster analysis is used to give us a full background of the partition of households according to the social-economic groups: low, medium and high.***

Keywords: principal components analysis, asset index, quintile, K means, cluster analysis

1. Introduction

Measuring standard of living has historically been problematic because of the difficulty of defining an aggregate measure that captures the notion of well-being [1].

The evaluation of economic status methods are basically: the first is based in the incomes or the expenditure of consumption and the second, that's not based in the mentioned elements, is called the non monetary poverty. According to INSTAT, the non monetary poverty is composed of some indicators that have no relation with the monetary aspect but with the possession of certain sustainable goods (such as television, washing machine, fridge ect.)

In general, the economists use the expenditure of consumption or the incomes as poverty indicators. Usually, the incomes are used in the developed countries, whereas the expenditure of consumption is used mostly in the developing countries [2]. In this study, the Principal Components Analysis (PCA) is used for the creation of an asset index which reflects the social-economic status of every household. There have been different points of view regarding the estimation of the economic status in the absence of incomes.

Carlo Azzari et al. (2005) study of poverty monitoring in the absence of incomes have used the Principal Components Analysis for the creation of an index asset for Albania dividing the data in rural and urban zones in order to observe the differences of poverty [3]. In one of the recent studies for Albania of Camilla Mastromarco et al. (2010), are analyzed different aspects of poverty using non linear principal components analysis [4].

Seema Vyas and Lilani Kumaranayake study, presents a summary of methods used for the creation of indexes that define the social-economic status of every household. In their study, it is described how in the lack of expenditure of consumption, the Principal Components Analysis is used and is supported in Demographic Health Survey for Ethiopia and Brazil [5]. Several studies have applied Factor Analysis methods to measure poverty. Among them, Whelan et al. (2006) used FA to identify five distinct dimensions of deprivation.

Filmer and Pritchett (2001) used Demographic and Healthy Survey data to show that the relationship between wealth and enrollment in school can be estimated without income or expenditure data, by using household asset variables. PCA provided acceptable and reliable weights for an index of asset to serve as a measure for wealth [6].

2. Methods

The data used to analyze the poverty is taken from the 2012 Living Standards Measurement Study (LSMS) for Albania. The survey covered both rural and urban populations. The survey collected information relating to demographic and detailed information for monthly expenditure per capita and on asset ownership, concerns with the possession of certain goods and housing characteristics. A household was defined as a person or a group of people related or unrelated to each other, who live together in the same dwelling unit and share a common source of food [7].

The survey includes a sample of 6671 household that constitute the observance unit. The sample is chosen by draw using the two selections. As selection basis the data of the Population Registration and Housing of October 2011 are used. In the first round, 834 Primary Unit of Election (PUE) are casually chosen in order to represent the whole territory of the country. Afterwards, 8 households are chosen to be interviewed in the second round for every PUE with the procedure of the systematical election. Four other households are also chosen for every PUE that will serve as replacement of the others in case of no responding or inability of contact making possible the objective target of 6671 interviews in the household [8].

3. Principal Components Analysis

Principal Components Analysis (PCA) is a technique used in the multidimensional statistical environment for the simplification of the original data without losing information. The purpose of this technique is the reduction of the variables in a data set into a smaller number of 'dimensions'.

The principal objective of using PCA in a poverty assessment is to extract the "poverty component" that can be used to compute a asset index for each household [9]. In mathematical terms, from an initial set of n correlated variables, PCA creates uncorrelated components, where each component is a linear weighted combination of the initial variables [10]. For example, from a set of variables X1, X2,…, Xp

$$PC_1 = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p$$
$$PC_2 = a_{21}X_1 + a_{22}X_2 + ...a_{2p}X_p \qquad (1)$$
$$PC_p = a_{p1}X_1 + a_{p2}X_2 + ...a_{pp}X_p$$

where a_{pp} represents the weight for the p-th principal component and the p-th variable. The coefficient of the first principal component $a_{11}, a_{12},..., a_{1p}$ are chosen in such a way that the variance of PC1 is maximized subject to the constraint that: $a_{11}^2 + a_{12}^2 + ...a_{1p}^2 = 1$

The second principal component is completely uncorrelated with the first component. This component explains additional but less variation in the original variable than the first component subject to the same constraint. Subsequent components are uncorrelated with previous components; while explaining smaller and smaller proportions of the variation of the original variables [11].

PCA works best when variables are correlated, but also when the distribution of variables varies across cases, or in this instance, households. It is the assets that are more unequally distributed between households that are given more weight in PCA [12].

The higher the a_{ip} (in absolute value), the higher the weight of the values of X_p will be in the determination of the *i* components. The coefficient which are equal to zero, correspond to the X variables that does not contribute in the determination of the PC components.

The results of the Principal Components Analysis depend on the measurement unit used on the variables. Usually, it happens to not have in disposal in the initial variables in the same unit of measurement. This is an important obstacle because the results would be totaly different. In these conditions, it is not possible to work with these data, but it is necessary standardizing them. Beginning with the initial data (X_1, X_2,…, X_p) by which the average and the standard deviation is found, the standardizing of the data is done.

$$Z_1 = \left(\frac{X_1 - \mu_1}{\sigma_1}\right), \quad Z_2 = \left(\frac{X_2 - \mu_2}{\sigma_p}\right),..., Z_p = \left(\frac{X_p - \mu_p}{\sigma_p}\right) \qquad (2)$$

where μ_p and σ_p are respectively the mean and the standard deviation of the p-th variable over all households.
So, the new variables Z_1, Z_2,...,Z_p are created which have zero average and variance one [11].

The criteria of election of the number of principal components are a few but in the study the Kaiser criteria is chosen (1960) which recommends keeping just the components that have a eigenvalue higher or equal to one. The other components are not to be taken in consideration. The measure of sampling adequacy is Kaiser-Meyer-Olkin (KMO) indicates whether the correlations between variables can be explained by other variables in the dataset. In general, scores above 0.60 are acceptable, above 0.70 are good, above 0.80 are commendable, and above 0.90 are exceptional [9]. Moreover, the Bartlett's test of sphericity can be used to test the null hypothesis that the correlation matrix is a diagonal matrix (that is, all non-diagonal elements are zero) in the sample. Since PCA requires high correlations, a small p-value will favor the rejection of the hypothesis [13].

4. Creating the Poverty Index

In this study, the data are divided in urban and rural zones and the weight of all variables are reflected. The data are fragmented even in 5 quintiles poorest, second, middle, fourth, richest and for each of them the index asset is calculated. Standard statistical software can be used and in this instance Spss was used. The data are obtained from LSMS which hold the information in relation to the household expenditure of consumption as well as the owning of some constant goods, the access to the basic services and the house characteristics.

At first, it has been calculated the standard deviation of the variables and the variables with low standard deviations would carry a low weight from the PCA, for example, an asset which all households own or which no households own (i.e. zero standard deviation) would exhibit no variation between households and would be zero weighted. So, a descriptive analysis is done in order to have a clear picture of which variables are to be kept and which are to be excluded.

Using the factor scores from the first principal component as weights, a dependent variable can then be constructed for each household (Y1) which has a mean equal to zero, and a standard deviation equal to one. This dependent variable can be regarded as the ``household asset score `` and the higher the "household asset score`` the higher the implied SES of that household. The end result of PCA is a single asset index that assigns to each sample household a specific value, called a ``household asset score``, representing that household's status in relation to all other households in the sample.

The variables in consideration to calculate the index asset:

- Number of tapes owned by the household (Tape);
- Number of videos owned by the household (Video);
- Number of electric (or gas) stove owned by the household (Gas/electric/stove)
- Number of cars owned by household (Car);
- Number of mobile phone owned by household (Mobile phone):
- Number of color televisions owned by the household (Color TV);
- Number of washing machines (Washing machine);
- Number of dish washers (Dish washer);
- Number of wood stove owned by household (Wood stove);
- Number of refrigerators owned by the household (Refrigerator);
- Number of water boiler owned by the household (Water boiler);
- Number of conditioner owned by the household (Conditioner);
- Number of computers owned by the household (Computer);
- Number of microwave owned by the household (Microwave);
- Number of decoder tv owned by the household (Decoder TV);
- Number of satellite antenna owned by the household (Antenna);

5. Results

5.1 Principal Components Analysis

The results of the Principal Components Analysis are listed in table 1 with the weight for each variable.

Table 1: Factor scores (weights) of the variables of the first principal component

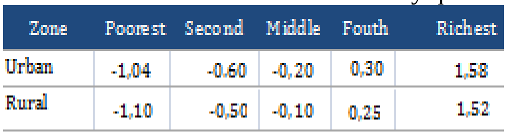
Variables	Component 1 (Total)	Component 1 (Urban)	Component 1 (Rural)
Tv Colour	0,145	0,155	0,134
Video	0,148	0,155	0,161
Tape	0,113	0,123	0,131
Refrigator	0,087	0,086	0,104
Washing mashine	0,118	0,093	0,151
Dishwasher	0,096	0,105	0,073
Gas/electric stove	0,128	0,101	0,152
Wood stove	-0,094	-0,047	-0,102
Conditioner	0,169	0,176	0,165
Water boiler	0,112	0,093	0,136
Computer	0,162	0,163	0,160
Car	0,145	0,151	0,165
Mobile phone	0,132	0,129	0,154
Microwave	0,157	0,163	0,147
Decoder tv	0,107	0,124	0,078
Antene	0,078	0,110	0,099

The factor scores show the different directions of the influence in the index asset for urban and rural zones.

Column Comp1 represents the vectors of weights that the original variables have in the determination of the principal components. This value represents how much the variable contributes in the determination of a component. The stronger is the relation between the variable and the component, the higher is the weight value. Generally, a variable with a positive weight is associated with higher SES, and conversely a variable with a negative weight is associated with lower SES. The first component is interpreted as an indicator of the economic status [14]. Its meaning depends in major part by variables Conditioner, Computer, Video, Microwave, and Car. This is explained by the fact that their coefficients are high and positive. The variables gas /electric stove, washing machine, mobile phone, color television have a positive weight even though lower, they are still variables related to the economic status and confirm the meaning of this components. The variable related with the negative aspect of the economic status is wood stove which has a negative weight. This implies, all things being equal, that a household with a wood stove will be ranked lower in terms of SES than a household that does not own a wood stove. The reason for such a result may be due to ownership of a wood stove being more strongly correlated with variables that are expected to be associated with lower SES. The variable that has a negative value in the economic status in the urban and rural zones is still wood stove.

The value of KMO in our model is 0, 86 and is relatively high, that means that the data are suitable for the Principal Components Analysis. The data are arranged in groups according to the urban and rural zones and each of urban zones data is divided in quintiles. For every quintile the average of household asset score is calculated.

Table 2: Mean household asset score by quintile



Zone	Poorest	Second	Middle	Fouth	Richest
Urban	-1,04	-0.60	-0,20	0,30	1,58
Rural	-1,10	-0,50	-0,10	0,25	1,52

The table 2 shows that between the fourth and richest of the urban region have big differences. The same thing is observed between the poorest and the second group of the same region. For the rural region the differences are noted between the fourth and the richest group.

4.2 Cluster Analysis

In the cluster analysis we search for patterns in a data set by grouping the (multivariate) observations into clusters. The goal is to find an optimal grouping for which the observations or objects within each cluster are similar, but the clusters are dissimilar to each other. To group the observations into clusters, many techniques begin with similarities between all pairs of observations. In many cases the similarities are based on some measure of distance. Other cluster methods use a preliminary choice for cluster centers or a comparison of within- and between-cluster variability [15].

The clustering procedures are:

a) The hierarchical procedure: Agglomerative (start from n clusters to get to one cluster) and divisive (start from one cluster to take n cluster).
b) The most known non hierarchical procedure is the K-means clustering.

c) MacQueen suggests the term K-means for describing an algorithm of his that assigns each item to the cluster having the nearest centroid (mean).
   1) Partition the items into K initial clusters.
   2) Proceed through the list of items, assigning an item to the cluster whose centroid (mean) is nearest. (Distance is usually computed using Euclidean distance with either standardized or unstandardized observations.) Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item.
   3) Repeat Step 2 until no more reassignments take place

Rather than starting with a partition of all items into K preliminary groups in Step 1, we could specify K initial centroids (seed points) and then proceed to Step 2 [11]. K-means clustering is thought to be superior to hierarchical methods as it is less affected by outliers and the presence of irrelevant variables. It is also suitable for applying to very large datasets, especially above sample size 500 [13]. However, unlike the other methods, the researcher has to specify the number of clusters to retain, which sometimes makes it less attractive. In our cluster analysis, the same variables used for PCA or the factor scores of PC1 can be used as inputs.

Table 3: Proportion of households in low, medium and high socio-economic group

	Low	Medium	High
Clusters based on PC_1	-0,58	0.62	2,45
Percentage of households	63,0%	29,5%	7,5%

From the results of the Clustering Analysis, it is noted that more than half of the household in the study have a low social-economic status. A small part of the households (about 7, 5%) has a high social economic status.

5. Conclusions

Compared with other statistical alternatives, PCA is computationally easier, can use the type of data that can be more easily collected in household surveys, and uses all of the variables in reducing the dimensionality of the data. From the Principal Components Analysis (PCA) it can be concluded that the household with high social economic status give high factor score and vice versa. K-means clustering the procedure aims at segmenting the data in such a way that the within-cluster variation is minimized. K-means clustering is thought to be superior to hierarchical methods as it is less affected by outliers and the presence of irrelevant variables. It is also suitable for applying to very large datasets, especially above sample size 500.

References

[1] Mazumdar, K.1999, Measuring the well-being of the developing countries: Achievement and Improvement indices, Social Indicators Research.

[2] Sahn D, Stifel D. 2003. Exploring alternative measures of welfare in the absence of expenditure data. Review of Income and Wealth 49,pp 463–89.

[3] Azzarri, C., G. Carletto, and A. Zezza (2006), “Monitoring Poverty Without Consumption Data. An Application Using the Albania Panel Survey”. Eastern European Economics, 44, pp. 59-82

[4] Mastromarco C, Peragine V, Russo F, Serlenga L, (2010). Poverty, Inequality and Growth in Albania: 2002 - 2005 Evidence, February 2010

[5] Vyas, S and Kumaranayake, L (2006). ``Constructing social-economic status indices``, How to use PCA. Health Policy and Planning

[6] Filmer, D. & Pritchett, L. 2001. Estimating Wealth Effects without Expenditure Data or Tears: An Application to Educational Enrollments in States of India.

[7] J. Haughton, Sh.R. Khandeker.'Handbook on Poverty and Inequality'. ISBN: 978-0-8213 - 7613-3, pp.11-12, 2009

[8] www.instat.gov.al

[9] Henry, C., Sharma, M., Lapenu, C., Zeller, M., 2003. Microfinance poverty assessment tool, Tech. T. S. 5, Consultative Group to Assist the Poor (CGAP) and The World Bank, Washington, D.C.

[10] George H.Dunteman. Principal Components Analysis. ISBN 0-8039-3104-2, pp 5-10

[11] Richard A Johnson, Dean W.Wichern. ``Applied Multivariate Statistical Analysis`` ISBN 0-13-187715-1

[12] McKenzie DJ. 2003. Measure inequality with asset indicators. BREAD Working Paper No. 042. Cambridge, MA: Bureau for Research and Economic Analysis of Development, Center for International Development, Harvard University

[13] MOOI, E. & SARSTEDT, M. 2011. A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics, Heidelberg, Springer. ISBN 978-3-642-12540-9

[14] Houweling TAJ, Kunst AE, Mackenbach JP. 2003. ‘Measuring health inequality among children in developing countries: does the choice of the indicator of economic status matter?’ International Journal for Equity in Health 2: 8.

[15] Alvin C. Rencher `` Methods of multivariate analysis`` second edition. ISBN 0-471-41889-7, pp 466- 467

Author Profile

Evgjeni Xhafaj received the B.S and M.S in the Mathematics in the Faculty of Natural Science in the University of Tirana during 2002-2007. Since 2011, she is a PhD candidate in probability-statistics. With an eight years experience in the academic sector, now she is a lector at the Faculty of Technology and Information at University ‘Aleksander Moisiu’ Durres (Albania).