УДК 330.59

L. Krasnikova, B. Povoroznyk

FACTOR ANALYSIS VERSUS PRINCIPAL COMPONENT ANALYSIS IN THE CASE OF "INFRASTRUCTURAL" POVERTY MEASUREMENT

We consider alternative way of looking on the problem of poverty measurement based on asset index method. Our decision is motivated by a number of measurement problems that prevent the use of monetary metrics (consumption and income) of welfare in developing countries. We substantiate the conception of "infrastructural" poverty that allows to define economic status of households in terms of assets of wealth (durables) rather than in terms of monetary units (income and consumption). Factor Analysis and Principal Component analysis are used for construction an asset index that is a refection of welfare profile in terms of assets. We analyze the advantages of both methods in the asset index construction.

Introduction

Social security programs in developing countries require detailed and precise analysis of the household survey data and of course the assessment of the factors that causes poverty. Traditionally, poverty profiles are constructed on the basis of various household income, consumption and expenditure surveys as a preferred indicators of living standards.

Researchers prefer to use data on income or consumption expenditures, thus relying on the money metric of utility. Also there is a common practice when income is used for measuring poverty in developed countries and consumption or expenditures for developing countries, and it is due to differences in the reliability and availability of data in these two categories of countries. Bollen indicates that Friedman's emphasis on the distinction between permanent and transitory income has led many researchers to reject proxy measures of permanent income and economic status such as current annual earnings, because income may vary greatly from year to year [3]. Behrman and Deolalikar propose to use average income over several years to get a better measure[2]. Fomenko argues that consumption is more preferable for measuring poverty in Ukraine, because of measurement bias - due to high taxation and black economy income is often underreported. Because a lot of workers in Ukraine get both official and unofficial salary, it is very difficult to collect precise data on the true income of the households [7]. Moreover, income of households in agricultural regions is comparatively poorly reflected in official statistics about income. Also, Friedman suggested that consumption behavior reflects permanent income because it is primarily driven by permanent income[8]. It is a well known fact that households tend to smooth their consumption from year to year Deaton considers expenditures to be less variable than income and more reflective of long - term economic status, on his mind annual household expenditures

may provide better permanent income proxies [5].

Though data on consumption expenditures is traditionally more preferable indicator of household's socio - economical status in developing countries it has a lot of disadvantages. In many developing countries data set does not contain any information on income or consumption, or is of poor quality. Proper and tailored use of consumption expenditures for construction of unified money metric requires precise and reliable information on the prices of consumed goods and services, data on nominal interest rates and depreciation rates of durable goods. Collection and consolidation of data on regional price indices and rental prices on housing requires considerable efforts and organization expenses due to regional diversity and disparity in economic development. There is also a purely data collection problem - recall bias, due to consumption expenditures surveys conducting on the basis of recall - several days[10]. The longer is the period of recall - the greater is the bias. All these problems involved in constructing monetary metric motivate researchers to use alternative approach for welfare assessment and designation, based on other data rather then consumption expenditures.

It is certain that researchers give relatively insufficient attention to the households' ownership of durables (assets) or to the inequality in possessing those assets among households or individuals. Sahn and Stifel believe that significant poverty alleviation is fundamentally predicated on the individual's ability to accumulate productive assets. Since income inequality will be reduced by addressing the unequal distribution of income generating assets, there is considerable merit in moving the process of poverty measurement away from solely expenditure - based measures towards a more assets - based form.

This idea of this research is based on the Sen's conception of "entitlements", defined as a set of alternative commodity bundles which person can operate and accumulate in society [10]. It allows us to move from the expenditures based idea of poverty towards assets conception of poverty. In addition it is often much more easier to collect data on ownership of different assets than on either income or consumption.

Problem Description

It was used Sen's conception of "entitlements" defining economic status of households in terms of assets of wealth (durables) rather than in terms of monetary units (income and consumption) and proposed assets based idea of "infrastructural" poverty. "Infrastructural" poverty assessment requires the use of asset index method as an alternative method of welfare estimation, based on the asset measurement.

Asset index method is another approach to poverty measures. With this technique the socio economic status of households is defined in terms of assets or wealth, rather than in terms of income or consumption. This method uses data on actual physical assets such as durables, human capital or housing characteristics. These variables arise "on an equal footing", so that there are no dependent variables and several explanatory variables as in multiple regression. So, asset index method deals with "multivariate" information on asset ownership of every household or individual from the sample. The idea of the method is to create uniform single - dimensional equivalent to multivariate vector of assets, called "asset index". Thus it will give us the possibility to provide wealth ranking among the households possessing varieties of assets. A number of different methods are used for this purpose. The most straightforward and easiest way is to assign equal weights to the ownership of each asset and to take a sum of these weights for every household, thus ranking households accordingly to the sum of weights. However such approach has some disadvantages and so it is not appropriate in many cases. For example, it assumes that having a radio has the same influence on the welfare of the household as having access to gas line. Another possible solution is to create our own set of weights, such as prices of different assets, that could be used for constructing an index of household wealth. Unfortunately this method involves various problems that deal with availability of the prices of those different assets.

It appears that there are two most appropriate methods designated for determination of the weights for the index of assets: factor analysis and principal components analysis.

Main results

Successful estimation of the weights for assets allows to solve main problem of asset index methodology: create uniform one-dimensional equivalent to the multidimensional vector of assets. Asset index method regards sets of assets correspondent to every household or individual from the sample and is aimed for providing welfare ranking among those households. Factor analysis and principal component analysis provide effective dimensionality reduction without loosing too much information.

According to Chatfield and Collins, principal component analysis consists of finding an orthog-

опаї transformation of the original variables (vector of assets correspondent to every household) to a new set of uncorrelated variables, called principal components, which are ranked in decreasing order of importance [4]. Researchers often hope that first few components will contain most of the variation in the original data so that the effective dimensionality of the data can be reduced. Principal component analysis was originated in work by Karl Pearson around the turn of the previous century, and was further developed in the 1930s by Harold Hotelling. According to this method each household is assigned a weight or factor score generated through principal components analysis (PCA). It is used for examining relationship among a set of *p* correlated. It is variable - directed technique that is appropriate when the variables arise 'equally', so, that we don't have dependent variable and several independent (explanatory) variables. Thus the advantage of such approach is that PCA technique allows the reduction of the number of variables (dimensionality) without losing too much information. And it is achieved by creation of smaller number of variables which explain most of the variation in the original variables. This newly created variables (principal components) are uncorrelated and are the linear combinations of old ones.

Many researchers tried to investigate whether asset index method and PCA approach as one of it's instruments is really an appropriate procedure for wealth ranking. Several studies tried to search the range to which the asset index is a nice proxy for household consumption expenditures. Filmer and Pritchett proposed a method for estimating the effect of economic status on educational outcomes without direct survey information on income or expenditures [6]. They constructed an index based on indicators of household assets, deriving them by the statistical procedure of principal components in order to solve so important problem of choosing the appropriate weights for the assets. Filmer and Pritchett used data from Indonesia, Nepal, and Pakistan which had both expenditures and asset variables. They showed that there is not only the correspondence between a classification of households based on the asset index and consumption expenditures but also that asset index is a better proxy for predicting enrollments than consumption expenditures. Bollen examined the performance of proxy for economic status based on the asset index method [3]. They found that there is a difference in outcomes while using proxies to direct estimation of poverty, but the choice of proxy variable using asset index for revealing influence on non-economic variables exhibit greater robustness than monetary proxies.

Methodology of PCA

An illustration of Principal Component Analysis (PCA) is provided upon basis of the Chatfield and Collins and the main idea shortly is presented below. Suppose $X^T=[X_1,...,X_p]$ is a p - dimensional random variable (in our case/) data on household asset) with mean μ and Σ covariance matrix The idea is to find a new set of variables, $Y_1,...,Y_p$ that are uncorrelated and whose variances decrease from first to last . Each Y_j *(j'-th* principal component) is taken to be linear combination of the X's:

$$Yj = a_{1j}X_1 + a_{2j}X_2 + ... + a_{pj}X_p = a^T_jX, \quad (1)$$

where $aT_j = [a_{1j},...,a_{pj}]$ is a vector of constants. Also, additional condition should be imposed:

$$a_j^T a_j = \sum_{k=1}^{p} a_{kj}^2 = 1.$$

This normalization procedure ensures that the overall transformation is orthogonal (distances in /'-space are preserved). The first principal component Y_1, is obtained by taking such a_1 that Y_1 has the largest possible variance: take such a_1, that variance $a^T_1 X$ is maximized. This approach is originally suggested by Harold Ho telling. The second principal component is found by choosing a_2 so that Y_2 has the largest possible variance for all combinations of the form of equation (1) which are uncorrelated with Y_1 Similarly, we derive $Y_3,...Y_p$ so as to be uncorrelated and to have decreasing variance.

Estimation of first principal component involves solving maximization problem with the help of Lagrange multipliers method as a standard procedure for maximizing a function of several variables subject to one or more constraints. Formula for the variance of first principal component is:

$$Var(Y_1) = Var(a^T_1X) = a^T_1\Sigma a_1$$

Applying the Lagrange multiplier method, we have: $L(a_1) = a^T_1\Sigma a_1 - \lambda(a^T_1a_1 - 1)$. First order condition is

$$\frac{\partial L}{\partial a_1} = 2\Sigma a_1 - 2\lambda a_1 = 0,$$

that is equivalent to $(\Sigma - \lambda I)a_1 = 0$. In order to have a solution for a_1 other than the null vector, then λ must be chosen so that $|\Sigma-\lambda I| = 0$.

Thus a non - zero solution exists if and only if λ is an eigenvalue of Σ. But Σ will generally have p eigenvalues, which all must be nonnegative as Σ is positive semidefinite. These eigenvalues are denoted as $\lambda_1,\lambda_2,...,\lambda_p$ and an assumption that they are distinct can be freely made, so that $\lambda_1>\lambda_2>...$ » $\lambda_p > 0$. It is easy to prove that $Var(a^T_1X)=\lambda$.

Since variance is equal to the eigenvalue, the largest eigenvalue should be chosen in order to maximize the variance. Then, the desirable value a_1 of must be the eigenvector of Σ corresponding to the largest taken eigenvalue. Vector a_1 is the vector of weights for assets in the asset index method.

The second principal component, $Y_2 = a^T_2X$ is obtained similarly but with one extension. In addition to the scaling constraint that $a^T_2a_2 = 1$ there is another constraint that Y_2 should be uncorrelated with Y_1 Continuing this argument, the jth principal component has to be associated with the jth largest eigenvalue. In case when some of the eigenvalues of Σ are equal there is no unique way of choosing the corresponding eigenvectors, but as long as the eigenvectors associated with multiple roots are chosen to be orthogonal, then the argument carries through.

The variables used in the poverty analysis are usually measured in different scales (some of the variables are binary, some other categorical and some other continuous). This can lead to one variable having an excessive influence on the principal components simply because of the scale of measurement. Researchers traditionally avoid this problem by standardizing original variables. So, that covariance of the standardized variables $X^*_1,X^*_2,...,X^*_p$ is simply the correlation matrix of the original variables. For the correlation matrix, the diagonal terms are all unity. Thus the sum of the diagonal terms (or the sum of the variances of the standardized variables) will be equal *top.* Thus the sum of the eigenvalues of correlation matrix P will also be equal to *p,* so that the proportion of the total variation accounted for by the j-th component is simply λ_j/p. Proportion of variance explained by the first principal components will depend on the number of variables included in the analysis, so it is crucial to include as much variables as possible.

Usually, only first principal component is considered due to sharp decrease in proportion of explained variance in asset index model. The corresponding eigenvector is the vector of weights ($\bar{a}_1 = (a_1,...a_{21})$). Vector $\bar{a}$ is taken such that Y_1 has the largest possible variance and it defines the weights of explanatory variables in forming the principal component. Having the corresponding weights for each explanatory variable gives us possibility to calculate asset index for each household from the sample.

Here is the formula that is used for calculating the asset index (A_i) for the i-th household:

where a_1 is the eigenvector for the first asset as determined by the procedure, xi_1 is the i^{th} household's value for the first asset and x_1 and s_1 are the mean and standard deviation of the first asset variable over all households. This formula shows the role of the assets characteristics in forming the level of welfare (asset index) computed according to our methodology. For asset variables which take only the values of zero or one, the weights have an easy interpretation. A move from 0 to I (if household does not own or owns first asset) changes index by a_1/s_1. Those scores are summed for every household, and indices are calculated by above formula. As a result, a welfare ranking can be done in single dimensional space according to corresponding indices. So, we have sample distribution of household's scores in descending order, that is used to create the breakpoint that defines wealth quintiles as follows. It is very important to provide poverty profile that characterizes the poor and distinguishes their attributes from the non-poor The distribution of household's indices is divided into population quintiles. In this case poverty line cant be defined in monetary terms. Moreover, there is no common agreement in the literature about the poverty lines for asset index method (poverty lines for infrastructural poverty). It is appropriate to set the relative poverty line, for example as an upper bound of the lowest 40 per cent quintile of the distribution of household's population indices for the whole sample.

Factor analysis

According to Chatfield and Collins, factor analysis (FA) has a similar aim to principal component analysis in that it is a variable-directed technique which is appropriate when the variables arise "on an equal footing" [4]. The idea of FA is to derive new variables called factors which will help to construct a single unified index from the multidimensional data, thus giving a better understanding of the data. While PCA provides an orthogonal transformation of the variables which does not depend on any model, FA requires proper statistical model and deals more with explaining the covariance structure of the variables than with explaining the variances.

The basic ideas of FA were originated from the works of Francis Galton and Charles Spearman, inspired by the efforts of psychologists to provide a better understanding of "intelligence". FA was developed to analyze test scores for intelligence which contained a large variety of questions (corresponding to verbal ability, mathematical ability, memory, etc). FA allowed to determine if intelli-

gence was made up of a single underlying general factor or of several more limited factors.

FA is usually applied in psychology and the social sciences. For example, in asset index method for welfare measurement information is collected from big number of households as to their ownership of various durables (assets) and living conditions. There is a question whether the concept of "infrastructural" poverty class is multidimensional or it is possible to construct a single index of class from the data. Sahn and Stifel applied FA technique while considering an asset - based alternative to the standard use of expenditures in defining poverty.

Methodology

Again using Chatfield and Collins, as in case of PCA we have observations on *p* variables (assets) $X^T = [X_1,...,X_p]$ with mean μ and covariance matrix Σ. The aim of factor analysis is to explain the covariance structure of the variables, so an assumption that and that matrix is of full rank *p* can be easily imposed. FA model imposes crucial assumption about existence of *m* underlying factors ($m < p$) which are denoted by $f_1,f_2,...f_m$ and that each observed variable is a linear function of these factors together with a residual variate, so that $X_j = \lambda_{j1}f_1 + ... + \lambda_{jm}f_m + e_j, j = 1,...,p$ (1).

The weights $\{\lambda_{jk}\}$ are usually called the factor loadings, so that they reflect the contribution of the factors to the variables. Specific residual variation is described by e_j. Traditionally, the factors $\{f_i\}$ are called common factors, while the residual variates $\{e_j\}$ are often called the specific factors. While using factor, it is necessary to make some assumptions about model (1). It is usually assumed that specific factors are independent of one another and of common factors. Also, common factors are assumed to be independent of one another.

As it was mentioned before, it is necessary to provide specific model for every concrete case. Sahn and Stifel proposed a model for asset index construction in poverty analysis consisting of only factor. This choice can be explained by ambiguous concept of the set of underlying unobservable variables that may be reasonable in some situations, particularly in psychological research, but not appealing in many other practical situations, such as welfare measurement. Thus, Sahn and Stifel imposed linear dependence of the ownership of every asset from an unobserved common factor for each household, which was labeled "household welfare". So, the structural model includes only one factor:

$$a_{ik} = \beta_k c_i + u_{ik} \text{ for } i=1,...,N \{households\}, \quad (2)$$
$$k=1,...,K \{household\ assets\}$$

The ownership of each observed asset (k) for each household *(i)*, represented by the variable a_{ik}, is a linear function of an unobserved common factor for each household, c*i,* which they designated as "household welfare". We should mention that main goal of FA is to estimate unobserved relationship between the asset and the unobserved common factor, β_k as well as the noise component ("unique element" or specific factor).

Following assumptions should be made for model identification:

(1) Households are distributed *iid*

$$(2)\ E(u_i / c_i) = 0_{Kx1}$$

$$(3)\ V(u_i) = Diag(\sigma^2_1,...,\sigma^2_K)$$

Structure on the variance- covariance matrix of the observed assets should be imposed. Equation (2) in the vector form is: $a_i = \beta c_i + u_i$. Assumptions on the common and specific factors gives us the variance - covariance matrix of the unique disturbances $E(u_i u_i') = Diag(\sigma^2_1,...,\sigma^2_K) = \psi$.

Without loss of generality, mean of the common factor (welfare) is assumed to be equal to zero, thus the variance of the common factor is $E(c_i c_i') = \sigma^2_c$.

The variance of the assets is

$$E(a_i a_i') = E[(\beta c_i + u_i)(\beta c_i + u_i)'].$$

Which gives: $\Omega = \beta\beta'\sigma^2_c + \psi$.

Traditionally, the variance of the unobserved factor is chosen ($\sigma^2_c = 1$). We should mention that this normalization makes it very difficult to interpret the coefficients on the common factors (P), though interpretation of these parameters is not crucial for asset index construction. Assuming the multivariate normality of c_i and u_i, gives possibility to use maximum likelihood techniques for estimating β and ψ. After estimating these parameters, a common factor (asset index) can be estimated for every household from the sample. Sahn and Stifel do it by defining the asset index as a projection of unobserved household wealth (c_i) on the observed household assets:

$$E^*(c_i/a_i) = \gamma_1 a_{i1} + ... + \gamma_k a_{ik},$$

$$\text{where } \gamma = \gamma(a_i)^{-1}\ cov(a_i, c_i).$$

Using ($\sigma^2_c = 1$), we can show that $cov(a_i,c_i) = \beta$, and thus $\gamma = \Omega^{-1}\beta$. Asset index for household j is estimated using the following formula

$$\text{where} = \hat{\gamma} = \hat{\Omega}^{-1}\beta\ \hat{\sigma}^2.$$

As a result, the goal of dimensionality cutting is achieved, an index depicting poverty profile of every household in single dimensional space can be

constructed that gives the possibility to provide poverty ranking among households. Poverty line should be imposed similarly to given for the PCA method.

Conclusions

Undoubtedly PCA and FA achieve the same goal in the asset index construction: reduce dimensionality and provide estimation of the weights involved in the asset indices construction. Nevertheless, whereas PCA finds an orthogonal transformation of the variables (assets), FA depends on a "proper" statistical model. These two methods are similarly useless if all the observed variables are approximately uncorrelated. For example, while PCA and FA are used for asset index construction on the same set of data it is obvious to analyze whether the first *m* principal components will be similar to factor loadings in an *m* factor model. There were showed that in some cases there can be agreement (spearman rank correlation between the principal components and factor analysis asset indices is about 0,98 for each of their sample), but usually there are no [9]. It depends on the number of significantly high values of principal components (proportion of explained variance) and on rotations in FA. If the sample of data were obtained using random perturbations, than the errors will influence on the variance of the first few components as well as those of the remaining components, although some researchers prefer to use PCA even when there is evidence of random errors. The main advantage of the FA over the PCA is that procedure of maximum likelihood estimation used in FA allows to overcome the scaling problem in PCA, and that FA provides proper statistical model with an error structure. Chatfield and Collins (1980) stress that researchers should be careful, because FA has various drawbacks. FA requires to impose a large number of assumptions, which are not always realistic in practice. Also, basic assumption on the existence of factors sometimes may be very controversial. In practice number of factors (*m)* is often unknown, and it is difficult to select the correct value of *m*. In PCA it is easy to calculate component scores for an every individual and use it in follow - up analyses, while FA model has no obvious inverse and it is difficult to estimate factor scores from observed data. The idea of superiority of PCA over FA is supported by Blackith and Reyment. Chatfield and Collins also recommended that FA should not be used in most practical situations.

While speaking about handling of one or another method in "infrastructural" poverty measurement, we should mention that it is difficult to impose some other factor influencing on the observed variables (assets) rather than "household welfare". That's why, researchers constructing asset indices using FA use only one - factor model. Similarly, scientists that use PCA technique regard only the highest first principal component (which they interpret as "household welfare") due to conceptual difficulties of interpretation of all the further components. The findings of the investigation convincingly show that contradictions between FA and PCA become more smoothed over in case of infrastructural poverty assessment. Further research will be provided using asset index method on the base of Ukrainian data from Household Budget Survey.

1. *Baschieri Angela, Craig Hutton.* Creating a Poverty Map for Azerbaijan // Programmatic Poverty Assessment. 2004.- P. 10-15.
2. *Behrman, J. R and A. B., Deolalikar.* The intrahousehold demand for nutrients in rural South India // Journal of Human Resources,- 1980.- № 25 (4).- C. 12-19.
3. *Bollen, K. A., J. Glanville, and G. Stecklov.* Economic Status Proxies in Studies of Fertility in Developing Countries: Do the Measure Matter? // Measure Evaluation Working Papers.- 2001.- WP - 01-38.
4. *Chatfield, C. and A.J. Collins.* Introduction to Multivariate Data Analysis. London: Chapman and Hall, 1980.
5. *Beaton, Angus. A* Microeconomic approach to Development Policy // The analysis of Household Surveys.- 1997.- The World Bank.
6. *Filmer D. and L. H., Prichett.* Estimating wealth effects without expenditure data - or tears: an application to educational enrollments in India // Demography.- 2001.- Vol. 38(1).
7. *Fomenko Hanna.* The Determinants of Poverty in Ukraine // EERC Thesis, 2004.
8. *Friedman, Milton. A* Theory of the Consumption Function.- Princeto, 1957.
9. *Gwatkin, D.R., S. Rutstein, K. Johnson.* Socio - Economic Differences in Health, Nutrition, and Population"// HNP/Poverty Thematic Group,- World Bank, 2000.
10. *Sen, A.K.* Development: Which way now? // The Economic Journal,- 1983.- Vol 83.
11. *Лук"яненко І. Г., Краснікова Л. І.* Економетрика: Підручник.- К.: Товариство «Знання», КОО, 1998.- 494 с.

Краснікова П., Поворозник Б.

ФАКТОРНИЙ АНАЛІЗ ТА АНАЛІЗ ГОЛОВНИХ КОМПОНЕНТІВ У ВИМІРЮВАННІ «ШФРАСТРУКТУРНОЇ» БІДНОСТІ

У статті розглянуто альтернативний підхід до проблеми вимірювання рівня бідності, базований на «методі активів». Це зумовлено цілою низкою проблем, що заважають використанню монетарних заходів (витрат на споживання та доходи) щодо добробуту у країнах, що розвиваються. Впроваджено концепцію «інфраструктурної» бідності, що дає змогу визначати економічний статус домогосподарств у термінах власності на товари тривалого користування (активів), а не в монетарних показниках (доходу і споживання). Методи факторного аналізу й аналізу головних компонентів використовуються для побудови індексу активів, який відображає рівень добробуту в монетарних показниках. Проаналізовано переваги обох методів багатовимірного аналізу у побудові індексу активів. Хоча більшість науковців надають перевагу аналізу головних компонентів і рекомендують уваж:но ставитись до використання факторного аналізу, розбіж:ності у методах зовсім незначні, якщо йдеться про побудову індексу активів.