C | E | D | L | A | S

Centro de Estudios
Distributivos, Laborales y Sociales

Maestría en Economía
Universidad Nacional de La Plata

Deprivation and the Dimnensionality of Welfare: A Variable-Selection Cluster-Analysis Approach

Germán Caruso, Walter Sosa-Escudero y Marcela Svarc

Documento de Trabajo Nro. 112
Enero, 2011

ISSN 1853-0168

Deprivation and the Dimensionality of Welfare: A Variable-Selection Cluster-Analysis Approach

Germán Caruso
Universidad de San Andres

Walter Sosa-Escudero
Universidad de San Andrés *

Marcela Svarc
Universidad de San Andrés

August 2010

Abstract

In this paper we tackle the problems of dimensionality of welfare and that of identifying the multidimensionally poor by first finding the poor using the original space of attributes, and then reducing the welfare space. The starting point is the notion that the 'poor' constitutes a group of individuals that are essentially different from the 'non-poor' in a truly multidimensional framwework. Once this group has been identified, we propose reducing the dimension of the original welfare space by solving the problem of finding the smallest set of attributes that can reproduce as accurately as possible the 'poor/non-poor' classification in the first stage.

JEL classification: D31, I32, C49
Keywords: Multidimensional welfare, poverty, factory analysis, clusters.

*Corresponding author: Walter Sosa-Escudero, Department of Economics, Universidad de San Andres. Address: Vito Dumas 284, Victoria, Buenos Aires, Argentina. wsosa@udesa.edu.ar. This work is part of a research started in Gasparini, L. Sosa Escudero, W., Marchionni, M. and Olivieri, S., 2008, 'Income, Deprivation, and Perceptions in Latin America and the Caribbean: New Evidence from the Gallup World Poll', prepared for the Latin American Research Network on Quality of Life in Latin America and the Caribbean. We are very grateful for the permission to use the data set from this paper. We thank Ricardo Fraiman for very useful insights and for kindly providing his computational routines. All errors and omisions are our responsibility.

1 Introduction

Well-being and its related notions, like deprivation or inequality, are elusive concepts, and the efforts leading to define them precisely cannot be disentangled from the practical need to quantify them to make valid comparisons, or to assess their importance. To complicate matters, a massive body of recent literature points towards the multidimensional nature of welfare (Kakwani and Silber, 2008). The mere notion of a concept being 'multidimensional' is elusive as well, but it clearly suggests the inability to measure it based on a single scalar dimension, like income or consumption in the case of welfare. Moreover, even when there is agreement on the multidimensionality of well-being, there remains the problem of deciding how many dimensions are relevant, and which attributes or variables should be considered for a more accurate assessment.

The multidimensionality of welfare translates almost directly into that of poverty or deprivation. A recent line of research has focused on first solving the problem of dimensionality of welfare, that is, to identify how many relevant dimensions must be considered to measure welfare, and then proceeding to identify the poor, based on this reduced set of variables. For example, Gasparini et al. (2009) and Ferro Luzzi et al. (2008) start with a rather large set of variables that can be seen as alternative measures of an underlying welfare space, and then use factor analytic methods in order to produce a small set or variables (factors) that appropriately capture the variability of welfare. The fact that more than one factor is needed to appropriately welfare is interpreted as evidence of its multidimensionality. After reducing the dimensionality of the problem, they proceed to find the poor based on this reduced set of factors. Gasparini et al. (2009) identify the poor along each of the relevant dimensions, whereas Ferro Luzzi et al. apply cluster techniques on all relevant dimensions, to find a group of individuals that can be safely labeled as 'poor', in a multidimensional sense.

In this paper we adopt and alternative route that first identifies the poor and then explores the dimensionality of welfare. The starting point is the notion that the 'poor' constitutes a group of individuals that are essentially different from the 'non-poor', in a multidimensional framework. Once this group has been identified, we propose reducing the dimension of the original welfare space by finding the smallest set of attributes that

can reproduce as accurately as possible the 'poor/non-poor' classification obtained in the first stage. More concretely, we start by applying cluster methods on a rather large set of attributes, in order to identify a group that can be reasonably be labeled as 'the poor'. Once this satisfactory classification has been produced, in order to reduce dimensionality, we use recent methods on variable selection for cluster analysis. We implement the *blinding* approach of Fraiman, et al. (2008) to find the smallest set of variables that is able to reproduce the 'poor/non-poor' classification of the first stage. In this context, the multidimensionality of welfare (and hence poverty) is related to the fact that this reduced set includes more than one variable.

A first important advantage of this approach is that by construction, cluster methods guarantee high similarity within groups and high dissimilarity between groups, and hence, if it exists, the poor is a coherent group, by construction. Reducing the dimensionality first may unnecessarily complicate the goal of finding the poor based on the similarity-dissimilarity requirements of the cluster based approach. For example, the usual 'single dimensional' classification based on poverty lines produces a sharp and unambiguous characterization of the 'poor/non-poor' status, but a well known drawback is that individuals close to the poverty line are indistinguishable among them, inducing a classification of poor-non/poor that does not satisfy the dissimilarity requirements imposed on the groups. An advantage of our approach is to allow *all* variables in the welfare space to contribute towards the goal of identifying the deprived. Second, factor methods have well known identification problems, that are usually by-passed by imposing orthogonality requirements, and/or the adoption of sometimes arbitrary 'rotations' (see Hardle and Simar, 2003, Ch.10). The usual output of standard factor analysis is a set of variables (factors) that are linear combinations of the original ones, and there is substantial controversy regarding the interpretability of these factors. Our approach is free from this ambiguities, since by construction, the reduced set of variables identified in the second stage is composed of variables originally in the welfare space.

The goals of this paper require the use of a data set that contains a large set of variables that together represent the relevant dimensions of welfare. This has usually been a hindrance in applied studies since available data usually focuses on some specific dimensions like those included in standard household surveys (typically income, expenditure and

other socioeconomic variables), but usually exluding aspects that the recent literature on multidimensional welfare emphasizes, in particular those related to subjective notions of welfare. In this paper we implement the proposed strategy using the Gallup World Poll, a comprehensive data set that includes questions on objetive and subjective attributes of welfare, that can appropriately provide a starting point for the goals of identifying the poor and studying the multidimensionality of welfare. In spite of being a very rich source of information, its use for research purpose is relatively new, see Gasparini et al. (2009) for a detailed review of this data set and a comparisson with other more standard sources like national household surveys.

The paper is organized as follows. The next section discusses with more detail the problems of multidimensional welfare and poverty and its empirical consequences. Section 3 describes the proposed methodology, based on recent cluster variable-selection methods. Section 4 describes the Gallup Poll data set, and section 5 presents the empirical results. Section 6 concludes and discusses further research

2 Multidimensional welfare and poverty

The seminal work by Sen (1985) and its related literature (see Kakwani and Silber (2008) for a recent collection of results) clearly points towards the multidimensionality of welfare, in the sense that it cannot be appropriately represented by a single dimensional notion like income or consumption. Consequently, the status of poor arises as a consequence of assessing all relevant dimensions involved in determining well being. Were these dimensions conceptually known in advance, the natural way to quantify welfare is to measure each of them empirically, in which case the number of variables coincides exactly with the dimension of the welfare space. The mere fact that welfare is multidimensional simply states that one variable is not enough to properly capture it, without clear signs of which variables to measure and map to each dimension. In this context, a large socioeconomic household survey can be seen as a collection of variables that *together* capture the variability of welfare. Two natural and related questions are the following: 1) how to find the poor based on the information provided by such a large set of attributes?, 2) which is the dimensionality of welfare? That is, how many underlying variables are relevant to capture

welfare and, eventually, if it is possible to represent the whole welfare space in terms of a few variables or indexes.

A recent line of research has relied on factor analytic methods to attack the problem of dimensionality. That is, welfare is thought as being appropriately represented by a few latent, not directly observed factors. Observed variables are then seen as being constructed as linear combinations of these factors, hence the empirical problem consists in recovering these latent factors based on the observed variables. The fact that welfare is multidimensional is linked to the relevance of more than one factor.

This is the approach adopted by recent papers by Ferro Luzzi et al. (2008) and Gasparini et al. (2010), with promising results. Gasparini et al. (2010) base their analysis on the Gallup World Poll, and their initial data set contains 15 variables, including income, and other monetary and non-monetary measures of welfare, as well as some indicators related to subjective welfare. They conclude that their initial space of 15 variables can be reasonable represented by three factors. The first one is based mostly on income. The second one is interpreted as related to subjective welfare, since it is composed mostly of questions related to this concept, and, finally, the third one is related to standard 'basic needs' measures, like water access. Ferro-Luzzi et al. (2008) start with 32 variables from the Swiss Household Panel, and conclude that they can be appropriately represented by four latent factors that they relate to financial, health, neighborhood and social exclusion dimensions.

To summarize, both papers find evidence that the original welfare space, composed of many relevant measures, can be drastically reduced to a few factors, and that more than one variable is needed to adequately represent it, even when income (in the case of Gasparini et al. (2009)) or variables closely related to it (the financial ones in the case of Ferro-Luzzi et al. (2008)) are included in their data sets. In spite of being strongly associated to a relevant factor, both studies point towards the inadequacy of income solely to capture the multidimensional nature of welfare.

Regarding the problem of finding the poor, both papers attempt to derive the poverty status based on the reduced welfare space, that is, on the factors obtained in the first stage. Gasparini et al. (2010) do not attempt to produce a single notion of poverty, instead, they compute a poverty status for each of the relevant factors, that is, they set poverty lines for

each of the factors separately, and produce poverty rates for each dimension, see Gasparini et al. (2010) for further details. Ferro Luzzi et al. (2008), on the other hand, produce a single notion of poverty by using cluster methods based on their reduced welfare space, that is, they use the factors produced in their initial stage as an input for standard clustering algorithms to identify coherent groups. They find that the scores obtained in the factor analysis stage can be reasonably clustered in three clusters for 1999, two for 2000 and 2001, and four for 2002 and 2003. In all cases, these authors find that one group presents substantially low values for all scores and hence this particular group is labeled as the 'multidimensional poor'.

There are several methodological concerns related to this approach, which basically consists in a first stage where the dimensionality of the original welfare space is reduced using factor methods, and then the poor are found based on this reduced set. First, though immensely popular in other disciplines (psychology, for example) covariance methods like factors or principal components are much scarce in Economics. This is mostly due to their well known identification issues which harm their direct interpretation. Basically, factors are linear combinations of the original variables, identified up to orthogonal transformations (see Elffers, Bethlehem and Gill (1978) for a detailed overview of these problems). The standard practice, and the one adopted in both Gasparini et al. (2009) and Ferro Luzzi et al. (2008), is to rely on 'rotations' or other algebraic transformations to produce interpretable results. Second, factors are not directly observable, and hence for practical reasons, new information must be constructed by sampling the whole set of initial variables. For example, suppose that the analysis must be repeated for a different period or region, then all the initial variables must be measured in order to construct the factors, even under the assumption that the underlying latent structure remains unchanged. Third, reducing the dimensionality first may unnecessarily complicate the identification of a coherent group (the poor) that can be safely distinguished from its complement (the non-poor). This is particularly relevant when most variables in the welfare space consist on categorical (in most cases, binary) variables. The aggregation process implicit in the factor analytic approach may smooth out relevant differences contained in the original welfare space. For example, standard income based poverty lines have serious troubles distinguishing the poor from the non poor when the distribution of income is densely populated around the poverty

line. Other categorical indicators may actually help separating the poor from the non-poor.

3 The Variable-selection cluster analysis approach

Based on the concerns of the previous sections, we will favor and approach that 1) preserves the original welfare space in order to identify the poor, and 2) can reduce its dimensionality by producing unambigously interpretable variables, that can be resampled or reconstructed easily.

Our strategy starts by applying cluster methods to the original welfare space. Once the poor is satisfactorily identified, the problem of dimensionality is solved by finding the smallest set of variables in the original welfare space, that can reproduce the poor/non-poor classification of the first stage, as accurately as possible. To this point, we will use recent results on variable selection for cluster analysis. As in the case of factor methods, 'multidimensionality' will be related to finding more than one variable in this reduced set of variables. Unlike factor approaches, our strategy produces immediately interpretable and reproducible variables, since the reduced set is a stric subset of variables sampled and contained in the original space. Additionally, and unlike latent-based strategies like factor analysis, further studies would require to collect information only in the optimal subset.

Before describing in detail our empirical strategy, we must comment on some limitations. First, the cluster approach is surely not free from identification and interpretation issues. Cluster methods cannot guarantee in advance that the optimal number of groups is necessarily two, moreover, the methods do not guarantee that even if two groups are found, these are economically different. Second, even if two groups are found, this does not necessarily mean that one of them is the poor and the other one the 'non-poor'. For example, one group might consist in the 'extremely rich' with the complement group containing all other individuals. The next subsections describe in detail the clustering methods used in this paper, and how they are exploited to deal with the aforementioned problems. In particular, to guarantee that there are actually two separate groups (instead of only one group or more than two) and that one of them can be safely regarded as containing the poor. The second subsection describes the variable selection approach.

3.1 Clustering methods and the poor

The underlying idea behind our empirical strategy is to understand the poor as a coherent group that can be conceptually and practically distinguished from its complement, the 'non-poor'. Cluster methods seem relevent since, by definition solve a within/between simmilarity trade-off, that is, they try to assign observations to groups so they are close to those in the same group and distant to those in other groups. Even though classical clustering algorithms have long been available in practice, recent advances in data mining and computer intensive methods has driven considerable attention on such techniques, see Cherkassky and Mulier (2007) or Bishop (2006) for a recent overview.

There are several difficulties that must be sorted out for the case of finding the poor. First, our data is of a mixed nature, that is, it contains categorical (mostly binary) as well as continuous variates (income, for example). This impacts in the choice of an appropriate clustering technique, since these methods are sensitive to the choice of distances, standardizations and initial conditions. Second, as previously discussed, the final goal is to guarantee that the process finds two essentially different groups, one of them containing the poor.

Regarding the choice of a clustering method for our mixed data, we started by standardizing all variables. This is common practice in this literature, to avoid scale effects. Each variable is divided by its range, i.e. for the observation x_{ij} we consider y_{ij}, the standardized observation,

$$y_{ij} = \frac{x_{ij}}{\max_i (x_{ij}) - \min_i (x_{ij})}.$$

This procedure is applied to all the variables except to monthly household income, a continuous and highly positive skewed variable. For this case we use standardized based on its natural logarithm. Consequently, all standardized varaibles have range $[0, 1]$ except for the monthly household income that range between $[-1, 1]$.

The literature is not clear about specific clustering techniques for mixed data. There are mainly two approaches: hierarchical and partitioning. The main difference between them is that in the second case a partitioning rule is obtained, while hierarchical clusters do not strictly define groups. Lately, several heuristic algorithms have been proposed but none of them has achieve acceptance in the literature. K-means (MacQueen (1967)) is the

most well known and widely applied clustering procedure, that yields a partition of the original space. Some preliminary results on the asymptotic distribution of this clustering behavior are given by MacQueen (1967) and Hartigan (1978). Pollard (1979) established conditions that ensure the almost sure convergence of the cluster centers the sample size increases. In addition, it has a good performance on many real data examples.

The k-means algorithm is sensitive to the choice of an appropriate distance. We have chosen an additive measure that can handle mixed as well as continous variables. The L_1-norm is a natural choice for our type of data. The distance between two observations y_i and y_j is given by

$$d_{ij} = \sum_{l=1}^{p} |y_{il} - y_{jl}| .$$

so it can be seen as being the standard L_1-norm for continuous or ordinal variables, and in the case of binary variables, as the number of points where the observations take different values), that is, the same information as in the standard Jaccard index (see Hand et al. (2001)), one of the most well known measures of similarity for binary variables.

K-means procedures are often very sensitive to the choice of initial conditions, that is, to the position of the initial centroids used to start the algorithm. Several proposals have been made to handle this effect (see Steinly and Brusco (2007)). We have followed the recommendations in this last reference and considered ten random initializations and kept the one with minimum within-cluster sum of squares.

Regarding the number of clusters, most methods produce forced partitions on any data set, either there is an endogenous structure or not. Hence, in order to find the poor we are interested in two null hypotheses. The first one is the null that no grouping exists versus the alternative that there is more than one group. The second one is the null that only two groups are relevant, against the alternative that more than two groups are needed. We use the standard Calinski and Harabasz (1974) statistic, the most frequently used method to find the optimal number of clusters, even though it is not designed to distinguish between one and more than one cluster. We will also use the more modern *gap statistic* introduced by Tibshirani et al. (2001). The intuition behind this statistic is that within cluster similarity decreases as the number of groups increase. However, partitioning a group with already high similarity reduces the within cluster similarity less

than partitioning a heterogeneous group. Then, a sharp decrease will be observed at the optimal number of groups.

Finally, even when the previous process leads to two significantly different groups, there is not guarantee that one of them can be safely labeled as containing the poor, i.e., the relevant partition might cluster the extremely rich in one group. We will implement some confirmatory tests, based on multivariate version of the Komogorov-Smirnov test, to explore the nature of the implied partition and to what extend one of them contains the poor.

3.2 Dimensionality through variable-selection

After having found an appropriate clusterization that divides the popuplation into poor/non-poor groups, the problem of reducing the dimensionality of the original welfare space will be handled as a variable-selection one. The main advantage of this approach is that, by construction, the resulting variables are directly interpretable since they are originally in the welfare space.

In the last years, and driven by the increased popularity of data mining methods, several proposals have been made on this field. In the majority of the cases, the proposals relate a clustering technique, a rule to determine the number of clusters, and a procedure to select the variables. We adopt the recent strategy in Fraiman et al. (2008) based on a 'blinding' process that eliminates unnecessary variables. These authors show that the process has good empirical performance, specially as compared to alternative ones like Tadesse, et al. (2005) and Raftery and Dean (2006).

Fraiman et al.'s procedure selects relevant variables after a satisfactory clustering procedure has been implemented. Their approach is based on the idea of blinding unnecessary non-informative or redundant variables. We will discuss the main intuitive ideas behind the procedure, details are provided in the Appendix. For simplicity, suppose there are only two variabes in the original space, X and Y. Given an appropriate clusterization based on X solely, Y is redundant if a) it is strongly related to X, so given X it adds little information to the clusterization, b) it is independent of X and non-informative about any clusterization (it only adds 'noise'). In these cases, the clusterization remains relatiely unaltered if Y is replaced by its best prediction based on X, its conditional expectation $E(Y|X)$. In

the extreme versions of the previous cases, Y will be replaced by X (X strongly related to Y or by a constant (Y just adding noise). Consequently, the goal is to find the smallest group of original variables that can reproduce the original clusterization as accurately as possible, by replacing redundant variables by their expectations conditional on this reduced subset. The algorithm is detailed in the Appendix and in the original paper by Fraiman et al. (2008). The variable selection procedure is shown to be strongly consistent under mild regularity conditions on the partitioning method, and on the (nonparametric) estimation of the conditional expectations in the blinding process. Though intuitively simple, the method can be computationally extremely expensive. Fraiman et al. introduce a forward-backward algorithm in order to find a subset of variables with the desired properties.

4 Empirical results

4.1 Data

The main input for our analysis is a set of variables that covers most relevant dimensions of welfare. To this purpose, the Gallup World Poll, collected by the Gallup Organization, provides a convenient framework. The Poll is based on a consistent and homogeneous questionnaire implemented on national samples of adults from 132 countries, providing and exceptional chance to make cross country comparisons. The Gallup World Poll contains an ample spectrum of questions related to welfare, including self-reported measures of quality of life, opinions and perceptions. It also incorporates fundamental questions on demographics, education, and family income. Respondents are adults (15 years or older), selected randomly within the household. In spite of its potential, the Gallup Poll is still relatively unexplored for research purposes. Gasparini et al. (2009) and Gasparini and Gluzman (2009) provide a detailed account on its adequacy and compare it with standard household surveys. They conclude that in many comparable dimensions, the information contained in the Gallup Poll is a valuable and reliable source for welfare analysis.

Consequently, our initial data set consists in the 15 variables used by Gasparini et al. (2008) as their initial welfare space. They classify variables in three main groups.

1. *Monetary welfare:* income is the most widely used measure of welfare. We use the income measure in the Gallup survey, which consists in monthly household income

before taxes. As in Gasparini et al. (2008), since the original question is posed in terms of brackets of income, we take a random value in the corresponding range of the original question in local currency unit, and then translated this value to US dollars using country exchange rates adjusted by purchasing power parity.

2. *Non-monetary welfare:* these variables capture alternative access to goods and services that impact directly on welfare, but are not necessarily well captured by incone. We access to running water, electricity, landline telephone, television, computer, internet or mobile phone.

3. *Subjective welfare:* the recent literature (Ravallion and Lokshin (2002) is a leading example) has emphasized the importance of complementing standard measures with self perceived notions of well-being, finding significant differences between self-rated and objective measures of welfare concepts like poverty. We include questions on how individuals perceive themselves regarding welfare.

A complete list of variables with more detailed description, is provided in the Appendix.

4.2 The poor as a cluster

As described in the previous section, the first step consisted in finding the optimal number of clusters using the k-means algorithm. Table 1 presents results of the Calinsky/Harabasz index and the the relevant information for the Gap statistics. The Calinsky/Harabasz index decreases monotonically, achieving a maximum at two clusters. The procedure based on the Gap statistic suggests that the optimal number of clusters is two.

The previous result and the nature of the clustering algorithm suggest that there are two essentially different groups, at least under the metric used to define similarty in the clustering procedure. Nevertheless, as a robustness check, we have explored an alternative confirmatory route. We have implemented a multivariate variant of the non-parametric Kolmogorov-Smirnov (KS) test, developed by Cuesta-Albertos et al.'s (2006, see also Opazo et. al (2009) for a recent application), which can be applied to either functional or multivariate data. Roughly speaking, it is based on performing a one dimensional KS test for the projections of the data on randomly selected directions. We proceeded as suggested by Cuesta-Albertos et al. (2006), by selecting 50 random projections, computing

the KS statistic for every case, and taking the maximum of these values. The corresponding p-value was less than 0.001, meaning that the distributions of both groups induced by clusterization are significantly different.

The previous results point towards the existence of two *statistically* different groups, but it remains to explore whether one of them can be seen as containing the poor. Table 2 presents basic statistics that explore this issue. We use the three optimal factors obtained by Gasparini et al. (2009), interpreted by these authors as representing monetary, subjective and non-monetary aspects of welfare. We have computed means for the two groups obtained in the clustering process. Group one contains 73.5% and group two the remaining 26.48% of the individuals in our sample. Group two present substantially lower values for the three dimensions of welfare, suggesting that this group contains those individual with low levels of welfare. Consequently, we will refer to this group as the 'cluster poor', that it, we see group two as a statistically and economically different entity with respect to its complement, in the sense that it contains individual with significantly low levels of welfare.

4.3 Dimensionality via variable-selection

After having found an acceptable clusterization, we have proceeded to solve the dimensionality problem by finding a reduced set of variables initially in the welfare space, that can reproduce the initial clusterization. As stressed in the previous section, the Fraiman et. al's procedure is computationally very expensive, with required computer time growing exponentially with the sample size and the number of variables. In our case, it is unfeasible to perform the procedure with the complete data set (a standard computer needs more than 100 days to attain the optimal data set) then we implement a subsample based strategy. We considered ten random subsamples proportional to the clusters sizes containing 85% of the observations of the complete data set.

Remarkably, in all cases the variables selected are: *monthly household income*; *not having had enough money to buy food over the last year in at least three opportunities* and *having a computer at your home or the place you live*. The correct cluster reallocation rate is always between 90% and 92%, that is almost all individuals classified as poor with the initial set of 12 variables are correctly classified as poor based on this much smaller set of three variables.

The fact that the reduced welfare space needs more than one variable to adequately reproduce the original welfare space is an indication of its multidimensionality. Nevertheless, income turns out to be one of the variables chosen in the reduced set. This result is consistent with those in the literature, that suggest that, though important, income is not enough to capture welfare. As a matter of fact, when the reduced set of variables is forced to keep only income, only 60% of the observations are reallocated on the correct cluster.

It is interesting to compare the results of our multidimensional approach, with a standard one based on income solely. Table 3 uses standard one and two dollars a day lines to identify the poor. Out of those identified as poor by our cluster approach, only 45% are labelled as poor by a poverty line set at two dollars a day, and only 25% when the line is lowered to one dollar a day. Though monotonically increasing with income, this result speaks about the severe discrepancies between a multidimensional notion of poverty (as implicity in our cluster analysis) and that based on income solely.

Table 4 offers another perspective. It shows the proportion of individuals in each income decile that belong to the 'cluster poor' group. For example, 54% of those in the first income decile are classified as poor. This proportion decreases monotonically with income, to the point that only 3% of those in the 10-th decile are classified as poor by the cluster method. This result is relevant since it suggest that even though income plays a relevant role in the cluster based multivariate notion of poverty, the relationship is rather weak, specially in low levels of income. That is, even tough more income reduces monotonically the chances of falling in the poverty cluster, low income is not necessary neither sufficient to explain the multivariate version of the poverty status, to the point that, for example, 46% of the individuals in the lowest decile are not rendered as poor by the cluster approach. This result, again, is compatible with the large literature that points towards the inadequacy of income as a sole factor to identify the poor.

Table 5 explores similarities by country, that is, after implementing the procedure in the original data base, we have computed cluster and income poor groups. As expected, the relationship between the two classifications is positive but weak. The cases of Honduras and Guatemala are interesting. Honduras has the higher proportion of cluster based poor, even though in terms of income, it ranks relatively in the bottom. Exactly the opposite occurs in the case of Guatemala. Uruguay and Argentina are cases where the aggregate

figures match, for example, in the latter, the cluster poor is 21% compared to 22.9% based on income.

5 Conclusion

The fact that welfare is progressively accepted as an essentially multidimensional notion implies many conceptual an practical challenges, which usually suggest a trade-off related to the desired degree of aggregation. On one hand, and for pragmatic and conceptual reasons as well, it seems reasonable to attempt to summarize welfare in a few readily available indexes that can help monitor social performance as well as implement valid comparisons. On the other hand, the complex nature of well being points toward retaining as many factors as possible in order to fully characterize it. In this context, this paper suggests a simple procedure that 1) treats the poor as a coherent, clearly identifiable group that can be economically and statistically distnguished from its complement, 2) fully exploits available information to detect it, 3) summarizes the initial welfare space into a few unambiguously interpretable variables.

The empirical implementation based on the Gallup Poll suggest that three variables can reproduce quite accurately the role of the original 15 variables in the goal of identifying the poor. From a practical perspective, once this 'cluster poor' group of individuals is successfully identified using a large data set, further classification or evaluations can be implemented by just assessing the variables in the reduced set.

From a methodological perspective, the use of multivariate methods in Economics is scarce, which is surprising in light of the massive acceptance these techniques have in closely related areas. For this reason we have tried to stay as close as possible to standard grouping techniques, relegating more modern and sophisticated approaches (like CART methods as in Keeley and Tan (2008)) for further research.

References

Bishop, C M. (2006). *Pattern Recognition and Machine Learning*, Springer, New York.

Calinsky, R.B. and Harabasz, J. (1974), A Dendrine Method for Cluster Analysis, *Communications in Statistics*, 3, 1-27.

Cherkassky, V. and Mulier, F. M. (2007). *Learning from Data: Concepts, Theory and Methods*, 2nd Edition, Wiley, New York.

Cuesta-Albertos J.A., Fraiman R. and Ransford T. (2006), "Random projections and goodness-of-fit tests in infinite-dimensional spaces", *Bulletin of the Brazilian Mathematical Society*, **New Series 37**, **Nro. 4**, 477–501.

Elffers, H., Bethlehem, J. and Gill, R. (1978), Indeterminacy problems and the interpretation of factor analysis results, *Statistica Neerlandica*, 32, 4, 181199.

Ferro Luzzi, G., Fluckiger, Y. and Weber, S. (2008), A cluster analysis of multidimensional poverty in Switzerland, in Kakwani and Silber (2008).

Fraiman R., Justel A. and Svarc M. (2008) Selection of Variables for Cluster Analysis and Classification Rules. *Journal of American Statistical Association*, 103, 1294-1303.

Gasparini, L. and Gluzman, P. (2009), Estimating Income Poverty and Inequality from the Gallup World Poll: The Case of Latin America and the Caribbean, CEDLAS Working Paper 83.

Gasparini, L. Sosa Escudero, W., Marchionni, M. and Olivieri, S. (2009), 'Income, Deprivation, and Perceptions in Latin America and the Caribbean: New Evidence from the Gallup World Poll'. *mimeo*, CEDLAS/UNLP.

Hand, D. Mannila, H. and Smyth P. (2001), *Principles of Data Mining*. MIT Press, Cambridge.

Hardle, W. and Simar, L. (2003), *Applied Multivariate Statistical Analysis*, Springer, New York.

Hartigan, J. A. (1978). Asymptotic distributions for clustering criteria, *The Annals of Statistics*, Vol. 6, pp. 117 131.

Kakwani, N. and Silber, J. (2008), *Quantitative Approaches to Multidimensional Poverty Measurement*, Palgrave Macmillan, New York

Keely, L and Tan, C. (2008), Understanding preferences for income redistribution, *Journal*

of Public Economics, 2008, vol. 92, issue 5-6, pages 944-961.

MacQueen, J.B., (1967), "Some methods for classification and analysis of multivariate observations". In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 281–297.

Opazo, L. Raddatz, C. and Smuckler, S. (2009), "The long and the short of emerging market debt", Working paper World Bank.

Pollard, D. (1979). Strong Consistency of K-Means Clustering, *The Annals of Statistics*, Vol. 9, No. 1, pp. 135-140.

Raftery, A.E., and Dean, N. (2006), Variable selection for model-based clustering. *Journal of the American Statistical Association*, **101**, 168–178.

Ravallion, M. and Lokshin, M. (2002), "Self-rated economic welfare in Russia". *European Economic Review*, **46**, 11453-1473.

Sen, A. (1985), *Commodities and Capabilities*, Oxford University Press, Oxford.

Steinly, D. and Brusco, M. J. (2007), Initializing K-means Batch Clustering: A Critical evaluation of Several Techniques. *Journal of Classification*, 24, 99-121.

Tadesse M.G., Sha N. and Vannucci M. (2005), Bayesian variable selection in clustering high-dimensional data. *Journal of American Statistical Association*, **100**, 602–617.

Tibshirani R., Walther G. and Hastie T. (2001), Estimating the Number of Clusters in a Data Set via the Gap Statistic. *Journal of the Royal Statistical Society. Serie B (Statistical Metodology)*, 63, 2, 411-423.

Appendix 1: Variable description

We use the set of 15 variables in Gasparini et. al. (2008) for their factor analytic approach. There are three types of variables in these sample: subjective, non-monetary, monetary variables.

Subjective variables: the first three variables are integers from 0 to 10, and the remaining four are binary indicators, resulting from answers to the following questions.

1. Please imagine a ladder/mountain with steps numbered from zero at the bottom to ten at the top. Suppose we say that the top of the ladder/mountain represents the best possible life for you and the bottom of the ladder/mountain represents the worst possible life for you. If the top step is 10 and the bottom step is 0, on which step of the ladder/mountain do you feel you personally stand at the present time?

2. Please imagine a ladder/mountain with steps numbered from zero at the bottom to ten at the top. Suppose we say that the top of the ladder/mountain represents the best possible life for you and the bottom of the ladder/mountain represents the worst possible life for you. On which step of the ladder/mountain would you say you stood 5 years ago?

3. Please imagine a ladder/mountain with steps numbered from zero at the bottom to ten at the top. Suppose we say that the top of the ladder/mountain represents the best possible life for you and the bottom of the ladder/mountain represents the worst possible life for you. Just your best guess, on which step do you think you will stand on in the future, say 5 years from now?

4. Are you satisfied or dissatisfied with your standard of living, all the things you can buy and do?

5. Have there been times in the past twelve months when you did not have enough money to buy food that you or your family needed?

6. Have there been times in the past twelve months when you did not have enough money to provide adequate shelter or housing for you and your family?

7. Have there been times in the past twelve months when you or your family have gone hungry?

Non-monetary variables: all variables in this group are binary indicators that are answers to the following questions.

8. Does your home or the place you live have running water?
9. Does your home or the place you live have electricity?
10. Does your home or the place you live have a landline telephone in working order?
11. Does your home or the place you live have television?
12. Does your home or the place you live have a computer?
13. Does your home or the place you live have access to the Internet?
14. Do you, yourself, have a cellular/mobile phone, or not?

Monetary variable:

15. Household income per capita, that income is expressed in local currency and converted in US$ adjusted for PPP with the aim of comparing the purchasing power of each household

Appendix 2: The Fraiman et al. (2008) algorithm

Let $X = (X_1, \ldots, X_p)$ be a random vector with distribution $\mathcal{P}$ and consider any statistical procedure whose output is a partition of the space $\mathbb{R}^p$, for instance many non-hierarchical clustering technics and classification procedures. In many cases, if p is large, there are dependence among several variables of X or some of them are not relevant. Then, if the information of the noisy or dependent variables is removed, the cluster allocation should not change, meaning that the data should be kept in the original partition. It is important to notice that the partition is defined in the original p-dimensional space and the input requires data from all the variables.

They propose to look for a subset of indices $I \subset \{1, \ldots, p\}$ for which the original partition rule applied to a new "less informative" vector $Z^I \in R^p$ built up from X. The variables whose indices are in I are the same as those in X, i.e. $Z_i = X_i$ while the rest of the variables will be replace by the optimal predictor based on X[I], the conditional expectation, $Z_i = E(X_i|X[I])$ (we denote X[I] the set of variables whose indices are in I), this is the blinding procedure. It is important to notice that in the case of noisy variables $E(X_i) = E(X_i|X[I])$. For a fixed integer $d < p$, the population target is the set $I \subset \{1, \ldots, p\}$, $\#I =$ d, for which the *population objective function*, given by

$$h(I) = \sum_{k=1}^{K} P\left(f(X) = k, f(Y^I) = k\right),$$

attains its maximum. The empirical version of this procedure follows these steps.

1. Given $X_1, \ldots, X_n \in \mathbb{R}^p$ i.i.d data, we consider the partitioning procedure $f_n : \mathbb{R}^p \to \{1, \ldots, K\}$.

2. For a fixed value of $d < p$, given a subset of indices $I \subset \{1, \ldots, p\}$, with $\#I = d$, fix an integer value r (the number of nearest neighbor to be used). For each $j = 1, \ldots, n$, find the set of indices C_j of the r-nearest neighbor's of $X_j[I]$ among $\{X_1[I], \ldots, X_n[I]\}$, where $X_j[I] = \{X_j[i] : i \in I\}$. And define,

$$Z_j^* [i] = \begin{cases} X_j [i] & \text{if } i \in I \\ \frac{1}{r} \sum_{m \in C_j} X_m [i] & \text{otherwise} \end{cases}$$

   where $X [i]$ stands for the i–coordinate of the vector X.

3. Calculate the empirical objective function

$$h_n(I) = \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n} \mathcal{I}_{\{f_n(X_j)=k\}} \mathcal{I}_{\{f_n(X_j^*)=k\}},$$

   where $\mathcal{I}_A$ stands for the indicator function of set A. The empirical objective function measures the proportion of observations that are reallocated on the same group as in the original partition after blinding the variables.

4. Look for a subset $I_{d,n} =: I_n$, with $\#I_n = d$, that maximizes the empirical objective function h_n.

In general, the goal is to find the smallest subset of variables that achieves a certain reallocation rate. The variable selection procedure is strong consistent under mild regular conditions on the partitioning method and on the nonparametric estimation of the conditional expectation.

Remark 1. As performing an exhaustive search can be computationally very expensive or even impossible Fraiman et al. introduce a forward-backward algorithm in order to find a subset of variables with the desired properties.

Remark 2. Estimating the conditional expectation can be very expensive computationally, therefor if one is interested in identifying only the noisy variables the mean could be found instead, in this way the procedure becomes much more faster.

Table 1: Optimal number of groups

Clusters (k)	CH Index	Gap(k)	S_k	Gap($k+1$) - S_{k+1}
1		0.368760	0.004639	0.553242
2	2087.42	0.586392	0.033150	0.381457
3	1441.01	0.391880	0.010423	
4	1570.45			
5	1305.30			
6	1394.86			
7	1283.33			

CH stands for the Calinski/Harabasz index. The Gap procedure chooses the optimal number of clusters (k) by finding the smallest k such that $Gap(k) \geq Gap(k+1) - S_{k+1}$.

Table 2: The poor as a cluster

Cluster	Monetary welfare	Subjective welfare	Non-Monetary welfare	Frecuency
1	232.5077	0.5940948	0.0147273	73.52%
2	96.1865	−1.900604	−0.4567784	26.48%

Columns 1 to 3 compute cluster means for the factors obtained in Gasparini et al. (2010). The last column computes the proportion of observations in each cluster.

Table 3: Intersection of Poverty Lines

International Poverty Line	Identification
Daily 2 USD	45
Daily 1 USD	25

Percentage of cluster poor individuals that are also classified by income poor

Table 4: Income and Cluster Poor

Decile	Cluster 1	Cluster 2
1	46%	54%
2	54%	46%
3	63%	37%
4	70%	30%
5	75%	25%
6	77%	23%
7	79%	21%
8	84%	16%
9	89%	11%
10	97%	3%

Percentage of individuals in each income decile, classified as cluster non-poor (cluster 1) and poor (cluster 2).

Table 5: Country Comparisson

Country	Cluster Poverty	Income Poverty
Honduras	47.89%	23.00%
Peru	44.50%	57.80%
Nicaragua	41.88%	59.50%
Paraguay	39.65%	54.9%
Bolivia	38.16%	58.80%
El Salvador	34.29%	60.50%
Ecuador	29.03%	45.80%
Uruguay	28.71%	25.60%
Chile	27.58%	22.00%
Panama	24.64%	32.60%
Colombia	21.77%	35.84%
Argentina	21.10%	22.90%
Costa Rica	20.97%	25.40%
Guatemala	16.89%	50.30%