Multidimensional Middle Class

María Edo *, Walter Sosa-Escudero * [1], and Marcela Svarc**

* *Departamento de Economía, Universidad de San Andrés and CONICET, Argentina.*

** *Departamento de Matemática y Ciencias, Universidad de San Andrés and CONICET, Argentina*

Abstract

Middle class studies have gained relevance in the economic literature. Nevertheless, a profound lack of agreement on conceptual and methodological issues for its identification remains. Furthermore, it has mostly relied on only one dimension: income. In this paper we present a new multidimensional approach for identifying the middle class based on multivariate quantiles. We provide an empirical application for the case of Argentina in the 2004-2014 period, characterizing its performance and main features.

Keywords: Argentina, distribution, feature extraction, multivariate quantiles, middle class.
JEL subject classification. Primary: D3; *secondary:* I3, D6.

1 Introduction

The study of the middle class has been traditionally part of the sociological realm. Going back at least as far as Max Weber [27] sociologists have dedicated a relevant space to the course of this group across changes in societies, generally defining it in terms of labor-market stratification, associated to human capital accumulation as well as general views, values and lifestyle.

Though still lagging behind, the economic literature has recently devoted increasing importance to the middle clase (Atkinson and Brandolini, [1],

[1]Corresponding author: Walter Sosa-Escudero, Departamento de Economía, Universidad de San Andrés, Vito Dumas 248, Victoria, Argentina. Email: wsosa@udesa.edu.ar

Ravallion, [23], Birdsall et al. [7]). This renewed attention derives fundamentally from the key role assigned to the middle class in contemporaneous societies. Indeed, much is expected from this group. Some authors go as far as claiming that they represent the foundation on which democracy and market economy may flourish (Birdsall et al., [7]). Others point to its capacity in terms of diminishing potential sources of conflict and polarization (Gigliarano and Muliere, [18]), as well as their central role in motorizing the economy through entrepreneurship and consumption (Banerjee and Duflo, [2]). The issue grew particular attention during the 1980s and early 1990s associated to the so-called middle-class decline that was claimed to occur in the US and other developed countries.

In spite of this renewed interest, empirical studies within the economic literature are still scarce, particularly when compared to other relevant socioeconomic groups such as the poor or even the upper class in recent years. This is likely related to the delicate conceptual and methodological difficulties embedded in its definition. In fact, the literature shows no clear consensus regarding how to identify and quantify the middle class. The problem mirrors the obstacles faced by the abundant literature on poverty measurement. Nevertheless, the definition of the middle class poses further challenges: even if it is possible to agree to on a lower threshold that separates the poor from the rest of the population (akin to the widespread use of lines in poverty analysis), agreement on how to set an upper bound that separates the middle class from the rich is far less obvious.

Conceptual and methodological concerns in defining the middle class as well as the poor- are not trivial. Different definitions may lead not only to differences in the level of wellbeing of that group at a certain point in time but may also to differences in the assessment of its evolution (see Edo and Sosa-Escudero, 2012, [13]).

In light of these difficulties, and mirroring the path followed by poverty and inequality measurment, pioneering quantitative studies on the middle class have generally favored a unidimensional, income based approach. Thus, the middle class has been identified as the group lying between a lower and an upper bound defined solely in terms of income.

This limits the analysis in at least two ways. On the one hand, even though certain agreement may be reached on establishing a lower bound in terms of income, it is far less clear that this should be the case for the upper limit. On the other hand, this implies disavowing the large and rich literature coming from the sociological and political theory realms that point to other dimensions as key in defining the middle class: the occupational structure, the level of education, wealth, etc.

Several authors claim to study the middle-class multidimensionally (see for instance Davis and Huston [10] and Gayo [17]) but they define the middle class in terms of the income and use several dimensions for the subsequent analysis. To the best of our knowledge the only attempt to define and describe the middle class multidimensionally has been done by Gigliarano and Mosler [19], they consider two different approaches. On the one hand, they define the middle class as a convex central region, typically a ball with center in the multidimensional mean of the attributes and a varying radius determining a region containing a given proportion (for instance, 50%) of the observations. On the other hand, the middle class is defined as the ellipsoid that covers at least a given proportion of population and has minimum volume among all such ellipsoids. In both cases they compare the evolution of the middle class analyzing the dispersion of the central regions defined. It is clear that the former approach will only be suitable if the variables are spherical and, as it is well, known such an assumption is likely to fail since incomes are asymmetrically distributed. This approach cannot guarantee to identify a subset in the central region of the distribution, an even when it does it will tend to capture the most dense region, and there is no reason to assume that it will contain the central observations.

Besides these drawbacks there are other important questions that remain without answer. A relevant one refers to the true dimensionality of the middle class, that is, after recongizing the multidimensional nature of well being, a relevant concern is how many welfare dimensions are appropriate to characterize the middle class and whether these dimensions can be appropriately summarized by observable variables.

The present paper contributes to the literature in several aspects. First,

it presents a new multidimensional approach to measure welfare through the construction of multivariate quantiles based on a growth direction of increasing wellbeing, based on principal components analysis. In particular, the growth direction is derived from the module of the first principal component. Thus, a truly multidimensional welfare index is produced. Second, the paper suggests a a new approach to reduce the dimensionality of welfare. A novel genetic algorithm is implemented to select variables from the original space ensuring that the resulting projection is similar enough to the one produced with the whole set of variables. This approach based on multivariate quantiles determined by a growth direction of increasing wellbeing allows for a truly multidimensional identification of the middle class. Moreover, we are able to identify how many and which are the dimensions relevant to define the middle class and distinguish this group from the poor and the upper class.

We apply the new approach to Argentina for the 2004-2014 period, a country that has experienced significant changes in its income distribution, providing relevant sampling variability to assess the middle class and its changes.

The rest of the article is organized in the following way: Section 2 describes the theoretical and empirical approach based on the α-quantile region definition orientated by a growth direction. In Section 3 numerical aspects are considered. Section 4 focuses on the empirical application, characterizing the middle class for Argentina under the 2004-2014 period while Section 5 concludes.

2 Middle class and multivariate quantiles

This section extends the univariate concept of α-quantile to the the multivariate setting. The middle class will be defined as the subset of observations within a *lower* bound that separates the poor from the middle class, and an *upper* bound that separates it from the rich, defined in terms of multivariate notion of quantiles.

We seek to define multivariate quantiles with two basic properties. On one

hand, we define the middle class as a given proportion of central population. Hence, the multivariate α-region, $C(\alpha)$ must have mass greater than or equal to α, i.e. $P(X \in C(\alpha)) \geq \alpha$. On the other hand, since our variables measure wellbeing, each of them has a natural increasing order, this order must be preserved by the definition stated, implying that the quantile function defined will not be equivariant.

Even though the concept of a multivariate quantile has been largely studied in the literature (see for instance, Chauduri [8], Serfling [24], Hallin et al. [20], Fraiman and Pateiro-Lopez [15] and Kong and Mizera [22]), none of these definitions are suitable for our analysis. There are two main drawbacks. First of all, quantile functions on $\mathbb{R}^p$ are desirably equivariant, that is the new quantile representation of a point x after affine transformation should agree with the original representation similarly transformed. Secondly, there are many definitions of multivariate α-quantile, most of them define α-quantiles orientated by a given direction, hence considering all the unitary directions a region in the space is determined, however there is no relation between the probability of these regions and the directional α-quantiles.

A proper definition that satifies the goals of identifying the middle class is the subject ot the next subsection.

2.1 The theoretical approach

Let X be a $p-$dimensional random vector with distribution P_X, representing aspects of social and economic wellbeing. The goal is to extend the univariate concept of α-quantile to the the multivariate setting.

As mentioned in the Introduction a first goal is to determine the $\alpha-$*upper* region of the distribution. A basic monotonicity assumption is that each random variable in the multidmensional welfare space is defined so as a natural increasing order can be assumed, that is higher levels of each of them correspond to increasing levels of wellbeing. The proposal is to project the data into the direction of g_D, which denotes the *growth direction*. To attain uniqueness this direction must have unitary norm and it should be positive coordinate wise. If there is no previous knowledge of the distribution, a

natural growth direction could be $g_D = (1, \ldots, 1)/\sqrt{p}$, which represents the mean of the welfare dimensions. In other cases different variables may have different weights and they could be determined for instance by the first principal component or the absolute value of the first principal component. Let $\mathbb{B} = \{X \in \mathbb{R}^p : \|X\| = 1\}$, then $g_D \in \mathbb{B}$.

Then we denote $Y_D = \langle X, g_D \rangle$, the projection of X respect to g_D. Following Fraiman and Pateiro-Lopez [15], let

$$\tilde{Q}(\alpha, g_D) = \inf_{t \in \mathbb{R}} \left\{ F_{\langle X - E(X), g_D \rangle}(t) \geq \alpha \right\}, \tag{1}$$

where

$$F_{\langle X - E(X), g_D \rangle}(t) = P(\langle X - E(X), g_D \rangle \leq t), \tag{2}$$

then the $\alpha-$quantile in the direction of g_D is given by,

$$Q(\alpha, g_D) = \tilde{Q}(\alpha, g_D) g_D + E(X). \tag{3}$$

Then we define the $\alpha-$quantile region as

$$C(\alpha, g_D) = \left\{ x \in \mathbb{R}^p : \ \langle x - E(X), g_D \rangle \leq \tilde{Q}(\alpha, g_D) \right\}. \tag{4}$$

It is clear that the $\alpha-$quantile region is bounded by the hyperplane orthogonal to g_D and that contains the point $\tilde{Q}(\alpha, g_D) g_D + E(X)$. Without loss of generality assume that $E(X) = 0$.. Proper coverage proability of the proposed definition is guaranteed by the folowing Lemma.

Lemma 1. $P(X \in C(\alpha, g_D)) \geq \alpha$.

Proof.

$$\begin{aligned}
P\left(X \in C(\alpha, g_D)\right) &= P\left(X \in \mathbb{R}^p : \langle X, g_D \rangle \leq \tilde{Q}(\alpha, g_D)\right) \\
&= P\left(\langle X, g_D \rangle \leq \inf_{t \in \mathbb{R}} \left\{ F_{\langle X, g_D \rangle}(t) \geq \alpha \right\}\right) \\
&= P\left(Y_D \leq \inf_{t \in \mathbb{R}} \left\{ F_{Y_D}(t) \geq \alpha \right\}\right) \\
&= F_{Y_D}\left(\inf_{t \in \mathbb{R}} \left\{ F_{Y_D}(t) \geq \alpha \right\}\right).
\end{aligned}$$

For every $t \in \mathbb{R}$ such that $F_{Y_D}(t) \geq \alpha$ if and only if $t \geq F_{Y_D}^{-1}(\alpha)$, if and only if $F_{Y_D}(t) \geq F_{Y_D}\left(F_{Y_D}^{-1}(\alpha)\right)$. Then $z = \inf_{t\in\mathbb{R}}\left\{F_{\langle X,g_D\rangle}(t) \geq \alpha\right\}$ if and only if $F_{Y_D}(z) \geq \alpha$. □

2.2 An empirical model

Let $X_1, \ldots, X_n$ be a random sample of vectors with distribution P_X and denote by P_n its empirical distribution. In order to define the empirical counterpart of the $\alpha-$quantile on the direction of g_D, we first need to define the empirical expression for (1)

$$\tilde{Q}_n(\alpha, g_D) = \inf_{t\in\mathbb{R}}\left\{F_{n,\langle X-\overline{X},g_D\rangle}(t) \geq \alpha\right\}, \tag{5}$$

where,

$$F_{n,\langle X-\overline{X},g_D\rangle}(t) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_{\left\{\langle X-\overline{X},g_D\rangle\leq t\right\}}. \tag{6}$$

Then the empirical expression for (3) is

$$\widehat{Q}_n(\alpha, g_D) = \tilde{Q}_n(\alpha, g_D)g_D + \overline{X}. \tag{7}$$

The empirical counterpart for the $\alpha-$ quantile region is

$$C_n(\alpha, g_D) = \left\{x \in \mathbb{R}^p,\ \langle x - \overline{X}, g_D\rangle \leq \tilde{Q}_n(\alpha, g_D)\right\}. \tag{8}$$

Remark 1. *If g_D is given by the first principal component, then in equations (5), (7) and (8) we should consider the empirical first principal component $g_{n,D}$. It is well known that under mild regular conditions $g_{n,D}$ converges almost surely to g_D. See Dauxois et al [11], Propositions 2 and 4. They establish that it is enough to show the convergence of the covariance matrix in the operator space norm. More specifically, let Σ be the covariance matrix of X and g_k the kth eigenvector, and Σ_n and $g_{n,k}$ the corresponding estimates. Then they prove that if*

$$\sup_{\|u\|=1}\|\left(\Sigma_n - \Sigma\right)(u)\| \to_{n\to\infty} = 0 \ a.s., \tag{9}$$

then

$$g_{n,k} \to g_k \ a.s. \tag{10}$$

The main goal is to establish the almost surely consistency of $C_n(\alpha, g_{n,D})$ to $C_(\alpha, g_D)$ under mild regular conditions. To attain this result some statements must be proved in advance.

Lemma 2. *Let $X_1, \ldots, X_n$ be a sample random of vectors in $\mathbb{R}^p$ with absolute continuous distribution and empirical distribution P_n. Let $g_{n,D}, g_D$ be unitary vectors in $\mathbb{R}^p$, such that $g_{n,D} \to g_D$ a.s. Then,*

$$\lim_{n\to\infty} \|F_{n,g_{n,D}} - F_{g_D}\|_\infty = \lim_{n\to\infty} \sup_{t\in\mathbb{R}} |F_{n,g_{n,D}}(t) - F_{g_D}(t)| = 0 \ a.s. \tag{11}$$

Where

$$F_{n,g_{n,D}}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}_{\left\{\langle X_i - \overline{X}, g_{n,D}\rangle \leq t\right\}}$$

and

$$F_{g_D}(t) = F_{\langle X - E(X), g_D\rangle}(t) = P\left(\langle X - E(X), g_D\rangle \leq t\right).$$

Proof. The proof follows the same ideas as in Fraiman and Paterio [15], Proposition 1 and Corollary 1. Given a probability measure in P and a function $f : \mathbb{R}^p \to \mathbb{R}$, denote by Pf the expected value of f under P, i.e. $Pf = \int f(x)dP(x)$. Let $f : \mathbb{R}^p \to \mathbb{R} : f(x) = \mathcal{I}_{\{\langle g_D, x\rangle \leq t\}}, t \in \mathbb{R}$. Then

$$\lim_{n\to\infty} \|F_{n,g_{n,D}} - F_{g_D}\|_\infty = \lim_{n\to\infty} |P_n f - Pf| \text{ a.s. },$$

where P_n is the empirical measure of $Z_{ni} = \langle g_{n,D}, X_i - \overline{X}\rangle,\ i = 1, \ldots, n$.

P_n converges weakly to P if for every $f \in C(\mathbb{R}^p)$, where $C(\mathbb{R}^p)$ is the set of all the continuous and bounded functions on $\mathbb{R}^p$, $\lim_{n\to\infty} P_n f = Pf$.

Let $f \in C(\mathbb{R}^p)$ and P_n^* be the empirical measure of $\langle g_D, X_i - E(X)\rangle$, for $i = 1, \ldots, n$. Then,

$$|P_n f - Pf| \leq |P_n f - P_n^* f| + |P_n^* f - Pf|.$$

If $P|f| < \infty$, P_n^* converges weakly to P. On the other hand, due to the continuity of f, the law of large numbers, the almost sure convergence of $g_{n,D}$ to g_D, we have that $|P_n f - P_n^* f|$ goes to zero a.s. The according to Billingsley- Topsøe [6] we have that (2) holds if and only if

$$\lim_{\epsilon \to 0} \left\{ x \in \mathbb{R}^p : \sup_{x_1, x_2 \in B(x,\epsilon)} \left| \mathcal{I}_{\langle g_D, x_1 \rangle \leq t} - \mathcal{I}_{\langle g_D, x_2 \rangle \leq t} \right| > \delta \right\} = 0,$$

where $B(x, \epsilon)$ is a p dimensional ball of center x and radius ϵ. Fraiman and Pateiro [15], in Corollary 1, proved that this statement holds. □

Lemma 3. *Let* $\alpha \in (0, 1)$,

(i) given $g_D \in B(0, 1)$, $F_{g_D} > \alpha$, *for all* $t > \tilde{Q}(\alpha, g_D)$.

(ii) $\lim_{n\to\infty} \|F_{g_{n,D}} - F_{g_D}\|_\infty = 0$ *a.s.*

Then $\lim_{n\to\infty} \tilde{Q}_n(\alpha, g_{n,D}) = \tilde{Q}(\alpha, g_D)$ *a.s.*

Proof. From assumption (i) is clear that $F_{g_D}(\tilde{Q}(\alpha, g_D) - \epsilon) < \alpha - \delta_0$ and $F_{g_D}(\tilde{Q}(\alpha, g_D) + \epsilon) > \alpha - \delta_0$. From assumption(ii), with probability 1 and for all $\delta > 0$, there exists n_0 such that for every $n \geq n_0$,

$$\left| F_{n,g_{n,D}}(\tilde{Q}_n(\alpha, g_{n,D})) - F_{g_D}(\tilde{Q}_n(\alpha, g_{n,D})) \right| < \delta.$$

Then $F_{g_D}(\tilde{Q}_n(\alpha, g_{n,D}) > \alpha - \delta$ and $F_{g_D}(\tilde{Q}_n(\alpha, g_{n,D}) < \alpha + \delta$. Thus, $\tilde{Q}(\alpha, g_{n,D}) > \tilde{Q}(\alpha, g_D) - \epsilon$. In an analogue way, it is clear that $\tilde{Q}(\alpha, g_{n,D}) < \tilde{Q}(\alpha, g_D) + \epsilon$. Then, $\tilde{Q}(\alpha, g_{n,D}) \to_{n\to\infty} \tilde{Q}(\alpha, g_D)$ a.s. □

Theorem 1. *Under the same conditions of Lemma 1, we have that*

$$\tilde{Q}_n(\alpha, g_{n,D}) \to_{n\to\infty} \tilde{Q}(\alpha, g_D) \text{ a.s.}$$

In addition, let $x \in \mathbb{R}^p$ *and denote*

$$A(x) =: < x - E(X), g_D > -\widetilde{Q}(\alpha, g_D)$$

and

$$A_n(x) =: < x - \overline{X}, g_{n,D} > -\widetilde{Q}_n(\alpha, g_{n,D}).$$

It is clear that

$$A_n(x) \to_{n\to\infty} A(x) \text{ a.s.} \tag{12}$$

Proof. It follows straight forward from Lemma 3, the almost sure convergence of $g_{n,D}$ to g_D and the Law of Large Numbers. □

Let K_1 and K_2 be compact sets in $\mathbb{R}^p$, the Hausdorff distance between K_1 and K_2 is given by

$$\rho(K_1, K_2) = \inf \{\epsilon | K_1 \subseteq K_2 + \epsilon, K_2 \subseteq K_1 + \epsilon\},$$

where $K + \epsilon = \{x | d(x, K) < \epsilon\}$.

The next theorem is the key result of this section, which establishes strong consistency of the empirical counterpart for the proposed $\alpha-$quantile region.

Theorem 2. *Under the same conditions of Lemma 3, let K be a compact set in $\mathbb{R}^p$, and denote*

$$\widetilde{C}^K(\alpha, g_D) = C(\alpha, g_D) \cap K$$

and

$$\widetilde{C}_n^K(\alpha, g_{n,D}) = C_n(\alpha, g_{n,D}) \cap K.$$

Then, $\rho(\widetilde{C}_n^K(\alpha, g_{n,D}), \widetilde{C}^K(\alpha, g_D)) \longrightarrow 0$ a.s.

Proof. Let $k_0 = \max_{x \in K} \|x\|$, then $K \subseteq B[0, k_0]$, ($B[c, r]$ where denotes the $p-$ dimensional closed ball of center c and radius r). Then,

$$\begin{aligned} \epsilon_n &= \rho(\widetilde{C}_n^K(\alpha, g_{n,D}), \widetilde{C}^K(\alpha, g_D)) \\ &\leq \rho(\widetilde{C}_n^{B[0,k_0]}(\alpha, g_{n,D}), \widetilde{C}^{B[0,k_0]}(\alpha, g_D)) \\ &\leq max_{\left\{x \in (C_n(\alpha, g_{n,D}) \cup C(\alpha, g_D)) \cap \partial B[0, 2k_0]\right\}} \{|A_n(x) - A(x)|\}. \end{aligned}$$

Since $(C_n(\alpha, g_{n,D}) \cup C(\alpha, g_D)) \cap \partial B[0, 2k_0]$ is compact, it attains a maximum, then from Theorem 12 it is clear that ϵ_n goes to zero a.s.

Hence, $C_n^K(\alpha, g_{n,D}) \subseteq C^K(\alpha, g_D) + \epsilon_n$, then

$$\limsup C_n^K(\alpha, g_{n,D}) \subseteq \cap_{n \geq n_0} C^K(\alpha, g_D) + \epsilon_n = C^K(\alpha, g_D) \text{ a.s.}$$

And also $C^K(\alpha, g_D) \subseteq C_n^K(\alpha, g_{n,D}) + \epsilon_n$, then

$$C^K(\alpha, g_D) \subseteq \liminf C_n^K(\alpha, g_{n,D}) + \epsilon_n = \liminf C_n^K(\alpha, g_{n,D}) \text{ a.s.}$$

Then $C_n^K(\alpha, g_{n,D}) \to C^K(\alpha, g_D)$ a.s.

□

2.3 Variable selection for multidimensional quantiles

Section 2.1 defines a multivariate quantile function for any arbitrary multivariate notion of welfare. An important question is whether all initial variables are equally important, since it might be the case that some variables only add noise or can be appropiately captured by other variables. To answer these questions we develop an *ad hoc* variable selection criterion, based on the *blinding* strategy introduced by Fraiman et al. [14].

We present the problem of variable selection in terms of the underlying distribution and then we apply the solution to the sample data using the empirical distribution in a plug-in way.

Let $\mathbf{X} \sim P \in \mathcal{P}_0$, be a random vector in $\mathbb{R}^p$, where $\mathcal{P}_0$ represents a subset of probability distributions on $\mathbb{R}^p$. The coordinates of the vector $\mathbf{X}$ are denoted $X[i], \quad i = 1, \ldots, p$.

Given a subset of indices $I \subset \{1, \ldots, p\}$ with cardinality $d \leq p$, we call $\mathbf{X}(I)$ the subset of random variables $\{X[i], i \in I\}$. With a slight abuse of notation, if $I = \{i_1 < \ldots < i_d\}$, we also denote the vector $(X[i_1], \ldots, X[i_d])$ as $\mathbf{X}(I)$, and define the *blinded* vector $\mathbf{Z}(I) := \mathbf{Z} = (Z[1], \ldots, Z[p])$, where

$$Z(I)[i] = \begin{cases} X[i] & \text{if } i \in I \\ E(X[i]|\mathbf{X}(I)) & \text{if } i \notin I. \end{cases} \tag{13}$$

$\mathbf{Z}(I) \in \mathbb{R}^p$, but it depends only on $\{X[i], i \in I\}$ variables. The distribution of $\mathbf{Z}(I)$ is denoted $Q(I)$. Finally, $\eta^i(z) = E(X[i]|\mathbf{X}(I) = z)$ for $i \notin I$ represents the regression function.

Suppose that we are satisfied with the multidimensional quantile function stated in the previous section. The goal is to find a minimal subset of variables from X that retains almost all the relevant information from the quantile function. That is, we seek to find the subset of variables, $I \in \{1, \ldots, p\}$, of cardinality q, $q < p$ that best explains the multidimensional quantile function stated in equation (14). Typically we are interested in the case where $d << p$. Given a fixed integer d, $1 \leq d << p$, we let $\mathcal{I}_d$ be the family of all subsets of $\{1, \ldots, p\}$ with cardinality d.

We seek a small subset, I, such that equation (14) is as close as possible

to

$$F_{\langle Z(I)-E(Z(I)),g_D\rangle}(t) = P(\langle Z(I) - E(Z(I)), g_D\rangle \leq t), \tag{14}$$

and $E(Z(I)) = E(X)$, since

$$E(Z(I)[i] = \begin{cases} E(X[i]) & \text{if } i \in I \\ E(E(X[i]|\mathbf{X}(I))) = E(X[i]) & \text{if } i \notin I. \end{cases} \tag{15}$$

Let

$$h(I) = \|F_{\langle X-E(X),g_D\rangle} - F_{\langle Z(I)-E(X),g_D\rangle}\|_\infty. \tag{16}$$

More precisely, $\mathcal{I}_0 \subset \mathcal{I}_d$ is defined as the family of subsets in which the minimum $h(I)$ is attained for $I \in \mathcal{I}_d$, i.e.,

$$\mathcal{I}_0 = argmin_{I\in\mathcal{I}_d} h(I). \tag{17}$$

We define the empirical version for our model. We require consistent estimates of the set I_0, $I_0 \subseteq I_d$ based on a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of iid random vectors, with a distribution $\mathcal{P}$.

Given a subset $I \in \mathcal{I}_d$, the first step is to obtain the blinded version of the sample of random vectors in $\mathbb{R}^p$, $\hat{\mathbf{X}}_1(I), \ldots, \hat{\mathbf{X}}_n(I)$, that only depend on $\mathbf{X}(I)$, estimating the conditional expectation (the regression function) non-parametrically.

For all $i \notin I$, $\hat{\eta}^i(z)$ is a uniform strongly consistent estimate of $\eta^i(z) = E(X[i]|\mathbf{X}(I) = z)$ for almost all z $(\mathcal{P})$. Conditions under which this holds can be found in Hansen [21].

First, we define the empirical version of the *blinded* observations. As an example, we consider the $r-$-nearest neighbour (r-NN) estimates. We fix an integer value r (the number of nearest neighbours used) and calculate the Euclidean distance restricted to the coordinates I among the observations $\mathbf{X}_1(I), \ldots, \mathbf{X}_n(I)$. For each $j \in \{1, \ldots, n\}$, we found the set of indices C_j of the r nearest neighbours of $\mathbf{X}_j(I)$ among $\{\mathbf{X}_1(I), \ldots, \mathbf{X}_n(I)\}$, where $\mathbf{X}_j(I) = \{\mathbf{X}_j[i], i \in I\}$.

Next we define the random vectors $\hat{\mathbf{X}}_j(I), 1 \leq j \leq n$ satisfying

$$\hat{X}_j(I)[i] = \begin{cases} X_j[i] & \text{if } i \in I \\ \frac{1}{r}\sum_{m\in C_j} X_m[i] & \text{otherwise,} \end{cases} \tag{18}$$

where $X_j[i]$ stands for the ith-coordinate of the vector $\mathbf{X}_j$.

$Q_n(I)$ stands for the empirical distribution of $\{\hat{\mathbf{X}}_j(I), 1 \leq j \leq n\}$. Given a subset of indices $I \in \mathcal{I}_d$, we define the empirical version of the objective function

$$h_n(I) = \|F_{n,\langle X-\overline{X},g_{n,D}\rangle} - F_{n,\langle \widehat{X}(I)-\overline{X},g_{n,D}\rangle}\|_\infty, \tag{19}$$

where

$$F_{n,\langle \widehat{X}(I)-\overline{X},g_{n,D}\rangle}(t) = \frac{1}{n}\sum_{j=1}^{n} \mathcal{I}_{\{\langle \widehat{X}_j(I)-\overline{X},g_{n,D}\rangle \leq t\}}. \tag{20}$$

Our aim is to find the optimal subsets of variables $I_0 \subset I_d$, which are the family of subsets in which the minimum of $h_n(I)$ is reached, i.e.,

$$\hat{\mathcal{I}}_n = argmin_{I\in\mathcal{I}_d} h_n(I). \tag{21}$$

Then the following strong consistency theorem can be stated.

Theorem 3. *Let $\{\mathbf{X}_j, j \geq 1\}$ be iid p dimensional random vectors. Given $d, 1 \leq d \leq p$, let I_d be the family of all the subsets of $\{1, \ldots, p\}$ with cardinality d and let $I_{d,0} \subset I_d$ be the family of subsets in which the minimum of equation (16) is reached. Then, under* **H1**, **H2** *we have that $\hat{I}_n \in \mathcal{I}_0$ eventually almost surely, i.e. $\hat{I}_n = I_0$ with $I_0 \in \mathcal{I}_0$ $\forall n > n_0(\omega)$, with probability one.*

Proof. In order to prove our result it is enough to show that for each fixed subset I the empirical objective function (21) converges almost surely to the theoretical objective function (16), which will hold if

$$h_n(I) \to h(I) \text{ a.s.}$$

It is clear that,

$$\begin{aligned} &|h_n(I) - h(I)| \\ &= \left| \|F_{n,\langle X-\overline{X},g_{n,D}\rangle} - F_{n,\langle \widehat{X}(I)-\overline{X},g_{n,D}\rangle}\|_\infty - \|F_{\langle X-E(X),g_D\rangle} - F_{\langle Z(I)-E(X),g_D\rangle}\|_\infty \right| \\ &\leq \left\| F_{n,\langle X-\overline{X},g_{n,D}\rangle} - F_{n,\langle \widehat{X}(I)-\overline{X},g_{n,D}\rangle} - F_{\langle X-E(X),g_D\rangle} + F_{\langle Z(I)-E(X),g_D\rangle} \right\|_\infty \\ &\leq \left\| F_{n,\langle X-\overline{X},g_{n,D}\rangle} - F_{\langle X-E(X),g_D\rangle} \right\|_\infty + \left\| F_{n,\langle \widehat{X}(I)-\overline{X},g_{n,D}\rangle} - F_{\langle Z(I)-E(X),g_D\rangle} \right\|_\infty \end{aligned} \tag{22}$$

The left hand–side of Equation (22) goes to zero almost surely because of Lemma 2, the right hand–side also vanishes almost surely the proof is analogue to the proof of Lemma 2.

□

3 Practical considerations

In this section, we present some practical considerations that will be useful for the implementation of the method, and will be helpful in the analysis of our real data.

The first point deals with the implementation of the nonparametric regression estimation. Given that multidimensional welfare analysis usually deals with large scale data–sets, the nonparametric estimation of the blinded variables is computationally expensive. To speed up nearest neighbors computation we used the R package "Fast Nearest Neighbor Search Algorithms and Applications" (FNN). Therein the results stated by Beygelzimer et al [5] are implemented. They show that the computing time can be speeded up over the brute force search in at least one order of magnitude.

Another important issue is how to select the number of variables that should be kept. Our aim is to find a subset of variables I where $F_{n,\langle X-\overline{X},g_{n,D}\rangle}$ and $F_{n,\langle \widehat{X}(I)-\overline{X},g_{n,D}\rangle}$ are close in infinity norm. All types of statistical software have built–in routines to measure the infinity norm between two empirical measures, by means of the Kolmogorov–Smirnov goodness–of–fit test (KS–GOF). Two empirical measures will be different if the corresponding p–value is large. In practice we may consider p–value > 0.2. It is important to note that since there is a bijection between the statistics estimation and the p-value of the KS–GOF it is analogue to consider either value. For the sake of simplicity, we choose to analyze the p–value since it is of common domain whether a p–value is large or not.

It is not feasible to visit all the $2^p - 1$ if p is large, for instance for $p = 15$ the number of subsets that should be analyzed is 3.2767, and this quantity grows exponentially with p. Implementing a Genetic Algorithm (GA) may yield a successful approximate solution of a discrete optimization prob-

lem. GA simulate some of the processes observed in biological evolution and natural selection. The basic idea is to consider a strings of $\{0, 1\}$ (chromosomes) representing the presence (1) or absence (0) of a feature, each feature is denoted gene. Each chromosome represents the potential solution of an optimization problem and a fitness score is assigned to each of them. In our setting the fitness score is the p–value of the the Kolmogorov–Smirnov goodness–of–fit test (KS-gof) corresponding to Equation (19). The fitness of each individual is evaluated and only the fittest one reproduce, passing the genetic information to their offspring. In addition, GA supports two phenomena, mutations and crossovers. Crossover forms new offsprings from two parents of chromosomes by combining part of the genetic information of each of them. Mutation randomly alters the value of genes in the parents a chromosome. In addition, a elitism strategy may be apply retaining the best chromosomes for the next generation. The R package GA [26] has been developed to solve optimization problems using genetic algorithms. In addition of the fitness function, which in our case is given by Equation (19), several parameters must be stated before hand. The crossover and mutation probabilities are 0.8 and 0.1 respectively, and the elitism is 2. Also a baseline p-value, $p - value_0$ for the corresponding KS-gof must be stated, big p-values are preferred, for instance $pvalue_0 = 0.2$. A set of initial solutions to be included in the initial population can be suggested, we decided to include all the solutions containing only one gene.

The output of the GA will contain several possible subsets of variables with p–values bigger very close to the $p - value_0$, these subsets of variables may still have many variables, one should retain $\mathcal{I}_G A$, the smallest the subset of variables as an initial solution, but there is no warranty that some of those variables could be dropped out. On a second stage we find a if there is a subset of $\mathcal{I}_G A$ with smaller cardinality than $\mathcal{I}_G A$. These search is done exhaustively. Finally, we retain the smaller subset of variables with p-value big enough so that we cannot reject the null hypothesis of equal distribution between the projection considering the original and the blinded variables.

4 The middle class in Argentina 2004-2014

4.1 Data

We now apply the new approach presented in the previous section to identify the Argentinean middle class over the 2004-2014 period. For this purpose we rely on micro data coming from the *Encuesta Permanente de Hogares* (EPH). The survey covers information on demographic aspects, education, employment and family income as well as characteristics of the dwelling for households across the country. Given the aim of the present analysis, we include a large set of variables in order to multidimensionally identify the argentinian middle class. In the first place we consider family income. Even though our objective is, precisely, to transcend this dimension, it remains one of the most relevant factors that will determine whether a family belongs or not to this particular socio-economic group. A second set of variables are incorporated, following the lines of classical economists who related class to the sources of income, property and wealth (Atkinson and Brandolini 2011 [1]). In this sense, we incorporate variables that identify whether the family has access to income from renting some other property, from profits of a business in which they do not actively participate as well as income from capital interests. Information on ownership of the dwelling is also incorporated. This is especially important in a country such as Argentina where access to mortgage credit is very expensive. Furthermore, data on whether the household receives subsidies is included. Finally, we also contemplate consumption strategies: we include a variable that identifies whether or not the household needs to rely on installments to acquire goods. A third set of variables addresses the concerns of the sociological and political theory literature that associated classes to labor market stratification. In the first place, we include data on whether the head of household is employed. We then move on to identify the occupation type of the household head, from unskilled employment to professional positions. In line with this, we incorporate variables on educational level of the head of household. We also include characteristics of the households dwelling. In particular, we concentrate on its construction materials, its access to basic services as well as whether it is

located in risky areas (flood zone, slums, etc). Finally, our analysis also incorporates one additional variable not traditionally included in middle-class studies: whether the household relies on a domestic employee to take care of household chores, a common practice among Latin American countries. The time span under analysis is 2004-2014. For each of these years, data for all four quarters are provided. Analysis are carried out independently for each quarter, implying more than forty data subsets . On average, each quarter contains around 16,500 households, summing up to around 712,000 observations for the whole period under consideration.

4.2 Multidimensional wellbeing in Argentina 2004-2014

As stated in the previous section, we have defined a multidimensional welfare space. In particular, we have included nineteen variables related to different aspects of well-being as suggested by the literature. In order to be able to assess the multidimensional welfare of these individuals we will proceed to apply the multidimensional quantile approach explained in the previous section.

In essence, our aim is to project the information contained in our original multidimensional space by way of establishing a growth direction that ensures a consistent ordering of the individuals in terms of well-being. In other words, we establish a sort of welfare index that may allow for a consistent ranking of welfare across-individuals, departing from multidimensional data that has not an obvious order.

As exposed in the previous section, the first step is to resort to principal components analysis. We apply a principal component factorization for all the quarters under analysis. Results suggest that our nineteen variables across more than forty quarters may appropriately summarized by four orthogonal factors, accounting for around 80% of total variability.

Our approach defines the growth direction by which the original space is projected as the module of the first principal component. Two results of the principal components analysis give strength to this procedure. On the one hand, the first principal component accounts for 30% of variability on average

across quarters, which is high relative to the magnitude of our original space. On the other hand, when zooming into the first four principal components which, as already said, account for around 80% of the variability- suggest that the variables that are relevant in terms of projecting the data are on average the same. This is surprising given the fact that we are repeating the analysis for over forty datasets, corresponding to the different quarters.

The final result of this procedure is a well-being indicator, that may allow for the consistent ranking of individuals using information coming from several dimensions of welfare. This projection allows for a multi-dimensional identification of the middle class which is the object of the following subsection.

However, before introducing the analysis of this multidimensional analysis of the middle class, a further exercise is proposed. Given the large set of variables contained in the original space, it is interesting to explore which of them are more relevant to assess multidimensional well-being. That is, we may wonder which is the minimum set of variables that define a welfare index that is practically the same than the one projected when departing from the original space. In order to proceed, we follow the variable selection approach explained in the previous section. In essence, this method goes through the variables in the original datasets (and their possible combinations) and leaves out variables that: i) contain redundant information (i.e., are highly correlated to others); or ii) only add noise. We then compare the cumulative distributions of the possible projections using the Genetic Algorithm (GA) described in Section 2.

Intuitively, the algorithm retains the smaller subset of variables that generate a projection as similar as possible as the one obtained by using the original space of variables. This is achieved by implementing hypothesis tests where the null hypothesis is that the projections do not differ. The GA yields the smaller subset of variables for which the p-value of the test is large enough so as not to reject the null hypothesis. As stated in Section 2, the GA is ran in two stages. The first one will yield several possible subsets of variables with large pvalues. We retain $\mathcal{I}_G A.$, the smallest subset of variables, as an initial solution. The second stage searches whether a smaller subset

exists, limiting the exhaustive search to subsets of $\mathcal{I}_G A$.. In our case this procedure must be carried out for each term and year, implying 43 different subsets (that correspond to each quarter) containing each of them 19 variables and more than 16,000 observations. Since, in almost every case IGA (the initial solution), has cardinality bigger than six, in the second stage the exhaustive search is limited to subsets of $\mathcal{I}_G A$., with cardinality smaller than six. Finally we retain a subset of IGA with cardinality four, given that p-values for four variables subsets are always large enough to not reject the null hypothesis of equal distribution between the projection considering the original and the blinded variables.

Therefore, the results of the GA imply that for all quarters we may retain only four of the original nineteen dimensions and still generate a welfare index that does not statistically differ from the one produced using the original space. Even though the subset of variables changes across quarters, on average the variables that seem to be more relevant to determine wellbeing are the following: consumption strategy (appears in 95% of quarters), per capita family income (72% of quarters), type of occupation (70% of quarters) and relying on a domestic employee for household chores (63% of quarters). In short, these four variables seem to sum up quite accurately wellbeing. It comes as no surprise that income is among them. Nevertheless, it does not seem to be the most important variable neither the only one. The presence of the other three dimensions also confirms that welfare is a truly multidimensional phenomena that can only be imperfectly captured by income.

4.3 The middle class in Argentina 2004-2014

By incorporating all of the information referred to in the previous section, the approach proposed is able to multidimensionally identify the middle class in Argentina. As stated before, multidimensional quantiles are built, by projecting the data in a growth direction. A further step, however, is required to finally identify the middle class: a lower and an upper bound need to be established to separate this group from the poor as well as from the upper class. For Argentina under the period of evaluation we established the 25 and

90 quantiles as the bound in between which the middle class is defined. This choice is certainly arbitrary, but consistent with previous work that, although unidimensionally, defined the middle class in what Wolfson [**?**] called "the people space" (see for instance, Beach 1989 [4], Birdsall et al. [7], Barro [3], Easterly [12] and Edo and Sosa-Escudero, [13]) .

With this definition at hand, three exercises are proposed to analysis the course of the Argentinean middle class across the 2004-2014 period. In the first place, we trace the economic performance of this group across time. Secondly, we go beyond income and attempt to characterize the middle class in terms of other indicators of wellbeing, contextualizing its features in terms of the other categories defined (the poor and the upper class). Finally, we explore which variables are most relevant in terms of distinguishing the middle class from the other two groups. This analysis allows us to shed light not only on how many and which variables are needed to do so, but also to assess whether distinguishing the middle class from the poor is relatively easier than distinguishing that same group from the upper class.

4.3.1 *Middle Class Economic Performance*

The analysis of the economic performance of the Argentinean middle class is limited by the definition chosen. In fact, the size of the group under analysis is fixed: by definition, 65% of the population will be identified as the middle class. Therefore, its evolution in economic terms may only be measured in terms of the path followed by some welfare indicators, such as the level of income, income share as well as its dispersion. Figure **??** shows all of the three indicators mentioned across the 2004-2014 period, for all groups identified by our method: the poor, the middle and the upper classes.

It can be observed that the middle class seems to have fared rather well across the period. On the one hand, mean income seems to have risen steadily from 2004, showing a slight decrease in 2013 (see Figure 1a) . It is worth noting, however that this is the case for both the poor and the upper class, which suggests that this indicator may be reflecting the path of the economy in general rather than the particular evolution of the middle class as a specific group. In terms of income share, it seems to have remain stable across the

Figure 1: Income Dispersion

Figure 2: Income Dispersion

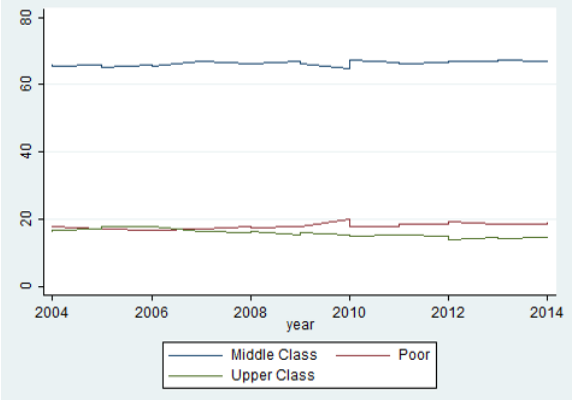

Figure 3: Income Share

period. Holding around 60% of the income share this seems a positive result for the middle class, given the fact that, by definition, the group holds 65% of the population. In terms of dispersion, we also detect signs of stability. This could also be interpreted in some sense in terms of internal cohesion, which should be viewed as positive in light of maintaining levels of polarization low and stable .

4.3.2 *Middle Class Features*

Beyond the economic evolution traced previously, it is also interesting to characterize the argentinian middle class in terms of its household characteristics, how these differ from other groups and how these features may have changed over time. Table 1 shows some of these characteristics for 2004 and 2014.

On average, the middle class in Argentina shows the largest household size as compared to the other two groups, even though it is not especially large (4.4 individuals per household across the period). As expected, they show a greater share of children (around 70% of middle class households have children under 18). The head of household is generally a male, and in this dimension it clearly differs from the poor for whom the head is female for around 50% of households. In terms of education, even though the group shows clear signs of having accumulated on average more human capital than the poor, they are far from the upper class standards: by 2014 almost half of middle class head of households had completed secondary school (or held even higher education levels) as compared to 33% of the poor and 78% of the upper class. It is worth noting that these indicator has improved for all three groups across the period.

Regarding employment of the head of household, this seems to be the most salient feature that separates the middle class from the poor. While the middle class shows rather high levels of employment almost 84%), for the poor less than 10% of the head of households are employed. This is also reflected in dependency ratios: almost 50% of individuals in a middle class households are employed compared to around 20% among the poor.

Subsidies and consumption strategies seem to be the characteristics that

differentiate the most the middle from the upper class. In fact, while in 2014 22.5% of middle class households received some kind of subsidy, only 8% of upper class households declared to receive one. In terms of consumption, by 2014 more than half of middle class household had to resort to installments to be able to access goods. Less than 9% of households in the upper class had to follow this strategy.

Difficulties in owning the dwelling is another salient feature. Around 60% of middle class households own their dwelling. Even though this might seem a large percentage in terms of other countries, in Argentina the access to mortgage credit is rather difficult. In contrast, on average 87% of those in the upper class own their dwellings, although it strikes that this percentage has been falling steadily as from 2004.

In terms of household access to services, all three groups seem to enjoy decent levels, although clearly the poor are worse-off.

	Poor		Middle Class		Upper Class	
	2004	2014	2004	2014	2004	2014
Household Size	4.11	4.07	4.51	4.4	4.33	4.15
Number of children < 18	1.33	1.32	1.76	1.66	1.53	1.42
%of HH with children < 18	0.52	0.55	0.72	0.7	0.69	0.68
% of HH with children < 10	0.4	0.44	0.55	0.55	0.5	0.5
% of Male HH head	0.54	0.45	0.74	0.68	0.8	0.77
% with completed secondary or higher	27.78	32.69	42.86	48.85	72.75	78.28
% of HH head employed	7.68	9.01	85.96	83.58	100	100
%of HH members employed	0.18	0.2	0.5	0.52	0.58	0.61
Ratio of women employed	0.36	0.34	0.54	0.57	0.61	0.65
% of HH receiving subsidies	21.86	25.49	16.56	22.46	5.17	7.97
% of HH buying in installments	77.72	63.3	76.11	56.1	41.42	8.7
% of HH owners of dwelling	63	60	66	66	91	83
% with solid floor	73.6	77.61	73.18	81	88.75	90.85
% with adequate sewage	82.35	87.33	84.2	89.21	92.85	94.07

Table 1: Middle Class in Argentina 2004-2014, Household Characteristics.

To summarize, the middle class in Argentina, during the 2004-2014 period seems to be characterized by families with children, where the household

head is usually an employed male and around 50% of them have completed secondary school or hold even higher levels of education. Their economic standing seems to be reasonable overall, but it must be noted that almost 1 every 4 families receive subsidies as of 2014 and that more than half of them need to resort to installments to access goods.

4.3.3 *Reducing dimensionality of the Middle Class*

The third exercise proposed is to explore which are the key variables that allow for distinguishing the middle class from the other two groups: the poor and the upper class. In particular, we would like to assess not only which and how many variables allow for this distinction but also whether distinguishing the middle class from the poor is relatively easier than doing so with respect to the upper class.

Our goal is to find a smaller subset of variables, of cardinality d, $d \ll 19$, which preserves the original grouping conformation on poor, middle and rich class as accurate as possible. We adopt the methodology introduced by Fraiman et al. [14], which is a variable selection criteria for cluster and classification, based on a 'blinding' process that eliminates unnecessary variables, analogue to the ideas presented in Section 2.3. Since the variable selection criteria may be computationally very expensive, we adopt the same numerical strategy introduced in section 3, we did not run an exhaustive search over all the $2^{19} - 1$, subsets of variables, instead on a first stage we carried out a genetic algorithm, al the parameters we set as stablished on section 3, and aditionally we had to state a threshold indicating the maximum percentage ofobservations that could be reallocated after the blinding procedure has been carried out, this threshold is 5%. On a second stage we pursue an exhaustive search, considering that the initial subset consider is the one given by the genetic algorithm on the first stage.

It is important to note that this procedure has been carried out, for each term and year, considering on each case two different problems. On one hand we want to select the variables that produce less grouping reallocation between the the poor and the middle class, and on the other hand we wanted to tackle the same problem considering clustering reallocations between the

rich and the middle class.

It is worth noting that the subset of variables selected does not necessarily need to coincide with those that were found relevant when performing the variable selection in terms of the multivariate quartiles. When doing so, we were focusing on the variables that best describe the whole cumulative distribution of the projected welfare index. In this analysis, instead, we are zooming into two particular points of that distribution, i.e., the cut-offs chosen to identify the poor, the middle class and the upper class. At this point we are interesting in obtaining almost the same group classification but reducing the original space to the minimum set of variables possible. Surprisingly enough, when focusing in each of both distinctions (middle class versus the poor; middle class versus the upper class) for all quarters analyzed the relevant dimensions seem to be the same. That is, for all quarters the relevant variables to distinguish the poor from the middle class seem to be the same. This also applies when looking at the upper and middle class, although the number of variables included in both cases is different.

We first focus on the poor-middle class divide. If we take into consideration variables that are selected in more than 20% of quarters by the Fraiman et al. [14] procedure, we find that on average we need two features to divide the poor from the middle class. This implies that departing from the original set of nineteen variables we may focus on only two dimensions of welfare still be able to classify quite accurately whether individuals are poor or middle class. The variables that the methodology employed founds to result relevant on average across quarters are the consumption strategy and whether the head of household is employed.

As expected, distinguishing the middle class from the upper class is relatively more difficult. Once again, we establish our standard of a maximum of approximately 5% of observations being re-classified, we need to focus at least on 4 variables on average to distinguish the middle from the upper class. Even if for each quarter the relevant variables differ, on average almost always 4 of them is the minimum set required. Which are these variables? The Fraiman et al. [14] procedure identifies that in more than 60% of the quarters the relevant variables equal to those found in the case of the poor and the

middle class: the consumption strategy as well as employment of the head of household. Nevertheless, we need to resort to two additional variables when looking at the middle-upper class distinction. The two following variables that appear in more than 20% of the quarters are whether the household relies on a domestic employee as well as its per capita family income.

This analysis confirms two hypothesis previously mentioned. On the one hand, wellbeing is a complex phenomena for which income is a rather imperfect proxy. It is true, however, that the relevant variables that may identify this socio-economic groups are tightly tied to income: consumption strategy and employment. Secondly, it confirms the common knowledge that distinguishing the middle from the upper class is far more difficult than separating that group from the poor.

5 Concluding Remarks

Middle class studies have gained relevance in recent years among economists, enlarging the scope traditionally lead by sociologists and political theorists. The economic literature, however, is far from conceptual and methodological agreements on how to identify this particular socio-economic group. As has been abundantly proven in the related poverty literature (Szekely et al. 2000), this is not trivial: different definitions of the middle class may lead to opposed conclusions regarding its characterization and evolution across time.

Furthermore, middle class studies have mainly concentrated the analysis on a unidimensional identification of this group through income levels. This poses an additional difficulty in terms of the middle class, since although it may seem rather reasonable to impose a lower bound based on income it is much more complicated to think of a threshold that divides the middle from the upper class. In fact, it seems natural to incorporate in its identification some of the salient features that classical economists, sociologists and political theorists have traditionally associated with the middle class: the role of property, wealth and occupational stratification.

In this paper we presented a new multidimensional approach to measuring welfare and identifying the middle class. Furthermore, we are able

to establish which are the most salient dimensions that allow to reduce the original space without substantive loss of information. We are thus able to multidimensionally identify the middle class, characterize its performance and main features as well as to establish which dimensions are most relevant to distinguish this group from the others (i.e., the poor and the upper class).

The approach is applied to Argentina for the 2004-2014 period. Several interesting results are found. On the one hand, departing from the nineteen original variables we are able to reduce the dimensionality of welfare to only four. Indeed, the households consumption strategy, its per capita family income, the type of occupation of the head of household and whether the family relies on a domestic employee for household chores seem to be the crucial characteristics that define welfare in Argentina under the period of analysis. The fact that income appears as a fundamental feature of welfare, but neither the only nor the most important confirms the need for a multidimensional approach to well-being.

On the other hand, we are able to multidimensionally identify the Argentinean middle class and thus characterize its evolution across the last two decades through three exercises. First of all, we focus on its economic performance, which seems to have been rather positive during this period. Mean income has risen, while its dispersion and share have remained rather stable. Secondly, we go beyond income evolution to characterize its main features. In this regard, we find that the middle class is characterized by families with children, where the head of household is usually an employed male and around 50% of them have completed secondary school or hold even higher levels of education. Although their economic standing seems to be reasonable overall, it must be noted that almost one in every four families receive subsidies as of 2014 and that more than half of them need to resort to installments or informal credit to access goods.

Finally, we find that distinguishing the middle class from the other groups depends to a great extent on two dimensions: its consumption strategy and whether the head of household is employed. However, it is worth noticing that in order to distinguish this group from the upper class two additional dimensions need to be taken into account: per capita family income and whether

the household relies on a domestic employee for household chores. This confirms the common knowledge that it is easier to establish the poor/non-poor divide than to establish an upper bound that identifies the middle class.

References

[1] Atkinson, A. and Brandolini A. (2011), "On the identification of the Middle Class". ECINEQ Working Paper No. 217.

[2] Banerjee, A. and Duflo, E.(2007), "What is Middle Class About the Middle Classes Around the World?" MIT Department of Economics Working Paper No. 07-29.

[3] Barro, R. (1999), "Determinants of Democracy." *Journal of Political Economy* 107(S6), 158-183.

[4] Beach, C.M., (1989), "Dollars and dreams: A reduced middle class? Alternative explanations." *Journal of Human Resources* 24, 162-193.

[5] Beygelzimer, A., Kakade, S. and Langford, J. (2006), "Cover trees for nearest neighbor." ACM Proc. 23rd international conference on Machine learning, 148, 97-104.

[6] Billingsley, P. and Topsøe, F. (1967). "Uniformity in Weak Convergence." *Zeitschrift fr Wahrscheinlichkeitstheorie und Verwandte Gebiete.* **7**(1), 1-16.

[7] Birdsall, N., Graham, C., Pettinato, S.(2000). "Stuck in the Tunnel: Is Globalization Muddling the Middle Class?" Center on Social and Economic Dynamics Working Paper No. 14.

[8] Chaudhuri, P. (1996). "On a Geometric Notion of Quantile for Multivariate Data." *Journal of the American Statistical Association.* **91**(434), 862-872.

[9] Cruces, G., López-Calva, L., Battiston, D. (2009). "Down and Out or Up and In? Polarization-Based Measures of the Middle Class for Latin

America." *Anales de la Asociacin Argentina de Economa Poltica; Lugar: Buenos Aires* vol. XLIV p. 1 - 31

[10] Davis, J. C., Huston, J. H. (1992). "The Shrinking Middle Class Income: A Multivariate Analysis." *Eastern Economic Journal.* **18** (3), 277–285.

[11] Dauxois, J. , Pousse, A. and Romain, Y. (1982). "Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference". *Journal of Multivariate Analysis* **12,** 136-154.

[12] Easterly, W. (2001). "The Middle Class Consensus and Economic Development." *Journal of Economic Growth.* 6(4), 317-35.

[13] Edo, M. and Sosa-Escudero, W. (2012). "Tracking the Evolution of the Middle Class in Argentina 1991-2012." *Anales de la Asociacin Argentina de Economa Poltica;* Lugar: Trelew; vol. XLVII.

[14] Fraiman, R., Justel, A. and Svarc, M. (2008). "Selection of variables for cluster analysis and classification rules." *J. Amer. Statist. Assoc.* **103** no. 483 1294–1303.

[15] Fraiman, R. and Pateiro-Lpez, B. (2012). "Quantiles for finite and infinite dimensional data". *Journal of Multivariate Analysis.* **108,** 1–14.

[16] Foster, J. and Wolfson, M. (2009). "Polarization and the decline of the Middle Class: Canada and the US." *Journal of Economic Inequality* 8(2), 247-273.

[17] Gayo, M. (2013). "Revisiting middle-class politics: a multidimensional approach evidence from Spain." *The Sociological Review.* **61**, 814-837.

[18] Gigliarano, C. and Pietro Muliere (2012). "Measuring Income Polairzation Using an enlarged Middle Class." Working Paper 271, ECINEQ.

[19] Gigliarano, C. and Mosler, K. (2009). "Measuring Middle-Class Decline in One and Many Attributes." Quaderno di Ricerca 333.

[20] Hallin, M., Paindavaine, D. and Siman, M. (2010). "Multivariate Quantiles and Multiple-Output Regression Quantiles: From L_1 Optimization to Halfspace Depth." *The Annals of Statistics.* **38** (2), 635-669.

[21] Hansen, B. E. (2008). "Uniform convergence rates for kernel estimation with dependent data," *Econometric Theory.* **24**, 726–748.

[22] Kong, L. and Mizera, I. (2012). "Quantile Tomography: Using Quantiles with Multivariate Data," *Statistica Sinica.* **22**, 1589-1610.

[23] Ravallion, M. (2009). "The Developing Worlds Bulging (but Vulnerable) Middle Class". World Bank Research Working Paper 4816.

[24] Serfling, R. (2010). "Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardization." *Journal of Nonparametric Statistics,* **22**, 915-936.

[25] Székely, M., Lustig, N., Cumpa, M. and Mejía, J.A. (2000). "Do we know how much poverty there is?", Working Paper 437, Inter American Development Bank.

[26] Strucca, L. (2013). "GA: a package for genetic algorithms in R." *Journal of Statistical Software*, **53(4)**, 1-37.

[27] Weber, M. (1905). "The Protestant Ethic and the Spirit of Capitalism." Germany.