Multidimensional Middle Class

María Edo [1] Walter Sosa Escudero [1] Marcela Svarc[2]
msvarc@udesa.edu.ar

[1]Universidad de San Andrés, Departamento de Economía and CONICET

[2]Universidad de San Andrés, Departamento de Matemática and CONICET

September 21, 2016

Existent proposals

Existent proposals

- Davies and Huston (1992) and Gayo (2013): definition based in terms of income, multidimensional benchmark.

Existent proposals

- Davies and Huston (1992) and Gayo (2013): definition based in terms of income, multidimensional benchmark.

Existent proposals

- Davies and Huston (1992) and Gayo (2013): definition based in terms of income, multidimensional benchmark.
- Girigliano and Mosler (2012):

Existent proposals

- Davies and Huston (1992) and Gayo (2013): definition based in terms of income, multidimensional benchmark.
- Girigliano and Mosler (2012):
  - Ball centered on the multidimensional mean.

Existent proposals

- Davies and Huston (1992) and Gayo (2013): definition based in terms of income, multidimensional benchmark.
- Girigliano and Mosler (2012):
  - Ball centered on the multidimensional mean.
  - Minimum Volume Ellipsoid.

Drawbacks

Drawbacks

- Non-spherical variables.

Drawbacks

- Non-spherical variables.

Drawbacks

- Non-spherical variables.
- Dense region vs. central region.

Drawbacks

- Non-spherical variables.
- Dense region vs. central region.

Drawbacks

- Non-spherical variables.
- Dense region vs. central region.
- True dimension of the problem?

Our goal...

Our goal...

- New Multidimensional approach to measure welfare through the construction of multivariate quantiles based on a growth direction g_D.

Our goal...

- New Multidimensional approach to measure welfare through the construction of multivariate quantiles based on a growth direction g_D.
- Tackles the problem of reduction of the dimension of welfare. Middle class dimension? Middle class features?

Our goal...

- New Multidimensional approach to measure welfare through the construction of multivariate quantiles based on a growth direction g_D.
- Tackles the problem of reduction of the dimension of welfare. Middle class dimension? Middle class features?
- The Argentinian Case, 2004-2014.

Middle Class

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.

Middle Class

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.

Properties:

Middle Class

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.

Properties:

- A given proportion of central population, the multivariate α-region, $C(\alpha)$, must $P(X \in C(\alpha)) \geq \alpha$.

Middle Class

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.
Properties:

- A given proportion of central population, the multivariate α-region, $C(\alpha)$, must $P(X \in C(\alpha)) \geq \alpha$.
- Our variables measure wellbeing, each of them has a natural increasing order, this order must be preserved by the definition stated.

Multivariate quantiles

Previous work? Non suitable!

Multivariate quantiles

Previous work? Non suitable!
Why?

Multivariate quantiles

Previous work? Non suitable!
Why?

- Affine or rotational equivariance.

Multivariate quantiles

Previous work? Non suitable!
Why?

- Affine or rotational equivariance.
- The multivariate α-region, $C(\alpha)$, do not satisfy $P(X \in C(\alpha)) \geq \alpha$.

Multivariate quantiles

Our definition...

Let $X \in \mathbb{R}^p$, 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.

First we want to determine the $\alpha-$*upper* region of the distribution.

Assumptions:

Multivariate quantiles

Our definition...
Let $X \in \mathbb{R}^p$, 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.
First we want to determine the $\alpha-$*upper* region of the distribution.
Assumptions:

- Wellbeing variables are increasing.

Multivariate quantiles

Our definition...
Let $X \in \mathbb{R}^p$, 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.
First we want to determine the $\alpha-$*upper* region of the distribution.
Assumptions:

- Wellbeing variables are increasing.
- Existence of a *growth direction*, g_D, $\|g_D\| = 1$.

Multivariate quantiles

Our definition...
Let $Y_D = \langle X, g_D \rangle$, the projection of X respect to g_D, and

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

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

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

Multivariate quantiles

Lemma

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

Multivariate quantiles: empirical counterpart

Let $X_1, \ldots, X_n$ be a random sample of vectors with distribution P_X and denote by P_n its empirical distribution.

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

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

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

Multivariate quantiles: empirical counterpart

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 $g_{n,D}$. Under mild regular conditions on the covariance matrix, it is well known that $g_{n,D} \to g_D a.s.$

Multivariate quantiles: consistency

Theorem

Let $X_1, \dots, X_n$ 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,

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

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

$$A_n(x) \to_{n\to\infty} A(x) \text{ a.s.} \tag{9}$$

Multivariate quantiles: consistency

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

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

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

Multivariate quantiles: Consistency

Theorem

Under the same conditions of Lemma 2, 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.*

Multivariate quantiles: Variable Selection

Are all the variables important?
We develop an *ad hoc* variable selection criterion, based on the *blinding* strategy introduced by Fraiman, Justel and Svarc (2008).
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 $X[i], \quad i = 1, \ldots, p$.

Multivariate quantiles: Variable Selection

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 $\{X[i], i \in I\}$.
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

$$Z(I)[i] = \begin{cases} X[i] & \text{if } i \in I \\ E(X[i]|\mathbf{X}(I)) & \text{if } i \notin I. \end{cases} \tag{10}$$

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

Multivariate quantiles: Variable Selection

The goal is to find a minimal subset of variables from X that retains almost all the relevant information from the quantile function.
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,

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

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

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

Multivariate quantiles: Variable Selection

$\mathcal{I}_0 \subset \mathcal{I}_d$ is defined as the family of subsets in which,

$$\mathcal{I}_0 = argmin_{I \in \mathcal{I}_d} \| F_{\langle X - E(X), g_D \rangle} - F_{\langle Z(I) - E(X), g_D \rangle} \|_\infty . \qquad (11)$$

Multivariate quantiles: Variable Selection

Empirical version, 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 $\mathcal{P}$.

Multivariate quantiles: Variable Selection

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 (Hansen, 2008). We estimate them by r–nearest neighbors.

Multivariate quantiles: Variable Selection

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

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

where $X_j[i]$ stands for the ith-coordinate of the vector $\mathbf{X}_j$.
$Q_n(I)$ stands for the empirical distribution of $\{\hat{\mathbf{X}}_j(I), 1 \le j \le n\}$.

Multivariate quantiles: Variable Selection

Our aim is to find the optimal subsets of variables $I_0 \subset I_d$,

$$\hat{\mathcal{I}}_n = argmin_{I \in \mathcal{I}_d} \| F_{n,\langle X - \overline{X}, g_{n,D} \rangle} - F_{n,\langle \widehat{X}(I) - \overline{X}, g_{n,D} \rangle} \|_\infty,$$

where

$$F_{n,\langle \widehat{X}(I) - \overline{X}, g_{n,D} \rangle}(t) = \frac{1}{n} \sum_{j=1}^{n} \mathcal{I}_{\{\langle \widehat{X}_j(I) - \overline{X}, g_{n,D} \rangle \leq t\}}. \tag{12}$$

Multivariate quantiles: Variable Selection

Theorem

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 $\|F_{\langle X-E(X), g_D\rangle} - F_{\langle Z(I)-E(X), g_D\rangle}\|_\infty$ is reached. Then, if the nonparametric estimator of the regression function is uniformly consistent a.s. and the covariance matrix is non-singular, 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.

Multivariate quantiles: Variable Selection

Practical considerations:

Multivariate quantiles: Variable Selection

Practical considerations:

- n large $\Rightarrow$ Fast non-parametric estimation,

Multivariate quantiles: Variable Selection

Practical considerations:

- n large $\Rightarrow$ Fast non-parametric estimation,
- p large $\Rightarrow$ Genetic algorithm, etc.

The Data

- Our goal is to identify the middle class over the 2004-2014 period.
- Micro data coming from the *Encuesta Permanente de Hogares* (EPH)
- Information: demographic aspects, education, employment, family income, characteristics of the dwelling, for households across the country.

The Data

Variables ($p = 19$):

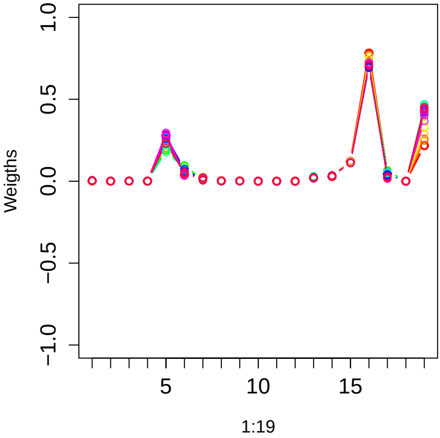
- Income.
- Access to renting other properties, profits of a business without active participation, ownership of dwelling, households receiving subsides, consumption strategy.
- Employment, occupation type of the household head, unskilled employment to professional positions, educational level of the household head, characteristics of the household dwelling.
- Domestic employee.

The time span under analysis is 2004-2014.
Data for all four quarters are provided. Analysis are carried out independently for each quarter, implying more than forty data subsets. Each quarter contains around 16500 households (n), summing up to around 712000 observations for the whole period under consideration.

Multidimensional wellbeing in Argentina 2004-2014

Our approach defines the growth direction by which the original space is projected as the module of the first principal component. The first principal component accounts for 30% of variability on average across quarters, which is high relative to the magnitude of our original space.
When zooming into the first four principal components which 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.

Multidimensional wellbeing in Argentina 2004-2014



Multidimensional wellbeing in Argentina 2004-2014

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.
We follow the variable selection approach explained the previously. 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 16000 observations. Two steps:

1. Genetic Algorithm.
2. Exhaustive selection.

Multidimensional wellbeing in Argentina 2004-2014

For each term we retained 4 variables (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).
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).
- relying on a domestic employee for household chores (63% of quarters).

The Middle Class in Argentina 2004-2014

The requirements to identify the middle class are a lower and an upper bound, to separate this group from the poor as well as from the upper class.
For Argentina, we established the 25 and 90 quantiles as the bound in between which the middle class is defined.

Economic Performance across time.

Figure : Mean Income

Economic Performance across time.

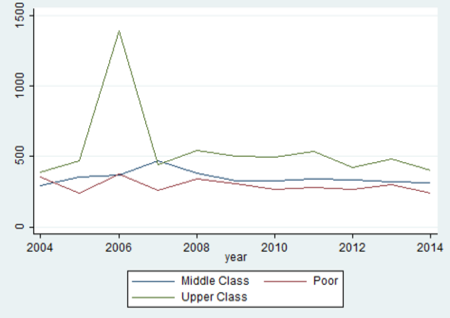


Figure : Income Dispersion

Economic Performance across time.

Figure : Income Share

Middle Class Features

	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

Reducing the Dimensionality of the Middle 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, Justel and Svarc (2008).

We carry out this procedure for each term and year.

We want to select the variables that produce less grouping reallocation between the the poor and the middle class.

We want to select the variables that produce less grouping reallocation between the the rich and the middle class.

Reducing the Dimensionality of the Middle Class Poor-middle class division

- Two features selected.
- Consumption strategy.
- Whether the head of household is employed.

Reducing the Dimensionality of the Middle Class Rich-middle class division

- Four features selected.
- Consumption strategy.
- Whether the head of household is employed.