# Report on "Polarization Measurement and Inference in Many Dimensions When Subgroups Cannot be Identified"

The paper proposes an extension of the Duclos, Esteban and Ray (2004)'s polarization index to the case of several dimensions, both continuous and discrete.

The author also provides an empirical application aimed at comparing the new multidimensional index with its univariate counterpart, by using data on Chinese urban households.

In my opinion, the contribution of the paper is potentially significant, since it deals with a relevant issue not enough explored in the well-being literature, that is measuring polarization in a multivariate approach. In particular, the interesting contribution of this manuscript consists of considering the joint distribution of the different dimensions of well-being rather than simply turning to an aggregating approach. Moreover, the paper proposes a general approach that allows for combining both discrete and continuous variables.

However, I think the manuscript is still in a form that is too preliminary to be ready for publishing; in particular it needs to be integrated with deeper discussions.

## General comments

1. In the Introduction the author carefully reviews the main axioms for univariate polarization indices, but then he does not discuss at all about how to generalize these axioms to the multivariate case. In my opinion these axioms cannot straightforwardly

extended to multivariate distributions, but they rather require more appropriate definitions.

2. The second part of the title "when subgroups cannot be identified" has not been discussed and analyzed enough in the manuscript; for example, it is not clear to me whether in equation (1a) $f_z$ and $F$ refer to a subgroup distribution or rather to the whole population distribution.

3. Since the distributions of the several well-being dimensions can be rather dissimilar one to the other, it seems to me that the Mahalanobis distance is more appropriate than the Euclidean distance, in order to mitigate the different variability of each dimension.

4. The author should provide a proof for the variance of the index's estimator.

5. In the sample counterpart of the index proposed, why does the author estimate $f(\cdot)^\alpha$ with the sample weights but $F(\cdot)$ simply with $1/n$?

6. Equation (1a) is not clear to me: first of all, the notation is not precise, since (i) the integrals are $k$-dimensional and (ii) the notation $z \subset x$ has not been defined. Also, I do not understand why the cdf $F$ refers only to the continuous variables. Why not considering instead the following index:

$$P_\alpha = \sum_{z \subset x} p(z) \int_{\mathbb{R}^k} \sum_{z \subset x} p(z) \int_{\mathbb{R}^k} (f_z(w_1|z)p(z))^\alpha \, ||y - x|| dF(w_1|z) dF(w_2|z),$$

where $dF(w_1|z)$ is the cdf of $w_1$ conditional to the discrete variables $z$?

# Minor comments

1. At page 1 the author states that Wang and Tsui (2000) index allows for many groups; however, it seems to me that the cited index is based only on two groups.

2. The list of references is incomplete. For example, Esteban and Ray (2007) cited at page 2 is not included in the References.

3. At page 4 it is not the index $P_\alpha$ in equation (1) that is asymptotically normally distributed but rather its estimator.

4. In note 3 I suggest the author to cite the Koshevoy and Mosler (1997)'s multivariate Gini index that seems quite similar to the author's index in case of $\alpha = 0$.

5. At page 6 line 2 there is a typo: $z_i$ is the vector of discrete (and not continuous) variables.

6. At page 7 the author should define more precisely the quantities used in the definition of estimator $\widehat{P}_\alpha$.