Economics
The Open-Access, Open-Assessment E-Journal

Discussion Paper
No. 2011-20 | June 27, 2011 | http://www.economics-ejournal.org/economics/discussionpapers/2011-20

Polarization Measurement and Inference in Many Dimensions When Subgroups Cannot be Identified

Gordon Anderson
Department of Economics, University of Toronto

Please cite the corresponding journal article:
http://dx.doi.org/10.5018/economics-ejournal.ja.2011-11

Abstract The most popular general univariate polarization indices for discrete (Esteban and Ray 1994), and continuous (Duclos, Esteban and Ray 2004) variables are combined and extended to describe the extent of polarization between agents in a distribution defined over a collection of many discrete and continuous agent characteristics. A formula for the asymptotic variance of the index is also provided. The implementation of the index is illustrated with an application to Chinese urban household data drawn from six provinces in the years 1987 and 2001 (years spanning the growth and urbanization period subsequent to the economic reforms). The data relates to household adult equivalent log income, adult equivalent living space, which are both continuous variables and the education of the head of household which is a discrete variable. For this data set combining the characteristics changes the view of polarization that would be inferred from considering the indices individually.

Paper submitted to the special issue
The Measurement of Inequality and Well-Being: New Perspectives

JEL C14, C30, I32
Keywords Multivariate polarization measurement

Correspondence Gordon Anderson, Department of Economics, University of Toronto, Max Gluskin House, 150 St George St, Toronto, Ontario M5S 3G7, Canada; e-mail: anderson@chass.utoronto.ca

Many thanks Oliver Linton for help with multivariate Kernel Estimation and to Jean-Yves Duclos and Joan Esteban for helpful comments on an earlier drafts of this paper.

Introduction.

The functionings and capabilities approach to wellbeing measurement (Sen (1992)) has given considerable impetus to multidimensional analyses of wellbeing (Grusky and Kanbur (2006)). The argument is that individual wellbeing is not just a matter of the incomes they have or could achieve, among other things it depends on individual health and educational status, their political freedoms and environmental factors. In the absence of a well specified wellbeing aggregator of these many sensibilities (i.e. some form of utility function) evaluation of wellbeing has to be evaluated over these many dimensions which of course could be measured discretely or continuously.

The multivariate polarization measure presented here is founded upon the notion of polarization within a distribution of individual characteristics across a population into many possible groups. Esteban and Ray (1994), Duclos, Esteban and Ray (2004), Wang and Tsui (2000) posited a collection of propositions with which a Polarization measure should be consistent and proposed a collection of univariate measures appropriate for a variety of circumstances that would reflect such polarization between potentially many groups. The propositions are based upon a so-called Identification-Alienation nexus wherein notions of polarization are fostered jointly by an agent's sense of increasing within-group identity or association and between-group distance or alienation.

There have been several proposed univariate polarization indices which focus on an arbitrary number of groups and a fortiori two groups (Esteban and Ray , 1994; Esteban, Gradin and Ray, 1998; Zhang and Kanbur, 2001; Duclos, Esteban and Ray 2004) and a similar number that focus on just two groups (Alesina and Spolaore 1997; Foster and Wolfson 1992; Wolfson 1994; Wang and Tsui, 2000). While much work has been done on extending one dimensional wellbeing measures to many dimensions in the context of poverty (Duclos, Sahn and Younger (2006)) and inequality measurement (Maassoumi (1987), (1999), Koshevoy and .Mosler (1997), Tsui (1995) and Anderson (2008))[1] little has been done in extending polarization measures to the many dimensioned case. While Gigliarano and Mosler, (2008) develop a family of multivariate polarization measures based upon measures of between and within group multivariate variation and relative group size which exploit notions of subgroup decomposability and Anderson (2010) and Anderson, Linton and Leo (2011) have developed a trapezoidal measure of polarization which can be applied to two identifiable groups or within a population distribution provided at least two modal points are identified, multivariate polarization measures have not been developed for the more general non-identified many group case, nor for the case where the joint distribution of sensibility indicators is a mixture of discrete and continuous variables[2].

An excellent summary of the properties of the univariate indices is to be found in (Esteban and Ray, 2007) wherein the properties of indices are evaluated in terms of their coherence with some basic axioms that reflect three broad notions, 1) When there is only one group there is little polarization, 2) polarization increases when within group

[1] all however confine themselves to continuous variables

[2] Furthermore extensions of the stochastic dominance techniques introduced in Anderson (2004), which really explore the anatomy of polarizing distributions, would prove cumbersome in many dimensions because it is not obvious how to define a sensible partition of the distribution across those many dimensions.

inequality is reduced, 3) polarization increases when between group inequality increases. The axioms are formed around a notional univariate density that is a mixture of kernels f(x, a) that are symmetric uni-modal on a compact support of [a,a+2] with $E(x) = \mu = (a+1)$ also representing the mode. However these axioms are readily extended to multivariate densities by thinking in terms of a notional multivariate density that is a mixture of multivariate kernels. The kernels are subject to slides (location shifts) $g(y) = f(y-x)$, which may be contemplated in terms of the Euclidean distance between y and x, and squeezes (shrinkages) of the form $f^{\lambda}(x) =f(\{x-[1-\lambda]\mu\}/\lambda)/\lambda$ $(0 < \lambda <1)$ and potential indices are evaluated in the context of such changes in terms of the extent to which they satisfy a set of axioms which reflect the following set of ideas. The squeeze of a uni-modal distribution cannot increase polarization and symmetric squeezes of the two kernels cannot reduce polarization. Sliding two kernels away from one another increases polarization and common population scaling preserves the polarization ordering. Polarization indices have to come from a family where if x and y are independently distributed with marginal distributions f(x) and f(y) then the index is the expected value of some function $T(f(x),|x-y|)$ which is increasing in its second argument. Symmetric squeezes of the sub distributions weakly increases polarization. The index should be non-monotonic with respect to outward slides of the sub distributions and .flipping the distribution around its support should leave polarization unchanged. Most of these ideas can be contemplated with respect to multivariate densities.

Here the most popular general univariate polarization indices for discrete (Esteban and Ray (1994)), and continuous (Duclos, Esteban and Ray (2004)) variables are combined and extended to describe the extent of polarization between agents in a distribution defined over a collection of many discrete and continuous agent characteristics. The univariate indices have been demonstrated to satisfy the aforementioned axioms. The implementation of the index is illustrated with an application to Chinese urban household data drawn from six provinces in the years 1987 and 2001 (years spanning the growth and urbanization period subsequent to the economic reforms). The data relates to household adult equivalent log income, adult equivalent living space, which are both continuous variables and the education of the head of household which is a discrete variable.

The Extension to Many Variables both Discrete and Continuous.

The multivariate generalization of the Duclos Esteban and Ray (2004) (DER) Polarization index is, like DER, based upon the sample equivalents of the population concepts. For scalar continuous x with distribution function F(x) the DER index is given by:

$$P_\alpha = \int\int f(x)^\alpha \mid y - x \mid dF(y) dF(x) \qquad [1]$$

Which DER show to be asymptotically normally distributed with an asymptotic variance V given by:

$$V_\alpha = \text{var}_{f(y)}\left((1+\alpha) f(y)^\alpha \int_0^\infty \| y-x \| dF(x) + y\int_0^\infty f(x)^\alpha dF(x) + 2\int_y^\infty \| y-x \| \, f(x)^\alpha dF(x) \right) \quad [2]$$

A similar discrete variable index is provided in Esteban and Ray (1994) and is given by:

$$P_\alpha = K\sum_{i=1}^{n}\sum_{j=1}^{n} | x_i - x_j | \pi_i^{1+\alpha}\pi_j$$

where π_i is the sample weight of the i'th observation and K is a normalizing factor.

Development of the polarization index was founded on a set of axioms that such an index should obey, the axioms concern changes (squeezes and slides) in the uni-modal sub distributions in the mixture distribution that is f(x). The resultant index reflects the two primary factors that underlay polarization, the alienation or distance between groups (given by |y-x|) and the association within a group (given by $f(x)^\alpha$). Indeed the intuitive interpretation of P_α is that it is the average value of the areas of all possible trapezoids that can be formed under f(x) whose average height is $f(x)^\alpha$ and whose base is |x-y|. As such it can be related to the trapezoidal index of polarization employed in Anderson (2010) to study multivariate poverty states and in Anderson, Leo and Linton (2011) to study multivariate convergence issues. Here α is a polarization sensitivity parameter[3] chosen by the investigator such that $0.25 \leq \alpha \leq 1$ with higher values of α corresponding to increased sensitivity. The same axioms can be applied when x is a vector and where ||x-y|| is the Euclidean distance between the vectors[4].

[3] Note when α = 0 the index is in essence twice the Gini coefficient thus a similar value in the following would provide a multivariate version of a Gini like coefficient and its variance.

[4] Anderson, Crawford and Leicester (2011) employ Euclidian distance in developing an non-parametric approach to multivariate welfare rankings.

Let w_i and z_i be jointly distributed vectors describing the status of the i'th agent with w_i being a k x 1 vector of continuous variables and z_i being an h x 1 vector of continuous variables with i =1,..,n being the elements of the sample. The continuous variables all reflect wellbeing positively and for convenience are defined on R^+ and the discrete variables are ordered integers reflecting positive wellbeing in the same fashion[5]. The joint density of the w's for a given configuration of z's is $f_z(w|z)$ and the joint probability of the z's is p(z) so that the joint density of the w's and z's for the i'th agent with discrete variables z_i is given by $f(w_i,z_i) = f_i(w_i\ |z_i\ )p(z_i)$. Let x_i be the stacked vector $w_i \mid z_i$ then the dimension normalized Euclidean distance between agents i and j $||x_i-x_j||$ is well defined and may be written as:

$$|| x_i - x_j || = \sqrt{\frac{\sum_{q=1}^{Q} (x_{iq} - x_{jq})^2}{Q}}$$

where x_{iq} is the q'th element of the vector x_i where Q = k+h. For notational convenience denote the first k continuous components of the vector x as $x_{\{c\}}$.

Then, retaining the trapezoidal intuition, a multivariate version of [1] is given by:

$$P_\alpha = \sum_{z\in x}\int \sum_{z\in y}\int (f_z(w\,|\,z)p(z))^\alpha\ ||\,y - x\,||\ dF(y_{\{c\}})dF(x_{\{c\}}) \quad [1a]$$

Here summation is over the domain of each element of the z vector and integration is over the domain of each element of the w vector. As in the univariate case the alienation

[5] For example the continuously measured variables may represent levels of consumption, leisure and housing stock whereas the discretely measured variables may reflect levels of educational, health or freedom status.

or distance between groups is given by ||y-x|| and the association within a group given by $f(w,z)^\alpha$ in exactly the same fashion[6]. By employing kernel estimates of the conditional multivariate distributions and sample estimates of the population proportions p(z) the sample equivalents, given n observations on Q variables in an n x Q matrix X with typical element x_{iq} i = 1,.., n, q = 1,..,Q and typical row x_i the index can be seen to be[7]:

$$\widehat{P_\alpha} = \frac{\sum_{i=1}^{n} (\widehat{f(w_i \mid z_i)}\widehat{p(z_i)})^{\alpha} \sum_{j=1}^{n} \sqrt{\sum_{q=1}^{Q} (x_{iq} - x_{jq})^2}}{n^2 \sqrt{Q}}$$

The multivariate version of [2], the variance of index is given by:

$$V_\alpha = \mathrm{var}_{f(y)} \begin{pmatrix} (1+\alpha)(f(w \mid z)P(z))^{\alpha} \sum_{z} \int_{0}^{\infty} \| y - x \| dF(w \mid z)P(z) + \\ \| y \| \sum_{z} \int_{0}^{\infty} (f(w \mid z)P(z))^{\alpha} dF(w \mid z)P(z) + 2\sum_{z} \int_{y}^{\infty} \| y - x \| (f(w \mid z)P(z))^{\alpha} dF(w \mid z)P(z) \end{pmatrix} \qquad [2a]$$

Where after ordering the vectors x_i on $||x_i||$ as x_i^o, the first, second and third terms of the i'th element of the variance vector may be respectively estimated in an obvious fashion as:

[6] Note that Esteban and Ray (1994) and DER respectively offer different ranges for α for discrete univariate and continuous univariate distributions this can be accommodated in the present context by considering the association component as $f(w|z)^{\alpha c}p(z)^{\alpha d}$ where αc is the polarization parameter for the continuous components and αd is the polarization parameter associated with the discrete components.

[7] In DER the columns of X are mean standardized and assumed to reside in the positive orthant.

$$\frac{\widehat{f(w_i \mid z_i)P(z_i)}^{\alpha} \sum_{j=1}^{n} \sqrt{\sum_{k=1}^{Q} (x_{ik} - x_{jk})^2}}{n\sqrt{Q}}$$

$$\sqrt{\frac{\sum_{k=1}^{Q} x_{ik}^2}{Q}} \frac{\sum_{j=1}^{n} \widehat{f(w_i \mid z_i)P(z_i)}^{\alpha}}{n}$$

$$2\frac{\widehat{f(w_i \mid z_i)P(z_i)}^{\alpha} \sum_{j=i}^{n} \sqrt{\sum_{k=1}^{Q} (x_{ik}^{o} - x_{jk}^{o})^2}}{n\sqrt{Q}}$$

Essentially the generalization simply involves employing the dimension normalized Euclidean norm for |y-x| and |y| when they are Q dimensioned vectors together with multivariate kernel estimates of f(w|z)p(z) for f(x) and f(y) raised to an appropriate power value of α, the polarization sensitivity index, which is of course the choice of the investigator.

An Application.

There is a suspicion that the economic reforms (including the one child policy) in China together with the massive urbanization over the period changed fundamentally the nature of urban households. Data on two independent surveys of urban households from three coastal and three interior provinces[8] in China for the years 1987 (for which there were 3651 observations) and 2001 (for which there were 4297 observations) a period over which the reforms took effect. The data were used to generate observations on log adult equivalent household income (at constant prices), adult equivalent[9] living space (in square meters) and an integer index of the education level of the head of household. Thus in this example the household is the agent. Table 1 presents the summary statistics. Considerable increases in both equivalent incomes and living space (due in part to growth and in part to reductions in family size) and educational attainment are evident. To calculate the polarization statistic the continuous multivariate mean standardized pdf's were estimated using a multivariate standard normal kernel with a window width h = $1.06*\sigma(x).*n^{-(1/(4+k))}$ (Silverman (1986)). The seven outcome educational scale was condensed to a three outcome scale, 1, 2 and 3 corresponding to high, medium and low educational attainments. Summary statistics are reported in Table 1, Table 2 reports the polarization indices and standard errors for the continuous univariate measures as per DER, Table 3 reports the paired multivariate measures and Table 4 reports the overall multivariate measures.

[8] The coastal provinces were Jilin, Shandong and Guangdong the interior, Sichuan, Shaanxi and Hubei .

[9] Equivalization was effected using the square root rule (Brady and Barber (1948)).

Table 1. Sumamary Statistics (1987 n=3651, 2001 n = 4297)

	1987	2001	1987	2001	1987	2001
	equivalized log income	equivalized log income at 1987 prices.	equivalized living space (sq meters)	equivalized house space(sq meters)	Education of household head	Education of household head
Mean	4.8227	8.8212	17.1178	24.7200	3.1035	3.4263
Median	4.8579	8.9074	15.2053	22.5167	4.0000	4.0000
Std Dev	0.4194	0.8489	9.4884	12.6218	1.5754	1.6137

Table 2. Univariate Measures (Standard Errors in brackets)

Sensitivity Parameter (α)	Income 1987	Income 2001	Housing 1987	Housing 2001	Edu 1987	Edu 2001
0.25	0.1173 (0.0003)	0.1321 (0.0003)	0.4182 (0.0007)	0.4314 (0.0007)	0.1784 (0.0016)	0.2166 (0.0021)
0.5	0.1524 (0.0003)	0.1661 (0.0003)	0.3634 (0.0004)	0.3745 (0.0004)	0.1339 (0.0011)	0.1641 (0.0015)
0.75	0.2041 (0.0003)	0.2136 (0.0003)	0.3271 (0.0003)	0.3349 (0.0002)	0.1019 (0.0007)	0.1244 (0.0012)
1.0	0.2789 (0.0004)	0.2793 (0.0003)	0.3007 (0.0002)	0.3051 (0.0002)	0.0785 (0.0005)	0.0942 (0.0009)

Table 2a. Univariate Polarization tests [H_0: Pol_{1987}-$Pol_{2001} \geq 0$, “t”, (P(t < “t”))]

Sensitivity Parameter (α)	Income	Housing	Education
0.25	-34.5502 (0.0)	-13.4451 (0.0)	-14.0846 (0.0)
0.5	-32.8214 (0.0)	-20.8722 (0.0)	-15.6281 (0.0)
0.75	-20.5224 (0.0)	-21.5707 (0.0)	-16.0812 (0.0)
1.0	-0.7099 (0.2389)	-15.9004 (0.0)	-15.3907 (0.0)

For all values of the polarization sensitivity parameter the index shows increased for all income, house space and education variables and, based upon the samples in the two years being independent of one another, the increase is seldom insignificant at usual levels of significance. The joint pair-wise distributions exhibit quite different effects to the univariate cases. At low levels of polarization sensitivity significant polarization is still the norm for all pair-wise comparisons with depolarization being the norm in almost all cases and significantly so at higher orders of polarization sensitivity.

Table 3 Multivariate Polarization Paired Comparisons

Sensitivity Parameter (α)	Income and Housing 1987	Income and Housing 2001	Income and Edu 1987	Income and Edu 2001	Housing and Edu 1987	Housing and Edu 2001
0.25	0.4136 (0.0011)	0.4170 (0.0007)	0.2984 (0.0007)	0.3325 (0.0009)	0.3578 (0.0008)	0.3697 (0.0007)
0.5	0.4900 (0.0011)	0.4735 (0.0007)	0.3146 (0.0007)	0.3310 (0.0008)	0.2574 (0.0006)	0.2545 (0.0005)
0.75	0.6122 (0.0011)	0.5600 (0.0008)	0.3419 (0.0006)	0.3340 (0.0008)	0.1904 (0.0005)	0.1778 (0.0003)
1.0	0.7913 (0.0012)	0.6810 (0.0009)	0.3806 (0.0006)	0.3407 (0.0008)	0.1439 (0.0003)	0.1255 (0.0002)

Table 3a. Bivariate Polarization tests [H_0: Pol_{1987}-$Pol_{2001} \geq 0$, “t”, (P(t < “t”))]

Sensitivity Parameter (α)	Income and Housing	Income and Education	Housing and Education
0.25	-2.5810 (0.0049)	-29.4864 (0.0000)	-10.9553 (0.0000)
0.5	12.4364 (1.0000)	-15.2877 (0.0000)	3.7932 (0.9999)
0.75	37.2647 (1.0000)	7.5415 (1.0000)	21.7465 (1.0000)
1.0	73.1129 (1.0000)	38.5854 (1.0000)	41.5799 (1.0000)

Table 4. Multivariate Measures (Standard Errors in brackets)

Sensitivity Parameter α	Income,Housing and Edu 1987	Income,Housing and Edu 2001	Polarization test [H_0: Pol_{1987}-$Pol_{2001} \geq 0$, “t”, (P(t < “t”))]
0.25	0.3993 (0.0010)	0.3994 (0.0009)	-0.0616 (0.4754)
0.5	0.3932 (0.0010)	0.3602 (0.0009)	25.1768 (1.0000)
0.75	0.4043 (0.0010)	0.3332 (0.0009)	53.2727 (1.0000)
1.0	0.4295 (0.0010)	0.3141 (0.0009)	83.9697 (1.0000)

Turning to the polarization measures across all three characteristics which, together with a test for depolarization, are reported in Table 4 note that the null of depolarization is never be rejected for all levels of polarization sensitivity.

Conclusions.

Many researchers have argued that, in the absence of a plausible aggregator of the many factors that affect wellbeing, its measurement needs to be pursued in the context of the several variables available rather than relying on just one of them. This applies to most aspects of wellbeing measurement. Here, by combining multivariate versions of the Polarization indices developed in Esteban and Ray (1994) and Duclos Esteban and Ray (2004) the polarization measurement toolkit has been extended to the case where the status of an agent is represented by many characteristics which can be both discretely and continuously measured and the agent subgroups in a population are not identified. The asymptotic variance of the statistic has been provided to facilitate inference.

As an example the statistic was applied to Data on two independent surveys of urban households from three coastal and three interior provinces in China for the years 1987 and 2001, a period over which the reforms took effect. The data reflected the log adult equivalent household income (at constant prices), adult equivalent living space (in square meters) and an integer index of the education level of the head of household each of which may be construed as contribution to the wellbeing of the household.

The results, while obviously specific to these particular data, were salutary with regard to the use of univariate as opposed to multivariate polarization indices. While the individual univariate indices all reflected significant increases in polarization between households over the period of the reforms, when they were combined the polarization result was attenuated. For pair-wise combinations of the variables significant polarization was detected at low levels of polarization sensitivity but at high levels of polarization sensitivity significant depolarization was detected. When all three variables were combined in an index, significant depolarization was detected at all levels of polarization sensitivity.

References.

Anderson, G.J. (2004) “Toward an Empirical Analysis of Polarization” Journal of Econometrics 122 1-26

Anderson, G.J. (2008) “The Empirical Assessment of Multidimensional Welfare, Inequality and Poverty: Sample Weighted Generalizations of the Kolmorogorov-Smirnov Two Sample Test for Stochastic Dominance.” Journal of Economic Inequality 6 73-87.

Anderson G.J. (2010) “Polarization of the Poor: Multivariate Relative Poverty Measurement Sans Frontiers” Forthcoming Review of Income and Wealth

Anderson G.J., I. Crawford and A Leicester (2011) “Welfare rankings from multivariate data, a nonparametric approach” forthcoming Journal of Public Economics

Brady D.S. and H.A. Barber (1948) “The Pattern of Food Expenditures” *Review of Economics and Statistics* 30 198-206.

Duclos J-Y, J.Esteban and D. Ray (2004) “Polarization: Concepts, Measurement, Estimation” Econometrica 72 pp. 1737-1772.

Duclos J-Y., Sahn, D., and S.D. Younger (2006) “Robust Multidimensional Poverty Comparisons” The Economic Journal 116 943-968.

Esteban, J.-M. and Ray, D. (1994). “On the measurement of polarization.” Econometrica 62, 819–51.

Gigliarano C. and K. Mosler “Constructing Indices of Multivariate Polarization.” Journal of Economic Inequality 2009.

Grusky, D., and R. Kanbur (eds.) (2006): Poverty and Inequality. Stanford University Press: Stanford, California.

Koshevoy G.A. and K.Mosler (1997) “Multivariate Gini Indices” Journal of Multivariate Analysis 60 252-276.

Maasoumi E. (1986) “The Measurement and Decomposition of Multidimensional Inequality” Econometrica 54 771-779

Maasoumi E. (1999) “Multidimensioned Approaches to Welfare Analysis” chapter 15 Handbook of Income Inequality Measurement J Silber editor Kluwer Academic Publishers Boston.

Sen, A. K. (1992): Inequality Reexamined. Harvard University Press: Cambridge, Massachusetts.

Silverman B.W. (1986) Density estimation for Statistics and Data Analysis. Chapman Hall.

Tsui K-Y (1995) “Multidimensional Generalizations of the Relative and Absolute Inequality Indices: The Atkinson-Kolm-Sen Approach. Journal of Economic Theory 67 pp 251-265.