University of Toronto
Department of Economics

Working Paper 414

Polarization Measurement and Inference in Many Dimensions: A Note

By Gordon Anderson

November 05, 2010

Polarization Measurement and Inference in Many Dimensions: A Note

Gordon Anderson

Department of Economics.
University of Toronto

October 2010

Very preliminary

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

Introduction.

The functionings and capabilities approach to wellbeing measurement (Sen (1992)) has given considerable impetus to multidimensional analyses (Grusky and Kanbur (2006)). 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) Anderson (2008)). While Anderson (2008) and Gigliarano and Mosler (2008) provide multivariate bi-polarization measures for two identified groups, multivariate polarization measures have not been developed for the more general non-identified many group case. Here the most popular general polarization index (Esteban and Ray (1994), Duclos, Esteban and Ray (2004)) is extended to many discrete and continuous dimensions. Its implementation 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 Extension.

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[1] is given by:

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

Which is asymptotically normally distributed with an asymptotic variance V given by:

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

Development of the polarization index was founded on a set of axioms that such an index should obey, the axioms[2] 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 $f(x)^{\alpha+1}$). 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.

---

[1] A similar discrete variable index is provided in Esteban and Ray (1994).

[2] Essentially concentrations (squeezes) of the sub distributions and increases in the distance between them should not reduce the polarization measure.

[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.

Let w_i and z_i be jointly distributed vectors describing the status of the i'th agent with w_i having dimension k x 1 and z_i having dimension h x 1 with i =1,..,n being the elements of the sample, where w are continuous variables and z are discrete variables. 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[4]. 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 $w_i | z_i$ 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 $x_{\{c\}}$.

From the above, a multivariate version of [1] is given by:

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

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 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. 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[5]:

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

The multivariate version of [2] is given by:

$$V_\alpha = \mathrm{var}_{f(y)} \left(\begin{array}{l} (1+\alpha)(f(w \mid z)P(z))^\alpha \sum_z \int_0^\infty \| y - x \| dF(w \mid z)P(z) + \\ \| y \| \sum_z \int_0^\infty (f(w \mid z)P(z))^\alpha dF(w \mid z)P(z) + 2 \sum_z \int_y \| y - x \| (f(w \mid z)P(z))^\alpha dF(w \mid z)P(z) \end{array} \right) \quad [2a]$$

---

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

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

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:

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

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

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

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).

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[6] 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[7] living space (in square meters) and an integer index of the education level of the head of household. 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. Table 2 reports the polarization indices and standard errors for the continuous univariate measures as per DER and Table 3 reports the multivariate measures.

---

[6] The coastal provinces were Jilin, Shandong and Guangdong the interior, Sichuan, Shaanxi and Hubei .
[7] 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	4.8227503	8.8212479	17.117842	24.720038	3.1035333	3.4263440
Median	4.8578656	8.9074170	15.205262	22.516660	4.0000000	4.0000000
Std Dev	0.4194296	0.8489056	9.4884068	12.621758	1.5753986	1.6137025

Table 2. Univariate Measures (Standard Errors in brackets)

Sensitivity Parameter (α)	Income 1987	Income 2001	Housing 1987	Housing 2001
0.25	0.11731620 (0.00031190748)	0.13209502 (0.00029271751)	0.41820690 (0.00072261910)	0.43142905 (0.00066703302)
0.5	0.15241276 (0.00031033800)	0.16609740 (0.00027844521)	0.36340574 (0.00039098378)	0.37452838 (0.00036208622)
0.75	0.20407695 (0.00035301313)	0.21356717 (0.00029870682)	0.32713093 (0.00026857448)	0.33492727 (0.00024186942)
1.0	0.27893908 (0.00043703932)	0.27933600 (0.00034867352)	0.30070666 (0.00020994170)	0.30512450 (0.00018199596)

For all values of the polarization sensitivity parameter the index shows an increase for both income and house space variables however, based upon the samples in the two years being independent of one another, it is seldom significant at usual levels of significance.

Table 3. Multivariate Measures (Standard Errors in brackets)

Sensitivity Parameter α	Income and Housing 1987	Income and housing 2001	Income,Housing and Edu 1987	Income,Housing and Edu 2001
0.25	0.41364256 (0.001067345)	0.41695109 (0.00070989495)	0.39928822 (0.00096876272)	0.39936785 (0.00085576967)
0.5	0.48998936 (0.001102657)	0.47347036 (0.00074058380)	0.39315119 (0.00097944440)	0.36022864 (0.00086639805)
0.75	0.61216969 (0.0011584829)	0.55999669 (0.00078619134)	0.40427540 (0.00099784998)	0.33319929 (0.00088564705)
1.0	0.79130432 (0.0012435013)	0.68103002 (0.00085357946)	0.42945411 (0.0010262811)	0.31408093 (0.00091355524)

The joint continuous distributions exhibit different effects to the univariate cases with depolarization being the norm in almost all cases and significantly so at higher orders of polarization sensitivity. This reflects the homogenization of urban households over the period of the reforms.

References.

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

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." Forthcoming Journal of Economic Inequality 2008.

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.