A general class of social distance measure: Algebraic derivation of existing measures within the $DIST$ class

In this appendix we provide algebraic proofs for the equivalence of existing measures of diversity and disparity within the $DIST$ general class. For general notation purposes:

- p_i designates the proportion of the total population in the ith groups
- d_{ij} designates the social distance between the ith and jth groups in the population comparison matrix $\mathbf{D}$
- y_i designates the income (or other continuous attribute) of group i, where the group is defined solely by that attribute (i.e. the within-group variance in income is by definition 0)
- r_i designates the income (or other continuous attribute) of groups i relative to the overall population average
- $\bar{r}_i$ designates the *average* income (or other continuous attribute) of group i relative to the overall population average where the group is defined by attributes in addition to y (i.e. where the within-group variance in income is ≥ 0)

1 Demographic fractionalization

The demographic fractionalization index—which we have termed $FRAC$—is given by the equation $FRAC = \sum_{i=1}^{n} p_i(1-p_i)$. It is expressed in the $DIST$ class as $DIST(1, 1, p_i p_j, \mathbf{D}^*)$.

$$DIST(1, 1, p_i p_j, \mathbf{D}^*) = 1 \cdot \left[\sum_{i=1}^{m} \sum_{j=1}^{m} q_{ij} d_{ij} \right] \tag{1}$$

where

$$\begin{aligned} q_{ij} &= p_i p_j \\ d_{ij} &= \begin{cases} 0 & \text{if} i = j \\ 1 & \text{if} i \neq j \end{cases} \end{aligned}$$

and, hence

$$q_{ij} d_{ij} = \begin{cases} 0 & \text{if} i = j \\ p_i p_j & \text{if} i \neq j \end{cases} \tag{2}$$

Substituting (2) back into (1) gives:

$$\begin{aligned} DIST(1, 1, p_i p_j, \mathbf{D}^*) &= 1 \sum_{i=1}^{m} \sum_{j \neq i} p_i p_j \\ &= \sum_{i=1}^{m} p_i \sum_{j \neq i} p_j \qquad (3) \\ &= \sum_{i=1}^{m} p_i (1 - p_i) \qquad (4) \\ &= FRAC \end{aligned}$$

where (4) follows from (3) because by definition $\sum_j p_j = 1$ and hence $\sum_{j \neq i} p_j = 1 - p_i$.

2 Demographic polarization

The demographic measure of polarisation development by Montalvo and Reynal-Querol, which we term MRQ, is given by the equation $MRQ = 1 - \sum_{i=1}^{n} \left(\frac{0.5-p_i}{0.5}\right)^2 = 4\sum_{i=1}^{n} p_i^2(1-p_i)$. This can be expressed in the $DIST$ class as $DIST(4, 1, p_1^2 p_j.\mathbf{D}^*)$. The proof of this follows similar logic to that of the demographic fractionalization index given above. The expression $q_{ij}d_{ij}$ in MRQ can be written as

$$q_{ij}d_{ij} = \begin{cases} 0 & \text{if} i = j \\ p_i^2 p_j & \text{if} i \neq j \end{cases} \tag{5}$$

Hence:

$$\begin{aligned} DIST(4, 1, p_i^2 p_j, \mathbf{D}^*) &= 4\sum_{i=1}^{m}\sum_{j\neq i} p_i^2 p_j \\ &= 4\sum_{i=1}^{m} p_i^2 \sum_{j\neq i} p_j \\ &= 4\sum_{i=1}^{m} p_i^2(1-p_i) \\ &= MRQ \end{aligned}$$

3 Gini index

There are various ways of representing and interpreting the Gini coefficient that measures 'vertical' inequality, but the most useful for our purposes here is as the 'relative mean difference'—the average difference in income (education, etc.) between every possible pair of individuals, divided by the overall mean income. Algebraically, where each individual has income (or educational attainment, etc.) y_i and the overall mean income (or educational attainment, etc.) is μ:

$$GINI = \frac{1}{2n^2\mu}\sum_{i=1}^{n}\sum_{j=1}^{n}|y_i - y_j|$$

If the variable is discrete (e.g. education years) or contiuous but clustered, it is possible to rewrite the Gini coefficient as:

$$GINI = \frac{1}{2\mu}\sum_{i=1}^{m}\sum_{j=1}^{m} p_i p_j |y_i - y_j|$$

Replacing the mean of each group y_i with the relative group income $r_i = y_i/\mu$ allows us to write:

$$GINI = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} p_i p_j |r_i - r_j|$$

Clearly, this fits the generalized form with $GINI = DIST(1/2, 1, p_i p_j, |r_i - r_j|)$. The equation for the 'horizontal' group-based Gini index $GINI^G$ is of the same form as $GINI$ but, as noted in the main text, groups are defined by attributes in addition to income. Replacing the relative group income r_i with the relative mean group income $\bar{r}_i$ allows us to express $GINI^G$ in $DIST$ form as

$$GINI^G = DIST(1/2, 1, p_i p_j, |\bar{r}_i - \bar{r}_j|)$$

4 Economic polarization

The economic polarization index development by Esteban and Ray likewise has a similar form as $DIST$. Their measure is given by the formula

$$ER = k\sum_{i=1}^{n}\sum_{j=1}^{n} p_i^{1+a} p_j |y_i - y_j|$$

for $k > 0$ and $a \in (0, a^*]$ where $a^* \simeq 1.6$. Like the Gini coefficient, the ER measure is invariant to overall population means if the normalizing constant k takes the form $k' \cdot 1/\mu$. For our purposes, it is hence useful to assume that k takes this form and replace the y_is in the equation with their value relative to the population mean, r_i. This form of the ER measure fits as a sub-class of $DIST$ where $ER(k, a) = DIST(k, 1, p_i^{1+a} p_j, |r_i - r_j|)$. In this form, ER measures 'vertical' polarization, where the groups are defined by the attribute across

which they are to be compared (income, etc.). Where groups are defined by additional attributes, following the same logic as for $GINI^G$ above, we can write a 'horizontal' version of ER with relative group income r_i replaced by relative group mean income $\bar{r}_i$, such that $ER^G(k,a) = DIST(k, 1, p_i^{1+a} p_j, |\bar{r}_i - \bar{r}_j|)$.

5 Group coefficient of variation

The derivation of the group coefficient of variation $GCOV$ in $DIST$ format is somewhat more complex, but all the more fun for that. The formula for $GCOV$ is given by

$$GCOV = \left(\sum_{i=1}^{n} p_i (\bar{r}_i - 1)^2 \right)^{1/2}$$

This can also be written in the $DIST$ format as

$$GCOV = DIST(1/2, 1/2, p_i p_j, (\bar{r}_i - \bar{r}_j)^2)$$

The equivalance is easiest to show taking the squared versions of each side thus:

$$\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} p_i p_j (\bar{r}_i - \bar{r}_j)^2 = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} (p_i p_j \bar{r}_i^2 - 2 p_i p_j \bar{r}_i \bar{r}_j + p_i p_j \bar{r}_j^2) \quad (6)$$

$$= \frac{1}{2} \left[\sum_{i=1}^{m} \sum_{j=1}^{m} p_i p_j \bar{r}_i^2 - 2 \sum_{i=1}^{m} \sum_{j=1}^{m} p_i p_j \bar{r}_i \bar{r}_j + \sum_{i=1}^{m} \sum_{j=1}^{m} p_i p_j \bar{r}_j^2 \right] \quad (7)$$

$$= \frac{1}{2} \left[\sum_{i=1}^{m} p_i \bar{r}_i^2 \sum_{j=1}^{m} p_j - 2 \sum_{i=1}^{m} p_i \bar{r}_i \sum_{j=1}^{m} p_j \bar{r}_j + \sum_{i=1}^{m} p_i \sum_{j=1}^{m} p_j \bar{r}_j^2 \right] \quad (8)$$

Now, by definition the total of the group proportions is 1 and the population-weighted sum of the mean relative distances $\bar{r}$ is 1, hence

$$\sum_{i=1}^{m} p_i = \sum_{j=1}^{m} p_j = \sum_{i=1}^{m} p_i \bar{r}_i = \sum_{j=1}^{m} p_j \bar{r}_j = 1 \quad (9)$$

Substituting (9) appropriately into (8) gives

$$\frac{1}{2} \left[\sum_{i=1}^{m} p_i \bar{r}_i^2 - 2 + \sum_{j=1}^{m} p_j \bar{r}_j^2 \right] = \frac{1}{2} \sum_{i=1}^{m} p_i \bar{r}_i^2 + \frac{1}{2} \sum_{j=1}^{m} p_j \bar{r}_j^2 - 1 \quad (10)$$

$$= \sum_{i=1}^{m} p_i \bar{r}_i^2 - 1 \tag{11}$$

where (11) follows from (10) because the summation terms are no longer nested and hence the is and the js can be taken as equivalent—each summation term simply rotates through each population group performing the same operation. Now from (9) we can observe

$$\sum_{i=1}^{m} p_i - \sum_{i=1}^{m} p_i \bar{r}_i = 0 \tag{12}$$

Adding (12) twice to (11) and then substituting in (9) appropriately gives

$$\begin{aligned}
\sum_{i=1}^{m} p_i \bar{r}_i^2 - 1 &= \sum_{i=1}^{m} p_i \bar{r}_i^2 - 1 + 2\sum_{i=1}^{m} p_i - 2\sum_{i=1}^{m} p_i \bar{r}_i && (13)\\
&= \sum_{i=1}^{m} p_i \bar{r}_i^2 - \sum_{i=1}^{m} p_i + 2\sum_{i=1}^{m} p_i - 2\sum_{i=1}^{m} p_i \bar{r}_i && (14)\\
&= \sum_{i=1}^{m} p_i \bar{r}_i^2 - \sum_{i=1}^{m} 2p_i \bar{r}_i + \sum_{i=1}^{m} p_i && (15)\\
&= \sum_{i=1}^{m} p_i (\bar{r}_i^2 - 2\bar{r}_i + 1) && (16)\\
&= \sum_{i=1}^{m} p_i (\bar{r}_i - 1)^2 && (17)\\
&= GCOV^2 && (18)
\end{aligned}$$