This paper is a draft submission to

Inequality—Measurement, trends, impacts, and policies

5–6 September 2014 Helsinki, Finland

Measuring Inequality by Asset Indices: The case of South Africa

Martin Wittenberg and Murray Leibbrandt
University of Cape Town

April 2014

Abstract

Asset indices have become widely used in a number of areas of social research, particularly in the analysis of Demographic and Health Surveys. Indeed the calculation of "wealth indexes" is now routine practice in the DHSs. Asset indices have been externally validated in a number of contexts. While these indices have been shown to work well as proxy measures of poverty, they are not suited to investigate inequality. In this paper we will show that, in fact, typical asset indices also fail an internal validity test: they frequently rank individuals in ways which violate the basic principle that individuals that have more (of anything) should be ranked higher than individuals that have less. We consider from first principle what sort of indexes might make sense, given the predominantly dummy variable nature of asset schedules. We show that there is, in fact, a way to construct an asset index which does not violate some basic principles and which also has the virtue that it can be used to construct "asset inequality" measures. When employing this measure on South African data we find that overall asset inequality has decreased markedly between 1993 and 2008. This contrasts with findings derived from income data which suggest that inequality has hardly changed at all.

1 Introduction

Asset indices have become widely used since Filmer and Pritchett (2001) described a simple way to calculate them. Their use really took off once the Demographic and Health Surveys incorporated the calculation of a "wealth index" with the release of each dataset (Rutstein and Johnson 2004). A recent Google Scholar search (April 18 2014) came up with 13 900 "hits" on "DHS wealth index"[1], 2 434 citations of the Filmer and Pritchett (2001) article and 591 citations of the Rutstein and Johnson (2004) paper documenting the creation of the DHS index. The main use of the indices in this vast literature is in creating wealth rankings, separating the "rich" from the "poor" as ingredients for more substantive analyses.

Several articles, including the original Filmer and Pritchett (2001) piece, have tried to validate these indices against **external** criteria, e.g. incomes or expenditures. A recent review (Filmer and Scott 2012) concludes that "the use of an asset index can clearly provide useful guidance to the order of magnitude of rich-poor differentials" (p.389), although it cautions that the asset indices measure a different concept than per capita consumption. Indeed the paper devotes attention to

[1] Many of them will obviously refer to different versions of the same paper.

the question under which circumstances the two measures will provide the most similar rankings, arguing that this will occur when per capita expenditures are well explained by observed household and community characteristics and when "public goods" are more important in household expenditures than "private ones" like food. In other work we have ourselves argued that asset indices do a good job of proxying for income differences (Wittenberg 2009, Wittenberg 2011).

None of this literature has examined whether the asset indices calculated in the traditional way make sense **internally**, i.e. according to a number of simple criteria such as that individuals that have more (of anything) should be ranked higher than individuals that have less. In particular little attention has been paid to the problems created by the predominantly dummy variable nature of asset schedules. We show that this is not just a theoretical issue but that DHS wealth indices in a number of cases exhibit anomalous rankings.

One additional issue that has been lamented in some contexts is that the way in which these indices are typically calculated precludes the use of traditional inequality measures. One might think that if it makes sense to talk about inequality in incomes or wealth that it would certainly make sense to think about inequality in asset holdings (Bhorat and van der Westhuizen 2013). Nevertheless manipulating traditional indices is not a viable strategy (Wittenberg 2013), a different approach is needed. As we show below, it is when we consider the particular problems of calculating inequality measures with dummy variables that many problems with the creation of asset indices crystallise. However we show that these problems are not insuperable. Indeed an approach due to Banerjee (2010) for dealing with multidimensional inequality can be used to create such asset indices, as we will show below.

We show that this approach is easy to implement and apply it to South African data. This provides a new perspective on the evolution of South African inequality which is somewhat at odds with the literature measuring inequality with money-metric approaches. We think it is likely that the asset approach reveals genuine improvements over time, although the reduction in inequality is unlikely to be as dramatic as the Gini coefficients calculated on the asset indices suggest. We think that more detailed asset inventories would moderate some of the conclusions. Indeed one of our key points is that asset indices need to be approached with some caution – churning out "wealth indices" in semi-automated ways without considering in detail what the individual scores suggests, is likely to be problematic.

The plan of the paper is as follows. In Section 2 we provide a very brief overview of the theoretical literature dealing with asset indices. We follow by enunciating several principles for the creation of such indices in section 3. We refer to these as "principles" since our approach is not fully axiomatic: we are not deriving indices by a process of deduction. Our approach is more heuristic – investigating what happens when we apply different approaches to simple data and considering whether the answers make sense. We do this in sections 4 through 6, where we consider first the case of a single binary variable and then progressively consider more complicated cases. Having set out what we consider to be a defensible approach, we turn to applying it to DHS data in section 7. Finally we consider what assets may tell us about the evolution of inequality in South Africa from 1993 to 2008.

The chief contributions of our paper to the literature are : (1) it describes how to construct an asset index that is internally coherent; (2) it shows that inequality measures on this index are well-defined and have reasonable interpretations; (3) it provides a fresh perspective on South African inequality; and (4) it provides some perspective on the "art" of index construction.

2 Literature review

McKenzie (2005, p.232) suggests that the idea of using the first principal component of a set of asset variables as an index for "wealth" has been around in the social science literature for a long time. Its use, however, has become common only after publication of the Filmer and Pritchett (2001) paper and the subsequent adoption of the methodology in the release of the DHS "wealth indices" (Rutstein and Johnson 2004). The basic idea of principal components is to find the linear combination of the asset variables that maximisises the variance of this combination. More formally, if we have k random variables $a_1, \ldots, a_k$, each standardised to be of mean zero and variance one, the objective is to rewrite these as

$$\begin{aligned} a_1 &= v_{11}A_1 + v_{12}A_2 + \ldots + v_{1k}A_k \\ a_2 &= v_{21}A_1 + v_{22}A_2 + \ldots + v_{2k}A_k \\ &\vdots \\ a_k &= v_{k1}A_1 + v_{k2}A_2 + \ldots + v_{2k}A_k \end{aligned} \quad (1)$$

where the A_i are unobserved components, created so as to be orthogonal to each other. Writing this in vector notation as

$$\mathbf{a} = \mathbf{VA}$$

it follows that the covariance matrix (here equal to the correlation matrix $\mathbf{R}$) is given by

$$\begin{aligned} E(\mathbf{aa}') &= E\left(\mathbf{VAA}'\mathbf{V}'\right) \\ \mathbf{R} &= \mathbf{V\Phi V}' \end{aligned}$$

where $\mathbf{\Phi} = E\left(\mathbf{AA}'\right)$. Note that $\mathbf{\Phi}$ is diagonal since the unobserved components are assumed to be orthogonal to each other. We need to impose some normalisation in order to get a determinate solution. Let $\mathbf{\Phi}$ be the matrix of eigenvalues, $\mathbf{V}$ the orthonormal matrix of eigenvectors, and assume that $\mathbf{V}$ is ordered so that the eigenvector associated with the largest eigenvalue is listed first. We can then solve for $\mathbf{A}$ to get

$$\mathbf{A} = \mathbf{V}'\mathbf{a}$$

in particular

$$A_1 = v_{11}a_1 + v_{21}a_2 + \ldots + v_{k1}a_k \quad (2)$$

We will refer to this as the PCA index. By assumption $var(A_1) = \lambda_1$, the first eigenvalue, and we can show that no other linear combination of the a_i variables will achieve a greater variance (Wittenberg 2009, pp.5–6).

If the asset variables a_i do not have unit variance and zero mean, they are first standardised, so that the equation for the first principal component will be given by

$$\begin{aligned} A_1 &= v_{11}\left(\frac{a_1 - \overline{a}_1}{s_1}\right) + v_{21}\left(\frac{a_2 - \overline{a}_2}{s_2}\right) + \ldots + v_{k1}\left(\frac{a_k - \overline{a}_k}{s_k}\right) \\ &= \frac{v_{11}}{s_1}a_1 + \frac{v_{21}}{s_2}a_2 + \ldots + \frac{v_{k1}}{s_k}a_k - c \end{aligned} \quad (3)$$

where the coefficients v_{i1} are the elements of the eigenvector $\mathbf{v}_1$ associated with the largest eigenvalue λ_1 of the correlation matrix $\mathbf{R}$ of the a_i variables. The constant c is the weighted sum of the means, which ensures that A_1 has a zero mean.

The use of the first principal component was defended by Filmer and Pritchett (2001) on a "latent variable" interpretation of the equations 1: A_1 is whatever explains most of what is common to $a_1, \alpha_2, \ldots, a_k$ and it makes most sense to think of this as "wealth". Other authors have taken this formulation more seriously and have suggested that other procedures, such as factor analysis, be used to retrieve the common latent variable (Sahn and Stifel 2003). Although the procedure produces a different index than the PCA one, in practice indices calculated by both approaches are highly correlated, particularly since authors using this approach seem to restrict themselves to extracting only one factor and eschew the "orthogonal rotations" that produce arbitrarily many solutions[2].

Reviews of the procedure have focussed on several issues. Firstly if the assets are measured mainly through categorical variables, then the index defined through equation 2 is intrinsically discrete. The more assets and the more integer-valued variables (e.g. number of rooms) are included in the index, the smoother the resulting index will be and the better its potential to differentiate finer gradations of poverty (McKenzie 2005). Secondly, if categorical variables with multiple categories are included (e.g. water access), then the resulting group of dummy variables will be internally negatively correlated with each other in ways that will influence the construction of the index. Some authors have used multiple correspondence analysis instead (Booysen, van der Berg, von Maltitz and du Rand 2008), which deals with the categorical variables more satisfactorily. Unfortunately it cannot accommodate continuous variables. In practice the PCA index is highly correlated with the MCA index also.

A third issue which has received some attention is whether or not the index should include infrastructure variables (like access to water and sanitation). Houweling, Kunst and Mackenbach (2003) tested the PCA index rankings for sensitivity to the assets included. They were concerned about the fact that the infrastructure assets might have independent effects on the outcome of interest, in particular child mortality. They show that the rankings change somewhat as some of the "assets" are stripped out.

Several authors have tried to validate asset indices against external benchmarks. We have already referred to the Filmer and Scott (2012) review article. They found that different techniques for constructing asset indices tended to get results that were highly correlated with each other, but in some cases differing from the rankings implied by per capita consumption. This is not thought to be a problem in principle, since it is possible that assets may be a more reliable indicator of long-run economic well-being. They may also be measured with less error (Filmer and Pritchett 2001, Sahn and Stifel 2003).

One noteworthy finding in the Filmer and Scott (2012) article is that urban-rural differences tend to be more marked when using asset indices than when using per capita expenditure. This may, however, be a result of the fact that many of the household durable goods that make up asset schedules (e.g. televisions, refrigerators) require electricity which tends to be more accessible in urban areas. Indeed we have argued that both principal components and factor analysis will tend to extract an index which is a hybrid of "wealth" and "urbanness" (Wittenberg 2009). This is an issue to which we will return in the empirical analysis.

[2] For a more detailed discussion of the factor analysis approach, see Wittenberg (2009).

3 Principles for the creation of asset indices

Intuitively all the justifications for the creation of an asset index rely on the idea that it a higher number on the index should imply ownership of "more stuff". This is a simple, yet obvious, internal consistency requirement. We shall refer to this as the **monotonicity** principle. In order to outline this more rigorously we first define what we mean by an asset index.

Definition 1 *Let $(a_1, a_2, \ldots, a_k) \in \Re^k$ be a vector of asset holdings. The function $A : \Re^k \to \Re$ defined for all possible asset holdings is called an asset index.*

Typically we will restrict attention to linear asset indices, i.e. indices that can be written in form $A(a_1, a_2, \ldots, a_k) = v_1 a_1 + v_2 a_2 + \ldots + v_k a_k$.

Principle 2 *Let $A(a_1, a_2, \ldots, a_k)$ be an asset index. The asset index is **monotonic** if, and only if*

$$(a_1, a_2, \ldots, a_k) \geq (a_1^*, a_2^*, \ldots, a_k^*) \Longrightarrow A(a_1, a_2, \ldots, a_k) \geq A(a_1^*, a_2^*, \ldots, a_k^*)$$

Note that this is a fairly weak condition. It does, not, for instance, rule out "inferior" assets. For instance if we had an asset schedule that listed different types of stoves: e.g. electric, paraffin, coal or gas, the corresponding "ownership" vectors might be recorded as $(1, 0, 0, 0)$, $(0, 1, 0, 0)$, $(0, 0, 1, 0)$ and $(0, 0, 0, 1)$ respectively. Since none of these vectors is numerically bigger than the other, there is no restriction on how the Asset index should rank them either. However if these are not recorded as mutually exclusive categories, then an individual that owned **both** an electric stove and a gas stove should receive a higher asset index than one that owns only an electric stove.

The second principle that we require is that the index must be ratio-scale, i.e. it must have an **absolute zero.** This is indispensable if we want to calculate inequality measures on the index, since it is not valid to calculate "shares" (required to construct, for instance, the Lorenz curve) if the variable is not ratio-scale. It implies in particular that the index must be able to recognise individuals or households that own nothing.

Principle 3 *Let $A(a_1, a_2, \ldots, a_k)$ be an asset index. The asset index has an **absolute zero** if, and only if*

$$A(0, 0, \ldots, 0) = 0$$

Obviously this principle is violated by all of the current asset indices, except those that simply sum up the number of assets. Nevertheless it is arguable that if the notion of "asset holdings" is to have any meaning it is only in relation to individuals that don't have any. Even for purely ranking exercises, it is conceptually necessary that it makes sense to define the "have-nots" and that they should rank at the bottom.

Assuming that the previous two principles hold, it then makes sense to consider inequality measures on the space of asset index measures. However we will investigate also inequality measures defined directly on vectors of asset holdings.

Principle 4 *We will say that the asset inequality measure I is **robust** if it can be applied to asset vectors of dummy variables as well as to continuous ones.*

Robustness is not a conceptual requirement, but it is desirable nonetheless, given that the asset information is typically dummy variable based. Theoretically there is no reason why one shouldn't construct different types of measures for different types of data. It is, however, much simpler if the approach can accommodate these differences. One big advantage of robust measures is that we know what the measures mean when the underlying data is of the continuous type that standard social welfare accounts treat. When these measures are applied to dummy variables, however, the interpretation becomes more complex. Robustness in this case means that the "standard" and "non-standard" treatments are part of the same continuum, so that if the measurement of the variable were to improve over time, we would only need to tweak our approach rather than switch completely. It is easier to see what this means by turning directly to the simplest case of all.

4 One binary variable

Consider first the case where we have precisely one binary variable, e.g. we know whether or not the respondent owns a television set. Note that in this case the only possible "asset index" is the variable itself. Note also that we cannot analyse these data "from first principles" according to the typical axioms of inequality measurement, since these type of data will not support the "principle of transfers" – it is impossible to take away an asset from person j and give it to person i without them changing places in the distribution. Furthermore such a "trade" (by the principle of anonymity) would leave the distribution precisely unchanged. Furthermore ratio-scale independence doesn't hold either, since rescaling of the variable does not provide a valid asset distribution.

4.1 Standard inequality measures

Furthermore many of the standard inequality measures (e.g. Atkinson indices) will not provide valid answers in the presence of zeroes. Nevertheless some, notably the Gini, do. It is instructive to consider what the Gini of such a variable would measure. Assume that there are n_0 observations with zeroes and n_1 ones. Let the proportion of ones be p, i.e. $p = \frac{n_1}{N}$ where $N = n_0 + n_1$. The Lorenz curve of this distribution is shown in Figure 1. The Gini coefficient is simply $1 - p$[3].

This is not an unattractive choice as a measure of inequality: if everyone has the asset then the Gini is zero; as $p \to 0$, i.e. the asset becomes concentrated in a smaller and smaller group, the index approaches one. It is obvious that given the paucity of information in the binary variable any "measure" of inequality must be, in some sense, a function of p.

There are some alternatives. For instance, the coefficient of variation applied to the binary variable would yield $\sqrt{\frac{1-p}{p}}$. This again yields a measure of zero when $p = 1$, but in this case the index of inequality approaches $+\infty$ as $p \to 0$.

Obviously both measures break down at $p = 0$. Indeed in a world in which nobody has the "asset" it seems hard to define what inequality in the possession of that asset would mean. It is also worth noting that both measures give meaningful results only if the variable records the possession of a "good". If the variable measures a deprivation it should be recoded first.

4.2 The Cowell-Flachaire measures

An alternative to the cardinally based measures is the approach for ordinal variables proposed by Cowell and Flachaire (2012). These require us to measure the status of everyone in the distribution

[3] Wagstaff (2005)provides a discussion of "concentration indices" for the case where the dependent variable is binary. This value for the Gini coefficient is a special case of his more general result.

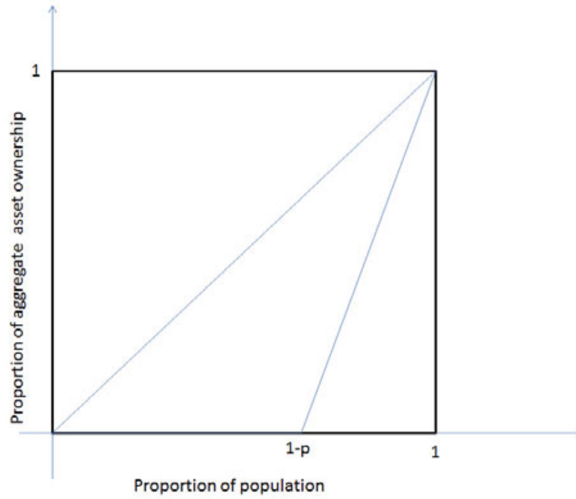

Figure 1: The Lorenz curve of a binary variable

where this is simply the count of everyone of equal rank or lower ("downward" measure) or alternatively everyone of equal rank or higher ("upward" measure). Both are expressed as proportions of the population. The vector of status measures $\mathbf{s} = (s_1, s_2, ..., s_N)$ is then used to calculate an inequality measure, relative to a "reference" status, which Cowell and Flachaire suggest should be set to 1. The inequality measures then become:

$$I_\alpha = \begin{cases} \frac{1}{\alpha(\alpha-1)}\left[\frac{1}{N}\sum_{i=1}^{N} s_i^\alpha - 1\right] & \text{if } \alpha \neq 0, 1 \\ -\frac{1}{N}\sum_{i=1}^{N} \log s_i & \text{if } \alpha = 0 \end{cases}$$

A virtue of this set of measures is that it is invariant to the way in which the ordinal variable is "cardinalised", since the cardinalisation will not affect the ranking of individuals in the distribution.

In the case of our binary variable we get the following status values:

Value	s_i ("downward")	s_i' ("upward")
1	1	p
0	$1-p$	1

Consequently the "downward" measure would be

$$I_\alpha = \begin{cases} \frac{(1-p)}{\alpha(\alpha-1)}\left[(1-p)^\alpha - 1\right] & \text{if } \alpha \neq 0, 1 \\ -(1-p)\log(1-p) & \text{if } \alpha = 0 \end{cases}$$

while the "upward" measure would be

$$I_\alpha = \begin{cases} \frac{p}{\alpha(\alpha-1)}\left[p^\alpha - 1\right] & \text{if } \alpha \neq 0, 1 \\ -p\log p & \text{if } \alpha = 0 \end{cases}$$

In Figure 2 we show the relationship between the "downward" I_α as α changes for three different values of p. We can see that low values of α emphasize inequality at the bottom of the distribution

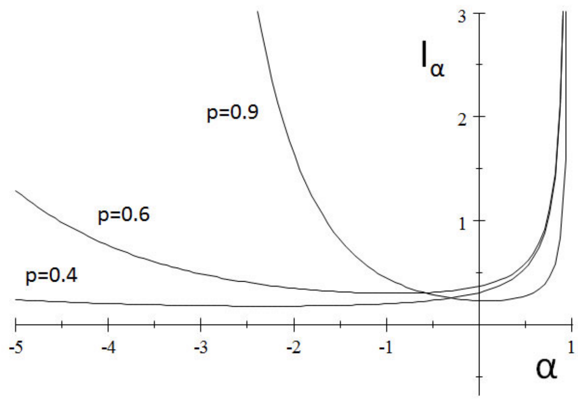

Figure 2: Curves for the Cowell-Flachaire I_α index for different values of α and selected values of p.

– hence the curves with higher proportions of asset owners show more inequality (the deprivation of those without the assets is felt more), while for α values close to one what happens at the top is more accentuated.

The pattern for a fixed value of α is an inverse "U" shaped curve as shown in Figure 3 for the case $\alpha = 0$.

It is worth noting that the variance of the distribution (which is also sometimes used as a measure of inequality) also exhibits this sort of pattern with a low measure of inquality near $p = 0$ and $p = 1$ respectively.

4.3 The meaning of asset inequality

The difference in the behaviour of the two groups of inequality "measures", viz. monotonic decrease in inequality as p goes from near zero to one versus inverse "U" shaped, raises fundamental questions about how we interpret the contrast between the "haves" and the "have nots". In the Gini and C.V. interpretation that gulf is the central feature of the distribution – so if 99% of the population are lacking the asset but 1% have it, that is the most salient fact about the distribution. In the Cowell-Flachaire view if most of the population shares the deprivation, then most outcomes are very similar to each other, i.e. there is not a lot of inequality.

Which of these perspectives is right? Consider a "satisfaction with life" variable that has been measured on a Likert scale ranging from 1 (very dissatisfied) to 5 (very satisfied). Let 99% of the population record a "3" (i.e. neutral) but 1% rate above that. This variable could be dichotomised

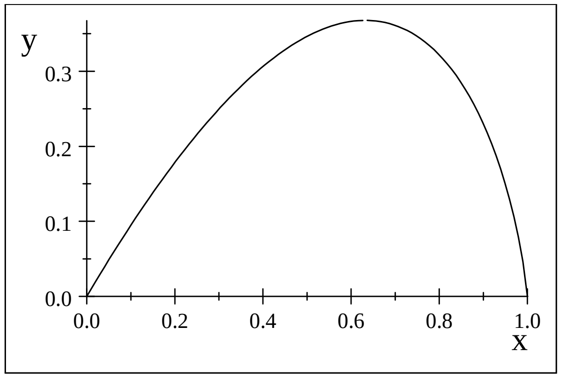

Figure 3: The Cowell-Flachaire I_0 index as a function of p.

as a 0/1 binary variable with the "satisfied" responses scored as 1 while those below recorded as zero. This distribution probably shouldn't rate as very unequal, so in this context the Cowell-Flachaire measure seems more reasonable than the equivalent Gini. Note, however, that the Cowell-Flachaire measure is invariant to linear translation – i.e. we would get the same measure whether 99% respond "3" and 1% "4" or whether 99% answer "4" and 1% "5" or indeed 99% "1" and 1% "5". Indeed the reason why these all give the same distributional measure is that the conversion of the underlying phenomenon into a cardinal measure is arbitrary.

The central question is therefore to what extent the binary variable is an **arbitrary** coding of the underlying distribution. The crucial difference is not so much what the "1" codes for (since that could stand for almost any value), but whether the 0 can be thought of as absolute. Indeed as we noted in the previous section, the Gini coefficient is sensible only if the variable is ratio-scale, i.e. if the zero is absolute.The reason why the Gini scores inequality so highly when p is low is that the gulf between having **nothing** and having **something** is enormous. This is true, however, only if the "0" is really nothing and "1" signals the real possession of an asset (e.g. car). Some of the variables typically used in the construction of asset indices need to be thought about very carefully in this context. For instance a dummy variable for "tiled roof" obviously really measures the presence of absence of a "tiled roof". Nevertheless the absence of a tiled roof does not imply the absence of all roofs; whether or not the gap between owning a thatched roof and a tiled roof is as vast as the gap between having nothing and having something is debatable.

Nonetheless many of the assets do measure material gaps – ownership of a car or of a television are examples. Some infrastructure variables arguably also satisfy this criterion. The presence or absence of water in the house may be such a salient difference that the "0" really denotes a key absence. For variables like these the Gini measure seems closer to our intuition of how we would think about "asset inequality".

We take two points away from this discussion. Firstly one needs to think quite carefully about what variables one wants to include in one's measure of "asset inequality". If the variables in question are, at best, ordinal quality of life measures (e.g "tiled roof"), then the appropriate "inequality measure" needs to be an ordinal one, like the Cowell-Flachaire approach. Secondly, if the binary variable really captures the presence or absence of a real asset, then the behaviour of the Gini coefficient accords more closely with our intuition of "asset inequality". Nevertheless we accept that this is a judgement issue and that different analysts might come to different conclusions. Our judgement call is embedded in Principle 3, which comes down on the "cardinal" rather than "ordinal" approach to inequality.

5 Two binary variables

We turn now to consider the case where we have two binary variables. We could obviously analyse both variables separately, but we might want to combine the information to arrive at some overall measure of "asset inequality". There are several potential ways of doing this. Firstly we could combine the two variables into one scale (an "asset index") and then apply some inequality measure to that scale. Depending on whether we think of the scale as giving us cardinal or ordinal values we could use either a standard inequality measure or the Cowell-Flachaire ordinal measures. Secondly we could utilise some of the approaches in the "multidimensional inequality" literature.

5.1 Some preliminaries

First, however, we will rehearse some of the issues that make the two variable case more complicated. To make the discussion more precise, let us presume that the empirical information on the two binary variables X_1 and X_2 is contained in the following matrix

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{n_1} & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} \\ \mathbf{0}_{n_3} & \mathbf{1}_{n_3} \\ \mathbf{1}_{n_4} & \mathbf{1}_{n_4} \end{bmatrix} \tag{4}$$

where $\mathbf{0}_{n_j}$ is the n_j null vector $[0, 0, ..., 0]'$ and $\mathbf{1}_{n_k}$ is the n_k vector of ones $[1, 1, ..., 1]'$. Let $N = \sum_j n_j$, and $p_1 = \frac{n_2+n_4}{N}$, $p_2 = \frac{n_3+n_4}{N}$ and $p_{12} = \frac{n_4}{N}$. Without loss of generality let us assume that $p_1 \geq p_2$.

Besides the general case we will also consider the polar cases:

- Special case 1:

  $n_2 = n_3 = 0$, i.e. $\mathbf{x}_1 = \mathbf{x}_2 = \begin{bmatrix} \mathbf{0}_{n_1} & \mathbf{1}_{n_4} \end{bmatrix}'$ with $p_1 = p_2 = p_{12}$; and

- Special case 2:

  $n_1 = n_4 = 0$ in which case $\mathbf{x}_1 = \mathbf{1} - \mathbf{x}_2$ with $p_1 = 1 - p_2$, and $p_{12} = 0$.

What distinguishes the cases is that the correlation between the two variables is positive in the former, while it is negative in the latter. The literature on multidimensional inequality measurement speaks about a "correlation increasing majorization" (e.g. Tsui 1999, p.150). Intuitively the second case, in which everyone has an asset should be less unequal than the first in which some people have nothing and some have everything. In general we would like a measure of inequality that is true to that intuition.

We turn now to the first method, that of combining the two variables into one scale.

Table 1: Value of Principal Components Asset Index for different Cases

x_1, x_2	General case $p_{12} \geq p_1 p_2$	$p_{12} < p_1 p_2$	Special case 1 $p_1 = p_2 = p_{12}$	Special case 2 $p_1 = 1 - p_2$, $p_{12} = 0$
0,0	$-\sqrt{\frac{p_1}{2(1-p_1)}} - \sqrt{\frac{p_2}{2(1-p_2)}}$	$\sqrt{\frac{p_1}{2(1-p_1)}} - \sqrt{\frac{p_2}{2(1-p_2)}}$	$-\sqrt{\frac{2p_1}{(1-p_1)}}$	n.a.
1,0	$\sqrt{\frac{1-p_1}{2p_1}} - \sqrt{\frac{p_2}{2(1-p_2)}}$	$-\sqrt{\frac{1-p_1}{2p_1}} - \sqrt{\frac{p_2}{2(1-p_2)}}$	n.a.	$-\sqrt{\frac{2(1-p_1)}{p_1}}$
0,1	$-\sqrt{\frac{p_1}{2(1-p_1)}} + \sqrt{\frac{1-p_2}{2p_2}}$	$\sqrt{\frac{p_1}{2(1-p_1)}} + \sqrt{\frac{1-p_2}{2p_2}}$	n.a.	$\sqrt{\frac{2p_1}{(1-p_1)}}$
1,1	$\sqrt{\frac{1-p_1}{2p_1}} + \sqrt{\frac{1-p_2}{2p_2}}$	$-\sqrt{\frac{1-p_1}{2p_1}} + \sqrt{\frac{1-p_2}{2p_2}}$	$\sqrt{\frac{2(1-p_1)}{p_1}}$	n.a.

5.2 Creating an asset index

As noted above, one of the most common ways of creating an asset index is by means of principal components. Applying the PCA formula mechanically the variables are demeaned and divided by the standard deviation. Let the resultant data matrix of standardised variables be $\widetilde{\mathbf{X}}$, i.e.

$$\widetilde{\mathbf{X}} = \begin{bmatrix} -\sqrt{\frac{p_1}{1-p_1}}_{n_1} & -\sqrt{\frac{p_2}{1-p_2}}_{n_1} \\ \sqrt{\frac{1-p_1}{p_1}}_{n_2} & -\sqrt{\frac{p_2}{1-p_2}}_{n_2} \\ -\sqrt{\frac{p_1}{1-p_1}}_{n_3} & \sqrt{\frac{1-p_2}{p_2}}_{n_3} \\ \sqrt{\frac{1-p_1}{p_1}}_{n_4} & \sqrt{\frac{1-p_2}{p_2}}_{n_4} \end{bmatrix}$$

Then the correlation matrix is given by:

$$\frac{1}{N}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} = \begin{bmatrix} 1 & \frac{p_{12}-p_1p_2}{\sqrt{p_1p_2(1-p_1)(1-p_2)}} \\ \frac{p_{12}-p_1p_2}{\sqrt{p_1p_2(1-p_1)(1-p_2)}} & 1 \end{bmatrix} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

The maximal eigenvalue of the matrix is $1+|\rho|$ with associated (normalised) eigenvector of $\frac{1}{\sqrt{2}} \begin{bmatrix} \operatorname{sgn}(\rho) & 1 \end{bmatrix}$, where $\operatorname{sgn}(\rho)$ is 1 or -1 depending on the sign of ρ[4]. Table 1 reports the resulting asset index values.

Several points flow from that table. Firstly, since the mean of the variables (by construction) is zero and they include positive and negative values we cannot use traditional inequality measures on these values. Secondly, the range of the index is a function of the ranges of the standardised variables $\widetilde{x}_1$ and $\widetilde{x}_2$. Those are of the form $\sqrt{\frac{1-p_1}{p_1}} + \sqrt{\frac{p_1}{1-p_1}}$. It is evident that these are unbounded near zero and one and follow a "U" shape, with minimum at $p_1 = \frac{1}{2}$. As an "inequality statistic" the range (and hence dispersion) of the asset index therefore works **inversely** to the Cowell-Flachaire statistic for the univariate case. It is unlikely to communicate useful information about real inequality in the distribution of assets.

The third point emerges from the fact that the "weight" assigned to asset 1 is $\operatorname{sgn}(\rho)$, which is negative if $p_{12} < p_1 p_2$. Indeed looking at the second column of Table 1 we observe that whenever

[4] When $\rho = 0$ $\operatorname{sgn}(\rho) = 0$. In this case there are two equally appropriate eigenvectors, i.e. $\begin{bmatrix} 0 & 1 \end{bmatrix}$ and $\begin{bmatrix} 1 & 0 \end{bmatrix}$ respectively. Note that $-\mathbf{v}$ would obviously also be a normalised eigenvector. We'll pick the sign of $\mathbf{v}$ so that the value at $x_1 = 1$ and $x_2 = 1$ is positive.

Table 2: An example where conventional asset indices produce perverse rankings

Example: $p_1 = 0.7$, $p_2 = 0.4$, $p_{12} = 0.2$, $p_{12} < p_1 p_2$

Assets	proportion	PCA	FA	MCA
$(1,0)$	0.5	-1.0403	-0.53318	-0.89322
$(1,1)$	0.2	0.40312	0.20661	0.34613
$(0,0)$	0.1	0.50277	0.25769	0.43170
$(0,1)$	0.2	1.9461	0.99748	1.6711

$p_{12} < p_1 p_2$ the asset scores give the following ranking: $(1,0) \prec (0,0)$ and $(1,1) \prec (0,1)$, i.e. a person who has the more common asset is always ranked **below** a person who does not have the asset. How can this possibly make sense? The problem arises from the fact that the principal components analysis correctly isolates the negative correlation between the two assets. But the PCA procedure is intended to isolate what is common to both; this quandary is resolved by interpreting x_1 as a "bad" instead of a genuine asset. Given the philosophy of the PCA approach this is understandable, but it is problematic in this context nonetheless.

Indeed it is not difficult to construct examples where the first asset becomes such an intense "bad" that a person having **no** assets gets a higher score than an individual with **both** assets. We show one example in Table 2. Indeed whenever $1 - p_1 < p_2 < p_1$ the PCA rankings will produce such a perverse outcome. Table 1 also shows that the principal components method is not unique in this regard: the most popular alternatives, viz. factor analysis (with one factor) and multiple correspondence analysis produce precisely the same perverse ranking. Indeed **any** "latent variable" approach to index construction is forced in this direction, since that approach can make sense of the negative correlation between the assets only if one of them is positively correlated with the latent variable and the other negatively so.

Is this case relevant for empirical analyses? There are, in fact, many practical examples where "assets" acquire negative weights in principal components procedures. In the South African case (as shown below) ownership of cattle is frequently negatively correlated with the ownership of other asset types, mainly because cattle are a typically "rural" asset while the other assets require a connection to the electricity grid.

Given these issues it would be wise to restrict the construction of "latent variable" asset indices to situations where the assets are positively correlated – although that would be a nontrivial limitation. Nevertheless even in these cases the question of how to deal with the negative values created by the estimation process remains. One possible response would be to add a positive constant to all index values big enough to ensure that only non-negative values remain. Linear translations of this sort have been used in some cases (Sahn and Stifel (2000, p.2126) in order to test for statistical dominance, Bhorat and van der Westhuizen (2013)). The problem is that while these shifts maintain the rankings, the Gini coefficients are not invariant to such transformations. Indeed Lorenz dominance on sub-groups can be reversed (see Wittenberg 2013).

5.3 Creating an ordinal scale

With two binary variables there are four possible outcomes, i.e. four possible values for any combined scale. If we can rank the value of the two assets, e.g. if having the x_1 asset is preferred to having the x_2 asset, then we know that $(1,0) \succ (0,1)$. But then we can rank all outcomes, in the obvious way i.e. $(1,1) \succ (1,0) \succ (0,1) \succ (0,0)$. Even if we cannot cardinalise these bundles, we could use the Cowell-Flachaire approach to create inequality measures. Indeed one of their examples is precisely of this form.

One problem with this approach is that it doesn't take into account the correlation between the two assets. Consider, for instance, two assets that can be ranked (e.g. ownership of car versus ownership of television) and assume that precisely half of the population have cars and the other half have televisions. Assume now that we "redistribute" the televisions to those who have cars, i.e. we now have half of the population that have cars and televisions and half that have nothing. The Cowell-Flachaire measures will report the **same** inequality measures before and after, despite the fact that the distribution has got a lot more unequal.

Additional problems arise if there are more than two binary variables, because we would then need to know how owning two lower ranking assets (e.g. a television and a cell phone) rank relative to owning only the most desirable asset (e.g. a car). There are 2^k possible bundles with k binary variables and we would need to be able to rank all of them. In practice, with more than two variables this approach is likely to be intractable. Consequently we will not pursue these possibilities further.

6 Multidimensional inequality indices

The literature dealing with multidimensional indices tends to approach the issue axiomatically, i.e. develop multidimensional analogues of properties such as that Pigou-Dalton transfers decrease inequality. Tsui (1999, Theorem 3) proves that an inequality measure that respects a set of these axioms must be a transformation of a multidimensional extension of the "Generalized Entropy" measures. In the two variable case these become (Tsui 1999, Corollary 1):

$$I = \begin{cases} \frac{\rho}{n}\sum_{i=1}^{n}\left\{\left(\frac{x_{1i}}{\mu_1}\right)^{c_1}\left(\frac{x_{2i}}{\mu_2}\right)^{c_2}-1\right\} \\ \frac{1}{n}\sum_{i=1}^{n}\frac{x_{1i}}{\mu_1}\left(a_{11}\log\frac{x_{1i}}{\mu_1}+a_{12}\log\frac{x_{2i}}{\mu_2}\right) \\ \frac{1}{n}\sum_{i=1}^{n}\frac{x_{2i}}{\mu_2}\left(a_{21}\log\frac{x_{1i}}{\mu_1}+a_{22}\log\frac{x_{2i}}{\mu_2}\right) \\ \frac{1}{n}\sum_{i=1}^{n}\left(\delta_1\log\frac{\mu_1}{x_{1i}}+\delta_2\log\frac{\mu_2}{x_{2i}}\right) \end{cases}$$

where $(c_1, c_2) \neq (0,0), (0,1), (1,0), (1,1)$, $\rho c_1 (c_1 - 1) > 0$, $c_1 c_2 (1 - c_1 - c_2) > 0$ and we haven't listed the parameter restrictions for the second equation onwards. Indeed all but the first version are not defined for vectors of binary variables. The first version yields an inequality index of

$$I = \rho\left[\frac{p_{12}}{p_1^{c_1} p_2^{c_2}} - 1\right]$$

This has certain attractive features, provided that ρ, c_1 and c_2 are all positive[5]. Inequality increases as p_{12} increases, so as the correlation between x_1 and x_2 increases measured inequality increases. As either p_1 or p_2. approaches zero the index becomes arbitrarily large and if $p_1 = p_2 = p_{12} = 1$ inequality is zero. But paradoxically the index is at a minimum when $p_{12} = 0$ even if $p_1 + p_2$ is far from one (e.g. if there are exactly two people who own assets, but each owns only one). The

[5] This is possible, however, only if we ignore the parameter restrictions – strictly speaking c_1 and c_2 have to be both positive, otherwise $\left(x_{ji}/\mu_j\right)^{c_j}$ is not defined for cases where $x_{ji} = 0$. This, however implies that $\rho < 0$ (from the first condition) which precisely reverses the implications.

Table 3: Index values for the uncentered principal components procedure of Banerjee

Bundle	Probability	Value (y)	Case 1: $p_1 = p_2$
$(0,0)$	$1-p_1-p_2+p_{12}$	0	0
$(1,0)$	p_1-p_{12}	$\frac{p_2-p_1+\sqrt{(p_2-p_1)^2+4p_{12}^2}}{2p_{12}p_1}$	$\frac{1}{p_2}$
$(0,1)$	p_2-p_{12}	$\frac{1}{p_2}$	$\frac{1}{p_2}$
$(1,1)$	p_{12}	$\frac{p_2-p_1+\sqrt{(p_2-p_1)^2+4p_{12}^2}}{2p_{12}p_1}+\frac{1}{p_2}$	$\frac{2}{p_2}$
Gini	$1-p_2-(p_1+p_2-2p_{12})(p_1-p_{12})u_1-p_{12}(p_1-p_2)u_1$ where $u_1=\frac{p_2-p_1+\sqrt{(p_2-p_1)^2+4p_{12}^2}}{p_1\left(p_2-p_1+\sqrt{(p_2-p_1)^2+4p_{12}^2+2p_{12}}\right)}$		$1-2p_2+2p_{12}-\frac{p_{12}^2}{p_2}$

problem, of course, is that the product $\left(\frac{x_{1i}}{\mu_1}\right)^{c_1}\left(\frac{x_{2i}}{\mu_2}\right)^{c_2}$ is zero whenever either x_{1i} or x_{2i} is zero, so any distribution in which nobody has all assets will score $-\rho$.

This is a major limitation of this approach. Of course we have taken liberties with Tsui's formulae, since they are designed to be used with positive x_{ji}.

6.1 The multidimensional Gini

An alternative "multidimensional" inequality measure is the multidimensional Gini proposed by Banerjee (2010). The procedure harks back to the first approach, i.e. creating a linear combination of the variables on which the Gini coefficient is then estimated. Again the weights on the components are given by the elements of an eigenvector of a cross-product matrix, but in this case the variables are not demeaned, so that the moments, as it were, are calculated around zero rather than the mean. As a result the weights are compelled to be positive. Banerjee proves that when applied to standard continuous non-negative variables this approach provides an inequality index that satisfies all of the key axioms, but also shows increasing inequality if a "correlation increasing transfer" occurs.

More concretely Banerjee suggests that variables should first be divided by their mean. In our case transforming the original data matrix $\mathbf{X}$ (given in equation 4) we get.

$$\mathbf{A} = \begin{bmatrix} \mathbf{0}_{n_1} & \mathbf{0}_{n_1} \\ \frac{\mathbf{1}}{\mathbf{p}_1}_{n_2} & \mathbf{0}_{n_2} \\ \mathbf{0}_{n_3} & \frac{\mathbf{1}}{\mathbf{p}_2}_{n_3} \\ \frac{\mathbf{1}}{\mathbf{p}_1}_{n_4} & \frac{\mathbf{1}}{\mathbf{p}_2}_{n_4} \end{bmatrix}$$

where $\frac{\mathbf{1}}{\mathbf{p}_j}_{n_k}$ should be interpreted in the obvious way as the n_k column vector containing $\frac{1}{p_j}$ in every element. It follows that

$$\frac{1}{N}\mathbf{A}'\mathbf{A} = \begin{bmatrix} \frac{1}{p_1} & \frac{p_{12}}{p_1p_2} \\ \frac{p_{12}}{p_1p_2} & \frac{1}{p_2} \end{bmatrix}$$

A (non-normalised) eigenvector associated with the maximal eigenvalue is $\begin{bmatrix} \frac{p_2-p_1+\sqrt{(p_2-p_1)^2+4p_{12}^2}}{2p_{12}} & 1 \end{bmatrix}'$ provided that $p_{12} \neq 0$. If $p_{12} = 0$ then the maximal eigenvalue is $\frac{1}{p_2}$ with associated eigenvector $\begin{bmatrix} 0 & 1 \end{bmatrix}$. The "index values" (up to a multiplicative constant) for this case are given in Table 3.

If $p_{12} \neq 0$ (and still assuming that $p_2 \leq p_1$) we can show that the index will order the asset bundles as $(0,0) \prec (1,0) \prec (0,1) \prec (1,1)$. However when $p_{12} = 0$, i.e. the vectors of asset holdings are completely orthogonal, then the first asset gets a weight of zero, i.e. $y(0,0) = y(1,0) = 0$. This does not create perverse rankings, but it does mean that asset one is completely ignored. This illustrates why Banerjee in his proofs requires that there be at least one individual who owns positive quantities of all assets. In this context this would require $p_{12} > 0$. This is undoubtedly a limitation, although one not nearly as severe as the requirement that the assets be positively correlated.

Table 3 also gives the Gini coefficient for this case. In fact, as we show in the appendix (section A.1) that formula is valid for any asset index which scores the assets as $y(0,0) = 0$, $y(1,0) = y_1$, $y(0,1) = y_2$, $y(1,1) = y_1 + y_2$ and where $u_1 = \frac{y_1}{p_1 y_1 + p_2 y_2}$. The formula is interesting, because it shows that $1 - p_2$ is an upper bound for the Gini – the expressions in brackets in the third and fourth term both have to be nonnegative. So the proportion of the less common asset is the key determinant for inequality overall. In the extreme case where $p_1 = p_2 = p_{12}$, i.e. where the society splits into two groups one which owns nothing and one which owns both assets, the upper bound is reached. Indeed it is also reached in the case we have ruled out, where $p_{12} = 0$, because then asset 1 is scored as having value zero, i.e. $u_1 = 0$. The other cases are not so straightforward.

It is easy to show that we will get the following partial derivatives, assuming for the moment that u_1 is exogenously given (i.e. is not itself a function of p_1, p_2 and p_{12})

- $\frac{\partial Gini}{\partial p_1} = -u_1 (2p_1 + p_2 - 2p_{12}) < 0$
- $\frac{\partial Gini}{\partial p_2} = -(p_1 - p_{12}) u_1 + u_1 p_{12} - 1 < 0$
- $\frac{\partial Gini}{\partial p_{12}} = 2u_1 (p_1 + p_2 - 2p_{12}) > 0$
- $\frac{\partial Gini}{\partial u_1} = -(p_1 + p_2 - 2p_{12})(p_1 - p_{12}) - p_{12}(p_1 - p_2) < 0$

But of course, as is also noted in the Table, u_1 is not exogenous in this case. Indeed for all but p_2 the effects operating through u_1 end up of opposite signs, so that the sign of the total derivative is ambiguous. We have the following partial derivatives

- $\frac{\partial u_1}{\partial p_1} < 0$
- $\frac{\partial u_1}{\partial p_2} > 0$
- $\frac{\partial u_1}{\partial p_{12}} > 0$

In Figure 4 we can see some of the complexities. At high levels of p_{12} the relationship between the Gini and p_1 is negative: higher values of p_1 yield lower Gini coefficients. This is reversed at low levels of p_{12}. For some levels of p_1 there is a limit below which p_{12} cannot fall – indicated in the figure by the fact that $p_{12} \geq 0.1$ if $p_1 = 0.7$ and $p_2 = 0.4$. Nevertheless we see clear reversals for the other cases plotted. At low levels of p_{12} the asset holdings are almost orthogonal and in this case the bigger the discrepancy between the size of p_1 and p_2 the more asset 1 becomes devalued. Indeed $\lim_{p_{12} \to 0} u_1 = 0$ and hence $\lim_{p_{12} \to 0} Gini = 1 - p_2$. Effectively the routine pays increasingly attention only to the rare asset and judges inequality accordingly.

The one obvious exception is if $p_1 = p_2$. In that case the Gini approaches $1 - 2p_2$, i.e. it treats the two assets equally and inequality gets measured according to who has any assets versus who has

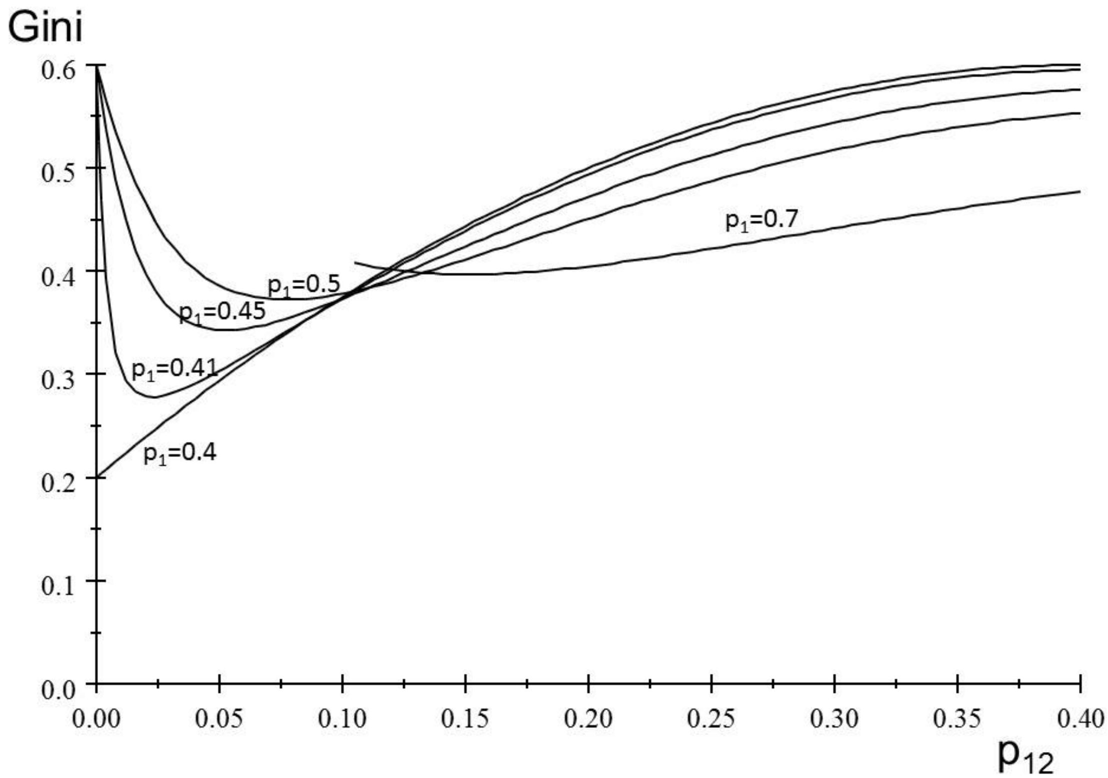

Figure 4: The relationship between the Gini coefficient and p_{12} for various values of p_1. In all cases $p_2 = 0.4$.

none. This is an attractive property although the probability of finding such a balanced relationship in any "real world" application is zero. The limiting value of $1-2p_2$ serves as a lower bound to the Gini coefficients that can be achieved. We observe that provided that p_{12} is small – but not too small – it is possible to get near that lower bound when p_1 is not very different from p_2.

6.2 Sum of assets

It is instructive to consider any "asset index" which fixes the value of each asset *ex ante*. The simple sum of assets (setting $y(1,0) = y(0,1) = 1$) is a particularly straightforward example. Note that u_1 is still a function of p_1 and p_2 (in this case $\frac{1}{p_1+p_2}$), but it is now easy to show that in this case there will be no reversals – the Gini will be decreasing in p_1, p_2 and increasing in p_{12}. Of course the reason why many analysts feel uncomfortable in simple summing up assets is that we are directly equating apples and pears. Why would we want to equate asset 1 and asset 2 in this way? In many practical cases (e.g. "car" and "radio") this is not appealing. The attraction of procedures like principal components (and related techniques) is that they decide on the weightings given the overall correlation structure in the data. Assets that are typically owned together receive a higher weight. The uncentered principal components index proposed by Banerjee also has this flavour – assets that are nearly orthogonal to the other assets will receive low weights. The reversals shown in Figure 4 above are due to the way in which the procedure shifts its opinion on how much the more common asset is actually worth.

If we could value the assets independently then we could create an "objective" asset value index. Any *ex ante* chosen values will lead to an index of the sort $y(0,0) = 0$, $y(1,0) = a$, $y(0,1) = b$ and $y(1,1) = a+b$. We assume that $a \leq b$. In this case the formula for the Gini given in Table 3 applies with $u_1 = \frac{a}{p_1 a + p_2 b}$. If the values of the two assets are far apart (we need $b > 2a$) and if p_2 is small, then it is possible for the Gini to **increase** with p_2[6]. Intuitively this is not all that surprising – an increase in the proportion of the super-affluent is likely to drive up inequality.

This shows, however, that the usefulness of the uncentered principal components index should be evaluated on whether the weightings make sense – not so much on what the comparative statics of the Gini coefficient are doing. **Any** index that has a lopsided valuation of the assets can potentially produce inequality measures that go up with increases in the proportions of people holding the assets.

6.3 Some provisional lessons

The key lesson is that the process of deriving weights for the asset index needs to be handled with care. The conventional PCA, factor analysis or MCA procedures can yield negative weights. Simply dropping these variables from the analysis (if they are genuine assets) is likely to skew the results in other ways. The uncentered PCA of Banerjee can handle these cases, provided that ownership of these assets is not completely orthogonal to that of the other assets. Nevertheless in situations where the overlap of asset holdings is relatively small these unconventional assets may be down-valued. Inspecting both the asset scores and the resulting rankings before doing any substantive analysis seems important.

[6] e.g. $a = 1$, $b = 8$, $p_1 = 0.5$, $p_2 = 0.01$ and $p_{12} = 0.005$

7 Application to the DHS wealth indices

The documentation released with the Demographic and Health Surveys doesn't give the scores according to which the wealth index is calculated. The general approach is outlined in the paper by Rutstein and Johnson (2004). As many assets as possible are used, including country-specific ones. Theoretically it should be possible to back out the coefficients by regressing the index on indicators for all asset variables that might be considered. The regression should fit perfectly. Doing this on the South African Demographic and Health Survey we managed to get an R^2 of 0.999 which is close, but not exactly equal to one. The coefficients on several key variables are shown in the first column of Table 1. The most important point for our purposes is the fact that the coefficients on the two livestock variables (possession of a donkey or horse, and possession of sheep or cattle) are both negative. It follows that individuals that have no assets will rank above individuals who have only donkey and/or cattle. Indeed if we search for the poorest individuals (according to the wealth index) they invariably own livestock.

In order to investigate this further we categorise individuals in terms of their possession (or otherwise) of "real" assets. We excluded building materials from the list and included only water piped inside the house and access to electricity. The list is shown in Table 1. The minimal possible asset holding corresponds to one room with nothing else. Households in the DHS with such minimal assets could have a large range of "wealth index" numbers, depending on what building material their accommodation was made of. Interestingly, however, 13% of individuals who had a **higher** asset holding (typically livestock with more rooms), nevertheless had a **lower** wealth index than the mean score among those with no moveable possessions. Indeed the richest person among those with no water in the house, no electricity, one room and no durables was better off (according to the wealth index) than 47% of individuals who had at least something on top of one room.

In order to explore the relationship between livestock ownership and other forms of assets further, we constructed a series of asset indices using our more restrictive list of assets. Besides the uncentered principal components index (labelled UC PCA in Table 1) we also constructed indices using PCA, MCA and Factor Analysis. The first point to note is that the negative weighting on livestock ownership is a feature of all the "latent variable" indices. The coefficients shown in Table 1 are those on the **untransformed** variables, i.e. in the case of PCA they are v_i/s_i (see equation 3).

The second point to note is that the uncentered PCA also has its bizarre feature: in this case it is the extremely large implied coefficient on ownership of a motor cycle. The reason for this is that the coefficient is v_i/μ_i where v_i is the score from the principal components calculation and μ_i is due to the standardisation suggested by Banerjee. As Table 2 shows, motor cycles are owned by very few South Africans and consequently the score becomes inflated in ways which are unlikely to reflect their real asset status. Consequently we decided to drop this variable and recalculate the index (the results are shown in column 3). Ownership of a personal computer now gets the highest score although its magnitude is not as outlandish as that for the motorcycle.

Similarly we also recalculated the PCA index without the livestock variables, to provide the fairest comparison between the two techniques. This, however, did not have much of an impact on the remaining coefficients, as can be seen by comparing columns 4 and 5 in Table 1. It will, of course, remove the anomalies noted earlier. Individuals owning livestock will now appear indistinguishable from individuals owning nothing. What is the impact of this for the identification of deprivation?

One simple check is to divide the population up into quintiles according to the two indices and to see how well they compare. Table 3 performs that analysis. We see that there are some key differences. The starkest contrast is provided by the 175 households which are rated in the bottom quintile according to the PCA index but are rated at the top of the UC PCA. Looking at the means of the asset variables it emerges that all of them owned horses/donkeys, 76% of them also owned sheep or cattle and 75% of them also owned a radio. Ownership of horses and/or donkeys is a significant asset according to the uncentered PCA. Perhaps the coefficient is on the large side, but it is unlikely that households that own both types of livestock should truly be ranked among the poorest of the poor (the bottom 20%). Of course the original PCA index would have ranked many of these households below the "poorest of the poor" (given the negative value on those assets).

In Table 4 we present the correlation matrix between the different asset indices. Although we have used fewer assets in our version of the principal components scores, they are still highly correlated with the wealth index released with the DHS. All the "latent variable" index formulations end up highly correlated. The two uncentered PCA indices show much lower correlations. The first of these has very low correlations with all the indices, since motorcycle owners receive such high scores that the entire distribution is highly skewed (95% of all scores are below 8, whereas motorcycle owners score above 50). The second shows correlations of .75 with the PCA index that doesn't weight livestock negatively – but correspondingly lower correlations with the others that maintained that negative weighting.

The obvious implication of all of this is that the standard asset indices will tend to find higher urban-rural contrasts in poverty than the uncentered PCA. This is shown clearly in Table 5. In each case we have classified the bottom 40% of individuals as "poor" according to the DHS wealth index, the PCA 2 index and the second uncentered PCA index. It is clear that there is a strong urban-rural poverty gradient. Nevertheless the DHS wealth index accentuates this contrast, while the uncentered PCA index finds more urban poverty and less rural poverty. This should not be surprising given the negative valuation of rural assets in the DHS wealth index and the strong positive valuations of urban infrastructure.

Interestingly calculating the Gini coefficient on the asset scores of the uncentered PCA we find (in Table 6) strong asset inequality in South Africa in 1998, not dissimilar to the magnitude of income inequality. Furthermore as that table also suggests, there were strong inequalities within rural areas, a finding that many South Africans will find plausible.

8 Asset inequality in South Africa 1993-2008

We turn now to consider the evolution of asset inequality in South Africa using two nationally representative surveys conducted under the auspices of SALDRU at the University of Cape Town. The first of these is the Project for Statistics on Living Standards and Development (PSLSD) conducted in 1993 and the second is the first wave of the National Income Dynamics Study (NIDS). These studies have already been used to investigate changes in money-metric income inequality over the period (Leibbrandt, Woolard, Finn and Argent 2010). It has been found that over this period money-metric inequality started at very high levels and remained at those high levels.

Figure 5 below benchmarks these discussions by presenting a picture of money metric inequality over the post-apartheid period by plotting 1993 and 2008 Lorenz curves of per capita income. It can be seen that inequality certainly did not improve. The point estimates for the Gini coefficients stayed at 0.69 over the period.

In order to look at asset inequality over time we need to calculate a pooled index for the two periods first, so that we are using the same scores for the assets in each period. This limits us to assets that were asked for in both periods. The descriptive statistics presented by Bhorat and

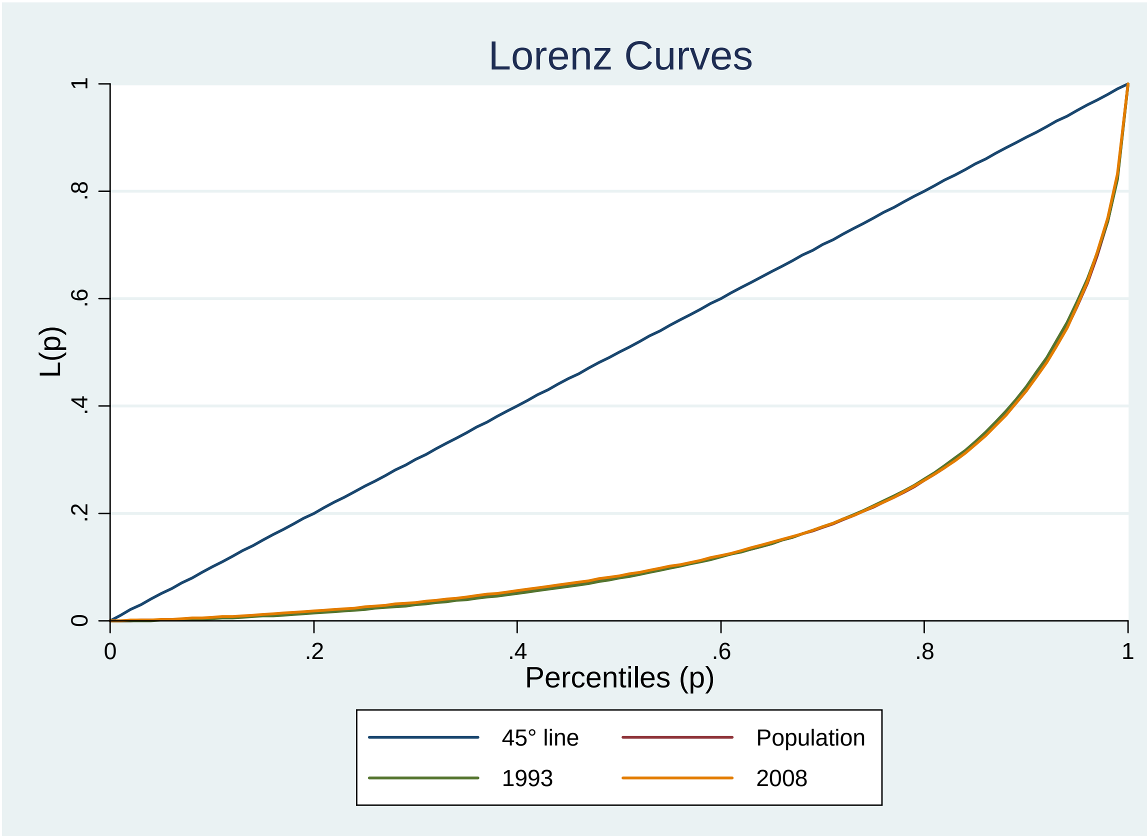

Figure 5:

van der Westhuizen (2013) suggest that there has been considerable progress over the period. Table 7 present the statistics as calculated on our data.

One immediately evident issue is that the prevalence of landlines has gone down as the availability of cell phones has become ubiquitous. If this measurement issue is not addressed it will result in a spurious decrease in assets over time. Indeed given the relative rarity of landlines in the later period these would become erroneously marked as valuable assets instead of as assets whose utility is actually in decline. Consequently we collapse landline and cell phone ownership into an omnibus "any phone" variable. The coefficients on the assets implied by our uncentered PCA asset index are given in Table 8.

When we use this asset index to construct Lorenz curves we get the result shown in Figure 6 below. The Lorenz curves show clear evidence that asset inequality fell considerably and this is confirmed by the Gini coefficients which fell markedly from 0.47 in 1993 to 0.29 in 2008. As a reflection of the fact that these Lorenz curves and Gini coefficients were estimated from the pooled UCPC measure, the pooled or Population Lorenz curve is plotted in the figure too.

Given South Africa's history, there is a great fascination with inequality how racial inequality has


Lorenz Curves
L(p)
0
.2
.4
.6
.8
1
Percentiles (p)
45° line
Population
1993
2008

Figure 6:

changed in South Africa over the post-apartheid period. This interest has results in many between-group versus within-group decompositions being run by race on South African survey data. It has not been possible to do such decompositions on asset indices before. However, the UCPC index can support such estimations. However, the UCPC can have zero values and, indeed, it does have such values in the South African case. Thus, the commonly used Theil entropy measures are not appropriate for such decompositions. In Table 9 below we use a decomposition of the Gini coefficient by racial group. As explained by Milanovic and Yitzhaki (2002), because of the overlap of Lorenz curves across racial groups, the decomposition leaves a residual or overlap term. Again we compare per capita income and the UCPC asset index for each year.

The aggregate Gini coefficients for UCPC and PCINC in 1993 and 2008 are the same as those reported in the two Figures above. In each year these aggregate figures are constituted by a combination of inequality with the racial groups and inequality across (between) the racial groups. The income figures show a rising within-group contribution as the income inequality within each racial group rises between 1993 and 2008. This is not true of the asset index as only the smaller white and Asian groups have rising within-group inequality between 1993 and 2008. Nevertheless the between group inequality in assets has declined so sharply that the within-group inequality becomes a larger portion of overall inequality.

The fact that asset inequality should have declined is not surprising given that the statistics shown in Table 7 show strong increases in access to assets between 1993 and 2008. This is not universally true - motor cars, for instance, remain relatively rare. Nevertheless the penetration of television, cell phones, refrigerators and electricity suggest that asset holdings have certainly increased. By contrast the money-metric measures suggest very little change. Part of the problem, of course, is that if the whole distribution shifts upwards by an equiproportionate amount measured inequality will remain static. Note, however, that dummy variables cannot be rescaled in this way. As everyone becomes better off asset holdings will increase across the board. In the way that we measure inequality through assets this will make all assets more common and will thus reduce asset inequality. The coarser asset inequality measure therefore allows us to see progress which is obscured by the more continuous measure.

It is also true, of course, that all measurements are contingent on the schedules that are employed. One note of caution in this regard is appropriate. The asset inequality measure for 1998 that we calculated for the DHS is significantly higher than either the 1993 or 2008 measures we have just considered. The main reason for this is that the asset schedule for 1998 included assets such as "personal computer" which allowed a better contrast to be drawn between high earners and the rest[7]. While we are sure that access to assets has spread and that in this sense asset inequality has decreased, the magnitude of the intial level of inequality and the size of the decrease are probably not as dramatic as suggested by the Gini coefficients that we have reported.

9 Conclusion

In this paper we have argued that asset indices can be interesting and powerful tools for analysing social trends. However doing so in an unreflective and automatic way is unlikely to provide useful insights. Some of the ways in which asset indices have been produced thus far, for instance, has obscured real asset holdings in rural areas in at least the South African case. Arguably this has led to an exaggerated sense of rural deprivation and a lack of appreciation for poverty in urban areas.

[7] Of course the case of the motor cycle should remind us that some of these contrasts can be overdrawn.

It may also have obscured real inequality within rural areas. Our analysis has also suggested that it is possible to create asset indices in ways that allow the calculation of Gini coefficients. To that end we have used the methodology suggested by Banerjee for the calculation of "multidimensional Gini coefficients" using continuous data. Our application suggests that the technique can work well, provided that care is taken in ensuring that some rare assets don't distort the index. In general we believe that such indices should not be used without scrutinising the implied coefficients.

More substantively our empirical work suggests that the money-metric approach to inequality measurement in South Africa may have obscured the real progress made in large portions of the population – mainly because the rich have progressed also. The issue of how to think about the place of the rich (and super-rich) in the new South Africa is, of course, highly interesting. But this should not detract from the perspective offered by our asset indices.

A Derivations

A.1 The Gini coefficient for the bivariate case

We make the following assumptions: The bundles $(0,0)$, $(1,0)$, $(0,1)$ and $(1,1)$ have index values 0, u_1, u_2 and u_1+u_2 respectively with $0 \leq u_1 \leq u_2$. Furthermore the values are normed so that $p_1u_1+p_2u_2=1$. Consequently the shares accruing to individuals having asset bundles $(0,0)$, $(1,0)$, $(0,1)$ and $(1,1)$ are, respectively, 0 $(p_1-p_{12})\,u_1$, $(p_2-p_{12})\,u_2$ and $p_{12}\,(u_1+u_2)$. Fig 7 depicts the Lorenz curve for this situation.

Now twice the area between the Lorenz curve and the 45° line is equal to the area of the unit square minus twice the area **under** the Lorenz curve. The area under the Lorenz curve consists of the three more darkly shaded triangles plus the two rectangular areas shaded in dark in Figure 7. Doubling the triangular areas will result in the removal of the lightly shaded triangles as well. Consequently the Gini coefficient will be the area of the unshaded rectangles minus the area of the darkly shaded rectangles. It follows that

$$\begin{aligned} Gini &= (1-p_1-p_2+p_{12})\,(p_1-p_{12})\,u_1+(1-p_2)\,(p_2-p_{12})\,u_2+ \\ &\quad (1-p_{12})\,(p_{12})\,(u_1+u_2)-p_2\,(p_1-p_{12})\,u_1-p_{12}\,(p_2-p_{12})\,u_2 \end{aligned}$$

Taking note of the fact that $p_2u_2=1-p_1u_1$ and simplifying we get the result that

$$Gini = 1-p_2-(p_1+p_2-2p_{12})\,(p_1-p_{12})\,u_1-p_{12}\,(p_1-p_2)\,u_1$$

References

Banerjee, A. K.: 2010, A multidimensional gini index, *Mathematical Social Sciences* **60**, 87–93.

Bhorat, H. and van der Westhuizen, C.: 2013, Non-monetary dimensions of well-being in South Africa, 1993–2004: A post-apartheid dividend?, *Development Southern Africa* **30**(3), 295–314.

Booysen, F., van der Berg, S., von Maltitz, M. and du Rand, G.: 2008, Using an asset index to assess trends in poverty in seven Sub–Saharan African countries, *World Development* **36**(6), 1113–1130.

Cowell, F. A. and Flachaire, E.: 2012, Inequality with ordinal data, mimeo, London School of Economics. available from http://darp.lse.ac.uk/pdf/IneqOrdinal.pdf.

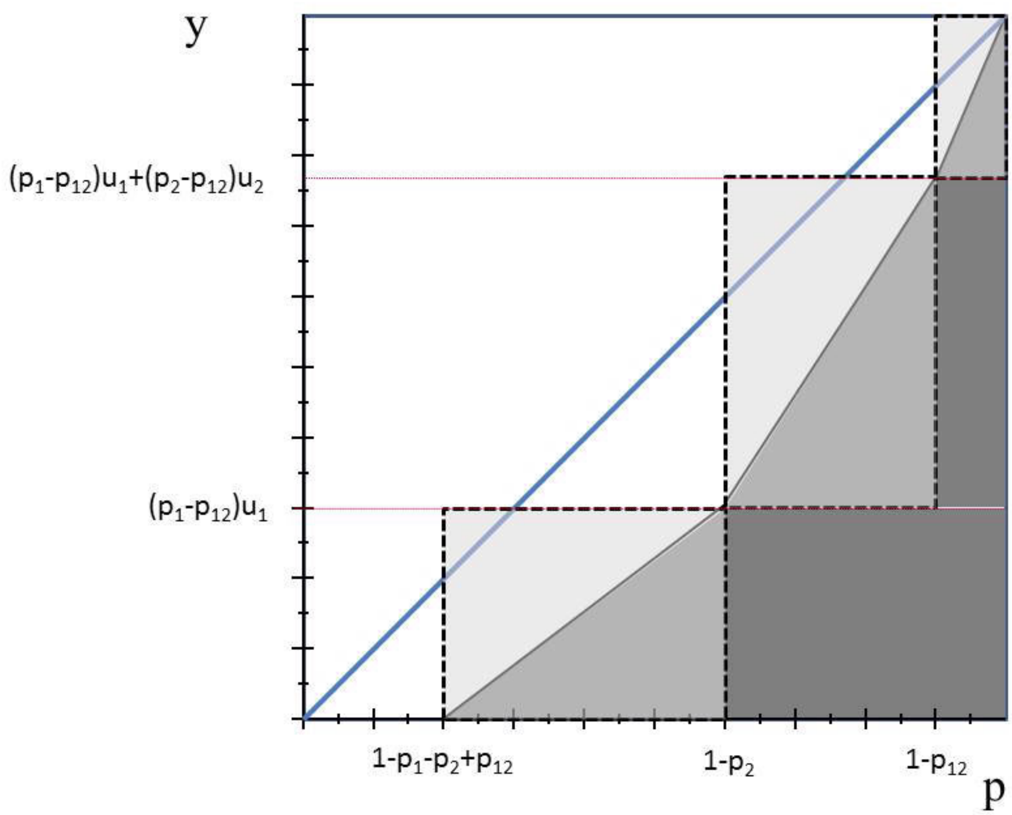

Figure 7: Calculating the Gini coefficient in the bivariate case, where $y\,(0,0) = 0$.

Filmer, D. and Pritchett, L. H.: 2001, Estimating wealth effects without expenditure data – or tears: An application to educational enrollment in states of India, *Demography* **38**(1), 115–132.

Filmer, D. and Scott, K.: 2012, Assessing asset indices, *Demography* **49**, 359–392.

Houweling, T. A., Kunst, A. E. and Mackenbach, J. P.: 2003, Measuring health inequality among children in developing countries: does the choice of the indicator of economic status matter?, *International Journal for Equity in Health* **2**(8).

Leibbrandt, M., Woolard, I., Finn, A. and Argent, J.: 2010, Trends in South African income distribution and poverty since the fall of Apartheid, *Social, Employment and Migration Working Papers 101*, OECD. http://dx.doi.org/10.1787/5kmms0t7p1ms-en.

McKenzie, D. J.: 2005, Measuring inequality with asset indicators, *Journal of Population Economics* **18**, 229–260.

Milanovic, B. and Yitzhaki, S.: 2002, Decomposing world income distribution: Does the world have a middle class?, *Review of Income and Wealth* **48**, 155–78.

Rutstein, S. O. and Johnson, K.: 2004, The DHS wealth index, *DHS Comparative Reports 6*, ORC Macro. Calverton, Maryland.

Sahn, D. E. and Stifel, D.: 2000, Poverty comparisons over time and across countries in Africa, *World Development* **28**(12), 2123–2155.

Sahn, D. E. and Stifel, D.: 2003, Exploring alternative measures of welfare in the absence of expenditure data, *Review of Income and Wealth* **49**(4), 463–489.

Tsui, K.: 1999, Multidimensional inequality and multidimensional generalized entropy measures: An axiomatic derivation, *Social Choice and Welfare* **16**(1), 145–157.

Wagstaff, A.: 2005, The bounds of the concentration index when the variable of interest is binary, with an application to immunization inequality, *Health Economics* **14**, 429–432.

Wittenberg, M.: 2009, Weighing the value of asset proxies: The case of the Body Mass Index in South Africa, *Working Paper 39*, SALDRU, University of Cape Town. Available at http://www.opensaldru.uct.ac.za/handle/11090/16.

Wittenberg, M.: 2011, Estimating expenditure impacts without expenditure data using asset proxies, *Economics Letters* **110**, 122–125.

Wittenberg, M.: 2013, Non-monetary dimensions of well-being: A comment, *Development Southern Africa* **30**(3), 826–829.