A graphical overview of social inequality in South Africa

Tsiresy Pierre Bernard*
Department of Statistical Sciences, University of Cape Town, Cape Town, South Africa
tsiresy.pb@gmail.com

Sugnet Lubbe
Department of Statistical Sciences, University of Cape Town, Cape Town, South Africa
sugnet.lubbe@uct.ac.za

Abstract

In this paper, social inequality is defined as differences in living conditions between households of different racial groups: differences in income levels, differences in housing conditions and access to basic services, etc. Because of its multifaceted nature, multivariate graphical tools are well suited to study this issue. Multiple Correspondence Analysis (MCA) is suitable in the analysis of datasets of categorical nature, as is the case here. To analyse household survey data, sampling weights must be included in the analysis to make the results representative of the South Africa population. The MCA technique is enhanced here with the inclusion of sampling weights. With the help of Canonical Variates Analysis (CVA) biplots, the differences in living conditions between the different racial groups are then graphically displayed using Canonical Variates Analysis (CVA) biplots. This paper uses sampling-weighted MCA and CVA biplots to obtain a concise overview of the evolution of social inequality over the post-Apartheid period. The datasets used here are the October Household Surveys (OHS) and the General Household Surveys (GHS) of Statistics South Africa.

Keywords: Biplot, household survey data, social inequality, sampling weights.

1 Introduction

Social inequality encompasses differences in the fulfilment of basic socio-economic human rights and access to basic services: access to education, housing, water, health, etc. In South Africa, social inequality has a strong racial feature and living conditions differ from one racial group to the other. Because of the issue's multifaceted nature, multivariate graphical tools are suitable to assess whether the overall differences in living conditions between the four major racial groups (Indian, Black, Coloured and White) have widened or narrowed since the beginning of the country's democratic era.
Six nationwide household surveys were chosen for this research: the October Household Surveys (OHS) 1996 and 1998 and the General Household Surveys (GHS) of 2002, 2006, 2009 and 2011, all released by Statistics South Africa. Only the variables that were common across the different datasets were kept. These variables - which were originally categorical or were recoded as categorical variables - were grouped depending on the aspect of living conditions which they measured: access to housing, access to education, access to energy, access to water and sanitation, access to healthcare and variables related to expenditure and employment.

2 Methodology

2.1 Sampling-weighted Multiple Correspondence Analysis

$\mathbf{X}$ is a $n \times Q$ categorical dataset with n observations and Q categorical variables. The number of categories of the q-th variable is J_q. The total number of categories is $J = \sum_{q=1}^{Q} J_q$. An indicator matrix $\mathbf{Z}$ of 0's and 1's of size $n \times J$ can be obtained from the original dataset $\mathbf{X}$. The symmetric Burt Matrix $\mathbf{B}$ of size $J \times J$ is

obtained as $\mathbf{B} = \mathbf{Z}^{\mathrm{T}}\mathbf{Z}$. $\mathbf{B}$ is made up of $Q \times Q$ contingency tables which concatenate variables against one another.

Sampling weights are used in the analysis of household surveys to make the analysis representative of the sampled population. Let $\mathbf{w} : n \times 1$ be the weight vector attached to the data matrix $\mathbf{X}$. The estimated population size is $n^\star = \sum_i w_i$. The sampling-weighted indicator matrix $\tilde{\mathbf{Z}}$ is computed by multiplying each row of the indicator matrix $\mathbf{Z}$ by its sampling weight: $\tilde{z}_{ij} = z_{ij} \times w_i$. The sampling-weighted Burt matrix $\tilde{\mathbf{B}}$ is the symmetric matrix computed as $\tilde{\mathbf{B}} = \tilde{\mathbf{Z}}^{\mathrm{T}}\mathbf{Z} = \mathbf{Z}^{\mathrm{T}}\tilde{\mathbf{Z}}$. Whereas the elements of each of the $Q \times Q$ subtables of $\mathbf{B}$ sum up to the sample size n, the elements of each subtable of $\tilde{\mathbf{B}}$ sum up to the population size $n^\star$.

Sampling-weighted MCA is a computation that can be applied to $\tilde{\mathbf{Z}}$ or to $\tilde{\mathbf{B}}$ as displayed in Table 1. In the MCA of $\tilde{\mathbf{Z}}$, row principal coordinates $\mathbf{F}$ are used to plot observations and column standard categories Γ^{Z} are used to plot variable categories. In the MCA of the Burt matrix $\tilde{\mathbf{B}}$, variable categories are plotted using column standard coordinates Γ^{B} (which are equal to Γ^{Z}). Sampling-weighted MCA is based on MCA. For more details on MCA, see Greenacre (2007).

Table 1: Computation of the MCA of $\tilde{\mathbf{Z}}$ and sampling-weighted Burt matrix $\tilde{\mathbf{B}}$

Analysis of $\tilde{\mathbf{Z}}$	Analysis of $\tilde{\mathbf{B}}$
$\mathbf{P}^{\mathrm{Z}} : n \times J = \frac{1}{\sum_i \sum_j \tilde{z}_{ij}} \tilde{\mathbf{Z}} = \frac{1}{Qn^\star} \tilde{\mathbf{Z}}$	$\mathbf{P}^{\mathrm{B}} : J \times J = \frac{1}{\sum_i \sum_j \tilde{b}_{ij}} \tilde{\mathbf{B}} = \frac{1}{n^\star Q^2} \tilde{\mathbf{B}}$
$\mathbf{r} : n \times 1$ where $r_i = \sum_j p_{ij}$ and $\mathbf{D}_r = diag(\mathbf{r})$	
$\mathbf{c} : J \times 1$ where $c_j = \sum_i p_{ij}^{\mathrm{Z}}$ and $\mathbf{D}_c = diag(\mathbf{c})$	$\mathbf{c}^{\mathrm{B}} = \sum_i p_{ij}^{\mathrm{B}} = \mathbf{c}$
$\mathbf{S}^{\mathrm{Z}} = \mathbf{D}_r^{-1/2}(\mathbf{P}^{\mathrm{Z}} - \mathbf{r}\mathbf{c}^{\mathrm{T}})\mathbf{D}_c^{-1/2} = \mathbf{U}^{\mathrm{Z}}\mathbf{D}_\alpha^{\mathrm{Z}}(\mathbf{V}^{\mathrm{Z}})^{\mathrm{T}}$	$\mathbf{S}^{\mathrm{B}} = \mathbf{D}_c^{-1/2}(\mathbf{P}^{\mathrm{B}} - \mathbf{c}\mathbf{c}^{\mathrm{T}})\mathbf{D}_c^{-1/2} = \mathbf{V}^{\mathrm{B}}\mathbf{D}_\alpha^{\mathrm{B}}(\mathbf{V}^{\mathrm{B}})^{\mathrm{T}}$
$\mathbf{V}^{\mathrm{B}} = \mathbf{V}^{\mathrm{Z}} = \mathbf{V}$	
$\mathbf{D}_\alpha^{\mathrm{Z}} = (\mathbf{D}_\alpha^{\mathrm{B}})^{1/2}$	
$\Gamma^{\mathrm{Z}} = \mathbf{D}_c^{-1/2}\mathbf{V} = \Gamma^{\mathrm{B}}$	$\Gamma^{\mathrm{B}} = \mathbf{D}_c^{-1/2}\mathbf{V} = \Gamma^{\mathrm{Z}}$
$\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}^{\mathrm{Z}}\mathbf{D}_\alpha^{\mathrm{Z}} = \mathbf{D}_r\mathbf{P}^{\mathrm{Z}}\Gamma$	$\mathbf{G}^{\mathrm{B}} = \mathbf{D}_c^{-1/2}\mathbf{V}\mathbf{D}_\alpha^{\mathrm{B}} = \Gamma^{\mathrm{B}}\mathbf{D}_\alpha^{\mathrm{B}}$

Figure 1 is the plot of the information related to access to healthcare in South Africa in 1996. Only the categories of the two relevant variables are plotted for sake of clarity. Variable *consultation* records which healthcare service is used by the household whenever a household member is sick and variable *medaidmember* indicates whether at least one household member is covered by medical insurance. Their categories are plotted as red triangles. Categories of the unique supplementary variable (the *race* variable) are plotted as green rhombuses. The position of a supplementary category indicates the average of the coordinates of households that belong to that category. On average, White households are more likely to consult private hospitals (*pr_hosp*), private clinics (*pr_clin*) and private doctors (*pr_doc*) and are often covered by medical insurance schemes. Black households on the other hand are often not covered by medical aid (*no_medaid*), consult public health centres (*pub_clin*, *pub_hosp*) or traditional healers (*trad_heal*) and are likelier to not consult health care practitioners whenever their household members get sick (*sick_nocins*). The first axis therefore opposes access to good healthcare services (high scores on axis one) with access to healthcare services of lower quality (low values on axis one).

In MCA, the total number of dimensions is equal to $J - Q$. The inertia λ_k of the k-th dimension is equal to the square of the k-th singular value (for the MCA of $\tilde{\mathbf{Z}}$) or k-th eigenvalue (for the inertia of $\tilde{\mathbf{B}}$): $\lambda_k = \alpha_k^2$. The inertias of the MCA of $\tilde{\mathbf{B}}$ are the square of that of $\tilde{\mathbf{Z}}$: $\lambda_k^{\mathrm{Z}} = (\lambda_k^{\mathrm{B}})^{1/2}$ for $k \in (1, J - Q)$. If l dimensions are retained, the matrix of retained components is the matrix $\mathbf{F}_{[l]}$, made up of the first l columns of the above-mentioned $\mathbf{F}$.

For each year of interest i where $i \in \{1996, 1998, 2002, 2006, 2009, 2011\}$ sampling-weighted MCA is used

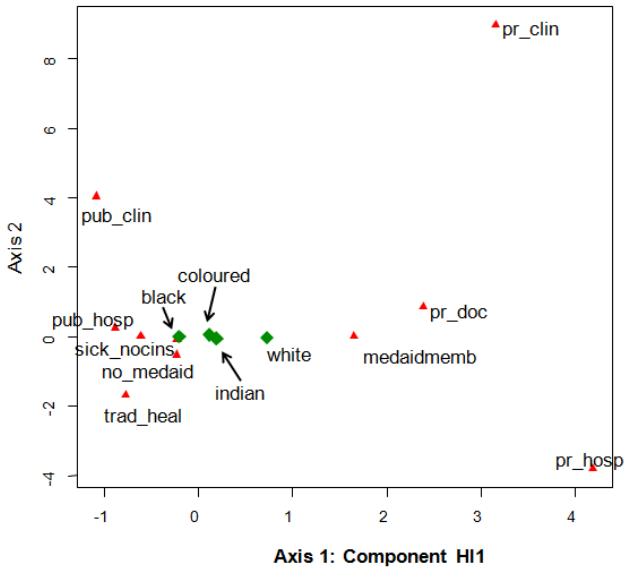

Figure 1: Monoplot for 1996 Healthcare information

to extract components from each aspect of living conditions. The number of components retained per aspect of living conditions were chosen such that at least 75 % of inertia were retained in the analyses of these groups of variables. A total of thirteen components were retained for each year: four components for analysis of the housing aspect, two for education, three for energy, two for expenditure, one for water and sanitation and one for healthcare. Data matrices of retained components are denoted $\hat{\mathbf{X}}_i : n_i \times 13$ where $i \in \{1996, 1998, 2002, 2006, 2009, 2011\}$.

2.2 Canonical variates analysis

In Canonical Variate Analysis (CVA), the observations of the dataset are grouped into groups or classes. The purpose of the CVA biplot is to optimally show the differences between the groups. The CVA computation displayed below comes from Gower et al. (2011).
Let $\mathbf{X}$ be a centred data matrix of size $n \times p$. The observations are grouped into g groups whose sizes are contained in the diagonal matrix $\mathbf{N}_g = diag(n_1, n_2, \dots n_g)$. The total variance can be decomposed into between group variance $\mathbf{B} = \bar{\mathbf{X}}^{\mathrm{T}}\mathbf{N}_g\bar{\mathbf{X}}$ and within group variance $\mathbf{W} = \mathbf{X}^{\mathrm{T}}\mathbf{X} - \bar{\mathbf{X}}^{\mathrm{T}}\mathbf{N}_g\bar{\mathbf{X}}$ where $\bar{\mathbf{X}}$ is the matrix of group means. In CVA, the distances are Mahalanobis. The distance between observations x_h and x_l is $\delta_{hl} = \sqrt{(x_h - x_l)'\mathbf{W}^{-1}(x_h - x_l)}$
The aim of the CVA computation is to find the linear combination $\mathbf{y} = \mathbf{X}\mathbf{v}$ which maximises the between-to-within variance ratio obtained as:

$$\frac{\mathbf{v}'\mathbf{B}\mathbf{v}}{\mathbf{v}'\mathbf{W}\mathbf{v}} \tag{1}$$

The set of solutions $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \dots \mathbf{v}_p)$ can be found by solving the two-sided eigendecomposition: $\mathbf{B}\mathbf{V} = \mathbf{W}\mathbf{V}\Lambda$. In a two-dimensional biplot, the coordinates of the samples are: $\mathbf{Y}_{[2]} = \mathbf{X}[\mathbf{v}_1\ \mathbf{v}_2]$ and the coordinates of the group means are: $\bar{\mathbf{Y}}_{[2]} = \bar{\mathbf{X}}[\mathbf{v}_1\ \mathbf{v}_2]$

2.3 α-bags

For univariate data, one can use a boxplot to visually assess the spread and location of the observations. For bivariate datasets, Rousseuw et al.(1999) developed the bagplot, a tool with the same properties as the boxplot. It is based on the concept of halfspace location depth (Tukey, 1975), a generalisation of ranks to

multivariate data. The main feature of the bagplot is the bag which contains the inner 50 % of the observation points. Based on the bagplot, the α-bag (Gower et al., 2011) contains the inner $\alpha \times 100\%$ of the points of a group, with the value of α ranging from 0 to 1. The default α-value is 0.95, a value which displays the clusters of data points fairly well (Gower et al., 2011).

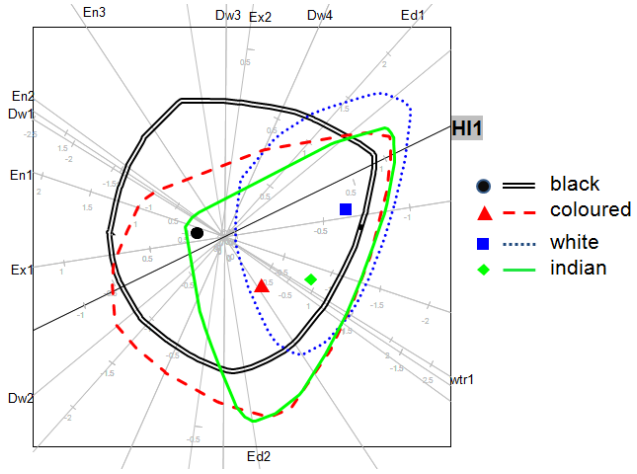

Figure 2: CVA biplot for 1996 data with α-bags at $\alpha = 0.95$

Figure 2 is the CVA biplot of the matrix of retained components for 1996 with α-bags at $\alpha = 0.95$. The mean point of the black group is represented by a black dot; that of the coloured group is a red triangle; a green rhombus for the indian group and a blue square for the white group. The shape and the orientation of the α-bags as well as the overlap between the α-bags indicates how the groups differ. For example, the size of the overlap between them as well as the orientation of the α-bags of the indian and white groups show that these two groups have some similar living conditions. One can see however that the black and the white groups are the most dissimilar.
The graph also shows on which variables the groups differ. In Figure 2 for example, the axis of Component **Hl1** is higlighted. This component corresponds to the first axis of Figure 1. High scores on **Hl1** (towards the upper-right quadrant) therefore indicate access to good healthcare services and low scores (lower-right quadrant) indicate a worse access to healthcare. The fact that the majority of White households have a fairly higher score on **Hl1** indicates that they have a better access to healthcare services as opposed to households of other racial groups.
Gower et al.(2011) suggest that α-bags can be used to quantify the overall differences between groups. A general overlap between α-bags is defined here as the space in a biplot contained in all the α-bags. A large general overlap indicates that the groups share many features. No overlap suggests that the groups are extremely different. To assess whether the differences between groups have narrowed or widened over time, one needs to see whether the size of the overlap has increased or decreased. For that purpose, Gower et al. (2011) indicate that the α-value should be decreased to a value $\alpha^\star$ where the overlap between the α-bags disappears. If the overlap at a value α=0.95 was genuinely large, the α value should be decreased a lot before reaching a value $\alpha^\star$ where the overlap disappears: a small $\alpha^\star$ indicates small differences. Inversely, a large $\alpha^\star$ value would suggest that differences between the groups were substantial. Figure 2 shows that the two most extreme groups are the black and white groups. The general overlap would therefore disappear when the α-bags of these two groups just touch each other. Figure 3 shows that this occurs when $\alpha^\star = 0.60$.

3 Results and conclusion

To assess whether socio-economic inequality has decreased over time, the $\alpha^\star$ values have been evaluated in the CVA of the matrices of retained components obtained in Section 2.1 for all the different years. These

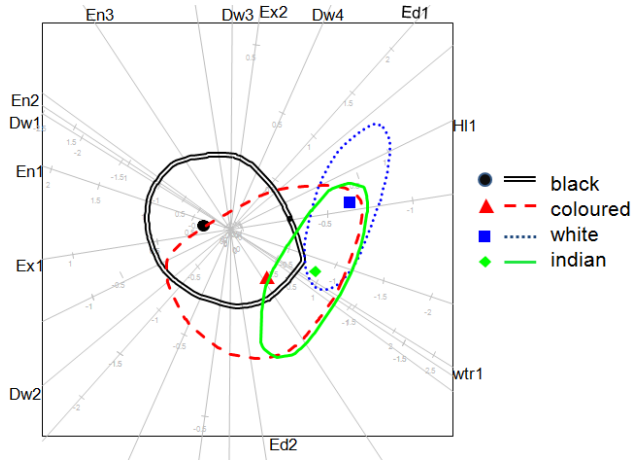

Figure 3: CVA biplot for 1996 data with α-bags at $\alpha^\star = 0.60$

values are shown in Table 2.

Table 2: Value of $\alpha^\star$ for the different years

Dataset	1996	1998	2002	2006	2009	2011
$\alpha^\star$	0.60	0.60	0.59	0.59	0.58	0.55

This table indicates that the $\alpha^\star$ values have decreased but not much over the period 1996 to 2011. There exists as yet no method to assess the significance of changes in $\alpha^\star$. The use of bootstrap or permutation tests to determine the significance in changes in $\alpha^\star$ are outside the scope of this paper. The trend displayed in Table 2 is however in the right direction. It suggests that overall differences in living conditions between households of different races have not narrowed much over the post-Apartheid period.

References

Gower, J., Lubbe, S., and Le Roux, N.J. (2011). *Understanding Biplots.* Wiley, Chichester.

Greenacre, M. (2007). *Correspondence Analysis in Practice.* Chapman & Hall/CRC, London.

Rousseeuw, P.J., Ruts, I. and Tukey., J.W. (1999). The bagplot: a bivariate boxplot. *The American Statistician*, **53** (4), 382-387.

Tukey., J.W. (1975). Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematics*, **2**,523-531.

Data sources

Statistics South Africa. October Household Survey 1996 [dataset]. Version 1. Pretoria: Statistics South Africa [producer], 1999. Cape Town: DataFirst [distributor], 2011.

Statistics South Africa. October Household Survey 1998 [dataset]. Version 1. Pretoria: Statistics South Africa [producer], 2000. Cape Town: DataFirst [distributor], 2011.

Statistics South Africa. General Household Survey 2002 [dataset]. Version 1.1. Pretoria: Statistics South Africa [producer], 2003. Cape Town: DataFirst [distributor], 2011.

Statistics South Africa. General Household Survey 2006 [dataset]. Version 1.2. Pretoria: Statistics South Africa [producer], 2007. Cape Town: DataFirst [distributor], 2011.

Statistics South Africa. General Household Survey 2009 [dataset]. Version 1.1. Pretoria: Statistics South Africa [producer] , 2010. Cape Town: DataFirst [distributor], 2011.

Statistics South Africa. General Household Survey 2011 [dataset]. Version 1. Pretoria: Statistics South Africa [producer], 2012. Cape Town: DataFirst [distributor], 2012.