A Comparison between Simple and Taxicab Correspondence Analysis and Examples

Sergio Camiz
Dipartimento di Matematica, Sapienza Università di Roma
E-mail: sergio.camiz@uniroma1.it
Gastão Coelho Gomes
Departamento de Métodos Estatísticos, Universidade Federal do Rio de Janeiro
E-mail: gastao@im.ufrj.br

Abstract

In this work Simple Correspondence Analysis, the exploratory method that aims at a graphical representation of the structure of a contingency table, based on L_2 Euclidean metrics, is compared to Taxicab Correspondence Analysis, aiming at the same purpose but based on both L_1 (Manhattan, Taxicab) and L_∞ metrics. Theoretical differences are outlined and two examples are proposed, in which Taxicab method seems suboptimal but appears to reduce the Guttman effect due to the studied data structure.

Keywords: Simple Correspondence Analysis, Taxicab Correspondence Analysis, L_2-metrics, Euclidean metrics, L_1-metrics, Manhattan metrics.

1 Introduction

The aim of this work is to compare both theoretically and in practice two exploratory methods whose aim is apparently the same, applied to a qualitative data: to represent character levels on the same graphical (reduced dimensional) space, in order to help interpretability. As interpretability we mean that the relations that underly the table structure may be seen graphically in terms of both absolute and relative position of the points-levels. The compared methods are *Correspondence Analysis* (*SCA*, Benzécri et al., 1973-82; Greenacre, 1983) and *Taxicab Correspondence Analysis* (*TCA*, Choulakian, 2006) for two characters crossing in a 2-way contingency table. Note that the extension to several characters crossing in multiple tables of both method exist (Benzécri et al., 1973-82; Greenacre, 1983; Choulakian, 2008).

Indeed, a major difference exists between the two methods, since they are based on two different metrics: *SCA* is based on the L_2 *Euclidean distance*, whereas *TSCA* is based on the L_1 *Taxicab*, a.k.a. *Manhattan*, distance. In practice, this means that, whereas the coordinates along corresponding factors may be comparable as well as the items distances along a factor, the distances in higher-dimensional spaces are larger in L_1, since they are obtained by the sum of those computed along each dimension, than in L_2, in which the measure along the straight line joining two items is by no means lower. Thus, we may say that the two methods address two different kinds of applications, depending on which way distances between items are considered in the different frameworks.

In the following, let $N = (n_{ij})$ an $r \times c$ contingency table, with $n = n_{..}$ its grand total, that is the number of units, $P = (p_{ij}) = (n_{ij}/n)$ the corresponding matrix of relative frequencies, $\boldsymbol{r} = (p_{1.}, ..., p_{r.})'$ the vector of row marginal profile $\boldsymbol{c} = (p_{.1}, ..., p_{.c})'$ the vector of column

marginal profile, and $D_r = diag(\boldsymbol{r})$, $D_c = diag(\boldsymbol{c})$ the corresponding diagonal matrices. We concentrate on matrix P, since n, the number of units, in all formulas is a scale factor and is relevant only in the statistical tests. It is well known that the matrix $\boldsymbol{r}\boldsymbol{c}'$ represents the matrix of independence among the crossing characters, so that we may be only interested to study, and thus to graphically represent, the matrix of deviations from independence $D = P - \boldsymbol{r}\boldsymbol{c}'$.

For this purpose, we must get pairs of unit vectors of coordinates $(\boldsymbol{c}_{\boldsymbol{r}}^{\alpha}, \boldsymbol{c}_{\boldsymbol{c}}^{\alpha})$, for the levels of the characters by row and column, respectively, with $\alpha = 1, \ldots, q$ the rank of P, $q \leq \min(r, c) - 1$, with the requirement of orthogonality. As the graphical representation aims at outlining these deviations, we may wish that these coordinates represent deviations and for that the additive model of data reconstruction is adopted, that is

$$d_{ij} = p_{ij} - p_{i\cdot}p_{\cdot j} = p_{i\cdot}p_{\cdot j}\sum_{\alpha=1}^{q} \iota_{\alpha} c_{ri}^{\alpha} c_{cj}^{\alpha\prime} \tag{1}$$

with the conditions

$$\begin{aligned} \sum_{ij}(p_{ij} - p_{i\cdot}p_{\cdot j}) &= 0 \\ \sum_{i} p_{i\cdot} c_{ri}^{\alpha} = \sum_{j} p_{\cdot j} c_{cj}^{\alpha} &= 0 \ \forall\alpha \\ \sum_{ik} p_{i\cdot}p_{k\cdot} c_{ri}^{\alpha} c_{rk}^{\alpha} = \sum_{jh} p_{\cdot j}p_{\cdot h} c_{cj}^{\alpha} c_{ch}^{\alpha} &= \delta_{ij} \ \forall\alpha \end{aligned} \tag{2}$$

The (2) are ordinary identification conditions on the deviations from expectation and on standardized coordinates. Essentially, the rationale of additive models is to decompose the table into independent additive unit-rank components, $P = \boldsymbol{r}\boldsymbol{c}' + \sum_{\alpha} L_{\alpha}$ that here will be named *layers*, each layer representing

$$L_{\alpha} = \iota_{\alpha}\, \boldsymbol{r}\boldsymbol{c}'\boldsymbol{c}_{r}^{\alpha}\boldsymbol{c}_{c}^{\alpha\prime}$$

an independent component of the deviation from the independence of the original table. Should the coordinates of both rows and columns be correlated with some other character, one may imagine to attribute to its influence the different levels of the characters crossed in the table (Orlóci, 1978).

2 The two methods

The two methods under examination adopt two different metrics in their spaces of representation. Consider two points A and B, whose coordinates are $A = (a_1, a_2, \ldots, a_n)$ and $B = (b_1, b_2, \ldots, b_n)$, and a vector $\boldsymbol{v}$, whose components are $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)$. We define the following metrics:

- L_2 metrics, also known as *Euclidean*, in which the distance between two points A and B is given by $d_2(A, B) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$ and the induced L_2 norm is thus $\| \boldsymbol{v} \|_2 = \sqrt{\sum_{i=1}^{n}(v_i)^2}$;
- L_1 metrics, also known as *Manhattan, City block*, or *Taxicab*, in which the distance between two points A and B is given by $d_1(A, B) = \sum_{i=1}^{n} |a_i - b_i|$ and the induced norm is thus $\| \boldsymbol{v} \|_1 = \sum_{i=1}^{n} |v_i|$;
- L_{∞} metrics, in which the distance between two points A and B is given by $d_{\infty}(A, B) = \max_{i\in(1,n)} |a_i - b_i|$ and the induced norm is thus $\| \boldsymbol{v} \|_{\infty} = \max_{i\in(1,n)} |v_i|$.

According to the first two metrics, two Correspondence Analyses are defined, in order to study a contingency data table:

1. *Simple Correspondence Analysis* (*SCA*, Benzécri et al., 1973-82; Greenacre, 1983), based on L_2 metrics and the *Singular Value Decomposition* (*SVD* Greenacre, 1983; Abdi, 2007);
2. *Taxicab Correspondence Analysis* (*TCA*, Choulakian, 2006), based on both L_1 and L_∞ metrics, and the *Taxicab Singular Value Decomposition* (*TSVD*, Choulakian, 2004).

2.1 Singular Value Decompositions

We may ground our further discussion on the well known Singular Value Decomposition (*SVD*, Greenacre, 1983; Abdi, 2007) theorem, that states

Theorem 1 (Singular Value Decomposition)**.** *Any real matrix X may be decomposed as $X = U\Lambda^{1/2}V'$, with Λ the diagonal matrix of the real non-negative eigenvalues of XX', U the orthogonal matrix of the corresponding eigenvectors, and V the matrix of eigenvectors of $X'X$ (with the same eigenvalues), with both constraints $U'U = I$ and $V'V = I$.*

This theorem corresponds to the reconstruction formula of a q-rank matrix

$$x_{ij} = \sum_{\alpha=1}^{q} \sqrt{\lambda_\alpha}\ \ u_{i\alpha}\ v_{j\alpha}$$

on which the Eckart and Young (1936) theorem is based:

Theorem 2 (Eckart and Young)**.** *The s-rank reconstruction of any real matrix X, with $s < q$, the rank of X, once its singular values are sorted in decreasing order,*

$$x_{ij} \approx \sum_{\alpha=1}^{s} \sqrt{\lambda_\alpha}\ \ u_{i\alpha}\ v_{j\alpha} \tag{3}$$

is the best one in the least-squares sense.

Choulakian (2004) proposes to build the *SVD* solution through a recursive optimization process. Indeed, it consists in finding the first vectors $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$ principal component of a matrix X as the solution of the equivalent optimization problems

$$\begin{aligned} &\max \parallel X\boldsymbol{u} \parallel_2, \text{ subject to } \parallel \boldsymbol{u} \parallel_2 = 1; \\ &\max \parallel X'\boldsymbol{v} \parallel_2, \text{ subject to } \parallel \boldsymbol{v} \parallel_2 = 1. \end{aligned}$$

The solution gives

$$\lambda_1 = \max_{\boldsymbol{u}} \frac{\parallel X\boldsymbol{u} \parallel_2}{\parallel \boldsymbol{u} \parallel_2} = \max_{\boldsymbol{v}} \frac{\parallel X'\boldsymbol{v} \parallel_2}{\parallel \boldsymbol{v} \parallel_2} = \max_{\boldsymbol{u},\boldsymbol{v}} \frac{\boldsymbol{v}'X\boldsymbol{u}}{\parallel \boldsymbol{u} \parallel_2 \parallel \boldsymbol{v} \parallel_2}$$

which is the largest singular value of X. The complete solution results by recursively applying the optimization problem on the residuals. Thus, the reconstruction formula holds:

$$X = \sum_{\alpha=1}^{q} \lambda_\alpha \boldsymbol{v}_\alpha \boldsymbol{u}'_\alpha$$

and it results

$$\sum_{\alpha} \lambda_\alpha^2 = \mathrm{Tr}(X'X).$$

Note that, if we consider the principal coordinates

$$\boldsymbol{f}_\alpha = X\boldsymbol{u}_\alpha, \text{ with } \boldsymbol{v}'_\alpha \boldsymbol{f}_\alpha = \| \boldsymbol{f}_\alpha \|_2 = \lambda_\alpha$$
$$\boldsymbol{g}_\alpha = X'\boldsymbol{v}_\alpha, \text{ with } \boldsymbol{u}'_\alpha \boldsymbol{g}_\alpha = \| \boldsymbol{g}_\alpha \|_2 = \lambda_\alpha$$

the reconstruction formula becomes

$$X = \sum_{\alpha=1}^{q} \frac{1}{\lambda_\alpha} \boldsymbol{f}_\alpha \boldsymbol{g}'_\alpha$$

Taxicab Singular Value Decomposition In analogy with what proposed for *SVD*, Choulakian (2004) proposes a recursive method in the Taxicab metrics too. The first vectors are the solution of the equivalent optimization problems

$$\max \| X\boldsymbol{u} \|_1, \text{ subject to } \| \boldsymbol{u} \|_\infty = 1;$$
$$\max \| X'\boldsymbol{v} \|_1, \text{ subject to } \| \boldsymbol{v} \|_\infty = 1.$$

using both L_1 and L_∞ norms: the latter is used since this way both $\boldsymbol{u}$ and $\boldsymbol{v}$ are vectors of signs, say only contain 1 and -1s. The solution

$$\lambda_1 = \max_{\boldsymbol{u}} \frac{\| X\boldsymbol{u} \|_1}{\| \boldsymbol{u} \|_\infty} = \max_{\boldsymbol{v}} \frac{\| X'\boldsymbol{v} \|_1}{\| \boldsymbol{v} \|_\infty} = \max_{\boldsymbol{u},\boldsymbol{v}} \frac{\boldsymbol{v}' X \boldsymbol{u}}{\| \boldsymbol{u} \|_\infty \| \boldsymbol{v} \|_\infty}$$

is a combinatorial problem Choulakian (2004). The complete solution results by recursively applying the optimization problem on the residuals, but it may be seen as a *TSVD*, *Taxicab Singular Value Decomposition*. The corresponding principal coordinates are

$$\boldsymbol{f}_\alpha = X\boldsymbol{u}_\alpha, \text{ with } \boldsymbol{v}'_\alpha \boldsymbol{f}_\alpha = \| \boldsymbol{f}_\alpha \|_1 = \lambda_\alpha$$
$$\boldsymbol{g}_\alpha = X'\boldsymbol{v}_\alpha, \text{ with } \boldsymbol{u}'_\alpha \boldsymbol{g}_\alpha = \| \boldsymbol{g}_\alpha \|_1 = \lambda_\alpha$$

In this case, since both $\boldsymbol{u}_\alpha$ and $\boldsymbol{v}_\alpha$ are vectors of signs ($\boldsymbol{u}_\alpha = \text{sgn}(\boldsymbol{g}_\alpha)$ and $\boldsymbol{v}_\alpha = \text{sgn}(\boldsymbol{f}_\alpha)$), the reconstruction formula becomes:

$$X = \sum_{\alpha=1}^{q} \frac{1}{\lambda_\alpha} \boldsymbol{f}_\alpha \boldsymbol{g}'_\alpha$$

Note that in L_1 metrics, the total inertia should be the sum of each layer's ones.

2.2 Simple Correspondence Analysis

Correspondence Analysis may be formulated according to different points of view. We try to ground it on *SVD*. We know that the relations between rows and columns of N are summarized by the χ^2 statistics, that measures the departure from the independence between rows and columns. Since the independence is estimated by $N_0 = nP_0 = n\boldsymbol{r}\boldsymbol{c}'$, the departure from independence is estimated by

$$\chi^2 = n\ \phi^2 = n \sum_i \sum_j \frac{(p_{ij} - p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}} = n \sum_i \sum_j \frac{d_{ij}^2}{p_{i\cdot}p_{\cdot j}} \tag{4}$$

with $(r-1)\times(c-1)$ degrees of freedom. Note that N and its grand total n are interesting only to evaluate the chi-square significance, so that interest may be concentrated most on the matrix P. Setting $\dot{S} = D_r^{-1/2} D D_c^{-1/2} = \left(\frac{d_{ij}}{\sqrt{p_{i\cdot}p_{\cdot j}}}\right)$, (4) may be written as

$$n\ \phi^2 = n\ \text{trace}(\dot{S}'\dot{S}) = n\ \text{trace}\left(\left(\frac{d_{ij}}{\sqrt{p_{i\cdot}p_{\cdot j}}}\right)'\left(\frac{d_{ij}}{\sqrt{p_{i\cdot}p_{\cdot j}}}\right)\right)$$

that is, in matrix form, simplifying:

$$\phi^2 = \text{trace}\left(D_c^{-1/2} D' D_r^{-1} D D_c^{-1/2}\right) \tag{5}$$

Our problem is to reduce the rank of P (and consequently of N) without losing relevant information. Indeed, we may formalize the problem, considering a suitable reduced rank matrix $\hat{P}$ that best approximates P in the sense of the weighed least squares, that is minimizing the residuals:

$$R = n \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(p_{ij} - \hat{p}_{ij})^2}{p_{i\cdot} p_{\cdot j}} = n \text{ trace}\left(D_c^{-1/2} (P - \hat{P})' D_r^{-1} (P - \hat{P}) D_c^{-1/2}\right) \tag{6}$$

where the weights are the inverse of the expected frequencies. Note that this formulation allows to check for significance of the residuals, since Ghomari (1983) showed that R is asintotically chi-square distributed. As the way the estimate was done does not matter, the test may be adopted for both *SCA* and *TSCA*: given a cell value $\hat{n}_{s,ij}$ issued by a s-dimensional partial reconstruction, the residuals

$$R_s = \sum_{ij} \frac{(n_{ij} - \hat{n}_{s,ij})^2}{n r_i c_j}$$

may be tested for significance as a chi-square with $(r - s - 1) \times (c - s - 1)$ degrees of freedom (Malinvaud, 1987).

Indeed, in the case of *SCA*,

$$R = \chi^2 - \sum_{\alpha=1}^{s} \chi_\alpha^2 = n \sum_{\beta=s+1}^{q} \lambda_\beta,$$

thus the test is quickly performed.

For our purpose, we may apply the SVD to $\dot{S} = D_r^{-1/2} D D_c^{-1/2} = U \Lambda^{1/2} V'$, with $U'U = I$, $V'V = I$, as it is known that the singular values $\Lambda^{1/2}$ are the square roots of the eigenvalues Λ of $\dot{S}'\dot{S}$. By setting the *principal coordinates* $F = D_r^{-1/2} U \Lambda^{1/2}$ and $G = D_c^{-1/2} V \Lambda^{1/2}$, we get $F D_r F' = \Lambda$ and $G D_c G' = \Lambda$, we get the reconstruction formula

$$nD = n(P - \boldsymbol{r}\boldsymbol{c}') = U \Lambda^{1/2} V' = n D_r F \Lambda^{-1/2} G' D_c. \tag{7}$$

Now, as $\dot{S}$ is centered, it has a zero singular value, thus its rank is $q \leq min(r, c) - 2$, and, by shifting $\boldsymbol{r}\boldsymbol{c}'$, for the elements of N (7) becomes

$$n_{ij} = n p_{ij} = n \; r_i c_j \left(1 + \sum_{\alpha=1}^{q} \frac{1}{\sqrt{\lambda_\alpha}} \; f_{i\alpha} \; g_{j\alpha}\right).$$

Thus, based on the Eckart-Young theorem, once sorted the λs in decreasing order: for any $s \leq q$ the partial s-rank reconstruction formula (3) becomes:

$$n_{ij} \approx \hat{n}_{ij,s} = n \; \hat{p}_{ij,s} = n \; p_{i\cdot} p_{\cdot j} \left(1 + \sum_{\alpha=1}^{s} \frac{1}{\sqrt{\lambda_\alpha}} \; f_{i\alpha} \; g_{j\alpha}\right), \tag{8}$$

whose residuals may be tested for nullity as said through the Malinvaud (1987) test.

2.3 Taxicab Correspondence Analysis

Taxicab Correspondence Analysis is defined as the Taxicab Singular Value Decomposition of the data table $D = P - \boldsymbol{r}\boldsymbol{c}'$, taking into account the table's *profiles*, respectively $R = D_r^{-1}D$ for the rows and $C = D_c^{-1}D$ for the columns. Unlike *SCA*, the solution is recursive, considering at each step the residuals from the previous factors. This leads to the reconstruction formula

$$P = \boldsymbol{p}_r\boldsymbol{p}_c' + \sum_{\alpha=2}^{q} \frac{1}{\lambda_\alpha}\, \boldsymbol{F}_\alpha\, \boldsymbol{G}_\alpha'.$$

since the first factor is shown to correspond to the independence, with λ_α the L_1-measure of dispersion along the α-th factor (note that $\lambda_1 = 1$). Expressed elementwise the formula becomes:

$$p_{ij} = p_{i.}p_{.j} + \sum_{\alpha=2}^{q} \frac{1}{\lambda_\alpha}\, F_{i\alpha}\, G_{j\alpha}.$$

Now, if we transform the coordinates $f_{i\alpha} = \frac{F_{i\alpha}}{p_{i.}}$ and $g_{j\alpha} = \frac{G_{j\alpha}}{p_{.j}}$ we get

$$n_{ij} = n\, r_i c_j \left(1 + \sum_{\alpha=2}^{q} \frac{1}{\lambda_\alpha}\, f_{i\alpha}\, g_{j\alpha}\right). \tag{9}$$

just as for *SCA*.

3 Examples

3.1 the Snee data

As a simple example, we take the Snee (1978) data table that crosses 592 students of the University of Delaware according to the color of the eyes and of the hair, both with 4 levels. The table N is thus:

	Hair				
Eyes	Black	Brown	Red	Blond	Total
Dark Brown	68	119	26	7	220
Light Brown	15	54	14	10	93
Green	5	29	14	16	64
Blue	20	84	17	94	215
Total	108	286	71	127	592

The table's chi-square equals $138.28984 = 592 \times 0.233598$ with 9 degrees of freedom and is thus highly significant. We may apply CA from the R package *FactoMineR*, thus obtaining 3 eigenvalues, whose statistics are reported in the following table: on the first row, the ϕ^2, that is the table's inertia, its chi-square, the degrees of freedom, the chi-square statistics p-value, a test-value (that is, its corresponding quantile in a standardized normal distribution), and the L_1-inertia, that is the sum of cells absolute values weighed by their expected values.

N	Iner	%	Cum%	CnCor	ResL2	df	p-val	v-test	ResL1
Ind	0.23360				138.28984	9	0.0000	Inf	6.4352
1	0.20877	0.89	0.89	0.4569	14.69643	4	0.0054	2.5508	2.9353
2	0.02223	0.10	0.99	0.1491	1.53828	1	0.2149	0.7896	0.8568
3	0.00260	0.01	1.00	0.0510	0.00000				

It results that the residual of the one-dimensional solution is significant, whereas the following is not. Thus, a two-dimensional solution is suitable. Note also that the first factor canonical correlation is .45, a medium value. In the table below, the coordinates, contributions and quality of representation of levels of both rows and columns are reported, as well as the sum of the qualities on both axes, a cumulate quality of representation:

Rows	Dim.1	ctr	cos2	Dim.2	ctr	cos2	cum
Dark Brown	-0.492	43.116	0.967	-0.088	13.042	0.031	0.998
Light Brown	-0.213	3.401	0.542	0.167	19.804	0.336	0.878
Green	0.162	1.355	0.176	0.339	55.910	0.773	0.949
Blue	0.547	52.128	0.977	-0.083	11.244	0.022	0.999

Columns	Dim.1	ctr	cos2	Dim.2	ctr	cos2	cum
Black	-0.505	22.246	0.838	-0.215	37.877	0.152	0.980
Brown	-0.148	5.086	0.864	0.033	2.319	0.042	0.906
Red	-0.130	0.964	0.133	0.320	55.131	0.812	0.945
Blond	0.835	71.704	0.993	-0.070	4.673	0.007	1.000

The first axis shows the opposition between *brown* and *blue eyes* as well as the one between *black* and *blond hair*: the medium value of canonical correlation lets imagine that these may be related. The second axis outlines the positions of both *green eyes* and *red hair* in respect to the said levels; in this case, a lower canonical correlation may prevent the interpretation in terms of a common factor, and suggest a possible Guttman effect. Based on the said results and applying the reconstruction formula (8) we get the partial 1-dimensional reconstruction and the corresponding residuals:

	Black	Brown	Red	Blond
Dark Brown	62	123	30	5
Light Brown	21	48	12	12
Green	10	29	7	18
Blue	16	85	22	92

	Black	Brown	Red	Blond
Dark Brown	6	-4	-4	2
Light Brown	-6	6	2	-2
Green	-5	0	7	-2
Blue	4	-1	-5	2

and the partial 2-dimensional reconstruction and the corresponding residuals:

	Black	Brown	Red	Blond
Dark Brown	67	121	25	7
Light Brown	17	50	16	11
Green	4	32	13	16
Blue	20	84	17	94

	Black	Brown	Red	Blond
Dark Brown	1	-2	1	0
Light Brown	-2	4	-2	-1
Green	1	-3	1	0
Blue	0	0	0	0

Indeed, an improvement results in the reconstruction in the 2-dimensional reconstruction, in particular in what concerns *green eyes* and *red hair*, but also all others.

Now, if we apply the *TCA* through the *R* package *TCA* to the same table, we obtain the decomposition of the inertia along each dimension as shown in the following table, with its percentage and cumulate percentage, the residual table chi-square statistics, with the corresponding p-value and test-values, and the said L_1-inertia:

N	Iner	%	Cum%	ResL2	df	p-val	v-test	ResL1
Ind	0.45913			138.28984	9	0.0000	Inf	6.4352
1	0.33883	0.74	0.74	16.06151	4	0.0029	2.7547	2.7858
2	0.08519	0.18	0.92	2.49843	1	0.1140	1.2057	0.9618
3	0.03510	0.08	1.00	0.00000				

Since no statistical test is available for *TCA*, we adopted the same Malinvaud (1987) test on the residuals: here we see that the first two dimension explain 92% of the total inertia instead of 99% of *SCA*, with a lower concentration along the first axis and a higher along the second. This is put in evidence by the little higher residual of both solutions, while the significance of the residuals of the 2-dimensional solution remains above the 5% threshold. In the table below, the coordinates on the first two dimensions are reported for both rows and columns:

	Axe_1	Axe_2		Axe_1	Axe_2
Dark Brown	-0.365	-0.065	Black	-0.480	-0.233
Light Brown	-0.214	0.153	Brown	-0.152	0.027
Green	0.071	0.172	Red	-0.069	0.248
Blue	0.445	-0.051	Blond	0.790	0.000

and we may see that the oppositions are the same as before. As well, we apply here the reconstruction formula (9) and we obtain the partial 1-dimensional reconstruction and the corresponding residuals:

	Black	Brown	Red	Blond		Black	Brown	Red	Blond
Dark Brown	61	124	28	7	Dark Brown	7	-5	-2	0
Light Brown	22	49	12	10	Light Brown	-7	5	2	0
Green	11	30	8	16	Green	-6	-1	6	0
Blue	14	83	23	94	Blue	6	1	-6	0

as well as the 2-dimensional ones. It is interesting to observe that here *blond hair* are perfectly reconstructed in the 1-dimensional solution, whereas *brown hair* are settled in the following one. whereas both *brown* and *red hair* remain with some residual of equal weight:

	Black	Brown	Red	Blond		Black	Brown	Red	Blond
Dark Brown	68	122	23	7	Dark Brown	0	-3	3	0
Light Brown	15	51	17	10	Light Brown	0	3	-3	0
Green	5	32	11	16	Green	0	-3	3	0
Blue	20	81	20	94	Blue	0	3	-3	0

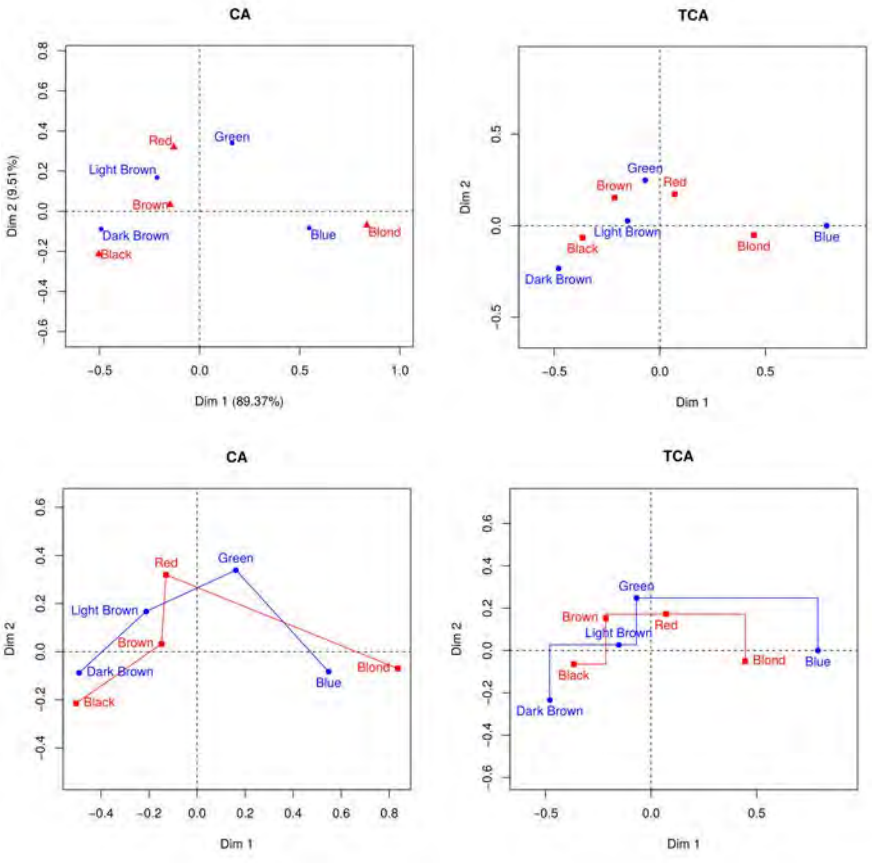

Figure 1: The scatter plot of both hair and eye colours, from Snee (1978), on the first factor plane issued by SCA correspondence analysis method (left) and that issued by TCA taxicab method (right). Below, each character levels are tied with segments showing a tentative unidimensional ordering (Guttman effect).

In Figure 1 are shown the scatter plots of both characters labels on the planes spanned by the first two factors of *SCA* (Figure 1 *left* and *right*, respectively).

In the graphics below, segments are drawn, in order to show a possible 1-dimensional ordering of levels, due to a Guttman effect. Note that on the *TCA* graphics the segments are drawn according to the $L1$ metrics, thus as paths parallel to the axes. It seems that the arch effect is reduced in the *TCA* representation in respect to *SCA*.

3.2 Ellenberg's data

As an example of the application of the polar coordinates method, we consider the Ellenberg grassland vegetation data table, taken from Müller-Dombois and Ellenberg (1974) and also used by Camiz (1994, 2005) where the whole table is reported. It concerns the presence/absence of 76 plant species in 25 rélevés, in which three types of plant communities were observed, namely *Bromus-Arrhenatherum, Geum-Arrhenatherum*, and *Cirsium-Arrhenatherum*. Previous analyses showed the existence of a main one-dimensional gradient, that appeared in the analyses in an arch-pattern according to the Guttman effect (Guttman, 1953; Camiz, 2005). We drop here showing all results, limiting attention to the pattern of rélevés: in Figure 2 they are represented on the plane spanned by the first two factors issued by both analyses. It is evident that the

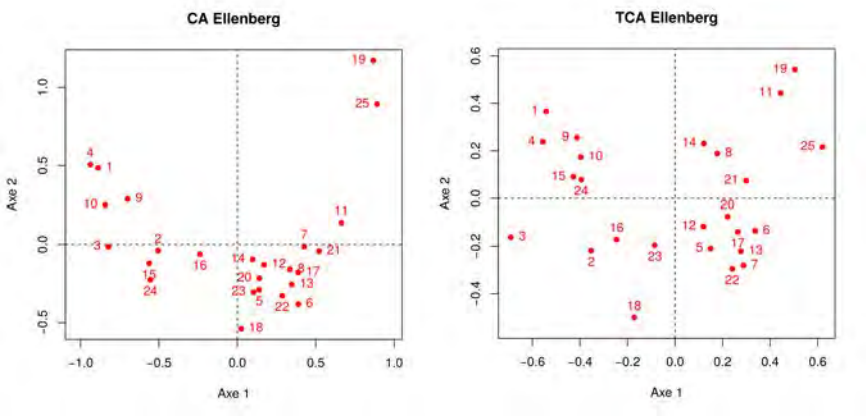

Figure 2: The scatter plot of Ellenberg's data, from Müller-Dombois and Ellenberg (1974), on the first factor plane issued by both SCA correspondence analysis (left) and TCA taxicab (right) methods.

Guttman effect, highly visible in the graphic issued by the classical *SCA* (Fig. 2 *left*) is somehow reduced in the one issued by the *TCA* (Fig. 2 *right*).

Acknowledgements

We are most indebted to Vartan Choulakian, who helped us in running the *TCA* procedure and gave suggestions for the results interpretation. This work was carried out during the reciprocal visits of both authors in the framework of the bilateral agreement between Sapienza Universitá di Roma and Universidade Federal do Rio de Janeiro, of which both authors are the scientific responsible.

References

Abdi, H. (2007). Singular Value Decomposition (SVD) and Generalized Singular Value Decomposition ($GSVD$). In: N. Salkind (Ed.), *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage.

Benzécri, J.P., et coll. (1973-82). *L'Analyse des données*, Tome 2. Paris: Dunod.

Camiz, S. (1994). A Procedure for Structuring Vegetation Tables, *Abstracta Botanica*, 18 (2): pp. 57-70.

Camiz, S. (2005). The Guttman Effect: its Interpretation and a New Redressing Method. *Tetradia Analushs Dedomenwn (Data Analysis Bulletin)*, 5: pp. 7-34.

Choulakian, V., (2004). A Comparison of two Methods of Principal Component Analysis. In J. Antoch (Ed.): *Proceedings of Compstat'2004 Symposium*. Berlin: Physica-Verlag/Springer, pp. 793-798.

Choulakian, V., (2006). Taxicab correspondence analysis. *Psychometrika*, 71(2): pp.333-345.

Choulakian, V., (2008). Multiple taxicab correspondence analysis. *Advances in Data Analysis and Classification*, 2(2): pp. 177-206.

Eckart, C., Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1, 211-218.

Ghomari, A. (1983). *Analyse Canonique d'une table de contingence. Étude asymptotique par échantillonnage*. Thèse de l'Université de Pau et des Pays de l'Adour (France).

Greenacre, M.J. (1983). *Theory and Application of Correspondence Analysis*. London: Academic Press.

Guttman, L. (1953). A Note on Sir Cyril Burt's Factorial Analysis of Qualitative Data, *British Journal of Statistical Psychology*, 6: pp. 21-24.

Malinvaud, E. (1987). *Data analysis in applied socio-economic statistics with special consideration of correspondence analysis*. Marketing Science Conference, Joy en Josas: HEC-ISA.

Müller-Dombois, D., Ellenberg, E. (1974), *Aims and Methods of Vegetation Ecology*, New York, John Wiley & Sons.

Orlóci, L. (1978). *Multivariate Analysis in Vegetation Research*, 2nd ed.. Den Haag: Junk.

R-project (2009), http://www.r-project.org/

Orlóci, L. (1978). *Multivariate Analysis in Vegetation Research*, 2nd ed.. Den Haag: Junk.

Snee, R. D. (1974). "Graphical display of two-way contingency tables". *The American Statistician*, 28: pp. 9–12.