doi.org/10.26398/IJAS.0029-008

TAXICAB CORRESPONDENCE ANALYSIS OF SPARSE TWO-WAY CONTINGENCY TABLES

Vartan Choulakian[1]

Département de mathématiques et de statistique, Université de Moncton, Moncton, Canada

Abstract. *Visualization and interpretation of contingency tables by correspondence analysis (CA), as developed by Benzécri, have a rich structure based on Euclidean geometry. However, it is a well established fact that, often CA is very sensitive to sparse contingency tables, where we characterize sparsity as the existence of relatively highvalued counts, rare observations and zero-block structure. Our main aim in this paper is to highlight the above mentioned three interrelated points in two ways: First, we propose a 7-number summary of sparsity based on the minimal size of an equivalent contingency table, where the invariance property of CA and TCA (a L1 variant of CA named taxicab) is used to construct the equivalence class of contingency tables; second, we compare the maps obtained by CA and TCA to explore under what conditions the CA and TCA maps produce similar, somewhat similar or dissimilar maps. Examples are provided.*

Keywords: *Sparse contingency tables, Correspondence analysis, Taxicab correspondence analysis, Interpretable maps, 7-number summary of sparsity.*

1. INTRODUCTION

Correspondence analysis (CA), developed by Benzécri (1973) since 1960s, as a statistical method for different kinds of data sets, in particular for contingency tables, is embedded both in theory and in practice. The theory is based on the chi-square distance between the profiles; parallel to this beautiful theory, the practice is entrenched in the joint interpretation of the graphical displays based on the Euclidean geometry. Seeing this extreme fondness of the use and interpretation of the maps by the users of CA, Nishisato (1998) suggested the replacement of the adage “seeing is believing“ with “graphing is believing“ and stressed the importance of interpretable graphs. Additionally, we recall the often cited quip “a picture is worth a thousand words“, and via geometric interpretation of maps CA offers much to the analysis of complex multivariate data sets. So the philosophical question asked by Schlick (2000, part 5) "Theory and Observation: Is seeing

[1] Corresponding author: Vartan Choulakian, email: vartan.choulakian@umoncton.ca

believing ?" is quite relevant here in the context of data analysis by CA.

It is well known that, CA is very sensitive to some particularities of a data set; further, how to identify and handle these is an open unresolved problem. Here, we enumerate three under the umbrella of sparse contingency tables: rare observations, zero-block structure and relatively high-valued cells. Rao (1995), among others, stressed the influence of rare observations (rows or columns that have relatively small marginal weights compared to others) and proposed an alternative to CA based on Hellinger distance (a square-root transformation of counts). Greenacre (2013) refuted Rao's assertion and argued that rare observations do not have an exaggerated influence in CA. Earlier Nowak and Bar-Hen (2005) developed a criterion based on the influence function to identify influential rare observations; and they arrived at the same conclusion as Greenacre in their analysis of a 207 × 15 abundance data in ecology; however they observed that "influential species are rare species that are concentrated in few plots". A similar observation is found in Greenacre (2013) "there is one exceptional situation where rare species would have a strong role in the solution, namely when a species is observed in a single sampling site and no or very few other species are observed there". We describe this particular situation as the existence of a large zero-block structure. Often few relatively high-valued cells, including outlier counts, have detrimental effect on the CA outputs by emphasizing some aspects of the data, even though apparently the interpretation of the CA maps seems meaningful to the researchers. Our main aim in this paper is to highlight the above mentioned three points by comparing the maps obtained by CA with the maps obtained by taxicab correspondence analysis (TCA), where TCA is a L_1 variant of CA; and to explore under what conditions the CA and TCA maps produce similar, somewhat similar or dissimilar maps. Our main conclusion is that: First, CA and TCA maps enrich each other; second, for sparse contingency tables, there is a positive probability that CA and TCA maps are partially similar or dissimalar. To do this we organize the paper in six sections.

In Section 2, we attempt to quantify the notion of sparsity in contingency tables by a 7-number summary based on the minimal size of an equivalent contingency table, where the invariance property of CA and TCA is used to construct the equivalence class of contingency tables. In Section 3, we present a brief mathematical comparison of CA and TCA; in Section 4 we present an empirical comparison using ten data sets; in Section 5 we consider sparsest contingency tables; and we conclude in Section 6.

The theory of CA can be found, among others, in Benzécri (1973, 1992), Greenacre (1984), Gifi (1990), Le Roux and Rouanet (2004), Murtagh (2005), and Nishisato (2007); the recent book, authored by Beh and Lombardi (2014), presents a panoramic review of CA and related methods. Since 2006, Choulakian and coauthors have studied mathematical properties of TCA applied to many kinds of non-negative data; in particular, TCA of contingency tables and their comparison with CA are studied in the following papers: Choulakian (2006), Choulakian et al. (2006), Choulakian (2008), and Choulakian, Simonetti and Gia (2014).

2. 7-NUMBER SUMMARY OF SPARSITY IN CONTINGENCY TABLES

Let $\mathbf{N} = (n_{ij})$ be a contingency table cross-classifying two nominal variables with I rows and J columns, where for $i = 1,...,I$ and $j = 1,...,J$, n_{ij} represents the frequency of statistical units having the ith category of the row variable and the jth category of the column variable. Thus, $n = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}$ represents the sample size. In the statistical literature, generally we see that the degree of sparsity of **N** are based on the following two quantities

$$ave(\mathbf{N}) = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}}{IJ},$$

the average value of counts; and

$$\%(0 \in \mathbf{N}) = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} 1_{n_{ij}=0}}{IJ} 100,$$

the percentage value of zero counts, where $1_{n_{ij}=0}$ is the indicator function: $1_{n_{ij}=0} = 1$ for $n_{ij} = 0$ and $1_{n_{ij}\neq 0} = 0$ for $n_{ij} \geq 1$.

According to Agresti and Yang (1987), **N** is *sparse* if $ave(\mathbf{N})$ is small such that the chi-squared approximations of the goodness-of-fit statistics are inaccurate. Radavicius and Samusenko (2012) characterize **N** as *very sparse* if the sample size (n) is less than the number of cells (IJ), that is, $ave(\mathbf{N}) < 1$. Greenacre (2013) uses $\%(0 \in \mathbf{N})$ as an index of sparsity.

Another qualitative definition of sparsity is used in the Ph.D thesis of Kraus (2012), based on Agresti (2002, p.391) "contingency tables having small cell counts are said to be sparse". A quantification of this definition will be given in subsection 2.3.

As we stated in the introduction, our concept of sparseness is broader, it also includes relatively large valued counts; to quantify this aspect of sparseness we consider the batch of nonzero counts of **N**, and following Tukey (1977, ch.2 or p.80), we summarize them by the 5-number summary,

$$MH1 = (min, Q1, Median, Q3, max);$$

where, *min* represents the lowest value in the batch of the positive counts, *max* the highest value, and, $Q1$, *Median* and $Q3$ are the three quartiles ($Q1$ and $Q3$ are the two hinges in Tukey's terminology). Thus, from the 7-number summary $(ave(\mathbf{X}), \%(0 \in \mathbf{X}), MH1)$, one gets an idea on the degree of sparsity concerning its different, but complementary, aspects in a contingency table $\mathbf{X}$.

2.1. EQUIVALENCE CLASS OF AN OBSERVED CONTINGENCY TABLE

An important property of CA and TCA is that columns or rows with identical profiles (conditional probabilities) receive identical factor scores. The factor scores are used in the graphical displays. Moreover, merging of identical profiles does not change the results of the data analysis: This is named the *principle of equivalent partitioning* by Nishisato (1984); it includes the famous invariance property named *principle of distributional equivalence*, on which Benzécri (1973) developed CA. Formally, Nishisato's *principle of equivalent partitioning* is based on the following

Definition 1: Let $\mathbf{N}$ be a contingency table of size $I \times J$, $\mathbf{x} = (x_k)$ and $\mathbf{y} = (y_k)$ are two rows or two columns of $\mathbf{N}$ such that they are proportional

$$\frac{\mathbf{x}}{\sum x_i} = \frac{\mathbf{y}}{\sum y_i} \quad \text{or} \quad (\sum y_i)\mathbf{x} = (\sum x_i)\mathbf{y}.$$

We construct a new contingency table, $N_{reduced}$, by replacing the two elements $\mathbf{x}$ and $\mathbf{y}$ in $\mathbf{N}$ by one element $\mathbf{x}+\mathbf{y}$, and keeping all the other columns and rows of $\mathbf{N}$ the same in $N_{reduced}$. Then we say that the contingency tables $\mathbf{N}$ and $N_{reduced}$ are equivalent, and we write $\mathbf{N} \sim N_{reduced}$. Thus the equivalence class of contingency tables of $\mathbf{N}$ is given by

$$\Omega(\mathbf{N}) = \{\mathbf{X} : \mathbf{X} \sim \mathbf{N}\}.$$

Given that, $\Omega(\mathbf{N})$ contains infinite number of contingency tables equivalent to a given $\mathbf{N}$, we define its representative element by the unique contingency table $\mathbf{M}$ of minimal size; that is, among all elements of $\Omega(\mathbf{N})$, $\mathbf{M}$ has minimum number of rows and columns. We can easily deduce the following inequalities: $ave(\mathbf{N}) \leq ave(\mathbf{M})$ and $\max(\mathbf{N}) \leq \max(\mathbf{M})$.

2.2. ARTIFICIAL EXAMPLE AND EXTREME SPARSITY

The following contrived example illustrates the idea. Let

$$\mathbf{N}=\begin{pmatrix} 1 & 2 & 0 & 0 \\ 2 & 4 & 0 & 0 \\ 0 & 0 & 1 & 2 \\ 3 & 6 & 0 & 0 \end{pmatrix}$$

be a two-way contingency table of size 4×4. Its 7-number summary of sparsity is

$$(ave(\mathbf{N})=1.3125,\ \%(0\in\mathbf{N})=50, MH1=(1,\ 1.5,\ 2,\ 3.5,\ 6)).$$

We note that the first, second and fourth rows of $\mathbf{N}$ are proportional to each other, so they can be lumped together into one row, and we obtain the equivalent contingency table

$$\mathbf{N}_1=\begin{pmatrix} 6 & 12 & 0 & 0 \\ 0 & 0 & 1 & 2 \end{pmatrix}$$

of size 2×4; its 7-number summary of sparsity is $(2.6250,\ 50,\ (1,\ 1.5,\ 4,\ 9,\ 12))$. Similarly, we see that the third and fourth columns of $\mathbf{N}_1$ are proportional, so they can be added together, and we obtain equivalent contingency table

$$\mathbf{N}_2=\begin{pmatrix} 6 & 12 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

of size 2×3. Similarly, we see that the first and second columns of $\mathbf{N}_2$ are proportional, so they can be added together, and we obtain the unique representative equivalent contingency table

$$\mathbf{M}=\begin{pmatrix} 18 & 0 \\ 0 & 3 \end{pmatrix}$$

of minimal size 2×2. The four contingency tables $\mathbf{N}$, $\mathbf{N}_1$, $\mathbf{N}_2$ and $\mathbf{M}$ are equivalent, because they belong to $\Omega(\mathbf{N})$: CA and TCA of $\mathbf{N}$, $\mathbf{N}_1$, $\mathbf{N}_2$ and $\mathbf{M}$ produce identical maps, because they have identical geometries within the mathematical framework of CA and TCA. However, we have four different 7-number summaries of the sparsity: we consider the one obtained from $\mathbf{M}$ the most representative.

We note that $\mathbf{M}$ is a diagonal contingency table and sparsest (most sparse), based on the following lemma, whose proof is given in the appendix.

Lemma 1: $\%(0\in\mathbf{M})\leq 100(1-\frac{1}{\min(I,J)})$.

Definition 2: A contingency table is named *sparsest* if $\%(0 \in \mathbf{M}) = 100(1 - \frac{1}{\min(I,J)})$, and extremely sparse if $\%(0 \in \mathbf{M})$ is very near to $100(1 - \frac{1}{\min(I,J)})$.

2.3. EXAMPLES OF SPARSE CONTINGENCY DATA SETS

Table 1 enumerates ten contingency tables and their 7-number summaries calculated on **N** and on **M**. Sections 4 and 5 provide further references to these data sets. The first data set is not sparse. For the last nine of them, which are considered to be sparse, we note that:

$$Q1 \leq 2 \quad \text{and} \quad Median \leq 5,$$

which is another quantification of sparsity describing "contingency tables having small cell counts are said to be sparse". Furthermore, comparison of $Q3$ and *max* values highlights very long tails for sparse contingency tables, which represents the existence of relatively high-valued counts. Concerning the equivalent tables **N** and **M,** we see noticeable changes in the 7-number summaries for the two data sets 6 (*Barents*) and 10 (*Synoptic Gospels*): these two contingency tables **N** and **M** are extremely tall: the number of columns is much smaller than the number of rows; so the merging of rows essentially happened for rows having very small marginal counts of 1 or 2. For these two data sets, **M** can be put in the following form

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_1 \\ \mathbf{D} \end{pmatrix},$$

where **D** is a square diagonal matrix.

We classify the data sets in Table 1 into three large groups according to our concept of sparsity:

Non sparse tables: Data set 1 (*TV programs*) belongs to this group.

Extremely sparse tables: Data set 3 (*Texel*) belongs to this group. Note that $\%(0 \in \mathbf{M}) = 96.3\%$ is very near to $100(1 - 1/220) = 99.5455$, the upper bound provided in Lemma 1.

Sparse tables: the remaining eight data sets belong to this group.

It is interesting to note that for Data set 10 (*Synoptic Gospels)*: The upper bound in Lemma 1, $100(1 - 1/7) = 85.7143$, is quite near to $\%(0 \in \mathbf{N}) = 78.2\%$, but quite far from $\%(0 \in \mathbf{M}) = 45\%$. For this reason, we characterized it as sparse and not extremely sparse.

Table 1: 7-number summary of sparsity of ten two-way contingency tables.

	size	ave	%(0)	MH1	map similarity
1) *TV programs*					yes
N=M	13 × 7	55.81	0%	(3 15 40 86 271)	
2) *Rodents*					no
N	28 × 9	3.96	66.7%	(1 2 5 12.3 78)	
M	21 × 9	5.3	58.7%	(1 2 4.5 14 78)	
3) *Texel*					no
N	285 × 220	0.26	96.6%	(1 1 1 4.8 97)	
M	266 × 220	0.28	96.3%	(1 1 1 7 97)	
4) *Macro*					partial
N	189 × 40	6.1	84.8%	(1 2 3 14 1848)	
M	161 × 40	7.47	81.9%	(1 2 3 14 1848)	
5) *Benthos*					partial
N=M	92 × 13	8.02	39%	(1 1 3 8 992)	
6) *Barents*					partial
N	446 × 10	2.91	78.4%	(1 1 2 8 798)	
M	221 × 10	5.87	67.5%	(1 1 3 10 903)	
7) *Seashore*					partial
N	126 × 68	0.14	88%	(1 1 1 1 5)	
M	106 × 65	0.17	86.4%	(1 1 1 1 12)	
8) *Punta Milazzese*					no
N=M	31 × 19	0.83	58.1%	(1 1 1 2 12)	
9) *Iversfjord*					no
N=M	37 × 14	2.643	60%	(1 1 2 6 64)	
10) *Synoptic Gospels*					no
N	7097 × 7	0.39	78.2%	(1 1 1 2 79)	
M	796 × 7	3.59	45%	(1 1 2 4 2740)	

3. CORRESPONDENCE ANALYSIS AND TAXICAB CORRESPONDENCE ANALYSIS: AN OVERVIEW

Let $\mathbf{P} = \mathbf{N}/n = (p_{ij})$ be the associated correspondence matrix of $\mathbf{N}$. We define as usual $p_{i*} = \sum_{j=1}^{J} p_{ij}$, $p_{*j} = \sum_{i=1}^{I} p_{ij}$, the vector $\mathbf{r} = (p_{i*}) \in \mathbf{R}^{I}$, the vector $\mathbf{c} = (p_{*j}) \in \mathbf{R}^{\mathbf{J}}$, and $\mathbf{D}_r = Diag(\mathbf{r})$ the diagonal matrix having diagonal elements p_{i*}, and similarly $\mathbf{D}_c = Diag(\mathbf{c})$. We suppose that $\mathbf{D}_r$ and $\mathbf{D}_c$ are positive definite metric matrices of size $I \times I$ and $J \times J$, respectively; this means that the diagonal elements of $\mathbf{D}_r$ and $\mathbf{D}_c$ are strictly positive. Let $k = rank(\mathbf{R}_0)$, where

$$\mathbf{R}_0 = (\mathbf{P} - \mathbf{r}\mathbf{c}^\top)$$

is the residual matrix with respect to the independence model. CA and TCA can be considered as principal components analysis for categorical data, where $\mathbf{P}$ or $\mathbf{R}_0$ is decomposed into a sum of bilinear terms shown in equation (1). Equation (1) is named the data reconstruction formula, and it is obtained by generalized singular value decomposition and its taxicab version with respect to the metric matrices $\mathbf{D}_r$ and $\mathbf{D}_c$, see in particular Choulakian, Simonetti and Gia (2014):

$$\mathbf{P} = \mathbf{D}_r(\mathbf{1}_I\mathbf{1}_J^\top + \sum_{\alpha=1}^{k} \mathbf{f}_\alpha \mathbf{g}_\alpha^\top/\sigma_\alpha)\mathbf{D}_c,$$

or elementwise

$$p_{ij} = p_{i*}p_{*j}\left[1 + \sum_{\alpha=1}^{k} f_\alpha(i) g_\alpha(j)/\sigma_\alpha\right], \qquad (1)$$

where $\mathbf{f}_\alpha$ and $\mathbf{g}_\alpha$ represent the principal coordinate scores of rows and columns, and σ_α is the associated dispersion measure for $\alpha = 1,...,k$. Note that in both methods $\mathbf{f}_\alpha$ and $\mathbf{g}_\alpha$ are $\mathbf{D}_r$ and $\mathbf{D}_c$ centered respectively; that is

$$\begin{aligned} \mathbf{f}_\alpha^\top \mathbf{D}_r \mathbf{1}_I &= \mathbf{g}_\alpha^\top \mathbf{D}_c \mathbf{1}_J \\ &= 0, \end{aligned} \qquad (2)$$

where $\mathbf{1}_I$ is a column vector of ones of size I.

In CA, $\mathbf{f}_\alpha$ and $\mathbf{g}_\alpha$ satisfy

$$\mathbf{f}_\alpha^\top \mathbf{D}_r \mathbf{f}_\alpha = \mathbf{g}_\alpha^\top \mathbf{D}_c \mathbf{g}_\alpha = \sigma_\alpha^2 \quad \text{for } \alpha = 1,...,k, \qquad (3)$$

$$\mathbf{f}_\alpha^\top \mathbf{D}_r \mathbf{f}_\beta = \mathbf{g}_\alpha^\top \mathbf{D}_c \mathbf{g}_\beta = 0 \quad \text{for } \alpha \neq \beta. \qquad (4)$$

Equation (3) says that the $\mathbf{D}_r$ weighted L_2 norm of $\mathbf{f}_\alpha$ is σ_α; likewise, equation (4) says that $\mathbf{f}_\alpha$ is $\mathbf{D}_r$ orthogonal to $\mathbf{f}_\beta$ for $\alpha \neq \beta$. In CA the standard coordinate scores are $\mathbf{f}_\alpha/\sigma_\alpha$ for column profiles and $\mathbf{g}_\alpha/\sigma_\alpha$ for row profiles.

In TCA, $\mathbf{f}_\alpha$ and $\mathbf{g}_\alpha$ satisfy

$$\mathbf{f}_\alpha^\top \mathbf{D}_r sgn(\mathbf{f})_\alpha = \mathbf{g}_\alpha^\top \mathbf{D}_c sgn(\mathbf{g}_\alpha) = \sigma_\alpha \quad \text{for } \alpha = 1,...,k, \qquad (5)$$

$$\mathbf{f}_\alpha^\top \mathbf{D}_r sgn(\mathbf{f}_\beta) = \mathbf{g}_\alpha^\top \mathbf{D}_c sgn(\mathbf{g})_\beta = 0 \quad \text{for } \alpha > \beta. \qquad (6)$$

where $sgn(\mathbf{g}_\alpha) = [sgn(g_\alpha(1)),...,sgn(g_\alpha(J)]^\top$, and $sgn(g_\alpha(j)) = 1$ if $g_\alpha(j) > 0$, $sgn(g_\alpha(j)) = -1$ otherwise. Equation (5) says that the $\mathbf{D}_r$ weighted L_1 norm of $\mathbf{f}_\alpha$ is σ_α; likewise, equation (6) says that $\mathbf{f}_\alpha$ is $\mathbf{D}_r$ orthogonal to $sgn(\mathbf{f}_\beta)$ for $\alpha > \beta$.

3.1. REMARKS

- a) CA of $\mathbf{P}$ is equivalent to CA of $\mathbf{R}_0$, with diagonal weight matrices $\mathbf{D}_r$ and $\mathbf{D}_c$. Analogously, TCA of $\mathbf{P}$ is equivalent to TCA of $\mathbf{R}_0$, with diagonal weight matrices $\mathbf{D}_r$ and $\mathbf{D}_c$.

- b) In CA, the principal coordinate scores $\mathbf{f}_\alpha$ and $\mathbf{g}_\alpha$ are functions of the eigenvectors of a similarity measure between the rows or columns and more importantly the similarity measure depends on the chosen metric $\mathbf{D}_r$ and $\mathbf{D}_c$. We describe the computation of $\mathbf{f}_\alpha$ and $\mathbf{g}_\alpha$ in four steps:

  Step 1: we calculate the matrix of Pearson residuals,

  $$\mathbf{S} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{rc}^\top)\mathbf{D}_c^{-1/2}. \tag{7}$$

  Step 2: we calculate the eigenvectors $\mathbf{x}_\alpha$ via the eigen-equation,

  $$\mathbf{S}^\top \mathbf{S}\mathbf{x}_\alpha = \sigma_\alpha^2 \mathbf{x}_\alpha \quad \text{with} \quad \mathbf{x}_\alpha^\top \mathbf{x}_\alpha = 1, \tag{8}$$

  where the (i, j)th element of $\mathbf{S}^\top\mathbf{S}$ represents a similarity measure between the two column categories i and j.

  Step 3: we calculate $\mathbf{f}_\alpha = \sigma_\alpha \mathbf{D}_r^{-1/2}\mathbf{x}_\alpha$.

  Step 4: we calculate $\mathbf{g}_\alpha$ via the transition formula (22).

- c) Compared to CA, TCA stays as close as possible to the original data: It directly acts on the correspondence matrix $\mathbf{P}$ or $\mathbf{R}_0$ in the largest sense that the basic taxicab decomposition is independent of the metrics $\mathbf{D}_r$ and $\mathbf{D}_c$: it is simply constructed from a sum of the signed columns or rows of the residual correspondence matrix, for further details see Choulakian (2006, 2016); only the relative direction of the rows or columns is taken into account without calculating a similarity (or dissimilarity) measure between the rows or columns.

  The optimization criterion is based on the famous Grothendieck problem, see Pisier (2012). The steps for the computation of the principal coordinate scores $\mathbf{f}_\alpha$ and $\mathbf{g}_\alpha$ are done iteratively for $\alpha = 1,...,k$:

Step 1: we compute the principal axis

$$\mathbf{u}_\alpha = \arg \max_{\mathbf{u}\in\{-1,1\}^J} ||\mathbf{R}_{\alpha-1}\mathbf{u}||_1,$$

where $\mathbf{R}_0 = \mathbf{P} - \mathbf{rc}^\top$ and $\mathbf{R}_\alpha = \mathbf{P} - \mathbf{rc}^\top - \sum_{\beta=1}^{\alpha} \mathbf{D}_r\mathbf{f}_\beta\mathbf{g}_\beta^\top\mathbf{D}_c/\sigma_\beta$ for $\alpha = 1,...,k.$

Step 2: we compute the principal coordinate scores $\mathbf{f}_\alpha = \mathbf{D}_r^{-1}\mathbf{R}_{\alpha-1}\mathbf{u}_\alpha$.

Step 3: we calculate $\mathbf{g}_\alpha$ via the transition formula (15),

$$\mathbf{g}_\alpha = \mathbf{D}_c^{-1}\mathbf{R}_{\alpha-1}^\top sgn(\mathbf{f}_\alpha)$$

.

Step 4: we update $\mathbf{R}_{\alpha+1} = \mathbf{P} - \mathbf{rc}^\top - \sum_{\beta=1}^{\alpha+1} \mathbf{D}_r\mathbf{f}_\beta\mathbf{g}_\beta^\top\mathbf{D}_c/\sigma_\beta$.

- d) An interesting and useful property of the taxicab dispersion measures, σ_α for $\alpha \geq 1$, is the following result well known in theoretical computer science, see Khot and Naor (2012):

Lemma 2:

$$\begin{aligned} \sigma_\alpha &= \frac{||\mathbf{R}_{\alpha-1}\mathbf{u}||_1}{||\mathbf{u}||_\infty} \quad \text{for} \quad \alpha \geq 1 \\ &= \max_{\mathbf{u}\in\{-1,1\}^J} ||\mathbf{R}_{\alpha-1}\mathbf{u}||_1 \\ &= 4\ ||\mathbf{R}_{\alpha-1}||_{cut}, \end{aligned}$$

where the cut norm of the matrix $\mathbf{R}_{\alpha-1}$ is defined as

$$||\mathbf{R}_{\alpha-1}||_{cut} = \max_{S\times T} |\sum_{(i,j)\in S\times T} \mathbf{R}_{\alpha-1}(i,j)| \text{ where } S \subseteq \{1,...,I\}$$

$$\text{and } T \subseteq \{1,...,J\}.$$

We know that taxicab principal axes have values ±1, that is, $\mathbf{u}_\alpha \in \{-1,1\}^J$ and $\mathbf{v}_\alpha \in \{-1,1\}^I$ for $\alpha \geq 1$. So we can represent $\mathbf{u}_\alpha = \mathbf{u}_{\alpha+} + \mathbf{u}_{\alpha-}$, and similarly, $\mathbf{v}_\alpha = \mathbf{v}_{\alpha+} + \mathbf{v}_{\alpha-}$, where

$$\begin{aligned} \mathbf{u}_{\alpha+} &= (\mathbf{1}_J + \mathbf{u}_\alpha)/2 \\ \mathbf{u}_{\alpha-} &= (\mathbf{u}_\alpha - \mathbf{1}_J)/2. \end{aligned}$$

Lemma 2 can be named 4-quadrants balancing property, because the taxicab dispersion measure σ_α for $\alpha \geq 1$ is divided into 4 equal parts having the common value of the cut norm of $\mathbf{R}_{\alpha-1}$:

$$\begin{aligned}
\sigma_\alpha/4 &= \mathbf{v}'_{\alpha+}\mathbf{R}_{\alpha-1}\mathbf{u}_{\alpha+} \\
&= \mathbf{v}'_{\alpha-}\mathbf{R}_{\alpha-1}\mathbf{u}_{\alpha-} \\
&= |\mathbf{v}'_{\alpha-}\mathbf{R}_{\alpha-1}\mathbf{u}_{\alpha+}| \\
&= |\mathbf{v}'_{\alpha+}\mathbf{R}_{\alpha-1}\mathbf{u}_{\alpha-}|.
\end{aligned}$$

As a corollary to this fact, we have: In TCA of **P** both principal coordinate scores $\mathbf{f}_\alpha$ and $\mathbf{g}_\alpha$ for $\alpha = 1,...,k$ satisfy the equivariability property, see Choulakian (2008b). This means that $\mathbf{f}_\alpha$ and $\mathbf{g}_\alpha$ are equally balanced in the sense that

$$\begin{aligned}
\frac{\sigma_\alpha}{2} &= \sum_{i\in I_{\alpha+}} p_{i*}f_\alpha(i), \\
&= -\sum_{i\in I_{\alpha-}} p_{i*}f_\alpha(i), \\
&= \sum_{j\in J_{\alpha+}} p_{*j}g_\alpha(j), \\
&= -\sum_{j\in J_{\alpha-}} p_{*j}g_\alpha(j),
\end{aligned} \tag{9}$$

where $I_{\alpha+} = \{i|f_\alpha(i) > 0\}$, $I_{\alpha-} = \{i|f_\alpha(i) < 0\}$, $J_{\alpha+} = \{j|g_\alpha(j) > 0\}$ and $J_{\alpha-} = \{j|g_\alpha(j) < 0\}$. This easily follows from the fact that the principal coordinate scores $\mathbf{f}_\alpha$ and $\mathbf{g}_\alpha$ are $\mathbf{D}_r$ and $\mathbf{D}_c$ centered, they satisfy equation (2). An informal illustrative interpretation of the equivariability property is that TCA pulls inside potential influential observations and pushes outside points around the origin, thus providing a more balanced and robust view of data.

We note that in CA the principal coordinate scores $\mathbf{f}_\alpha$ and $\mathbf{g}_\alpha$ do not satisfy the equivariability property, because they are unequally balanced in the sense that

$$\begin{aligned}
A &= \sum_{i\in I_{\alpha+}} p_{i*}f_\alpha(i), \\
&= -\sum_{i\in I_{\alpha-}} p_{i*}f_\alpha(i), \\
B &= \sum_{j\in J_{\alpha+}} p_{*j}g_\alpha(j),
\end{aligned}$$

$$= \; -\sum_{j \in J_{\alpha-}} p_{*j} g_{\alpha}(j),$$

and in general,

$$A \neq B;$$

furthermore, A and B are not related to the dispersion measure σ_{α}, because CA maximizes the variance of the principal coordinate scores.

- e) Given that the approach in CA and TCA is geometric, influence measure of a point (a column or a row) to the αth factor is provided by the contribution of that point to the dispersion measure of the αth factor in per 1000 units.

  In CA, based on (3), this corresponds to:

$$C_{\alpha}(i) = 1000 \frac{p_{i*} f_{\alpha}^{2}(i)}{\sigma_{\alpha}^{2}} \quad \text{and} \quad C_{\alpha}(j) = 1000 \frac{p_{j*} g_{\alpha}^{2}(j)}{\sigma_{\alpha}^{2}}. \tag{10}$$

  In TCA, based on (5), we have the signed contribution

$$SC_{\alpha}(i) = 1000 \frac{p_{i*} f_{\alpha}(i)}{\sigma_{\alpha}} \quad \text{and} \quad SC_{\alpha}(j) = 1000 \frac{p_{j*} g_{\alpha}(j)}{\sigma_{\alpha}}. \tag{11}$$

  It is important to note that, in CA,

$$0 < C_{\alpha}(point) < 1000; \tag{12}$$

  while in TCA, from (9) we get,

$$-500 \leq SC_{\alpha}(point) \leq 500. \tag{13}$$

- f) In both methods the maps or joint displays are obtained by plotting $(\mathbf{f}_{\alpha}, \mathbf{f}_{\beta})$ and $(\mathbf{g}_{\alpha}, \mathbf{g}_{\beta})$ for $\alpha \neq \beta$. Both CA and TCA have common residual transition formulas, see Choulakian (2006),

$$f_{\alpha}(i) = p_{i*}^{-1} \sum_{j=1}^{J} R_{\alpha-1}(i,j) u_{\alpha}(j)) \quad \text{for} \quad \alpha = 1, ..., k, \tag{14}$$

  and

$$g_{\alpha}(j) = p_{*j}^{-1} \sum_{i=1}^{I} R_{\alpha-1}(i,j) v_{\alpha}(i) \quad \text{for} \quad \alpha = 1, ..., k, \tag{15}$$

  where $\mathbf{R}_{\alpha}$ is the residual correspondence matrix, and $\mathbf{u}_{\alpha}$ and $\mathbf{v}_{\alpha}$ for $\alpha = 1, ..., k$ are the normed principal axes and related to the principal coordinate scores $\mathbf{g}_{\alpha}$ and $\mathbf{f}_{\alpha}$ for $\alpha = 1, ..., k$ in the following way. In both methods

$$\mathbf{R}_\alpha = \mathbf{P} - \mathbf{r}\mathbf{c}^\top - \sum_{\beta=1}^{\alpha} \mathbf{D}_r \mathbf{f}_\beta \mathbf{g}_\beta^\top \mathbf{D}_c / \lambda_\beta. \tag{16}$$

In TCA

$$\mathbf{u}_\alpha = sgn(\mathbf{g}_\alpha) \text{ and } \mathbf{v}_\alpha = sgn(\mathbf{f}_\alpha) \text{ for } \alpha = 1, ..., k, \tag{17}$$

so equations (14) and (15) become

$$f_\alpha(i) = p_{i*}^{-1} \sum_{j=1}^{J} R_{\alpha-1}(i,j) sgn(g_\alpha(j)) \text{ for } \alpha = 1, ..., k, \tag{18}$$

and

$$g_\alpha(j) = p_{*j}^{-1} \sum_{i=1}^{I} R_{\alpha-1}(i,j) sgn(f_\alpha(i)) \text{ for } \alpha = 1, ..., k. \tag{19}$$

Equations (18) and (19) help us to interpret the joint TCA maps in the following way: $f_\alpha(i)$, the coordinate of point i on the αth axis is the signed centroid of the residual correspondence matrix within the p_{i*}^{-1} constant. Analogous interpretation applies to $g_\alpha(j)$, the coordinate of point j on the αth axis.

In CA

$$\mathbf{u}_\alpha = \mathbf{g}_\alpha / \sigma_\alpha \text{ and } \mathbf{v}_\alpha = \mathbf{f}_\alpha / \sigma_\alpha \text{ for } \alpha = 1, ..., k. \tag{20}$$

The joint interpretation of column and row categories in the CA map is based on the well known transition formulas

$$f_\alpha(i) = \sum_{j=1}^{J} \Pr(j|i) g_\alpha(j) / \sigma_\alpha \text{ for } \alpha = 1, ..., k, \tag{21}$$

and

$$g_\alpha(j) = \sum_{i=1}^{I} \Pr(i|j) f_\alpha(i) / \sigma_\alpha \text{ for } \alpha = 1, ..., k, \tag{22}$$

where $\Pr(j|i) = p_{ij}/p_{i*}$, the conditional probability of observing j given i. Note that (21, 22) can be obtained from (14, 15) via (3, 4, 20). In (21), the principal coordinate score $f_\alpha(i)$ is the weighted average (centroid) of the principal coordinate scores $g_\alpha(j)$ within the λ_α^{-1} constant. Analogous interpretation applies to $g_\alpha(j)$, the coordinate of point j on the αth axis.

4. DATA ANALYSIS

Here we carry out CA and TCA on three data sets mentioned in Table 1, and comment on the remaining. The visual comparison of CA and TCA maps shows that we can have three distinct cases: similar maps, dissimilar maps, and partially similar maps.

4.1. TV PROGRAMS DATA SET

Table 2 presents a contingency table of size 13 × 7 taken from Benzécri (1976), where a sample of 400 individuals evaluate 13 TV programs on a Likert scale from *1(excellent)* to *5 (bad)*; also two other categories of response are included *noopinion* on the program and *dontknow* the program. The data set is not sparse.

Table 2: TV progams data

programs	excellent	verygood	good	average	bad	noopinion	dontknow
1	9	28	89	124	51	19	71
2	31	87	165	63	24	4	17
3	7	21	65	103	83	8	103
4	3	26	121	142	45	11	43
5	17	40	117	111	83	16	7
6	8	35	115	119	78	6	28
7	4	22	73	56	77	12	147
8	15	44	102	83	32	25	90
9	5	18	63	61	15	9	219
10	8	15	40	37	8	12	271
11	5	16	64	54	15	17	220
12	29	87	140	62	24	9	40
13	12	18	89	95	41	9	127
total	153	457	1243	1110	576	157	1383
C1	24	83	106	45	40	1	700
C2	128	285	63	181	330	2	11
SC1	-28	-96	-165	-137	-73	-2	500
SC2	-82	-235	-173	222	278	-10	0

Figure 1 displays CA and TCA maps, where we see that both maps are similar and produce the same interpretation. The % of explained variation for CA (resp. for TCA) of the first two dimensions are 70.7 (resp. 78%) and 21.6 (resp (16.7); with almost equivalent cumulative value of 92.4% for CA and 94.7 for TCA.

The interpretation of the first two dimensions in Figure 1 will be based on two principles: principle of dichotomy and principle of graduation.

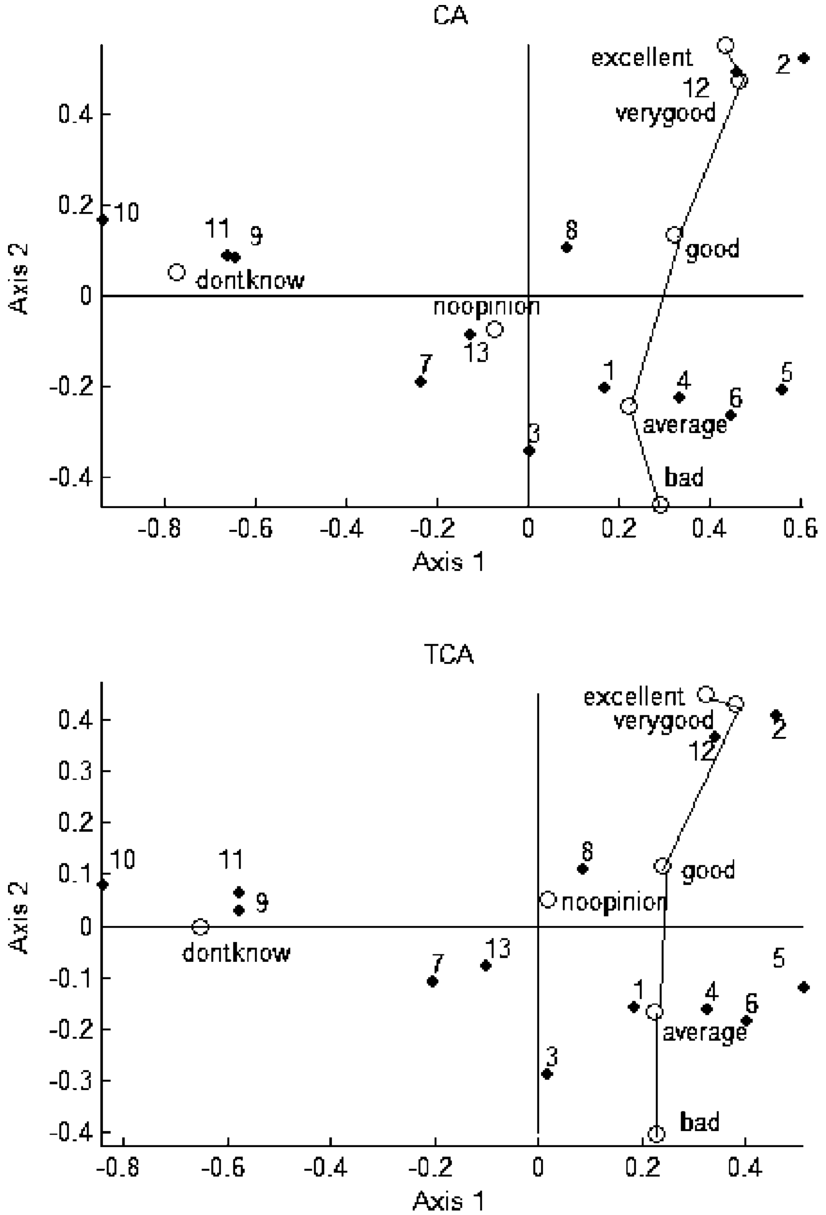

Figure 1: CA and TCA biplots of TV Programs data.

4.1.1. INTERPRETATION OF THE 1ST AXIS

We note that $C_1(dontknow) = 700$ and $SC_1(dontknow) = -500$ as given at the bottom of Table 2; so the first axis represents the dichotomy between ignorance and knowledge: where the response category *dontknow* opposes to the 5 Likert response scales; *noopinion* is near the origin.

4.1.2. INTERPRETATION OF THE 2nd AXIS

The 5 Likert response categories are ordered from *excellent* to *bad.*

4.1.3. INTERPRETATION OF THE TV PROGRAMS

Programs 2 and 12 are considered *excellent* and *verygood*; programs 10, 11 and 9 are mostly *unknown,* and so on.

It is important to note that, the response category *dontknow* is very influential in both methods CA and TCA, and it reveals a central important feature of the data: in TCA, the category *dontknow* contributes only to the first axis, because it attains the maximum value of its contribution, $|SC| = 500$, see equation (9); but this is not the case in CA.

4.2. RODENT SPECIES ABUNDANCE DATA SET

Table 2 displays abundance data (**N**) of size 28×9 (equivalent to **M** of size 21×9), where 9 species of rodents have been counted at each of 28 sites in California. For the interested reader, we identify the 9 rodents by their scientific names: *rod1=Rt.rattus, rod2=Mus.musculus, rod3=Pm.californicus, rod4=Pm.eremicus, rod5=Rs.megalotis, rod6=N.fuscipes, rod7 =N.lepida,*
rod8=Pg.fallax, and *rod9=M.californicus*. Genus abbreviations are: *Rt* (Rattus), *Rs* (Reithrodontomys), *Mus* (Mus), *Pm* (Peromyscus), *Pg* (Perognathus), *N* (Neotoma) and *M* (Microtus). Rattus and Mus, rodents 1 and 2, are invasive species, whereas the others are native. This data set is very interesting, because we see that it has, in particular, three specificities which characterize our concept of sparsity: rare observations, a zero-block structure and relatively high-valued cells. It is sparse based on the 7-number summary calculated in Table 1. It was proposed in 2014 as an exercise in a course on an ecology workshop in UBC in Canada; the workshop site mentions that the data set is downloaded from the web site of Quinn and Keough (2002), and it can be found at

https://www.zoology.ubc.ca/~bio501/R/workshops/workshops-multivariate-methods/

Table 3: Rodent species abundance data.

Sites	rod1	rod2	rod3	rod4	rod5	rod6	rod7	rod8	rod9
1		13	3	1	1	2			
2		1	57	65	9	16	8	2	3
3		4	36		2	9			
4		4	53	1	5	30		18	3
5		2	63	21	11	16			
6		1	48	35	12	8	12	2	2
7		**11**							
8		**16**							
9	**3**	**8**							
10	**1**	**2**							
11		**9**							
12		3	1		5	16		7	
13		4	39		4	12			
14	**1**	**3**							
15		**11**							
16		**4**							
17	**3**								
18		2	78		10	14		4	
19			1						
20	3		27	1					
21	**2**	**1**							
22		**3**							
23					2	8		2	
24	**1**								
25		**5**							
26			22			11		2	
27			29		10	9		1	
28			10	1		1			
total	*14*	**107**	**467**	**125**	71	**152**	20	38	8
σ_α^{TCA}	0.478	0.422	0.347	0.138	0.120	0.091	0.061	0.010	
σ_α^{CA}	0.864	0.678	0.536	0.391	0.189	0.157	0.107	0.045	
C_1	127	**750**	59	29	9	15	5	4	2
C_2	**854**	140	3	0	0	2	0	1	0
SC_1	-23	-196	298	-221	22	**135**	-51	44	-8
SC_2	-26	**-238**	**202**	**224**	32	**-139**	42	-95	-1

The instructor of the course suggested the analysis of this data set by CA in two rounds:

In round 1 the invasive species dominate, see the CA map in Figure 2. This fact is confirmed by looking at the contibutions in Table 3, where we find $C_1(rod2) = 750$, so the first axis in the CA map is dominated by rodent 2; $C_2(rod1) = 854$, so the second axis is dominated by rodent 1. The highlighted subset of sites, 7-11, 14-17, 21-22, 24-25, which are completely associated only with the invasive rodents 1 and 2, characterizes the CA solution; their total weight is $84/1002 = 8.38\%$. The minimal matrix $\mathbf{M}$ has a zero-block structure and it can be reexpressed as

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_1 & \mathbf{M}_2 \\ \mathbf{M}_3 & \mathbf{0} \end{pmatrix},$$

where

$$\mathbf{M}_3 = \begin{pmatrix} 0 & 59 \\ 3 & 8 \\ 1 & 2 \\ 1 & 3 \\ 4 & 0 \\ 2 & 1 \end{pmatrix},$$

and the submatrix $(\mathbf{M}_3 \quad \mathbf{0})$ represents the 13 highlighted sites. We conclude that the combination of the high count cell 59 (representing 7 rare sites), the last five rows in $\mathbf{M}_3$ (representing 6 rare sites), and the large zero-block in $\mathbf{M}$, created this particular CA solution.

In round 2 the instructor suggested eliminating rodents 1 and 2 and their associated sites, and carrying out a second application of CA on the reduced data set representing only the native species (the CA map is not shown).

Figure 2 displays the principal maps produced by CA and TCA, where the two invasive species and their associated sites are fenced by linear segments: they are completely different. It is evident that for this data the TCA map is much more informative than the corresponding CA map; further, one TCA map is as informative as two CA maps obtained from the two rounds.

4.2.1. INTERPRETATION OF THE TCA MAP

Let us interpret the TCA biplot, the lower diagram in Figure 2. We note that the two invasive species are grouped and found in the first quadrant of the TCA map; further, they are associated with the 13 highlighted subset of sites that we enumerated above. The contibutions to dimensions 1 and 2 of the rodents, SC_1

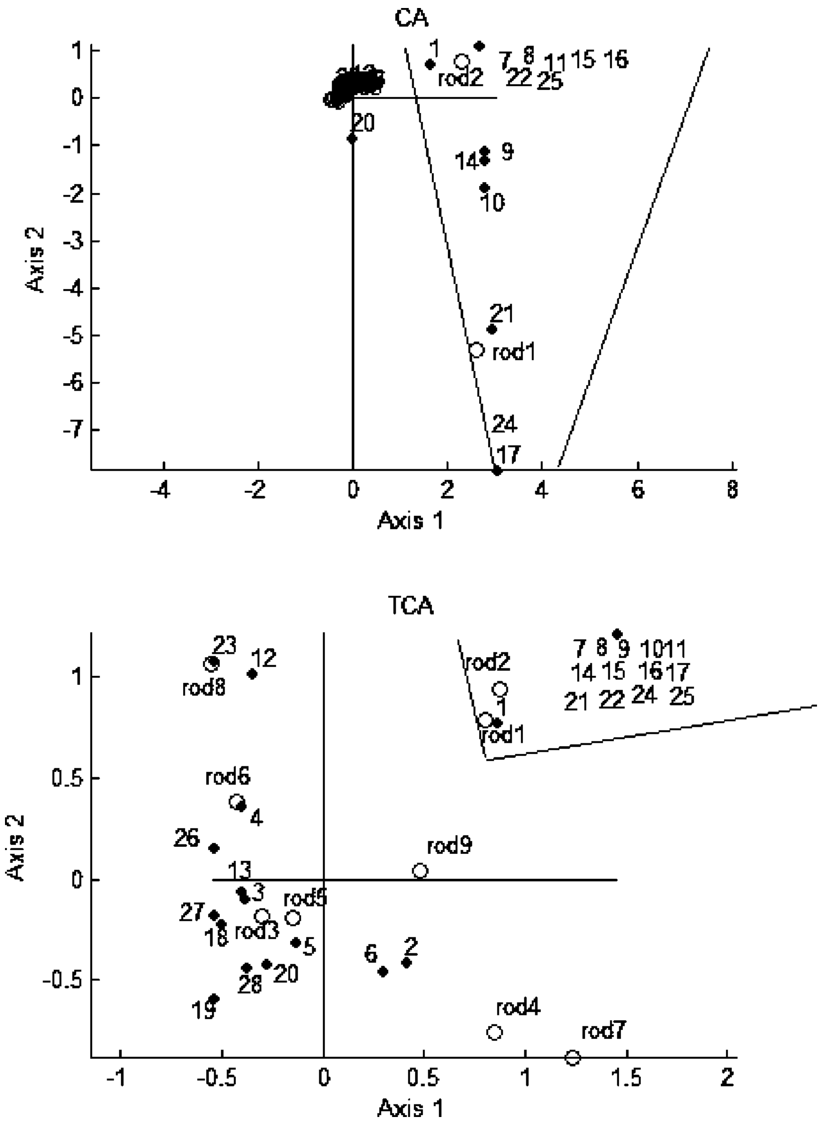

Figure 2: CA and TCA biplots of Rodents data.

and SC_2 displayed in Table 2, show that the TCA map is dominated by the four most frequent species: rodents 2, 6, 3 and 4; and each of them occupy a quadrant in the TCA map.

Let us interpret the most frequent ($freq$) species rodent 3 ($freq = 467$ out of 1002): it dominates the third quadrant, and it is associated with site3 ($freq = 36$), site5 ($freq = 63$), site13 ($freq = 39$), site18 ($freq = 78$), site27 ($freq = 29$) and site28 ($freq = 10$). Rod3 is also associated with rod5, their positions are quite near on Figure 2; looking at the entries in the column of rod5, we see that its high frequency sites are site27 ($freq = 10$), site18 ($freq = 10$), site6 ($freq = 12$), site5 ($freq = 11$); further these high frequency sites 27, 18, 6 and 5 also characterize rod3. We also note that site6 is associated with both species rod3 ($freq = 48$) and rod4 ($freq = 35$), and its position is in between rod3 and rod4; however it is found in quadrant 3, because $(35/125) > 48/467$.

4.2.2. COMPARISON

By comparing the first two principal dimensions in CA and TCA, we note the following two facts:

a) Rodent 2, with nonnegligeable weight (around 10%), has a very high influence on the first principal axis in CA ($C_1(rod2) = 750$); but it distributes its influence onto the first two principal axes in TCA ($SC_1(rod2) = -196$ and $SC_2(rod2) = -238)$.

b) Rodent 1 is a rare influential point in CA ($weight = 1.4\%$ and $C_2(rod1) = 854$); but is no more influential in TCA ($SC_1(rod1) = -23$ and $SC_2(rod1) = -26)$.

We conclude that in Figure 2, the CA map emphasized some particular aspects of the data set; while the TCA map revealed the central abundances in Table 2.

4.3. MACRO ABUNDANCE DATA SET

This macroinvertebrate sparse abundance data set of size 197×40 was also considered by Greenacre (2013). Figures 3 and 4 display the CA and TCA maps: The first dimension has almost the same separation of the 40 sites, so the same interpretation; while the second dimension seems somewhat different. For this reason we labeled the map similarity partial in Table 1.

4.4. REMAINING DATA SETS

Here, we just give references for the remaining data sets listed in Table 1.

The five ecology abundance data sets numbered 3 to 7 are available in Greenacre (2013) and he discussed them in his essay.

Mallet-Gauthier and Choulakian (2015) analyse *Punta Milazesse* and *Iverjsford* abundance data in archeology; they reproduce the data sets, provide CA and TCA maps and their interpretations.

Choulakian et al. (2006) is the reference for the *Synoptic Gospels* textual count data; they provide CA and TCA maps and discuss their stability.

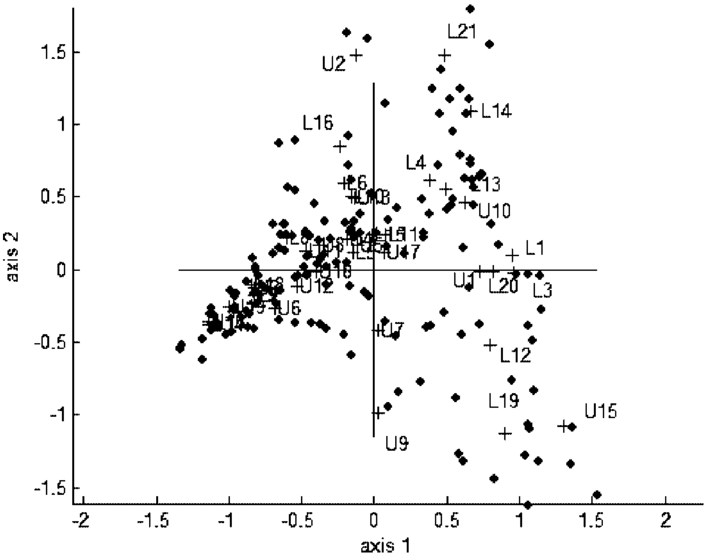

Figure 3: CA map of Macro data.

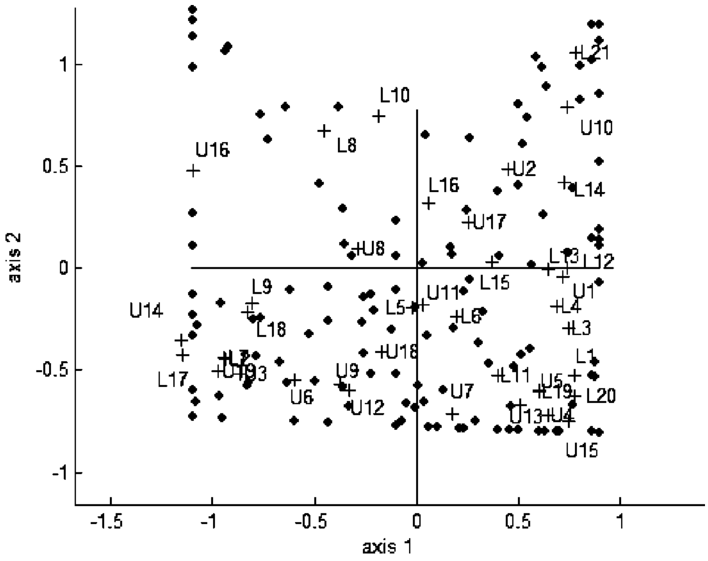

Figure 4: TCA map of Macro data.

5. SPARSEST CONTINGENCY TABLES

Sparsest contingency tables are diagonal contingency tables, as we defined in section 2. Here, we show that CA and TCA results are completely different.

5.1. CA OF SPARSEST CONTINGENCY TABLES

Let **N** be a diagonal contingency table of size *I*, then it is well known, see for instance Benzécri (1973, p. 188) that CA of **N** produces $(I-1)$ dispersion measures of 1; that is, $\sigma_\alpha^{CA} = 1$ for $\alpha = 1,...,I-1$. So, CA shows that **N** is composed of *I* diagonal blocks. A similar result is known in spectral clustering as Fiedler's theorem, see Choulakian and de Tibeiro (2013).

5.2. TCA OF SPARSEST CONTINGENCY TABLES

Let $\mathbf{P_N} = diag(p_1,...,p_I)$ be the correspondence matrix of a diagonal contingency table **N**. Then we have the following easily proven result:

Corollary to Lemma 2: $\sigma_1^{TCA} = 1$ if and only if there is a subset $S \subset \{1,...,I\}$ such that $\sum_{i\in S} p_i = 0.5$.

proof: By Lemma 2,

$$\begin{aligned}
\sigma_1^{TCA} &= 4\ \mathbf{v}'_{1+}\mathbf{R}_0\mathbf{u}_{1+} \\
&= 4\ \mathbf{u}'_{1+}\mathbf{R}_0\mathbf{u}_{1+},\ \text{ for } \mathbf{R}_0 \text{ is symmetric} \\
&= 4\ \sum_{i\in S} p_i(1-\sum_{i\in S} p_i),
\end{aligned}$$

where $S = \{i : \mathbf{u}_{1+}(i) = 1 \text{ for } 1,...,I\}$, and the required result follows.

5.3. EXAMPLES

We present three exemples, two contrived and one real.

5.3.1. EXAMPLE 1

Let $\mathbf{N} = \mathbf{Diag}(1,\ 2,\ 3,\ 4,\ 6)$; then CA produces identical singular values $\sigma_1^{CA} = \sigma_2^{CA} = \sigma_3^{CA} = \sigma_4^{CA} =\ 1$; while TCA produces $\sigma_1^{TCA} = 1$ for $2+6 = 1+3+4$, and the remaining dispersion measures are

$$\sigma_\alpha^{TCA} = 0{:}875;\ 0{:}85714\ \text{ and }\ 0.18750.$$

5.3.2. EXAMPLE 2

Let $\mathbf{N} = \mathbf{Diag}(1,\ 2,\ 3,\ 4,\ 5)$; then CA produces $\sigma_1^{CA} = \sigma_2^{CA} = \sigma_3^{CA} = \sigma_4^{CA} = 1$; while TCA produces dispersion measures

$$\sigma_\alpha^{TCA} = 0.99556,\ 0.95714,\ 0.95522\ \text{ and }\ 0.17778.$$

5.3.3. TEXEL ABUNDANCE DATA SET

Greenacre (2013) described this data set of size 285 × 220 "as large and very sparse table of vegetation abundances on a coastal sand dune area on the island of Texel, the Netherlands". CA is of no help for the analysis of this table, because according to Greenacre CA needs as much as 71 dimensions. This is evident by looking at the sequence of CA singular values: The first five singular values are: $\sigma_\alpha^{CA} =$ 1, 0.9932, 0.9908, 0.9798 and 0.9761; $\sigma_\alpha^{CA} =$ 1 means that by permuting rows and columns of the data set, the data can be represented in two diagonal blocks. The corresponding TCA dispersion measures are: $\sigma_\alpha^{TCA} =$ 0.8026, 0.7768, 0.7331, 0.7106 and 0.7012. Figures 5 and 6 display the CA and TCA maps, which are completely different. We leave the interpretation of the TCA map, if there is any, to the ecologists.

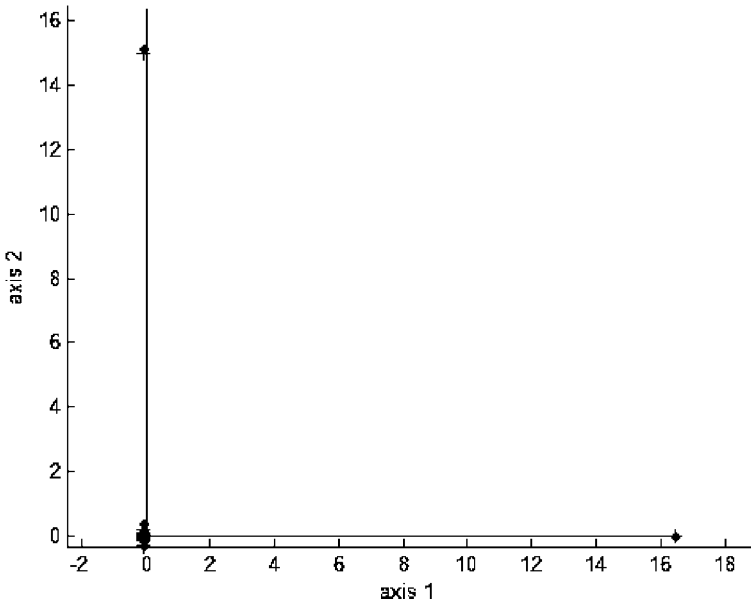

Figure 5: CA map of Texel data.

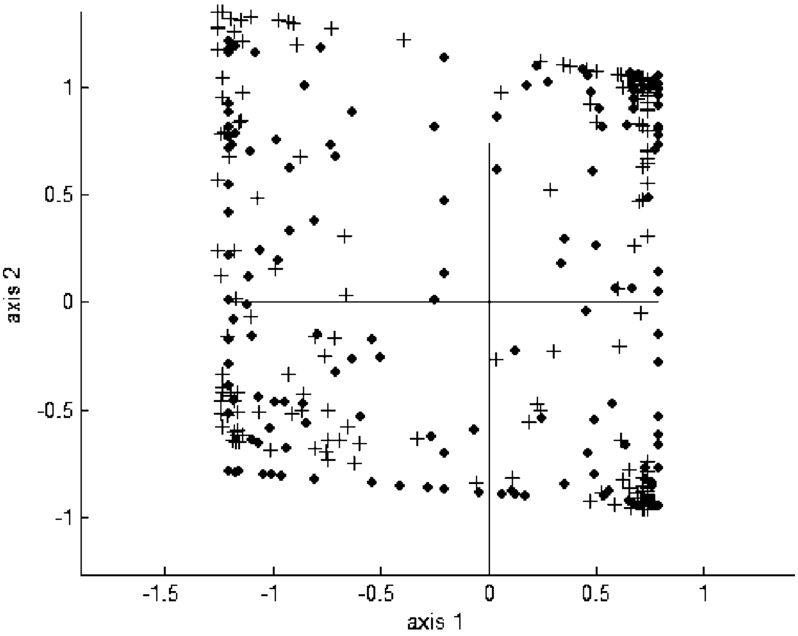

Figure 6: TCA map of Texel data.

6. CONCLUSION

The fundamental aim of CA and TCA is to produce interpretable maps that reflect central contents in a data set. In this paper, first we provided a 7-number quantification of sparsity; then we showed that for sparse contingency tables CA and TCA maps can differ with positive probability, because a map produced by CA or TCA is dependent on the underlying geometry, Euclidean or Taxicab. Based on our experience, we suggest the analysis of a data set by both methods CA and TCA: Like a cubist painting where an object is painted from different angles, sometimes the views are similar, and at other times dissimilar or partially similar.

ACKNOWLEDGEMENTS

This research has been supported by NSERC of Canada. The detailed comments of M. Greenacre and two anonymous reviewers significantly improved the presentation of this essay.

APPENDIX

Lemma 1: $\%(0 \in \mathbf{N}) \leq 100(1 - \frac{1}{\min(I,J)})$.

The proof is very easy by considering 2 distinct cases of data sets, square $(I = J)$ and rectangular $(J > I)$.

Case 1: If $\mathbf{M}$ is diagonal and has exactly I nonzero cells, then by permuting some rows and columns, it can be rearranged into a diagonal contingency table; thus

$$\begin{aligned} \%(0 &\in \mathbf{M}) = 100(I^2 - I)/I^2 \\ &= 100(1 - 1/I), \end{aligned}$$

which is the upper bound. If $\mathbf{M}$ is a square contingency table but not diagonal, by permuting some rows and columns, it can be rearranged into a square contingency table with all diagonal cells nonzero plus some, say, α number of nondiagonal nonzero cells. Then it is evident that

$$\begin{aligned} \%(0 &\in \mathbf{M}) = 100(I^2 - I - \alpha)/I^2 \\ &\leq 100(1 - 1/I). \end{aligned}$$

Case 2: $\mathbf{M}$ is rectangular and $(J > I)$. Then $\mathbf{M} = (\mathbf{M}_1|\mathbf{M}_2)$, where $\mathbf{M}_1$ is square with nonzero diagonal elements of size $I \times I$ and has α number of nondiagonal nonzero cells ; $\mathbf{M}_2$ is rectangular of size $I \times (J - I)$, such that each column of $\mathbf{M}_2$ has exactly β_i nonzero cells for $i = 1, ..., (J - I)$ and $2 \leq \beta_i \leq I$. Then

$$\begin{aligned} \%(0 &\in \mathbf{M}) = \frac{100\left[(IJ - I - \alpha - \sum_{i=1}^{J-I} \beta_i)\right]}{IJ} \\ &\leq 100(1 - 1/I), \end{aligned}$$

because $\sum_{i=1}^{J-I} \beta_i \geq 2(J - I)$.

REFERENCES

Agresti, A. (2002). *Categorical Data Analysis.* John Wiley & Sons, New York, 2nd edition.

Agresti, A. and Yang, M.C. (1987). An empirical investigation of some effects of sparseness in contingency tables. In *Computational Statistics & Data Analysis.* 5: 9–21.

Beh, E. and Lombardo, R. (2014). *Correspondence Analysis: Theory, Practice and New Strategies.* John Wiley & Sons, New York.

Benzécri, J.P. (1976). Sur le codage réduit d'un vecteur de description en analyse des correspondances. In *Les Cahiers de l'analyse des données.* 1(2): 127-136.

Benzécri, J.P. (1973). *L'Analyse des Données: Vol. 2: L'Analyse des Correspondances.* Dunod, Paris.

Benzécri, J.P (1992). *Correspondence Analysis Handbook.* Marcel Dekker, New York.

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

Choulakian, V. (2008). Taxicab correspondence analysis of contingency tables with one heavyweight column. In *Psychometrika.* 73: 309-319.

Choulakian, V., Kasparian, S., Miyake, M., Akama, H., Makoshi, N. and Nakagawa, M. (2006). A statistical analysis of synoptic gospels. In J.R. Viprey, editor, *Proceedings of 8th International Conference on Textual Data, JADT'2006,* Besançon, Presses Universitaires de Franche-Comté: 281-288.

Choulakian, V. and de Tibeiro, J. (2013). Graph partitioning by correspondence analysis and taxicab correspondence analysis. In *Journal of Classification.* 30: 397-427.

Choulakian, V., Allard, J. and Simonetti, B. (2013). Multiple taxicab correspondence analysis of a survey related to health services. In *Journal of Data Science.* 11(2): 205-229.

Choulakian, V., Simonetti, B. and Gia, T.P. (2014). Some new aspects of taxicab correspondence analysis. In *Statistical Methods and Applications.* 23: 401-416.

Choulakian, V. (2016). Matrix factorizations based on induced norms. In *Statistics, Optimization and Information Computing.* 4: 1-14.

Gifi, A. (1990). *Nonlinear Multivariate Analysis.* John Wiley & Sons, New York.

Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis.* London, Academic Press.

Greenacre, M. (2013). The contributions of rare objects in correspondence analysis. In *Ecology.* 94(1): 241-249.

Khot, S. and Naor, A. (2012). Grothendieck-type inequalities in combinatorial optimization. In *Communications in Pure and Applied Mathematics.* 65 (7): 992-1035.

Kraus, K. (2012). *On the Measurement of Model Fit for Sparse Categorical Data.* Ph.D thesis, Acta Universitatis Upsaliensis, Uppsala.

Le Roux, B. and Rouanet, H. (2004). *Geometric Data Analysis. From Correspondence Analysis to Structured Data Analysis.* Dordrecht: Kluwer–Springer.

Mallet-Gauthier, S. and Choulakian, V. (2015). Taxicab correspondence analysis of abundance data in archeology: three case studies revisited. In *Archeologia e Calcolatori.* 26: 77-94.

Murtagh, F. (2005). *Correspondence Analysis and Data Coding with Java and R.* Chapman & Hall/ CRC, Boca Raton, FL.

Nishisato, S. (1984). Forced classification: A simple application of a quantification method. In *Psychometrika.* 49(1): 25-36.

Nishisato, S. (2007). *Mutidimensional Nonlinear Descriptive Analysis.* Boca Raton, Chapman & Hall/CRC.

Nishisato, S. (1998). Graphing is believing: interpretable graphs for dual scaling. In J. Blasius and M. Greenacre, editors, *Visualization of Categorical Data,* NY, Academic Press: 185-196.

Nowak, E. and Bar-Hen, A. (2005). Influence function and correspondence analysis. In *Journal of Statistical Planning and Inference.* 134: 26-35.

Pisier, G. (2012). Grothendieck's theorem, past and present. In *Bulletin of the American Mathematical Society.* 49 (2): 237-323.

Quinn, G. and Keough, M. (2002). *Experimental Design and Data Analysis for Biologists.* Cambridge, UK, Cambridge University Press.

Radavicius, M. and Samusenko, P. (2012). Goodness-of-fit tests for sparse nominal data based on grouping. In *Nonlinear Analysis: Modelling and Control.* 17 (4): 489–501.

Rao, C.R. (1995). A review of canonical coordinates and an alternative to correspondence analysis. In *Qüestiió.* 19: 23-63.

Schlick, Th. (2000). *Readings in the Philosophy of Science: From Positivism to Post Modernism.* Mayfield Publishing Company, Mountain View, California.

Tukey, J. (1977). *Exploratory Data Analysis.* Addison-Wesley, Massachusetts.