Journal of Data Science **11**(2013), 205-229

Multiple Taxicab Correspondence Analysis of a Survey Related to Health Services

Vartan Choulakian[1]*, Jacques Allard[1] and Biagio Simonetti[2]
[1] *Université de Moncton and* [2] *University of Sannio*

Abstract: We present an analysis of a health survey data by multiple correspondence analysis (MCA) and multiple taxicab correspondence analysis (MTCA), MTCA being a robust L_1 variant of MCA. The survey has one passive item, gender, and 22 active substantive items representing health services offered by municipal authorities; each active item has four answer categories: this service is used, never tried, tried with no access, non response. We show that the first principal MTCA factor is perfectly characterized by the sum score of the category this service is used over all service items. Further, we prove that such a sum score characterization always exists for any survey data.

Key words: First factor success, multiple correspondence analysis, multiple taxicab correspondence analysis, non response, outliers, sum score.

1. Introduction

The data, that will be discussed in this paper, represent a survey of 3530 individuals residing in downtown eastside Vancouver with high incidence of AIDS/ HIV related diseases. Table 1 displays the marginal distribution of 22 active or substantive response variables or items filled by the 3530 respondents, where each item describes a health related service offered by municipal authorities; for instance, the first question asks whether the service offered on needle exchange, coded by NXCHG, was used or not. Each item represents a polytomous qualitative variable having four categories: (1) = used this service, (2) = never tried, (3) = tried with no access, (N) = non response or missing. In Europe, particularly in France, multiple correspondence analysis (MCA) is a popular method to describe and visually explore complex relationships among items in such a questionnaire survey. MCA is the application of correspondence analysis (CA) to the super indicator 0/1 matrix $\boldsymbol{Z}$ of size 3530×88. The number of columns 88 comes from

*Corresponding author.

4×22, which represents the total number of categories of the 22 items. To see how the matrix $\boldsymbol{Z}$ is constructed, refer to Section 3. An advantage of coding the data as in $\boldsymbol{Z}$ is that the missing values are incorporated in data analysis naturally without imputation, just like any other category value. Imputation for missing categorical survey data is discussed quite in detail by Finch (2010). The aim of this paper is to compare the MCA results with the multiple taxicab correspondence analysis (MTCA) results, MTCA being a robust L_1 version of MCA developed by Choulakian (2006; 2008a; 2008b). Because of its robustness, MTCA will reveal that there is a clear structure in this data set based on a simple sum score statistic. Further, we show that such a sum score characterization always exists for any survey questionnaire data; and this will help the researcher to see if the active items are broadly similar in objective and point to the same direction.

Table 1: The marginal distribution of frequencies of the categories of 22 health related service items, with symbols used for their representations

Categories	Used this service (1)	Never tried (2)	Tried with no access (3)	Missing (N)
Needle exchange(NXCHG)	1832	1592	61	45
Food bank(FB)	1404	2021	58	47
Pharmacy(PH)	790	2647	69	24
Methadone treatment(MET)	2733	666	24	107
HIV medications(HIVM)	3119	268	18	125
A&D counselling(ADC)	2597	815	30	88
Nursing care(NUR)	2353	1055	48	74
Doctor care(DOC)	802	2631	79	18
Mental health unit(MHU)	2890	498	23	119
Mental health worker(MHW)	2888	504	19	119
Outreach worker(OWU)	2653	743	28	106
Detox-residential(DETR)	2931	470	45	84
Day-tox day program(DETD)	3295	95	7	133
Recovery house(RH)	3123	291	8	108
Other drug treatment centre(ODTC)	3182	227	6	115
Ambulance pick-up(APU)	2588	867	26	49
Emergency Department - sph(EDSPH)	2396	1050	25	59
Emergency Department - vgh(EDVGH)	3007	412	12	99
Emergency Department - other(EDO)	3117	292	6	115
Hospital admission - sph(HASPH)	2874	542	18	96
Hospital admission - vgh(HAVGH)	3167	233	11	119
Hospital admission - other(HAO)	3218	184	4	124

First we present the underlying mathematics, then we discuss the case study. This paper is organized as follows. In Section 2 we present an overview of taxicab correspondence analysis of a contingency table; Section 3 presents the main

theoretical results; in Sections 4 and 5 we present the analysis of the survey data by MCA and MTCA, respectively; and we conclude in Section 6.

We suppose that the theory of multiple correspondence analysis (MCA) is known, which can be found, among others, in Benzecri (1973; 1992), Greenacre (1993), Gifi (1990), Nishisato (1994), Le Roux and Rouanet (2004). Note that MCA is also known as homogeneity analysis, reciprocal averaging, dual scaling or third method of quantification.

2. Taxicab Correspondence Analysis: An Overview

2.1 Introduction

In a series of papers Choulakian (2003; 2005; 2006a; 2006b) developed principle component analysis (PCA) based on matrix norms, thus generalizing the classical PCA, or equivalently generalizing the well known singular value decomposition (SVD). This led to the development of taxicab principal component analysis (TPCA) based on the most robust matrix norm named taxicab matrix norm, and on which taxicab correspondence analysis (TCA) is based.

To see that TPCA is similar to and has the same mathematical framework of classical PCA, we start with an overview of classical PCA, which can be described in many ways, see Jolliffe (2002) for a comprehensive account. However, TPCA is similar to only one of the ways, that we present it in the next subsection to make the paper self contained and reader friendly.

2.2 Classical Principal Component Analysis

Let $\boldsymbol{T}$ be a centered or standardized data set of dimension $I \times J$, where I observations are described by the J variables, that is, $\boldsymbol{T}'\boldsymbol{T}/I$ is the covariance or the correlation matrix. For a vector $\boldsymbol{u} \in \boldsymbol{R}^J$, we define its Euclidean or L_2-norm to be $||\boldsymbol{u}||_2 = (\boldsymbol{u}'\boldsymbol{u})^{\frac{1}{2}}$. Let $k = \text{rank}(\boldsymbol{T})$. The classical principal component analysis (PCA) consists of successive maximization of the variance or the square of the L_2-norm of the linear combination of the variables of the matrix $\boldsymbol{T}$ subject to a quadratic constraint; that is, it is based on the following optimization problem

$$\max ||\boldsymbol{T}\boldsymbol{u}||_2 \quad \text{subject to } ||\boldsymbol{u}||_2 = 1; \tag{1}$$

or equivalently, PCA can also be described as maximization of the square of the L_2-norm of the linear combination of the observations of the matrix $\boldsymbol{T}$

$$\max \left|\left|\boldsymbol{T}'\boldsymbol{v}\right|\right|_2 \quad \text{subject to } ||\boldsymbol{v}||_2 = 1. \tag{2}$$

Equation (1) is the dual of (2), and they can be reexpressed as matrix norms

$$\begin{aligned}\lambda_1 &= \max_{\boldsymbol{u}\in\mathbb{R}^J} \frac{||\boldsymbol{T}\boldsymbol{u}||_2}{||\boldsymbol{u}||_2} \\ &= \max_{\boldsymbol{v}\in\mathbb{R}^I} \frac{||\boldsymbol{T}'\boldsymbol{v}||_2}{||\boldsymbol{v}||_2} \\ &= \max_{\boldsymbol{u}\in\mathbb{R}^J,\, \boldsymbol{v}\in\mathbb{R}^I} \frac{\boldsymbol{v}'\boldsymbol{T}\boldsymbol{u}}{||\boldsymbol{u}||_2\,||\boldsymbol{v}||_2}.\end{aligned} \tag{3}$$

The solution to (3), λ_1, is the square root of the greatest eigenvalue of the matrix $\boldsymbol{T}'\boldsymbol{T}$ or $\boldsymbol{T}\boldsymbol{T}'$. The first principal axes, $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$, are defined as

$$\boldsymbol{u}_1 = \arg\max_{\boldsymbol{u}} ||\boldsymbol{T}\boldsymbol{u}||_2 \quad \text{such that} \quad ||\boldsymbol{u}_1||_2 = 1, \tag{4}$$

where $\boldsymbol{u}_1$ is the eigenvector of the matrix $\boldsymbol{T}'\boldsymbol{T}$ associated with the greatest eigenvalue λ_1; and

$$\boldsymbol{v}_1 = \arg\max_{\boldsymbol{v}} \left|\left|\boldsymbol{T}'\boldsymbol{v}\right|\right|_2 \quad \text{such that} \quad ||\boldsymbol{v}_1||_2 = 1. \tag{5}$$

Let $\boldsymbol{f}_1$ be the vector of the first principal component (pc) scores, and $\boldsymbol{g}_1$ the vector of the first pc loadings defined as

$$\boldsymbol{f}_1 = \boldsymbol{T}\boldsymbol{u}_1 \quad \text{and} \quad \boldsymbol{g}_1 = \boldsymbol{T}'\boldsymbol{v}_1; \tag{6}$$

then

$$||\boldsymbol{f}_1||_2 = \boldsymbol{v}_1'\boldsymbol{f}_1 = ||\boldsymbol{g}_1||_2 = \boldsymbol{u}_1'\boldsymbol{g}_1 = \lambda_1. \tag{7}$$

Equations (6) and (7) are named transitional formulas, because $\boldsymbol{v}_1$ and $\boldsymbol{f}_1$, and, $\boldsymbol{u}_1$ and $\boldsymbol{g}_1$, are related by

$$\boldsymbol{u}_1 = \boldsymbol{g}_1/\lambda_1 \quad \text{and} \quad \boldsymbol{v}_1 = \boldsymbol{f}_1/\lambda_1. \tag{8}$$

To obtain the second pc scores $\boldsymbol{f}_2$, loadings $\boldsymbol{g}_2$, and axes $\boldsymbol{u}_2$ and $\boldsymbol{v}_2$, we repeat the above procedure on the residual dataset

$$\boldsymbol{T}_2 = \boldsymbol{T}_1 - \boldsymbol{f}_1\boldsymbol{g}_1'/\lambda_1, \tag{9}$$

where $\boldsymbol{T}_1 = \boldsymbol{T}$. We note that $\text{rank}(\boldsymbol{T}_2) = \text{rank}(\boldsymbol{T}_1) - 1$, because by (6) and (7)

$$\boldsymbol{T}_2\boldsymbol{u}_1 = \boldsymbol{0} \quad \text{and} \quad \boldsymbol{T}_2'\boldsymbol{v}_1 = \boldsymbol{0}. \tag{10}$$

Classical PCA can be described as the sequential repetition of the above procedure for $k = \text{rank}(\boldsymbol{T})$ times till the residual matrix becomes $\boldsymbol{0}$; thus, using $\alpha = 1, \cdots, k$ as indices, the matrix $\boldsymbol{T}$ can be written as

$$\boldsymbol{T} = \sum_{\alpha=1}^{k} \boldsymbol{f}_\alpha\boldsymbol{g}_\alpha'/\lambda_\alpha, \tag{11}$$

which, by (8), can be rewritten in a form known as singular value decomposition (SVD)

$$\boldsymbol{T} = \sum_{\alpha=1}^{k} \lambda_\alpha \boldsymbol{v}_\alpha \boldsymbol{u}'_\alpha. \tag{12}$$

Further, we have

$$\lambda_\alpha = ||\boldsymbol{f}_\alpha||_2 = ||\boldsymbol{g}_\alpha||_2 \quad \text{and } \lambda_\alpha\text{'s are decreasing for } \quad \alpha = 1, \cdots, k; \tag{13}$$

and

$$\begin{aligned} Tr(\boldsymbol{T}'\boldsymbol{T}) = Tr(\boldsymbol{T}\boldsymbol{T}') &= \sum_{\alpha=1}^{k} \lambda_\alpha^2 \\ &= \sum_{\alpha=1}^{k} ||\boldsymbol{f}_\alpha||_2^2 = \sum_{\alpha=1}^{k} ||\boldsymbol{g}_\alpha||_2^2, \end{aligned} \tag{14}$$

which represents, by the Pythagorean theorem, I times the sum of the variances of the J variables or the sum of the squared Euclidean distances of the I rows from the origin, because we assumed that $\boldsymbol{T}$ is centered or standardized. Also the relative cumulative explained variability by the first α axes is

$$CEV(\alpha) = \sum_{\gamma=1}^{\alpha} \lambda_\gamma^2 / \sum_{\beta=1}^{k} \lambda_\beta^2 \quad \text{for} \quad \alpha = 1, \cdots, k. \tag{15}$$

2.3 Taxicab Principal Component Analysis (TPCA)

The L_1 norm of a vector $\boldsymbol{v} = (v_1, \cdots, v_m)'$ is defined to be $||\boldsymbol{v}||_1 = \sum_{i=1}^{m} |v_i|$ and $||\boldsymbol{v}||_\infty = \max_i |v_i|$ is the L_∞ norm. TPCA consists of maximizing the L_1 norm of the linear combination of the variables of the matrix subject to L_∞ norm constraint; more precisely, it is based on the following optimization problem

$$\max ||\boldsymbol{T}\boldsymbol{u}||_1 \quad \text{subject to} \quad ||\boldsymbol{u}||_\infty = 1; \tag{16}$$

or equivalently, TPCA can also be described as maximization of the L_1 norm of the linear combination of the rows of the matrix $\boldsymbol{T}$

$$\max \left|\left|\boldsymbol{T}'\boldsymbol{v}\right|\right|_1 \quad \text{subject to} \quad ||\boldsymbol{v}||_\infty = 1. \tag{17}$$

Equation (17) is the dual of (16), and they can be reexpressed as matrix norms

$$\begin{aligned} \lambda_1 &= \max_{\boldsymbol{u} \in \mathbb{R}^J} \frac{||\boldsymbol{T}\boldsymbol{u}||_1}{||\boldsymbol{u}||_\infty} \\ &= \max_{\boldsymbol{v} \in \mathbb{R}^I} \frac{||\boldsymbol{T}'\boldsymbol{v}||_1}{||\boldsymbol{v}||_\infty} \\ &= \max_{\boldsymbol{u} \in \mathbb{R}^J,\, \boldsymbol{v} \in \mathbb{R}^I} \frac{\boldsymbol{v}'\boldsymbol{T}\boldsymbol{u}}{||\boldsymbol{u}||_\infty ||\boldsymbol{v}||_\infty}, \end{aligned} \tag{18}$$

which is a well known and much discussed matrix norm related to Grothendieck problem, see for instance, Alon and Naor (2006). The solution to (18), λ_1, is a combinatorial optimization problem given by

$$\max ||\boldsymbol{T}\boldsymbol{u}||_1 \text{ subject to } \boldsymbol{u} \in \{-1,+1\}^J. \tag{19}$$

Equation (19) characterizes the robustness of the method, in the sense that, the weights affected to the variables (similarly to the individuals by duality) are uniform ± 1. The first principal axes, $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$, are defined as

$$\boldsymbol{u}_1 = \arg\max_{\boldsymbol{u}} ||\boldsymbol{T}\boldsymbol{u}||_1 \quad \text{such that} \quad ||\boldsymbol{u}_1||_\infty = 1, \tag{20}$$

and

$$\boldsymbol{v}_1 = \arg\max_{\boldsymbol{v}} \left|\left|\boldsymbol{T}'\boldsymbol{v}\right|\right|_1 \quad \text{such that} \quad ||\boldsymbol{v}_1||_\infty = 1. \tag{21}$$

Let $\boldsymbol{f}_1$ be the the vector of the first principal component (pc) scores, and $\boldsymbol{g}_1$ the vector of the first pc loadings. These are defined as

$$\boldsymbol{f}_1 = \boldsymbol{T}\boldsymbol{u}_1 \text{ and } \boldsymbol{g}_1 = \boldsymbol{T}'\boldsymbol{v}_1; \tag{22}$$

then

$$||\boldsymbol{f}_1||_1 = \boldsymbol{v}_1'\boldsymbol{f}_1 = ||\boldsymbol{g}_1||_1 = \boldsymbol{u}_1'\boldsymbol{g}_1 = \lambda_1. \tag{23}$$

Equations (22) and (23) are named transitional formulas, because $\boldsymbol{v}_1$ and $\boldsymbol{f}_1$, and, $\boldsymbol{u}_1$ and $\boldsymbol{g}_1$, are related by

$$\boldsymbol{u}_1 = \text{sgn}(\boldsymbol{g}_1) \quad \text{and} \quad \boldsymbol{v}_1 = \text{sgn}(\boldsymbol{f}_1), \tag{24}$$

where $\text{sgn}(\boldsymbol{g}_1) = (\text{sgn}(g_1(1)), \cdots, \text{sgn}(g_1(J))'$, and $\text{sgn}(g_1(j)) = 1$ if $g_1(j) > 0$, $\text{sgn}(g_1(j)) = -1$ otherwise. Note that (24) is completely different from (8).

To obtain the second pc scores $\boldsymbol{f}_2$, loadings $\boldsymbol{g}_2$, and axes $\boldsymbol{u}_2$ and $\boldsymbol{v}_2$, we repeat the above procedure on the residual dataset

$$\begin{aligned} \boldsymbol{T}_2 &= \boldsymbol{T}_1 - \boldsymbol{T}_1\boldsymbol{u}_1\boldsymbol{v}_1'\boldsymbol{T}_1/\lambda_1 \\ &= \boldsymbol{T}_1 - \boldsymbol{f}_1\boldsymbol{g}_1'/\lambda_1, \end{aligned} \tag{25}$$

where $\boldsymbol{T}_1 = \boldsymbol{T}$. We note that $\text{rank}(\boldsymbol{T}_2) = \text{rank}(\boldsymbol{T}_1) - 1$, because by (22), (23) and (24)

$$\boldsymbol{T}_2\boldsymbol{u}_1 = \boldsymbol{0} \quad \text{and} \quad \boldsymbol{T}_2'\boldsymbol{v}_1 = \boldsymbol{0}; \tag{26}$$

which implies that

$$\boldsymbol{u}_1'\boldsymbol{g}_\alpha = 0 \quad \text{and} \quad \boldsymbol{v}_1'\boldsymbol{f}_\alpha = 0 \quad \text{for} \quad \alpha = 2, \cdots, k. \tag{27}$$

TPCA is described as the sequential repetition of the above procedure for $k = \text{rank}(\boldsymbol{T})$ times till the residual matrix becomes $\mathbf{0}$; thus the matrix $\boldsymbol{T}$ can be written as

$$\boldsymbol{T} = \sum_{\alpha=1}^{k} \boldsymbol{f}_\alpha \boldsymbol{g}'_\alpha / \lambda_\alpha. \tag{28}$$

It is important to note that (28) has the same form as (11), but it can not be rewritten as (12), because (24) is completely different from (8).

Further, similar to (13), we have

$$\lambda_\alpha = ||\boldsymbol{f}_\alpha||_1 = ||\boldsymbol{g}_\alpha||_1 \quad \text{for} \quad \alpha = 1, \cdots, k. \tag{29}$$

But the dispersion measures λ_α's in (29) will not satisfy (14), because the Pythagorean theorem is not satisfied in L_1. Given that for the classical PCA (14) is used, so for both methods we define the total variability to be

$$TotD = \sum_{\alpha=1}^{k} \lambda_\alpha^2, \tag{30}$$

and the relative cumulative explained variability by the first α axes to be

$$CEV(\alpha) = \sum_{\gamma=1}^{\alpha} \lambda_\gamma^2 / \sum_{\beta=1}^{k} \lambda_\beta^2 \quad \text{for} \quad \alpha = 1, \cdots, k. \tag{31}$$

In TPCA, the optimization problem (16), (17) or (18) can be accomplished by two algorithms. The first one is based on complete enumeration (19); this can be applied, with the present state of desktop computing power, say, if $\min(I, J) \simeq 25$. The second one is based on iterating the transitional formulas (22), (23) and (24), similar to Wold's (1966) NIPALS algorithm, also named criss-cross regression by Gabriel and Zamir (1979). It is easy to show that this is also an ascent algorithm. The criss-cross algorithm can be summarized in the following way, where $\boldsymbol{g}$ is a starting value:

Step 1: $\boldsymbol{u} = \text{sgn}(\boldsymbol{g})$, $\boldsymbol{f} = \boldsymbol{T}\boldsymbol{u}$ and $\lambda(\boldsymbol{u}) = ||\boldsymbol{T}\boldsymbol{u}||_1$;

Step 2: $\boldsymbol{v} = \text{sgn}(\boldsymbol{f})$, $\boldsymbol{g} = \boldsymbol{T}'\boldsymbol{v}$ and $\lambda(\boldsymbol{v}) = ||\boldsymbol{T}'\boldsymbol{v}||_1$;

Step 3: If $\lambda(\boldsymbol{v}) - \lambda(\boldsymbol{u}) > 0$, go to Step 1; otherwise, stop.

This is an ascent algorithm; that is, it increases the value of the objective function λ at each iteration. The convergence of the algorithm is superlinear (very fast, at most two iterations); however it could converge to a local maximum; so we restart the algorithm I times using each row of $\boldsymbol{T}$ as a starting value. The

iterative algorithm is statistically consistent in the sense that as the sample size increases there will be some observations in the direction of the principal axes, so the algorithm will find the optimal solution.

For the survey dataset, the computations are done by the iterating algorithm.

2.4 Taxicab Correspondence Analysis of A Contingency Table

Often correspondence analysis (CA) is identified as categorical PCA; that is, it is considered an adaptation of PCA to contingency tables. Similarly we consider TCA an adaptation of TPCA to contingency tables. Here we introduce TCA of a contingency table $\boldsymbol{N} = (n_{ij})$ of two nominal variables with I rows and J columns. Let $\boldsymbol{P} = \boldsymbol{N}/n$ be the associated correspondence matrix with elements p_{ij}, where $n = \sum_{j=1}^{J} \sum_{i=1}^{I} n_{ij}$ is the sample size. We define $p_{i\cdot} = \sum_{j=1}^{J} p_{ij}$, $p_{\cdot j} = \sum_{i=1}^{I} p_{ij}$, the vector $\boldsymbol{r} = (p_{i\cdot}) \in \mathbb{R}^I$, the vector $\boldsymbol{c} = (p_{\cdot j}) \in \mathbb{R}^J$, and $\boldsymbol{D}_r = \mathrm{Diag}(\boldsymbol{r})$ a diagonal matrix having diagonal elements $p_{i\cdot}$, and similarly $\boldsymbol{D}_c = \mathrm{Diag}(\boldsymbol{c})$.

The application of TPCA algorithm to $\boldsymbol{P}$, described in the previous subsection, is named TCA of the contingency table $\boldsymbol{N}$. We put $\boldsymbol{P}_0 = \boldsymbol{P}$ and denote by $\boldsymbol{P}_\alpha$ be the residual correspondence matrix at the α-th iteration. That is, in the calculations described in the previous subsection, we replace $\boldsymbol{T}$ by $\boldsymbol{P}$ and the numbering of the iterations α varies from 0 to k, where $k = \mathrm{rank}(\boldsymbol{P}) - 1$.

For $\alpha = 0$, $\boldsymbol{P}_0 = \boldsymbol{P}$. Row and column profiles with their masses play an important role in both CA and TCA. Let $\boldsymbol{R}_0 = \boldsymbol{D}_r^{-1}\boldsymbol{P}_0 = (r_{ij}) = (p_{ij}/p_{i\cdot})$ designate the row profiles, that is for each i, $\sum_{j=1}^{J} r_{ij} = 1$. The cloud of row profiles with their masses is the set $\{(\boldsymbol{r}_{0i}, p_{i\cdot})|$ for $i = 1, \cdots, I\}$, where $\boldsymbol{r}_{0i}$ is the ith row of $\boldsymbol{R}_0$; and the cloud of column profiles with their masses is the set $\{(\boldsymbol{c}_{0j}, p_{\cdot j})|$ for $j = 1, \cdots, J\}$, where $\boldsymbol{c}_{0j}$ is the jth row of $\boldsymbol{C}_0 = \boldsymbol{D}_c^{-1}\boldsymbol{P}_0'$. We shall interpret the steps of TCA using the row profiles; however, we remind the reader that similar interpretation can be done using the column profiles.

For $\alpha = 0$, the optimization problem (16) is

$$\begin{aligned} \max ||\boldsymbol{P}_0\boldsymbol{u}||_1 &= \max ||\boldsymbol{D}_r\boldsymbol{R}_0\boldsymbol{u}||_1 \ \text{ subject to } \ ||\boldsymbol{u}||_\infty = 1; \\ &= \max \sum_{i=1}^{I} p_{i\cdot} \, |\boldsymbol{r}_{0i}\boldsymbol{u}| \ \text{ subject to } \ ||\boldsymbol{u}||_\infty = 1. \end{aligned} \tag{32}$$

The objective function in (32) is the weighted L_1 dispersion of the projection of the row profiles $\boldsymbol{r}_{0i}$ on the axis $\boldsymbol{u}$. The 0-th principal axes are, see (20) and (21),

$$\boldsymbol{u}_0 = \arg \max_{\boldsymbol{u} \in \{-1,+1\}^J} ||\boldsymbol{P}_0\boldsymbol{u}||_1 \quad \text{and} \quad \boldsymbol{v}_0 = \arg \max_{\boldsymbol{v} \in \{-1,+1\}^I} \left|\left|\boldsymbol{P}_0{}'v\right|\right|_1, \tag{33}$$

which can be seen to be trivially $\boldsymbol{u}_0 = \mathbf{1}_J$, the J component vector with coordinates of 1's, and $\boldsymbol{v}_0 = \mathbf{1}_I$. The 0-th principal factor scores are

$$\boldsymbol{f}_0 = \boldsymbol{D}_r^{-1}\boldsymbol{P}_0\boldsymbol{u}_0 = \boldsymbol{R}_0\boldsymbol{u}_0 \quad \text{and} \quad \boldsymbol{g}_0 = \boldsymbol{D}_c^{-1}\boldsymbol{P}_0'\boldsymbol{v}_0 = \boldsymbol{C}_0\boldsymbol{v}_0, \tag{34}$$

which can be seen to be trivially $\boldsymbol{f}_0 = \mathbf{1}_I$ and $\boldsymbol{g}_0 = \mathbf{1}_J$; these are related to the corresponding principal axes by (24); that is,

$$\boldsymbol{u}_0 = \operatorname{sgn}(\boldsymbol{g}_0) = \mathbf{1}_J \quad \text{and} \quad \boldsymbol{v}_0 = \operatorname{sgn}(\boldsymbol{f}_0) = \mathbf{1}_I. \tag{35}$$

And, the 0-th taxicab dispersion measure can be represented in many different ways as

$$\begin{aligned}\lambda_0 &= \left|\left|\boldsymbol{P}_0'\boldsymbol{v}_0\right|\right|_1 = ||\boldsymbol{p}_c||_1 = ||\boldsymbol{D}_c\boldsymbol{g}_0||_1 = \boldsymbol{u}_0'\boldsymbol{D}_c\boldsymbol{g}_0 \\ &= ||\boldsymbol{P}_0\boldsymbol{u}_0||_1 = ||\boldsymbol{p}_r||_1 = ||\boldsymbol{D}_r\boldsymbol{f}_0||_1 = \boldsymbol{v}_0'\boldsymbol{D}_r\boldsymbol{f}_0 \\ &= 1.\end{aligned} \tag{36}$$

The first residual correspondence matrix is, by (25),

$$\begin{aligned}\boldsymbol{P}_1 &= \boldsymbol{P}_0 - \boldsymbol{P}_0\boldsymbol{u}_0\boldsymbol{v}_0'\boldsymbol{P}_0/\lambda_0 \\ &= \boldsymbol{P}_0 - \boldsymbol{D}_r\boldsymbol{f}_0\boldsymbol{g}_0'\boldsymbol{D}_c/\lambda_0 \\ &= \boldsymbol{P} - \boldsymbol{p}_r\boldsymbol{p}_c'.\end{aligned} \tag{37}$$

Note that $\boldsymbol{p}_r\boldsymbol{p}_c'$ represents the correspondence matrix under the assumption that the row and column variables are independent. This solution is considered trivial both in CA and in TCA.

For $\alpha = 1$, we define the residual row and column profiles to be: $\boldsymbol{R}_1 = \boldsymbol{D}_r^{-1}\boldsymbol{P}_1$ and $\boldsymbol{C}_1 = \boldsymbol{D}_c^{-1}\boldsymbol{P}_1$. The cloud of the residual row profiles with their masses is the set $\{(\boldsymbol{r}_{1i}, p_{i\cdot})|$ for $i = 1, \cdots, I\}$, where $\boldsymbol{r}_{1i}$ is the ith row of $\boldsymbol{R}_1$; and the cloud of residual column profiles with their masses is the set $\{(\boldsymbol{c}_{1j}, p_{\cdot j})|$ for $j = 1, \cdots, J\}$, where $\boldsymbol{c}_{1j}$ is the jth row of $\boldsymbol{C}_1 = \boldsymbol{D}_c^{-1}\boldsymbol{P}_1'$. We repeat steps (20) through (25), or (32) through (37), where $\boldsymbol{P}_0$ is replaced by $\boldsymbol{P}_1$. Note that the maximization problem is NP hard and not trivial. So in general, the α-th taxicab dispersion measure can be represented in many different ways

$$\begin{aligned}\lambda_\alpha &= ||\boldsymbol{P}_\alpha u_\alpha||_1 = ||\boldsymbol{D}_r\boldsymbol{f}_\alpha||_1 = \boldsymbol{v}_\alpha'\boldsymbol{D}_r\boldsymbol{f}_\alpha \\ &= \left|\left|\boldsymbol{P}_\alpha' v_\alpha\right|\right|_1 = ||\boldsymbol{D}_c\boldsymbol{g}_\alpha||_1 = \boldsymbol{u}_\alpha'\boldsymbol{D}_c\boldsymbol{g}_\alpha.\end{aligned} \tag{38}$$

And the $(\alpha+1)$-th residual correspondence matrix is

$$\begin{aligned}\boldsymbol{P}_{\alpha+1} &= \boldsymbol{P}_\alpha - \boldsymbol{D}_r\boldsymbol{f}_\alpha\boldsymbol{g}_\alpha'\boldsymbol{D}_c/\lambda_\alpha \\ &= \boldsymbol{P}_0 - \sum_{\beta=1}^{\alpha}\boldsymbol{D}_r\boldsymbol{f}_\beta\boldsymbol{g}_\beta'\boldsymbol{D}_c/\lambda_\beta.\end{aligned} \tag{39}$$

From which one gets the data reconstitution formula both in TCA and CA

$$p_{ij} = p_{i.}p_{.j}[1 + \sum_{\alpha=1}^{k} f_\alpha(i)g_\alpha(j)/\lambda_\alpha]. \tag{40}$$

Similar to the classical CA, the total dispersion is defined to be $\sum_{\alpha=1}^{k} \lambda_\alpha^2$, and the proportion of the explained variation by the α-th principal axis is $\lambda_\alpha^2/\sum_{\beta=1}^{k} \lambda_\beta^2$, and the cumulative explained variation is

$$CEV(\alpha) = \sum_{\gamma=1}^{\alpha} \lambda_\gamma^2 / \sum_{\beta=1}^{k} \lambda_\beta^2 \quad \text{for} \quad \alpha = 1, \cdots, k. \tag{41}$$

The visual maps are obtained by plotting the points $(f_\alpha(i), f_\beta(i))$ for $i = 1, \cdots, I$ or $(g_\alpha(j), g_\beta(j))$ for $j = 1, \cdots, J$, for $\alpha \neq \beta$.

An important property of TCA and CA is that columns (or rows) with identical profiles (conditional probabilities) receive identical factor scores. One important advantage of TCA over CA is that it stays as close as possible to the original data: It directly acts on the correspondence matrix $\boldsymbol{P}$ without calculating a dissimilarity (or similarity) measure between the rows or columns. TCA does not admit a distance interpretation between profiles; there is no chi-square like distance in TCA. Fichet (2009) described it as a scoring method.

More technical details about TCA and a deeper comparison between TCA and CA is done in Choulakian (2006a). Further results can be found in Choulakian *et al.* (2006), Choulakian (2008a), and Choulakian and de Tibeiro (2012).

3. Main Theoretical Results

3.1 Multiple Taxicab Correspondence Analysis

Let n individuals fill out a questionnaire survey consisting of Q items, and each item has J_q number of answer categories. Let j_q be the value of the jth category in the qth item for $q = 1, \cdots, Q$ and $j_q = 0, \cdots, J_q - 1$. Let $\boldsymbol{Z}$ be the super indicator 0/1 matrix of order $n \times \sum_{q=1}^{Q} J_q$. An example of a matrix $\boldsymbol{Y}$ and its 0/1 indicator matrix $\boldsymbol{Z}$ are shown below with $n = 4$, $Q = 3$ and $J_1 = 3$, $J_2 = 3$ and $J_3 = 2$.

$$\boldsymbol{Y} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 1 & 1 \\ 2 & 2 & 0 \end{pmatrix} \Longrightarrow \boldsymbol{Z} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}.$$

CA of $\boldsymbol{Z}$ is named MCA of $\boldsymbol{Y}$ and TCA of $\boldsymbol{Z}$ is named MTCA of $\boldsymbol{Y}$.

Theorem 1. (Choulakian, 2008b): Along the first principal axis, the projected response patterns in MTCA of $\boldsymbol{Y}$ will be clustered and the number of cluster points is less than or equal to $1+Q$.

This theorem shows that MTCA automatically clusters the response patterns, that is the individuals, into at most $1+Q$ clusters. This is an important feature of the method, and an important help to the researcher. Note that some clusters can be empty.

3.2 Characterization of the First MTCA Principal Factor as Sum Score Statistic

The next theorem, which is new, characterizes completely the $1+Q$ clusters as a sum score statistic, more precisely as total number of "first factor successes" over all the items. So the crucial point is how to define first factor success of an item, and and its complement "first factor failure". It is important to note that the sum score statistic of items makes sense only when the nature of the items are similar, which in Section 5 we will see is the case for the data set considered in this paper.

First we consider the case of dichotomous items, when $J_q = 2$ for $q = 1, \cdots, Q$; then generalize the result to polytomous items.

Theorem 2. (The first MTCA factor property for dichotomous items): Let $\boldsymbol{Y} \in R^{n\times Q}$, where $Y_{ij} = 0$ if the response of the ith individual on the jth dichotomous item is a failure, and $Y_{ij} = 1$ if the response of the ith individual on the jth dichotomous item is a success, and consider MTCA of $\boldsymbol{Y}$. Then the first principal factor scores $f_1(i)$ and subject sum scores $Y_{i.}$, for $i = 1, \cdots, n$, are linearly related (i.e., $\text{corr}(f_1(i), Y_{i.}) = \pm 1$) if and only if the first principal factor item weights is $\boldsymbol{u}_1 = (\boldsymbol{u}'_{11} \mid \boldsymbol{u}'_{12})' = (\mathbf{1}'_Q \mid -\mathbf{1}'_Q)'$ and, when it is the case, $f_1(i) = 2(Y_{i.} - Y_{..}/n)/Q$.

Proof: Let $\boldsymbol{Y} \in R^{n\times Q}$, where $Y_{ij} = 0$ or 1, $\boldsymbol{Z} = (\boldsymbol{Y} \mid \mathbf{1}_n\mathbf{1}'_Q - \boldsymbol{Y})$ of size $n \times 2Q$ is the 0/1 indicator matrix of $\boldsymbol{Y}$, and $\boldsymbol{P} = \boldsymbol{Z}/(nQ)$ is the correspondence matrix of size $n \times 2Q$. Then $p_{i\cdot} = \sum_{j=1}^{2Q} p_{ij} = 1/n$, $p_{\cdot j} = \sum_{i=1}^{n} p_{ij} = Y_{\cdot j}/(nQ)$ if $j = 1, \cdots, Q$ and $p_{\cdot j} = \sum_{i=1}^{n} p_{ij} = (n - Y_{\cdot j})/(nQ)$ if $j = Q+1, \cdots, 2Q$. Equation (37), the first residual correspondence matrix is

$$\boldsymbol{P}_1 = \frac{((Y_{ij} - Y_{\cdot j}/n) \mid -(Y_{ij} - Y_{\cdot j}/n))}{nQ}. \tag{42}$$

The second matrix block in $\boldsymbol{P}_1$ $(-(Y_{ij} - Y_{\cdot j}/n))$ is the negative of the first matrix block $(Y_{ij} - Y_{\cdot j}/n)$, so $\boldsymbol{u}_1 = (\boldsymbol{u}'_{11} \mid \boldsymbol{u}'_{12})' = (\boldsymbol{u}'_{11} \mid -\boldsymbol{u}'_{11})'$ that maximizes λ_1 in

(38). By (34), we get

$$f_1(i) = \frac{2}{Q}\sum_{j=1}^{Q} u_{11j}(Y_{ij} - Y_{.j}/n) \quad \text{for} \quad i = 1, \cdots, n. \tag{43}$$

It is evident that (43) equals

$$f_1(i) = \frac{2}{Q}(Y_{i.} - Y_{..}/n), \tag{44}$$

if and only if $u_{11j} = 1$, which is the required result. □

Since the orientation of $\boldsymbol{f}_1$ is arbitrary, if the condition of Theorem 2 holds, we will choose $\boldsymbol{f}_1$ so that $\text{corr}(f_1(i), Y_{i.}) = 1$. In this case, the points $(f_1(i), Y_{i.})$ will lie on a straight line by (44).

To see what happens if some $u_{11j} = -1$, we consider the case when only one, say, $u_{11Q} = -1$. Then by (43), we have

$$\begin{aligned} f_1(i) &= \frac{2}{Q}\left[\sum_{j=1}^{Q-1}(Y_{ij} - Y_{.j}/n) - (Y_{iQ} - Y_{.Q}/n)\right] \\ &= \frac{2}{Q}\left[\sum_{j=1}^{Q}(Y_{ij} - Y_{.j}/n) - 2(Y_{iQ} - Y_{.Q}/n)\right] \\ &= \frac{2}{Q}\left[(Y_{i.} - Y_{..}/n + 2Y_{.Q}/n)\right] \quad \text{if} \quad Y_{iQ} = 0 \qquad (45) \\ &= \frac{2}{Q}\left[(Y_{i.} - Y_{..}/n - 2(1 - Y_{.Q}/n)\right] \quad \text{if} \quad Y_{iQ} = 1. \qquad (46) \end{aligned}$$

Equations (45) and (46) show that the points $(f_1(i), Y_{i.})$ will locate on two parallel lines defined by success or failure of the ith respondent on item Q.

Definition: a) For a dichotomous item q for $q = 1, \cdots, Q$, we define the first factor success of the item q to be the category of the item q with first MTCA factor score $g_1(j_q) > 0$ for $j_q = 0, 1$.

b) For a polytomous item q for $q = 1, \cdots, Q$, we define the first factor success of the item q to be the category set $\{j_q | g_1(j_q) > 0 \quad \text{for} \quad j_q = 0, \cdots, J_q - 1\}$.

Now, we can interpret Theorem 2 in the following way:

a) All the success (coded as 1 in $\boldsymbol{Y}$) categories of the Q items, $\boldsymbol{u}_{11} = \boldsymbol{1}_Q$, oppose all the failure categories (coded as 0 in $\boldsymbol{Y}$) of the Q items, $\boldsymbol{u}_{12} = -\boldsymbol{1}_Q$; that is, the first principal axis is $\boldsymbol{u}_1 = (\boldsymbol{u}'_{11} \mid \boldsymbol{u}'_{12})' = (\boldsymbol{1}'_Q \mid -\boldsymbol{1}'_Q)'$.

b) A success of item q is identical to the first factor success of item q; that is, for each item success and first factor success coincide. If for an item, success and first factor success are different then, depending on the subject matter, either we delete this item from analysis or swap success by first factor success (failure).

If the condition of Theorem 2 holds, then the above two points imply that the Q items are broadly similar in objective and point to the same direction towards one general latent variable; further, principal dimensions of order higher than one will reveal specific local factors conditioned by the first general latent variable sum score, as will be seen in the analysis of the health survey data set.

The case of polytomous data follows easily from Theorem 2, if we define success of a polytomous item to be identical to the first factor success as given in the above definition; thus by Theorem 2 each cluster will be perfectly characterized by the raw sum score of the first factor successes in the response patterns belonging to that cluster.

For some theoretical and empirical comparisons of the sum score statistic for binary data that point to one underlying latent variable with parametric and non parametric models, see in particular Cox and Wermuth (2002).

4. Multiple Correspondence Analysis of the Health Survey Data

The second column in Table 2 displays the first five dispersion measures, the standard deviations, of the first five important factors resulting from CA of $\boldsymbol{Z}$; in CA terminiology λ_α^2 represents the inertia (variance) of the αth factor. We see that the first three values are clearly singled out: $\lambda_1 = 0.8974$ being close to 1, implies that the dataset $\boldsymbol{Z}$ has quasi 2 blocks structure; and, $\lambda_2 \approx \lambda_3$ implies that the principal plane 2-3 should be looked at. We did not present the percentage of the variance explained by each principal factor, because they are misleading; further many adjusted values have been proposed in the litterature, see for instance Greenacre (1993).

Table 2: The first five dispersion measures

	CA of $\boldsymbol{Z}$	TCA of $\boldsymbol{Z}$
α	λ_α	λ_α
1	**0.8974**	**0.3014**
2	**0.4290**	**0.1910**
3	**0.4218**	0.1759
4	0.3003	0.1703
5	0.2925	0.1647

Figures 1 and 2 show the MCA maps of the principal planes 1-2 and 2-3, respectively. In Figure 1, we clearly see that the missing (N) and tried with no access (3) categories dominate the map by forming two different clusters far away from the center; and the remaining column points representing used this service (1) and never used (2) category values are clustered around the origin; further, the second dimension separates the missing (N) categories from the tried with no access (3) categories. Figure 2 shows the complete separation of the four category values (1), (2), (3) and (N). Table 1 shows that the two categories, (N) and (3), for each of the 22 items have small weights; and it is a well known fact that often rare elements disturb the graphical displays in CA or MCA. Another way of interpreting Figure 1 is that, the categories (3) and (N) can be considered as outliers, and their harmful influence should be eliminated. Different approaches have been proposed to handle missing or outlier categories by Michailidis and de Leeuw (1998), Le Roux and Rouanet (2004, Chapter 5), Greenacre (2006) and Greenacre (2009). Figures 3 and 4, which display the projection of the individuals on the principal planes 1-2 and 2-3 have the same form as Figures 1 and 2, and they admit the same interpretation.

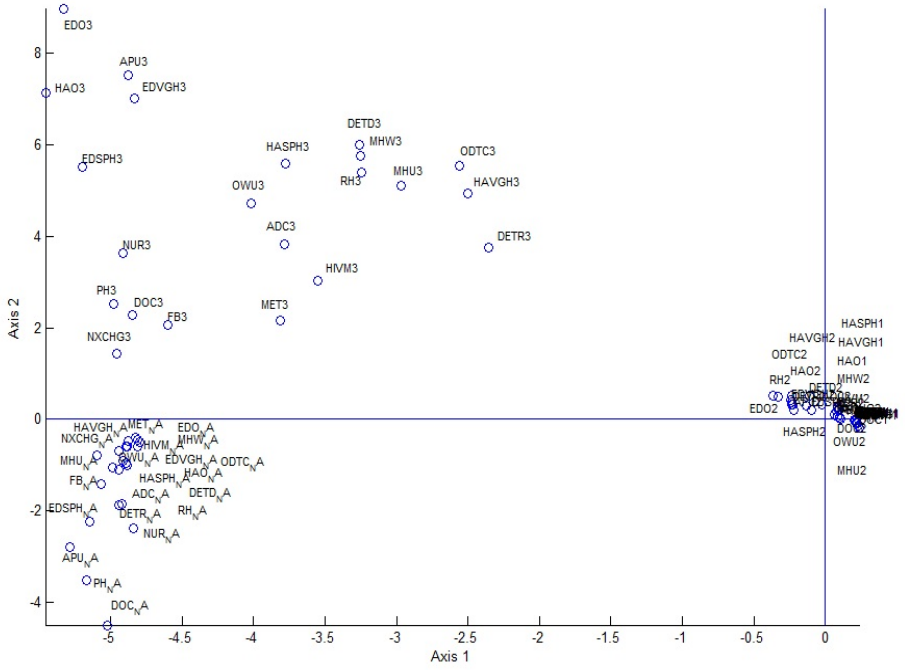

Figure 1: MCA map of the 88 categories

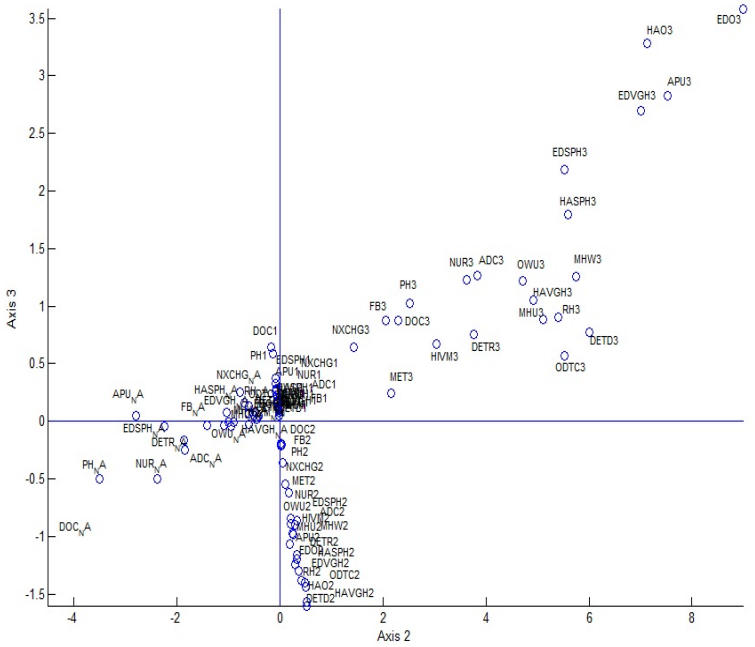

Figure 2: MCA map of the 88 categories

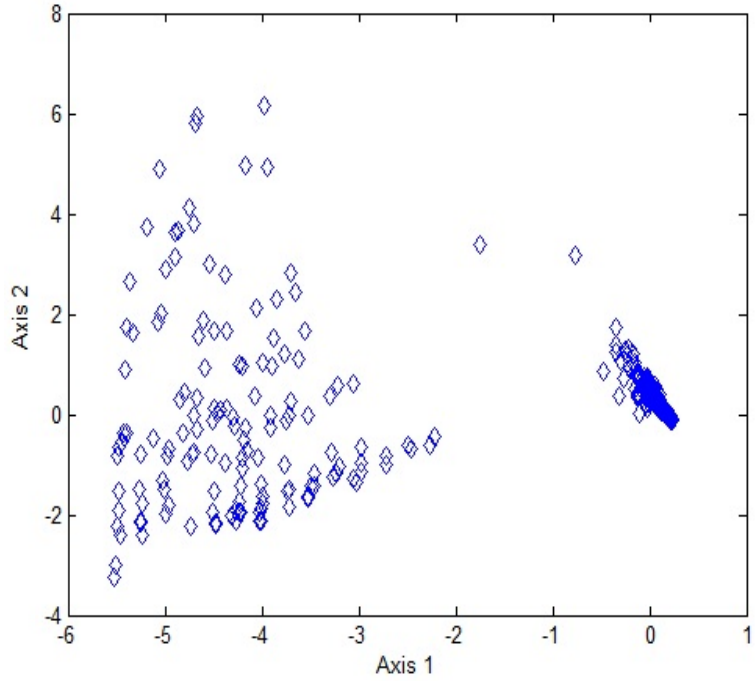

Figure 3: MCA map of the 3530 respondents

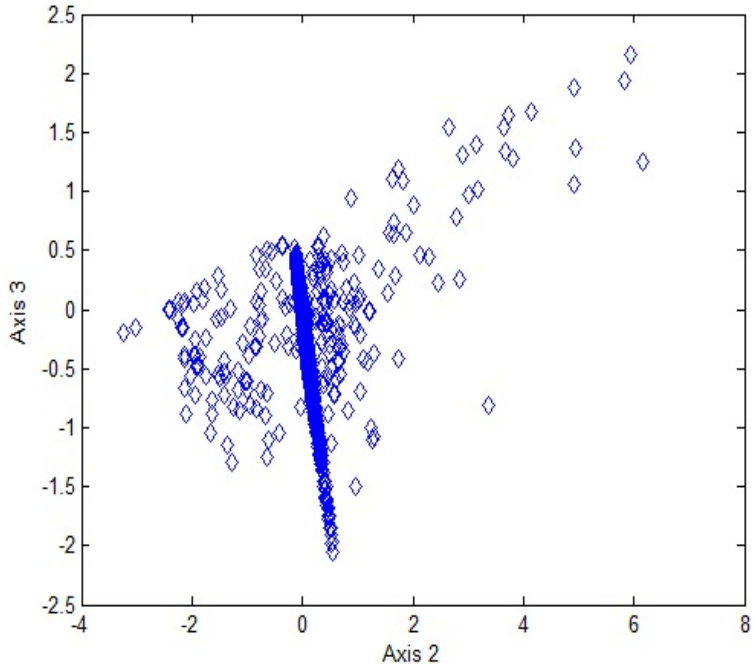

Figure 4: MCA map of the 3530 respondents

5. Multiple Taxicab Correspondence Analysis of the Health Survey

5.1 The 22 Substantive Items

The third column in Table 2 displays the first five dispersion measures, the mean deviations, of the first five factors resulting from TCA of $\boldsymbol{Z}$. We see that the first dimension is very important, $\lambda_1 = 0.3014$, and probably the second, $\lambda_2 = 0.1910$. The remaining dimensions were not interpretable.

Figure 5 shows the MTCA map of the principal plane 1-2, where we see the four groups of categories are clearly separated, and the image that they form looks like a curved horseshoe or a parabola; which implies that there is one underlying latent variable. For recent interesting discussions of horseshoes in multivariate analysis, see Diaconis, Goel and Holmes (2008) and De Leeuw (2007). Also we note that, the first principal axis clearly separates the categories used this service (1) from the rest, (2), (3) and (N). By comparing Figure 5 with Figure 1, we see that in Figure 5 there is no evidence to characterize the categories (3) and (N) of all the questions as outliers: In fact all the 22 (N) categories are clustered in one point at the extreme corner of the third quadrant in Figure 5, and the 22 tried with no access (3) categories are clustered at the corner of the 3rd quadrant in Figure 5. Looking at the categories used this service (1), we see that

the second principal axis opposes medical services [DOC1 (doctor care), NUR1 (nursing care), PH1 (pharmacy), NXCH1 (needle exchange), MET1 (methadone treatment), EDSPH1 (emergency department St. Paul Hospital)] to mental services [(MHU1 mental health unit), MHW1 (mental health worker), DETR (detox residential), DETD1 (day-tox day program), OWU1 (outreach worker)].

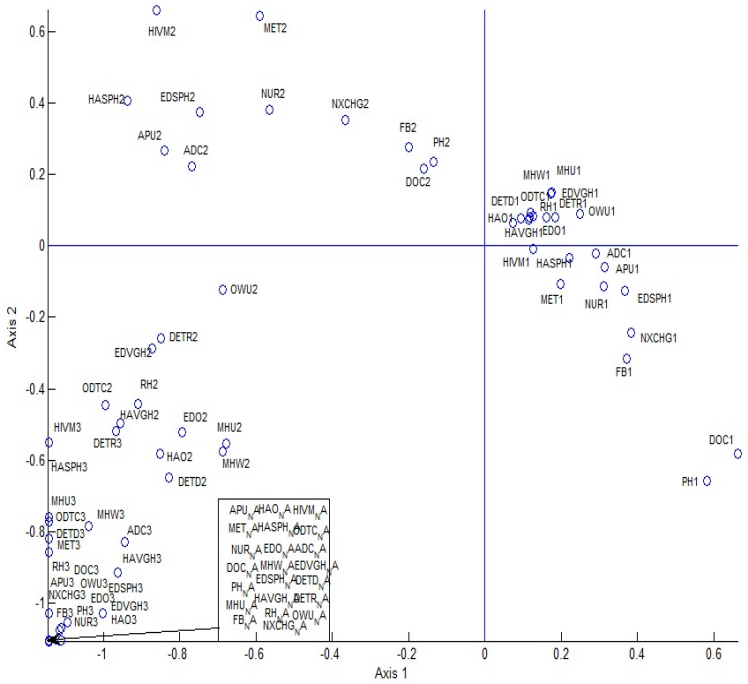

Figure 5: MTCA map of the 88 categories

Figure 6, which should be compared with Figures 3 and 4, shows the projection of the respondents on the first principal plane. We see a very clear pattern: the 3530 individuals are clustered, and on the first axis there are 22 clusters. Theorem 1 in Section 4 states that the maximum number of clusters of respondents on the first principal axis is $23 = (22 + 1) = (Q + 1)$, where Q is the number of questions. What is the interpretation of the 22 clusters? Theorem 2 of Section 3 states that the 22 clusters of respondents can be completely characterized by a discrete variable S, the simple sum score statistic of used this service (1) over all items, because the 22 categories used this service (1) have positive first principal factor scores. We name the category used this service (1) to be first factor success category for each item. The complement of "first factor success" will be "first factor failure"$=\{(2), (3), \mathrm{N}\}$. Table 3 provides some summary statistics of the clusters that we describe in steps:

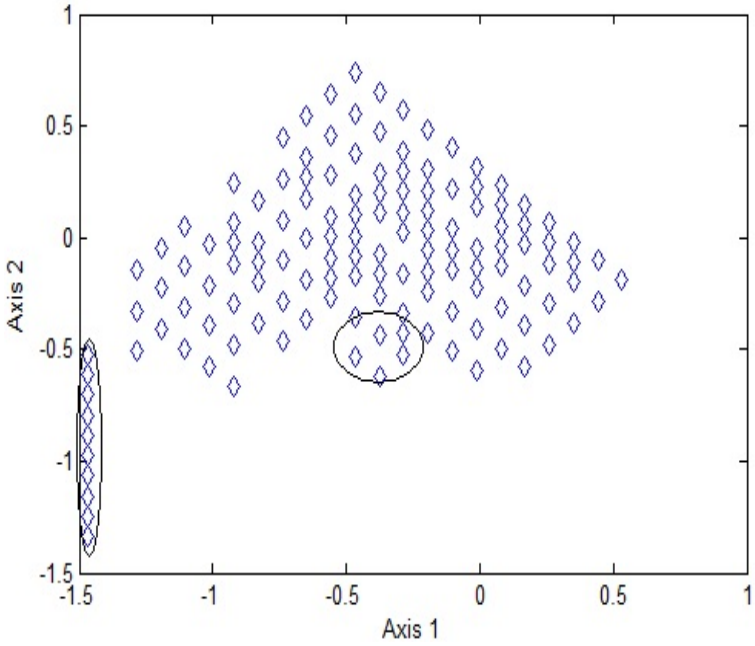

Figure 6: MTCA map of the 3530 respondents

a) The first column provides the first principal factor scores of the respondents, where we see 16 clusters of respondents with negative first principal factor scores and 7 clusters of respondents with positive first principal factor scores. The distance between two consecutive clusters on the first principal factor is constant and equals 0.09091 except for the first two clusters which is equal to $|-1.4669+1.2851| = 0.18159 \approx 2 \times 0.09091$.

b) The third column provides the frequency of each cluster of respondents; for example, there are 143 individuals in the first cluster whose first principal factor score is -1.4669, and 3 individuals in the second cluster whose first principal factor score is -1.2851.

c) We introduce some notation to formulate mathematically the calculations done in columns 4 to 7. Let $Q = 22$ be the number of items or questions, $C = 22$ be the number of clusters; n_c be the frequency of individuals in cluster c, for example $n_1 = 143$. We can express the 0/1 matrix $\boldsymbol{Z}$ as a three-way array, z_{iqv} for $i = 1, \cdots, 3530$, $q = 1, \cdots, Q$ and $v = 1, 2, 3, N$. Consider the matrix $\boldsymbol{W} = (w_{iv})$ of size 3530×4, where $w_{iv} = \sum_{q=1}^{Q} z_{iqv}$, and which represents the number of times that the respondent i chose the category value v across all items. Let $\boldsymbol{W}_c$, of size $n_c \times 4$ be the subset of the matrix $\boldsymbol{W}$ whose individuals belong to the cluster c; for instance, $\boldsymbol{W}_4$ is of size 7×4, whose elements are given in Table 4. In Table 4 the row identified by min = (4 17 0 0) provides the minimum values in the four columns of the matrix $\boldsymbol{W}_4$; and the row identified by

Table 3: The 22 clusters of individuals on the first principal axis

1st principal factor score	Sum score of (1)'s	frequency	Categories (min, max)			
			Used this service (1)	Never tried (2)	Tried with no access (3)	Missing (N)
−1.467	0	143	(0,0)	(0,14)	(0,14)	(0,22)
−1.285	2	3	(2,2)	(20,20)	(0,0)	(0,0)
−1.194	3	3	(3,3)	(19,19)	(0,0)	(0,0)
−1.103	4	7	(4,4)	(17,18)	(0,1)	(0,0)
−1.012	5	7	(5,5)	(15,17)	(0,2)	(0,0)
−0.921	6	12	(6,6)	(14,16)	(0,2)	(0,0)
−0.840	7	20	(7,7)	(13,15)	(0,2)	(0,0)
−0.740	8	18	(8,8)	(14,14)	(0,0)	(0,0)
−0.649	9	29	(9,9)	(10,13)	(0,3)	(0,0)
−0.558	10	51	(10,10)	(10,12)	(0,2)	(0,0)
−0.467	11	82	(11,11)	(10,11)	(0,1)	(0,0)
−0.376	12	11	(12,12)	(8,10)	(0,2)	(0,0)
−0.285	13	18	(13,13)	(4,9)	(0,5)	(0,0)
−0.194	14	228	(14,14)	(5,8)	(0,3)	(0,1)
−0.103	15	282	(15,15)	(6,7)	(0,1)	(0,0)
−0.012	16	327	(16,16)	(5,6)	(0,1)	(0,0)
0.079	17	358	(17,17)	(3,5)	(0,2)	(0,0)
0.169	18	440	(18,18)	(2,4)	(0,2)	(0,0)
0.260	19	461	(19,19)	(2,3)	(0,1)	(0,0)
0.351	20	416	(20,20)	(1,2)	(0,1)	(0,0)
0.442	21	224	(21,21)	(0,1)	(0,1)	(0,0)
0.533	22	120	(22,22)	(0,0)	(0,0)	(0,0)

max = (4 18 1 0) provides the maximum values in the four columns of the matrix $\boldsymbol{W}_4$. So we see that the variable sum score of (1)'s is constant and equals $S = 4$ in cluster 4. Finally, the last row in Table 4 represented by (min, max) is reproduced in the fourth row of Table 3 in the columns 4 to 7.

d) In Table 3 the discrete variable S = sum score of (1)'s completely characterizes the 22 clusters of respondents. Note that there is no individual who has used exactly 1 service among the 22 services, that is why the cluster with $S = 1$ is missing in Table 3.

Table 4: The W_4 matrix, where the seven respondents have first principal factor score of -1.1032

respondents in cluster 4	Categories			
	Used this service (1)	Never tried (2)	Tried with no access (3)	Missing (N)
1	4	17	1	0
2	4	18	0	0
3	4	18	0	0
4	4	18	0	0
5	4	18	0	0
6	4	18	0	0
7	4	18	0	0
min	4	17	0	0
max	4	18	1	0
(min, max)	(4,4)	(17,18)	(0,1)	(0,0)

So we see that the first principal factor of MTCA revealed that the data set has a very clear structure based on the simple sum score statistic of first factor success categories over all items. Further, the 22 health items are broadly similar in objective and point to the same direction.

The second principal factor has a simple interpretation: For a fixed sum score S = sum score of (1)'s, it will show the intravariability of the response patterns for that sum score. Which clusters have the most variability and what is the nature of the variability? Going to Table 3, we check the $(\min, \max)$ values for each cluster: it is evident that the first cluster characterized by $S = 0$ is the most heterogenous, followed by clusters $S = 13$ and $S = 14$.

The cluster $(S = 0)$ has $(\min, \max) = (0, 14)$ for the categories never tried (2) and tried with no access (3), and $(\min, \max) = (0, 22)$ for the category missing (N). We also note that all the missing non response values, save one, are found in this cluster. Further, Figure 6 confirms this fact, where we see 8 points aligned vertically in the third quadrant. We also note that the relative frequency of this group is very small, $143/3530 = 0.04051$. In fact, we recall that the units in this group were designated as outliers in MCA.

The cluster $(S = 13)$ has $(\min, \max) = (4, 9)$ for the categories never tried (2) and $(\min, \max) = (0, 5)$ for the categories tried with no access (3), and $(\min, \max) = (0, 0)$ for the category missing (N). This is natural variability, because the sum score statistic being a sum of successes has the most variability around its centeral values. Similar interpretation is given to the cluster $(S = 14)$, which has relative frequency of $228/3530 = 0.0646$.

In Figure 5, we saw that the categories of the variables have a parabolic curved shape. Is there a parabola in Figure 6? The answer is yes: By suppressing the circled respondents which make less than 6% of the data, we see an inverted V shaped band of points, which represents a taxicab parabola with a lot of dispersion, see for instance Krause (1986).

We conclude our analysis by the following remark: The 22 health items are broadly similar in objective and point to the same direction towards one general latent variable (because of the parabolic shape of the projected categories in Figure 5). Further, it is completely described by the simple sum score statistic of used this service (1) over all items.

5.2 Gender as a Passive Variable

Usually in a survey, in addition to substantive response questions, concomitant personal information data about the respondents are gathered. These variables are named passive or exogenous. In this analysis we include one such qualitative variable, gender having three categories male (m), female (f) and transgendered (t). Now we describe the clusters by computing the log-odds ratio of males to females for each cluster with respect to the marginal distribution. For example

$$\mathrm{LOR}(S=0) = \ln(\frac{94/49}{2406/1097}) = -0.13391,$$

and

$$\mathrm{LOR}(S=22) = \ln(\frac{103*1097}{17*2406}) = 1.0161.$$

The interpretation of LOR $(S = s)$ is:

a) $\mathrm{LOR}(S = s) = 0$, then the proportion of males in cluster s equals the proportion of females in the sample.

b) $\mathrm{LOR}(S = s) > 0$, then the proportion of males in cluster s is greater than the proportion of females in the sample. That is, the cluster s is positively associated with males, and negatively associated with females.

c) $\mathrm{LOR}(S = s) < 0$, then the proportion of females in cluster s is larger than the proportion of males in the sample. That is, the cluster s is positively associated with females, and negatively associated with males.

We did not compute the standard errors of the LOR's, because our sample almost exhausts the population, which is a closed well located community. Looking at Table 5, we discern three groups of clusters: clusters with $(S \geq 21)$ are positively associated with males (the LOR values are positive); no association of the clusters $(14 \leq S \leq 20)$ with gender (the LOR values are around 0); and finally

the clusters ($S \leq 13$) are positively associated with females (the LOR values are negative).

Table 5: The distribution of gender in each cluster

Sumscore of (1)'s	frequency	gender			100× LOR
		male	female	transgender	
0	143	94	49	0	−13.4
2	3	2	1	0	−9.2
3	3	1	2	0	−147.9
4	7	2	5	0	−170.2
5	7	3	4	0	−147.9
6	12	7	5	0	−44.9
7	20	7	12	1	−132.4
8	18	10	8	0	−56.2
9	29	14	14	1	−78.5
10	51	27	23	1	−62.5
11	82	51	30	1	−25.5
12	118	78	39	1	−9.2
13	181	113	66	2	−24.8
14	228	153	71	4	−1.8
15	282	194	87	1	1.7
16	327	218	107	2	−7.3
17	358	252	105	1	9.0
18	440	301	136	3	0.9
19	461	319	136	6	6.7
20	416	275	139	2	−10.3
21	224	182	41	1	70.5
22	120	103	17	0	101.6
Total	3530	2406	1097	27	

6. Conclusion

MCA is a popular well established method since 1970 to analyze questionnaire surveys of qualitative variables; but it is sensitive to the presence of outliers, which usually form a small fraction of the data. MTCA is a robust L_1 variant of MCA.

MCA and MTCA can produce different results, because the geometry underlying these two methods are different. We suggest the analysis of a data set by both methods: each method sees the data from its point of view, and sometimes the views are similar and other times not similar. So MCA and MTCA complement and enrich each other.

Cox (2006) titled his talk "In praise of simple sum score". We showed that the first MTCA factor scores can always be interpreted as simple sum score of

the first factor successes over all items.

Acknowledgements

Choulakian's research is financed by NSERC of Canada. The authors thank Pr. Cynthia Patton of Simon Fraser University and Dr. Mark Tyndall of University of Ottawa (formerly at University of British Columbia), who made the data available.

References

Alon, N. and Naor, A. (2006). Approximating the cut-norm via Grothendieck's inequality. *SIAM Journal on Computing* **35**, 787-803.

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

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

Choulakian, V. (2003). The optimality of the centroid method. *Psychometrika* **68**, 473-475.

Choulakian, V. (2005). Transposition invariant principal component analysis in L_1 for long tailed data. *Statistics & Probability Letters* **71**, 23-31.

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

Choulakian, V. (2006b). L_1 norm projection pursuit principal component analysis. *Computational Statistics & Data Analysis* **50**, 1441-1451.

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

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

Choulakian, V. and De Tibeiro, J. (2012). Graph partitioning by correspondence analysis and taxicab correspondence analysis. *Journal of Classification.* Accepted to appear.

Choulakian, V., Kasparian, S., Miyake, M., Akama, H., Makoshi, N. and Nakagawa, M. (2006). A statistical analysis of synoptic gospels. *JADT'2006*, pp. 281-288.

Cox, D. R. (2006). In praise of the simple sum score. http://www.stat.unipg.it/forcina/shlav/Thursday%207/Cox2.pdf

Cox, D. R. and Wermuth, N. (2002). On some models for binary variables parallel in complexity with the multivariate Gaussian distribution. *Biometrika* **89**, 462-469.

De Leeuw, J. (2007). A horseshoe for multidimensional scaling. Technical report, UCLA.

Diaconis, P., Goel, S. and Holmes, S. (2008). Horseshoes in multidimensional scaling and local kernel methods. *Annals of Applied Statistics* **2**, 777-807.

Fichet, B. (2009). Metrics of L_p-type and distributional equivalence principle. *Advances in Data Analysis and Classification* **3**, 305-314.

Finch, W. H. (2010). Imputation methods for missing categorical questionnaire data: a comparison of approaches. *Journal of Data Science* **8**, 361-378.

Gabriel, K. R. and Zamir, S. (1979). Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics* **21**, 489-498.

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

Greenacre, M. J. (1993). *Correspondence Analysis in Paractice.* Academic Press, London.

Greenacre, M. J. (2009). Canonical correspondence analysis in social research. http://ssrn.com/abstract=1435256.

Greenacre, M. J. and Pardo, R. (2006). Subset correspondence analysis. *Sociological Methods & Research* **35**, 193-218.

Jolliffe, I. T. (2002). *Principal Component Analysis*, 2nd edition. Springer, New York.

Krause, E. F. (1986). *Taxicab Geometry*: *An Adventure in Non-Euclidean Geometry.* Dover, New York.

Le Roux, B. and Rouanet, H. (2004). *Geometric Data Analysis*: *From Correspondence Analysis to Structured Data Analysis.* Kluwer Academic Publishers, Dordrecht, the Netherlands.

Michailidis, G. and de Leeuw, J. (1998). The Gifi system of descriptive mutivariate analysis. *Statistical Science* **13**, 307-336.

Nishisato, S. (1994). *Elements of Dual Scaling*: *An Introduction to Practical Data Analysis.* Hillsdale, Lawrence Erlbaum Associates, New Jersey.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis* (Edited by P. R. Krishnaiaah), 391-420. Academic Press, New York.

Received May 16, 2012; accepted October 15, 2012.

Vartan Choulakian
Department of Mathematics and Statistics
Université de Moncton
New Brunswick, E1A 3E9, Canada
vartan.choulakian@umoncton.ca

Jacques Allard
Department of Mathematics and Statistics
Université de Moncton
New Brunswick, E1A 3E9, Canada
jacques.allard@umoncton.ca

Biagio Simonetti
Department of Economic and Social Systems Analysis
University of Sannio
Via delle Puglie, 82, 82100, Benevento, Italy
simonetti@unisannio.it