MTISD 2008 – Methods, Models and Information Technologies for Decision Support Systems
Università del Salento, Lecce, 18-20 September 2008

Taxicab Non Symmetrical Correspondence Analysis

Biagio Simonetti[1]
Università del Sannio
simonetti@unisannio.it

Abstract. *Non Symmetrical correspondence analysis (NSCA) is a variant of the classical Correspondence Analysis (CA) for analyze two-way contingency table with a structure of dependence between two variables. In order to overcome the influence due to the presence of outlier, in this paper, it is presented Taxicab Non Symmetrical Correspondence Analysis (TNSCA), based on the taxicab singular value decomposition. It will show that TNSCA it is more robust than the ordinary NSCA, because it gives uniform weights to all the points. The visual map constructed by TNSCA offers a clearer perspective than the map obtained by correspondence analysis. Examples are provided.*

Keywords: Non symmetrical correspondence analysis; taxicab singular values decomposition; L_1 norm; robustness.

1. Introduction

Non Symmetrical Correspondence Analysis (NSCA, D'Ambra, Lauro 1989) is a method for visualizing contingency tables, with an asymmetric relationship between two variables. For our purposes we consider it as a particular kind of reduced rank matrix approximation method derived from generalized singular value decomposition (SVD).

Generalized SVD of a data set can be derived in a stepwise manner and is based on the Euclidean matrix norm. The aim of this paper is to use a particular SVD based on the taxicab norm, named taxicab singular value decomposition (TSVD). Choulakian (2006) presented taxicab correspondence analysis (TCA) based on TSVD, showing that TCA produced more interpretable results than the classical Correspondence analysis. In this paper we shall present a variant of NSCA based on TSVD of a contingency table and it will be named Taxicab Non Symmetrical Correspondence Analysis (TNSCA).

The paper is organized as follows. In section 2, we present a technical review of NSCA; in section 3 we, present TNSCA; in section 4, we present the analysis of two data sets by TNSCA, where we show that TNSCA is more robust than NSCA and provides more interpretable results than NSCA.

2. Non Symmetrical Correspondence Analysis

Consider a two-way contingency table N of dimension $I \times J$ according to I and J categories of variables **Y** (response) and **X** (predictor), respectively.

Denote the matrix of joint relative frequencies by $\mathbf{P} = (p_{ij})$ so that $\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1$.

[1] The participation of Biagio Simonetti was funded by a FIRB 2003 grant held by a local research unit of Benevento lead by Prof Filippo Bencardino.

Let $p_{i\bullet} = \sum_{j=i}^{J} pij$ and $p_{\bullet j} = \sum_{i=i}^{I} pij$ be the ith marginal row proportion and the jth marginal column proportion respectively.
Suppose that the relationship between these two variables is such that the J columns are predictor variables and are used to predict the outcome of the I rows response.

categories. Furthermore, let $\mathbf{\Pi} = \left(\pi_{ij} = \frac{p_{ij}}{p_{\bullet j}} - p_{i\bullet}\right)$ be the matrix of differences between the unconditional marginal prediction $p_{i\bullet}$ and the conditional prediction $\frac{p_{ij}}{p_{\bullet j}}$.
If, for all of the (i, j)th cells, there is a perfect lack of predictability of the rows given the column categories then $\pi_{ij} = 0$. This is equivalent to concluding that there is complete independence between the two variables. A more formal, and more global, measure of predictability can be made by calculating the tau index (Goodman and Kruskal, 1954; Light and Margolin, 1971):

$$\tau = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} p_{\bullet j}\left(\frac{p_{ij}}{p_{\bullet j}} - p_{i\bullet}\right)^2}{1-\sum_{i=1}^{I} p_{i\bullet}^2} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} p_{\bullet j}\pi_{ij}^2}{1-\sum_{i=1}^{I} p_{i\bullet}^2} = \frac{N_\tau}{1-\sum_{i=1}^{I} p_{i\bullet}^2} \qquad (1)$$

Here the tau numerator, N_τ, is the overall measure of predictability of the rows given the columns, while the denominator measures the overall error in prediction, which does not depend on the predictor categories. When all distributions of the predictor variable are identical to the overall marginal distribution, there is no relative increase in predictability, and so τ is zero. Therefore the NSCA of two-way contingency tables involves decomposing N_τ to obtain optimal measures of dependence. This is achieved by applying singular value decomposition (SVD) on π_{ij} so that:

$$\pi_{ij} = \sum_{s=1}^{S} \lambda_s a_{is} b_{js} \qquad (2)$$

where $S = \min(I, J) - 1$. The SVD (2) involves obtaining the measure λ_s which is the sth singular value of π_{ij}. Similarly, $\mathbf{a}_s$ and $\mathbf{b}_s$ are the orthonormal singular vectors in an unweighted and weighted metric, respectively:

$$\sum_{i=1}^{I} a_{is} a_{is}^{'} = \begin{cases} 1 & s = s^{'} \\ 0 & s \neq s^{'} \end{cases} \qquad \sum_{j=1}^{J} p_{\bullet j} b_{is} b_{is}^{'} = \begin{cases} 1 & s = s^{'} \\ 0 & s \neq s^{'} \end{cases}$$

Furthermore, it follows that N_τ can be expressed in terms of the singular values such that $N_\tau = \|\mathbf{\Pi}\|^2 = \sum_{s=1}^{S} \lambda_s^2$.

For the graphical representation of the asymmetric dependence between the variables, define the coordinates of the ith response category (row) and jth predictor category (column) for the sth dimension of a correspondence plot by

$$f_{is} = a_{is}\lambda_s \qquad g_{is} = b_{is}\lambda_s$$

Therefore, a predictor profile coordinate, g_{is} that is situated close to the origin indicates that the jth predictor category does not contribute to the predictability of the response variables. Similarly a predictor coordinate that lies at a distance from the origin will indicate that that category is important for predicting the row categories. For more details on the theory and application of the classical technique of NSCA, refer to D'Ambra and Lauro(1989) and Kroonenberg and Lombardo (1999).

3. Taxicab Non Symmetrical Correspondence Analysis

The Taxicab Singular Values Decomposition (TSV D, Choulakian 2006) is a particular orthogonal decomposition based on L1-norm distance, called also Manhattan or City-Block or more colorfully taxicab norm, because it is the distance a car would drive in a city laid out in square blocks (if there are no one-way streets).

Taxicab geometry, is essentially the study of an ideal city with all roads running horizontal or vertical. The roads must be used to get from point A to point B; thus, the normal Euclidean distance function in the plane needs to beJ modified. The shortest distance from the origin to the point (1,1) is

now 2 rather than $\sqrt{2}$. So, taxicab geometry is the study of the geometry consisting of Euclidean points, lines, and angles in $\Re^2$ with the taxicab metric: $d[(x_1,y_1),(x_2,y_2)] = |x_2 - x_1| + |y_2 - y_1|$.

A nice discussion of the properties of this geometry is given by Krause(1986) and Kay (2001).

We shall apply TSVD to the matrix $\mathbf{\Pi}$ containing the differences between the unconditional marginal prediction $p_{i\bullet}$ and the conditional prediction $\frac{p_{ij}}{p_{\bullet j}}$, as defined before. Let be the rank of $\mathbf{\Pi} = k$.

Denoting by $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_I)'$ an *I*-dimensional vector, than the quantity $\mathbf{\Pi v}$ the projection of the *I* row points of $\mathbf{\Pi}$ on **v.** Let $\mathbf{T}_J$ be the collection of all vector of length *J* with coordinates ± 1.

The first principal axis $\mathbf{v}_1$ of the row points of $\mathbf{\Pi}$ is an element of $\mathbf{T}_J$ such that the taxicab norm of $\mathbf{\Pi v}$ is maximized:

$$\max_{v \in T_J} \|\mathbf{\Pi v}\|_1 = \|\mathbf{\Pi v}_1\|_1$$

The first row factor score is

$$f_1 = \mathbf{\Pi v}_1 \tag{3}$$

to which is related the first taxicab measure of dispersion $\lambda_1 = \|\mathbf{\Pi v}_1\|_1$

Following the same procedure, it be obtain the first column scores

$$\mathbf{g}_1 = \mathbf{\Pi}'\mathbf{u}_1 \tag{4}$$

defining with $\mathbf{u}_1 = \operatorname{sgn}(\mathbf{f}_1) \in \mathbf{T}_I$ where $\operatorname{sgn}(\cdot)$ is the coordinateswise sign function that assigns 1 if $(\cdot) > 0$ and -1 if $(\cdot) \le 0$ and $\mathbf{T}_I$ is the collection of all vector of length *I* with coordinates ± 1.

The relationship between $\mathbf{v}_1$ and $\mathbf{u}_1$ allow to express the measure of dispersion λ_1 in the following way:

$$\lambda_1 = \|\mathbf{\Pi v}_1\|_1 = \|\mathbf{f}_1\|_1 = \mathbf{u}_1\mathbf{f}_1 = \|\mathbf{\Pi}'\mathbf{v}_1\|_1 = \|\mathbf{g}_1\|_1 = \mathbf{v}_1\mathbf{g}_1$$

To compute the second axis, a sequential procedure may be applied considering the residual data set:

$$\mathbf{\Pi}^{(1)} = \mathbf{\Pi} - \mathbf{f}_1\mathbf{g}_1'/\lambda_1$$

where $\mathbf{f}_1\mathbf{g}_1'/\lambda_1$ is the rank 1 reconstruction of $\mathbf{\Pi}$ matrix on the first principal axis.

The described procedure can be applied for *k* times in order to compute the *k* principal axes.

Can be shown that after *k* iterations, the residual data matrix $\mathbf{\Pi}^{(k)}$ becomes zero.

To quantify the robustness of TNSCA and NSCA, will used Benzécri et al. measure of influence as adapted and previously used by Choulakian (2006).

4. Examples

In this section, we present the analysis of two data sets. The first data set that does not contain any outliers, that cross-classifies the daily consumption of wine with the attained level of education for liver patients.

The data are based on the findings of a 2003 survey of 826 patients suffering from liver sickness which was conducted by the Department of Clinic Medicine and Infectious Disease, Second University of Naples.

The second data set contains some influential points or outliers and is found in Bradley, Katti and Coons (1962) concerning a sample of 210 individuals that were asked to reflect their impression of five foods on a five point scale.

It will be seen that for these two data sets TNSCA produces more interpretable results than the ordinary NSCA.

References

Benzécri, J.-P. et al. (1973). Analyse des Données (vol. 2). Paris: Dunod.

Bradley, R. A., Katti, S. K., and Coons, I. J. (1962), Optimal scaling for ordered categories. Psychometrika, 27, 355-374.

Choulakian, V., (2006). Taxicab correspondence analysis. Psychometrika, 71, 2, 1-13.

D'Ambra, L., Lauro, N.C., 1989. Non-Symmetrical Correspondence Analysis for three-way contingency table. In Multiway Data Analysis. Eds. R. Coppi and S. Bolasco, Amsterdam, Elsevier. 301-315.

D'Ambra, L., Lombardo, R., (1993). Normalized Non Symmetrical Correspondence Analysis for three-way data sets. Bulletin of The international Statistical Institute, 49th session book 1. 301-302.

Goodman, L.A., Kruskal, W.H., (1954). Measures of association for cross classifications.

In Journal of the American Statistical Association. 49, 732-764.

Kay, David C., (2001). College Geometry: A Discovery Approach, Second Edition. Addison Wesley Longman Inc. Boston MA.

Krause, E.F., (1986). Taxicab geometry: An adventure in non-Euclidean geometry. New York: Dover.

Kroonenberg, P., Lombardo, R., (1999). Non symmetric correspondence analysis: A tool for analysing contingency tables with a dependence structure. Multivariate Behavioral Research Journal. 34, 367-397.