

Multiple Correspondence Analysis in Marketing Research

Yangchun Du

Advisor: John C Kern II

Department of Mathematics and Computer Science
Duquesne University

25 April, 2003

Abstract

Multiple Correspondence Analysis (MCA) is a data mining tool used to display graphically the relationships among the categories of several categorical variables. Data collected across such variables are used by MCA algorithms to assign to each category of each categorical variable a two-dimensional coordinate in a special manner: categories whose coordinates are close (in Euclidean distance) share a greater association than those categories whose coordinates are relatively further apart. This research dissects the MCA algorithm currently used by *SAS* software as well as the algorithm proposed by Greenacre (1988). Features and properties of MCA are highlighted through application to simulated data. We then apply MCA to a brand preference data set provided by Management Science Associates, Inc. Comparison of standard MCA with that of Greenacre in these applications reveal little meaningful difference.

1 Introduction

The primary goal of correspondence analysis (CA) is to convert the numerical information from a contingency table into a two-dimensional graphical display. Such a graphical display offers greater insight to the relationships among the categorical variables than does the raw data, especially in cases where the number of categorical variables represented in a contingency table is greater than two. Simple correspondence analysis analyzes a contingency table made up of only two categorical variables; multiple correspondence analysis (MCA) examines the relationship among several (more than two) categorical variables. Greenacre, (1984) and Greenacre and Balasius, (1994) are two texts that provide helpful background and details to correspondence analysis theory.

Today there exist several CA and MCA algorithms. These vary in complexity, amount of required computational resources, and data assumptions. Common to all these algorithms, however, is their applicability to many fields. In the area of market research, MCA can describe the relationships between, for example, brand preference, gender, and store location. It is common for such a description to take the form of a two-dimensional scatter plot with points representing the categories of each categorical variable. Euclidean distance is then interpreted as the measure of similarity between any two points.

It is our intention to apply competing MCA algorithms to simulated data as well as a market research data set of current interest. This project documents the application of multiple MCA algorithms and provides the accompanying results. Management Science Associates Inc. (MSA) provides the data set relevant for correspondence analysis. We compare the effectiveness of the varying algorithms through their application to this data.

2 Correspondence Analysis

In this section we give a step-by-step description of the computations involved in performing simple correspondence analysis (CA). We begin by presenting relevant notation.

Let \mathbf{N} be a $I \times J$ matrix representing a contingency table of two categorical variables. We use the following definitions/notation:

- Row mass r_i : the row sums n_{i+} of \mathbf{N} divided by the grand total n ,

$$r_i = \frac{n_{i+}}{n} .$$

We denote the vector of row masses by r .

- Column mass c_j : the column sums n_{+j} of \mathbf{N} divided by the grand total n ,

$$c_j = \frac{n_{+j}}{n} .$$

We denote the vector of column masses by c .

- Correspondence Matrix \mathbf{P} : the original table \mathbf{N} divided by the grand total n , $\mathbf{P} = \frac{\mathbf{N}}{n}$.
- Row profiles: the rows of the original table \mathbf{N} divided by their respective row totals, which also can be written as $\mathbf{D}_r^{-1}\mathbf{P}$, where \mathbf{D}_r is the diagonal matrix of row masses.
- Column profiles: the columns of the original table \mathbf{N} divided by their respective column totals, which also can be written as $\mathbf{P}\mathbf{D}_c^{-1}$, where \mathbf{D}_c is the diagonal matrix of column masses.
- Let \mathbf{A} be the matrix of standardized residuals: $\mathbf{A} = \mathbf{D}_r^{-1/2}(\mathbf{P} - rc^T)\mathbf{D}_c^{-1/2}$. Notice that \mathbf{A} is an $I \times J$ matrix with elements $a_{ij} = \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}}$.
- The Singular Value Decomposition (SVD) of the $I \times J$ matrix \mathbf{A} yields the three matrices:

$$\mathbf{A} = \mathbf{U}\mathbf{\Gamma}\mathbf{V}^T ,$$

Where the matrix $\mathbf{\Gamma}$ is a diagonal matrix of positive numbers in descending order $\gamma_{11} \geq \gamma_{22} \geq \dots \geq \gamma_{kk} > 0$ (these are the singular values). The columns of the matrix \mathbf{U} are the left singular vectors, while the columns of \mathbf{V} are the right singular vectors.

- Chi-square statistic for the contingency table: $\chi^2 = n \sum_i \sum_j a_{ij}^2$.

- The sum of the squares of elements of \mathbf{A} is the total inertia of the contingency table. Total inertia = $\sum_i \sum_j a_{ij}^2$. Note that the total inertia is also equal to the chi-square statistic divided by n .
- There are $K = \min(I - 1, J - 1)$ dimensions for graphical display in CA. The squares of the singular values of \mathbf{A} also decompose the total inertia; these are called the principal inertias and denoted by $\chi_1 \dots \chi_K$.
- The principal coordinates of the rows are obtained as $\mathbf{D}_r^{-1/2} \mathbf{U} \mathbf{\Gamma}$. This matrix has number of rows equal to the number of categories in the row variable of \mathbf{N} , and number of columns equal to K . The first two columns of this matrix provide the coordinates of the row variable categories in two dimensions.
- The principal coordinates of the columns are obtained as $\mathbf{D}_c^{-1/2} \mathbf{V} \mathbf{\Gamma}$. This matrix has number of rows equal to the number of categories in the column variable of \mathbf{N} , and number of columns equal to K . The first two columns of this matrix provide the coordinates of the column variable categories in two dimensions.

3 Multiple Correspondence Analysis

CA is normally generalized to the case of Q categorical variables by analyzing the “indicator matrix” of the data. When Q exceeds two, we use the term Multiple Correspondence Analysis (MCA). Let this indicator matrix \mathbf{Z} be defined as $\mathbf{Z} = [\mathbf{Z}_1, \dots, \mathbf{Z}_Q]$, where \mathbf{Z}_q is a $(n \times J_q)$ matrix referring to the q th categorical variable with J_q categories. Here n represents the total number of observations in the data set. For example, if \mathbf{Z}_1 represents the categorical variable gender, then in the i^{th} row of \mathbf{Z}_1 would be $[0,1]$ if person i is female, and $[1,0]$ if person i is male.

With this notation, we can now define the main component of any MCA—the Burt matrix. The Burt matrix \mathbf{B} is given by $\mathbf{B} = \mathbf{Z}^T \times \mathbf{Z}$. \mathbf{B} is a symmetric matrix, and contains all pairwise cross-tabulations of Q variables, including the cross-tabulations of each variable with itself.

It is through this Burt matrix that we obtain the coordinates (in a maximum of K dimensions) of all categories under investigation. The manner in which these coordinates are obtained is what separates the various MCA algorithms. Here is a sample of some MCA algorithms:

- Currently, most statistical software packages can perform MCA. One such package, *SAS*, offers a built-in MCA procedure (`corresp`) which decomposes the Burt matrix using a singular value decomposition (SVD).
- Greenacre (1988) proposed using a modified Burt matrix which provides the advantage (over standard MCA analysis) of a greater percentage of explained variation by the two-dimensional solution for some categorical datasets.. He defines an iterative algorithm which requires the construction of a modified Burt matrix \mathbf{B}^* . This modified matrix

is just the original Burt matrix with modified sub-matrices on its diagonal. For the weighted least-squares approximation of a Burt matrix,

$$\mathbf{B} \approx nrr^T + n\mathbf{D}\mathbf{X}\mathbf{D}_\beta\mathbf{X}^T\mathbf{D}.$$

Note here that n is the grand total of the Burt matrix, r is the row mass of the Burt matrix (equivalent to the column mass of \mathbf{B} because the Burt matrix is always symmetric), and \mathbf{D} is the diagonal matrix of the row masses r . We use this notation throughout the remainder of this paper.

3.1 Standard MCA

Let $\mathbf{S} = n^{-1/2}\mathbf{D}^{-1/2}\mathbf{B}\mathbf{D}^{-1/2}$. Standard MCA takes the SVD of \mathbf{S} as $\mathbf{S} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^T$. With \mathbf{D}_α and \mathbf{U} , we find the matrices \mathbf{D}_β and \mathbf{X} as follows:

$$\begin{bmatrix} 1 & 0 \\ 0 & \mathbf{D}_\beta \end{bmatrix} = n^{-1/2}\mathbf{D}_\alpha$$

$$[\mathbf{1} \quad \mathbf{X}] = \mathbf{D}^{-1/2}\mathbf{U}$$

Finally, define the matrix of coordinates Ξ as:

$$\Xi = \mathbf{X}\mathbf{D}_\beta^{1/2}.$$

The first two columns of Ξ yield the coordinates for each of the $\sum_q J_q$ categories. The maximum number of dimensions for MCA analysis—and hence the number of columns of Ξ —is $\sum_q J_q - 1$.

3.2 Greenacre MCA

Greenacre (1988) builds a model for the whole matrix $\mathbf{B} - nrr^T$, namely

$$\mathbf{B} - nrr^T \approx n\mathbf{D}\mathbf{X}\mathbf{D}_\beta\mathbf{X}^T\mathbf{D} + \mathbf{C}$$

where \mathbf{C} is a block diagonal matrix with sub-matrices \mathbf{C}_{qq} ($q = 1, \dots, Q$) down the diagonal and zeros elsewhere. So, the new sub-matrix \mathbf{N}_{qq}^* on the diagonal of \mathbf{B}^* , which is given by

$$\mathbf{N}_{qq}^* = nr_q r_q^T + nD_q X_q D_\beta X_q^T D_q, \tag{1}$$

has the same row and column margins as \mathbf{N}_{qq} . where the vector of \mathbf{J}_q masses for variable q is denoted by r_q . Moreover, the $\mathbf{J}_q \times \mathbf{J}_q$ diagonal matrix formed from the elements of r_q is now denoted by \mathbf{D}_q . Meanwhile, the diagonal matrix \mathbf{D}_β contains a scale parameter for each dimension. The parameter \mathbf{X} is partitioned two-wise according to the variable as $\mathbf{X}_1 \cdots \mathbf{X}_Q$. Thus \mathbf{X}_q is a $\mathbf{J}_q \times \mathbf{K}$ sub matrix.

The following iterative algorithm is then used to produce the modified Burt matrix \mathbf{B}^* :

1. Start with a solution for \mathbf{X} and \mathbf{D}_β based on the MCA.

2. Replace the sub matrices on the diagonal of \mathbf{B} with those “estimated” by X and \mathbf{D}_β given by (1).
3. Perform a correspondence analysis on the modified matrix \mathbf{B}^* , setting \mathbf{X} equal to the first K vectors of optimal row or column parameters (as usual) and the diagonal of \mathbf{D}_β equal to the square roots of the first K principal inertias respectively.
4. Go back to 2 and repeat until the iterations converge; that is, when the decrease in the discrepancy function from iteration to iteration is practically zero.

Details of the computational efficiency of this algorithm are not provided. In our experience, however, convergence was obtained in no more than $n = 10$ iterations for every data set we considered.

Section 4 investigates the application of both of these MCA procedures to simulated data, while Section 5 investigates their application to actual market research data. Before these applications, however, we pause to examine an interesting property of standard MCA.

3.3 Sample Size Consideration

In many inferential statistical methods, a larger sample size provides better estimates of unknown population quantities (usually in the sense of smaller standard errors of the estimators). MCA, a descriptive procedure, does not share this property. We show now that duplicating the data used in MCA will not change the MCA output. Notation in this section is not used only in this section.

Lemma 1: The Burt matrix of a duplicated data set is exactly 2 times that of the original data.

Proof:

Let \mathbf{Z} be a $m \times n$ indicator matrix, having binary entries representing the data with n categorical variables and m cases (observations).

$$\mathbf{Z} = \begin{bmatrix} Z_{11} & Z_{12} & \cdots & Z_{1n} \\ Z_{21} & Z_{22} & \cdots & Z_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{m1} & Z_{m2} & \cdots & Z_{mn} \end{bmatrix}$$

Therefore, the transpose of \mathbf{Z} is $\mathbf{Z}^T = \begin{bmatrix} Z_{11} & Z_{21} & \cdots & Z_{m1} \\ Z_{12} & Z_{22} & \cdots & Z_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{1n} & Z_{2n} & \cdots & Z_{mn} \end{bmatrix}$

The Burt matrix is a $n \times n$ matrix, $\mathbf{B} = \mathbf{Z}^T \times \mathbf{Z} = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1n} \\ B_{21} & B_{22} & \cdots & B_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ B_{n1} & B_{n2} & \cdots & B_{nn} \end{bmatrix}$

where

$$\begin{aligned} B_{11} &= Z_{11}Z_{11} + Z_{12}Z_{12} + \cdots + Z_{1n}Z_{1n} \\ B_{12} &= Z_{11}Z_{12} + Z_{21}Z_{22} + \cdots + Z_{m1}Z_{1n} \\ &\vdots \\ B_{nn} &= Z_{1n}Z_{1n} + Z_{2n}Z_{2n} + \cdots + Z_{mn}Z_{mn} \end{aligned}$$

If we duplicate the data, that means \mathbf{Z}^* is $2m \times n$ indicator matrix as follows.

$$\mathbf{Z}^* = \begin{bmatrix} Z_{11} & Z_{12} & \cdots & Z_{1n} \\ Z_{21} & Z_{22} & \cdots & Z_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{m1} & Z_{m2} & \cdots & Z_{mn} \\ Z_{11} & Z_{12} & \cdots & Z_{1n} \\ Z_{21} & Z_{22} & \cdots & Z_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{m1} & Z_{m2} & \cdots & Z_{mn} \end{bmatrix}$$

According to the matrix multiplication rules, \mathbf{B}^* is still a $n \times n$ symmetric matrix.

$$\mathbf{B}^* = \mathbf{Z}^{*T} \times \mathbf{Z}^* = \begin{bmatrix} B_{11}^* & B_{12}^* & \cdots & B_{1n}^* \\ B_{21}^* & B_{22}^* & \cdots & B_{2n}^* \\ \vdots & \vdots & \ddots & \vdots \\ B_{n1}^* & B_{n2}^* & \cdots & B_{nn}^* \end{bmatrix} = \begin{bmatrix} 2B_{11} & 2B_{12} & \cdots & 2B_{1n} \\ 2B_{21} & 2B_{22} & \cdots & 2B_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 2B_{n1} & 2B_{n2} & \cdots & 2B_{nn} \end{bmatrix} = 2 \times \mathbf{B}$$

where

$$\begin{aligned} B_{11}^* &= Z_{11}Z_{11} + Z_{12}Z_{12} + \cdots + Z_{1n}Z_{1n} + Z_{11}Z_{11} + Z_{12}Z_{12} + \cdots + Z_{1n}Z_{1n} = 2B_{11} \\ B_{12}^* &= Z_{11}Z_{12} + Z_{21}Z_{22} + \cdots + Z_{m1}Z_{1n} + Z_{11}Z_{12} + Z_{21}Z_{22} + \cdots + Z_{m1}Z_{1n} = 2B_{12} \\ &\vdots \\ B_{nn}^* &= Z_{1n}Z_{1n} + Z_{2n}Z_{2n} + \cdots + Z_{mn}Z_{mn} + Z_{1n}Z_{1n} + Z_{2n}Z_{2n} + \cdots + Z_{mn}Z_{mn} = 2B_{nn} \end{aligned}$$

The proof of this lemma is now complete.

Before stating our theorem, consider this second (and final) lemma:

Lemma 2 : The MCA for Burt Matrix \mathbf{B} is identical to MCA for Burt matrix $\mathbf{B}^* = k \cdot \mathbf{B}$ for any $k > 0$.

Proof:

Let \mathbf{B} be an $I \times I$ Burt matrix, with row and column totals B_{i+} ($i = 1, \dots, I$) and grand total n . Let r be the vector of row masses $r_i = \frac{B_{i+}}{n}$. \mathbf{D} is the diagonal matrix of these masses. Recall that for MCA, the expected frequencies $e_{ii} = \frac{B_{i+}B_{+i}}{n} = nr_i r_i$. We also recall $\mathbf{S} = n^{-1/2} \mathbf{D}^{-1/2} \mathbf{B} \mathbf{D}^{-1/2}$ with $S_{ii} = \frac{B_{ii}}{\sqrt{e_{ii}}}$. So the SVD of \mathbf{S} is $\mathbf{S} = \mathbf{U} \mathbf{D}_\alpha \mathbf{V}^T$.

$$\begin{bmatrix} 1 & 0 \\ 0 & \mathbf{D}_\beta \end{bmatrix} = n^{-1/2} \mathbf{D}_\alpha \quad (2)$$

$$\begin{bmatrix} 1 & \mathbf{X} \end{bmatrix} = \mathbf{D}^{-1/2} \mathbf{U} \quad (3)$$

Actually, in MCA, the element in the diagonal of \mathbf{D}_β is principal inertias; and the $\mathbf{X} \sqrt{\mathbf{D}_\beta}$ is principal coordinates. In this lemma, we need to prove that \mathbf{D}_β and \mathbf{X} of \mathbf{B}^* are the same as that of \mathbf{B} .

For \mathbf{B}^* ,

row mass $r_i^* = \frac{B_{i+}^*}{n^*} = \frac{kB_{i+}}{kn} = r_i$ So, $e_{ii}^* = knr_i r_i = ke_{ii}$.

for each element S_{ii}^* in \mathbf{S}^* , $S_{ii}^* = \frac{B_{ii}^*}{\sqrt{e_{ii}^*}} = \frac{kB_{ii}}{\sqrt{ke_{ii}}} = \sqrt{k} S_{ii}$

once again, the SVD of \mathbf{S} is $\mathbf{S}^* = \mathbf{U}^* \mathbf{D}_\alpha^* \mathbf{V}^{*T}$, we also know $\mathbf{U}^* = \mathbf{U}$, $\mathbf{D}_\alpha^* = \sqrt{k} \mathbf{D}_\alpha$, and $\mathbf{V}^{*T} = \mathbf{V}$.

As we can see, in the MCA of \mathbf{B}^* , the right of equation [2] keeps the same since $(kn)^{-1/2} \sqrt{k} \mathbf{D}_\alpha = n^{-1/2} \mathbf{D}_\alpha$, and the right side of equation [3] ($\mathbf{D}^{-1/2} \mathbf{U}$) also keeps the same.

Therefore, we conclude that $\mathbf{D}_\beta^* = \mathbf{D}_\beta$, and $\mathbf{X}^* = \mathbf{X}$, and the proof is complete.

We are now ready to state and prove the following theorem:

Theorem: Duplicating the data does not change the MCA. (Or tripling, etc.)

Proof: Combine the results of lemmas 1 and 2.

We point out again that MCA is non-inferential procedure. The result of this theorem is analogous to comparing histograms of a data set and the duplicated version of that data set—the relative frequencies would remain unchanged, yielding two identical histograms.

4 Application of Simulated Data

4.1 Data

We simulated $n = 1000$ observations from the following four categorical variables (with categories following in parentheses):

- Gender (M, F).
- Education level (NHS, HS, U). This variable denotes highest degree obtained: NH represents No High School; HS represents High School; U represents University.
- Brand (X,Y,Z).
- Location (A,B,C).

The data were simulated in such a way as to impose built-in relationships that will appear in MCA. The simulation steps are described as follows:

- 1. Generate Gender with equal probability.
- 2. Generate Education given Gender:
 $P\{NHS|F\}=0.1$, $P\{HS|F\}=0.4$, $P\{U|F\}=0.5$
 $P\{NHS|M\}=0.4$, $P\{HS|M\}=0.5$, $P\{U|M\}=0.1$
- 3. Generate Brand given Education:
 $P\{X|NHS\}=0.9$, $P\{Y|NHS\}=0.05$, $P\{Z|NHS\}=0.05$
 $P\{X|HS\}=0.7$, $P\{Y|HS\}=0.15$, $P\{Z|HS\}=0.15$
 $P\{X|U\}=0.5$, $P\{Y|U\}=0.25$, $P\{Z|U\}=0.25$
- 4. Generate Location given Brand:
 $P\{A|X\}=0.76$, $P\{B|X\}=0.12$, $P\{C|X\}=0.12$
 $P\{A|Y\}=0.12$, $P\{B|Y\}=0.76$, $P\{C|Y\}=0.12$
 $P\{A|Z\}=0.12$, $P\{B|Z\}=0.12$, $P\{C|Z\}=0.76$

From these steps notice the following: Females are much more likely to have a HS or U degree, while men are much more likely to have a NHS or HS degree. Brand X is most popular across all education levels, in decreasing strength from NHS to HS to U. Brand X is most likely to be purchased at location A; brand Y at location B, and brand Z at location C.

4.2 Analysis

The Burt matrix and modified Burt matrix are shown in Table 1 and 2 respectively. The summary of MCA is shown in Table 3. For this dataset, Greenacre method doesn't increase the total percentage of inertia in the first two dimensions. Figures 1 and 2 reflect the relationship described above. For instance, given the non high school education (NHS) level, brand Z is more strongly associated with NHS than others. This relationship is also explained by the distance among brand X and three education levels. Although those two graphical displays

may look different, they are similar if we just invert the second one. This indicates the MCA of modified Burt matrix is very similar to standard MCA. Ultimately, the imposed relationships in the simulated data are reflected well in both the standard MCA and Greenacre MCA.

5 Application of MSA Data

5.1 Data

In this section we analyze $n = 1537$ observations from the market research data set obtained from Management Science Associates, Inc. (MSA). It contains appropriately-masked brand-preference information for the following three categorical variables (with categories following in parentheses):

- Gender (M,F).
- Education level (College, High School).
- Brand (1,2).

5.2 Analysis

The result of Burt matrix and Modified Burt Matrix analyses are shown in Tables 4 and 5 respectively. As we can see, the Burt Matrix (Table 4) with its diagonal were replaced by some optimal estimates (Table 5), where the replacement estimates on the diagonal have been rounded to the nearest integers. We apply MCA to this modified matrix in *SAS* or *S-plus*—see the appendix for the appropriate code. Here we should mention that, in *SAScorresp* procedure, it can accept the Burt matrix, however, the sum of all elements in each diagonal partition of the Burt matrix must equal to the grand-total divided by the square of number of variables. Hence modified Burt matrices are not accepted by *SAS*; *S-plus* code in the appendix was created to perform standard MCA on the modified Burt matrix. The results of the application of both methods are summarized in Table 6. In addition, the two-dimensional graphical display by principal coordinate visualizes how these three categorical variables associate together.

The interpretation of MCA as follows.

For the Burt matrix, we see that 41.57% of the total inertia can be explained with a single dimension; that is, the relative frequency values that can be reconstructed from a single dimension can reproduce 41.57% of total Chi-square value. Two dimensions allow you to explain 74.78%. The Greenacre algorithm on the other hand, shows a dramatic improvement: the first dimension explains 85.6% of inertia, while two dimensions explain 100%.

Secondly, what is important are the distances of the points in the two-dimensional display, which are informative in that row points that are close to each other are similar with regard to the pattern of relative frequencies across the columns. Despite the differing scales Figures 3 and 4, both show that people who have a college education are more likely to purchase Brand 1. High education individuals prefer the Brand 2. Even though Greenacre's method out-performed standard MCA in terms of total inertia, Figures 3 and 4 appear to communicate exactly the same relationships among these data.

6 Conclusion and Future Work

Categorical data are common products of marketing research. However, the analysis of such data often is hindered by the requirements and limitations of many familiar research tools. Correspondence analysis is an easily implemented analytical method that can help researchers in detecting and explaining relationships among complex marketing phenomena. Moreover, CA and MCA provide a graphical representation of the structure in the data, and it is flexible in terms of data requirements. Of course, it CA or MCA can be shown in 3 or high-order dimensions.

The generalization of CA to multivariate categorical data by analyzing the indicator matrix \mathbf{Z} , or equivalently the Burt matrix \mathbf{B} , offers a choice among various existing solution algorithms. Standard MCA and Greenacre MCA are two that we considered in this analysis. Indeed, Greenacre's method does explain more variability in the first two dimensions for some data sets (see Section 5), but not for all data sets (see Section 4). Future work in this area includes identifying the characteristics of a data set that make it more suitable for standard MCA or Greenacre MCA. In the meantime, since MCA is fast and easily implemented, we suggest both methods be applied when MCA is needed.

We recognize that duplication a data set does not change the MCA analysis. Investigators should not feel that a larger sample size may provide a more meaningful MCA, assuming this larger sample size yields data with similar characteristics of the original data. And finally, we note that MCA is not ideal for identifying interactions among categories. For example, it is difficult from MCA to discern the following relationship: Male college graduates strongly prefer brand X, while male high school drop-outs strongly prefer brand Y. In this case, males may well appear (from a two dimensional MCA plot) to have no preference between brands X and Y. Enhancing MCA algorithms to the point where such interactions can be readily identified is another area for future research.

7 References

- Beh, E.J. (1997). Simple correspondence analysis of ordinal cross-classifications using orthogonal polynomials. *Biometrical Journal*, 39, 589-613.
- Bendixen, M (1996). A practical guide to the use of correspondence analysis in marketing research. Marketing Research on-line Vol 1.
- Greenacre, M.J. (1984). Theory and Applications of Correspondence Analysis. London: Academic Press.
- Greenacre, M.J. (1988). Correspondence Analysis of Multivariate categorical data by weighted least squares. *Biometrika*, 75, 457-467.
- Greenacre, M.J. and Balasius, J. (1994). Correspondence Analysis in the Social Sciences. London: Academic Press.

8 Appendix: S-plus code

```
# Function to caculate the Square Root of Matrix A

pdsroot <- function(A)
{
  p <- dim(A)[1] # dimension of matrix
  eigenv <- eigen(A, symmetric = T) # spectral decomposition of A
  rootA <- eigenv$vectors %*% diag(sqrt(eigenv$values)) %*% t(eigenv$vectors)
  rootA
}

# Perform MCA for Burt matrix in S-plus

r_apply(N,1,sum) # N is the i by i Burt Matrix
n_sum(r)
r_apply(N,1,sum)/n
c_apply(N,2,sum)/n
Dr_diag(r)
Dc_diag(c)
S_(1/sqrt(n))*solve(pdsroot(Dr))%*%N%*%solve(pdsroot(Dc))
U_svd(S)$u
V_svd(S)$v
Dalph<-diag(svd(S)$d)
Dmu_(sqrt(1/n)*Dalph)[2:(i-1),2:(i-1)] # Dmu is the principal inertia
X_pdsroot(Dr) #sqrt(Dmu) is the singular value
X_solve(X)%*%U
X_X/X[1,1]
X_X[1:(i-1),2:(i-1)]
Y_pdsroot(Dc)
Y_solve(Y)%*%V
Y_Y/Y[1,1]
Y_Y[1:(i-1),2:(i-1)]
dyc_n*(r%*%t(c)+Dr%*%X%*%Dmu%*%t(Y)%*%Dc) # dyc is the Weight
```

```

# least-squares approximation of Burt matrix
fs<-X%*%sqrt(Dmu) # this is the principal coordinate
#for row and column ( note: row and column is same in this case)

# Iterative Alogrithm to obtain the Modified Burt Matrix
# for MSA Data (number of observation is 1537), B is 6 by 6 Burt Matrix,
# and have three categorical variables, each has 2 variables.

N<-B
for (i in 1:5)
{
r_apply(N,1,sum)
n_sum(r)
r_apply(N,1,sum)/n
c_apply(N,2,sum)/n
Dr_diag(r)
Dc_diag(c)
S_(1/sqrt(n))*solve(pdsroot(Dr))%*%N%*%solve(pdsroot(Dc))
U_svd(S)$u
V_svd(S)$v
Dalph<-diag(svd(S)$d)
Dmu_(sqrt(1/n)*Dalph)[2:6,2:6]
X_pdsroot(Dr)
X_solve(X)%*%U
X_X/X[1,1]
X_X[1:6,2:6]
Y_pdsroot(Dc)
Y_solve(Y)%*%V
Y_Y/Y[1,1]
Y_Y[1:6,2:6]
dyc_n*(r%*%t(c)+Dr%*%X%*%Dmu%*%t(Y)%*%Dc)

Dbeta<-sqrt(Dmu)
N<-B

N11<-N[1:2,1:2]
R11<-apply(N11,1,sum)
D11<-diag(R11)
X11<-X[1:2,1:5]
N11star<-round(R11%*%t(R11)/1537)
N11star<-N11star+round(D11%*%X11%*%Dbeta%*%t(X11)%*%D11/n)

```

```

N22<-N[3:4,3:4]
R22<-apply(N22,1,sum)
D22<-diag(R22)
X22<-X[3:4,1:5]
N22star<-round(R22%*%t(R22)/1537)
N22star<-N22star+round(D22%*%X22%*%Dbeta%*%t(X22)%*%D22/n)

N33<-N[5:6,5:6]
R33<-apply(N33,1,sum)
D33<-diag(R33)
X33<-X[5:6,1:5]
N33star<-round(R33%*%t(R33)/1537)
N33star<-N33star+round(D33%*%X33%*%Dbeta%*%t(X33)%*%D33/n)

N[1:2,1:2]<-N11star
N[3:4,3:4]<-N22star
N[5:6,5:6]<-N33star
}

# plot two-dimensional graphical display

fsname<-c('Brand1','Brand2','Male','Female','College','High School')

corrplot<-function(fs,fsname)
{
  xlabes<-fsname
  plot(fs[,1],fs[,2],pch="*",xlim=range(fs),ylim=range(fs),xlab=paste("First
    Principal Axis "),ylab=paste("Second Principal Axis"))
  text(dycfs[,1]-0.01,dycfs[,2]-0.03,labels=xlabes,adj=0)
  title(main="MCA Graphical display of Modified Burt Matrix for MSA Data")
  abline(h=0,v=0)
  return(fs=fs[,c(1,2)])
}

```

| | Male | Female | nonHS | HS | College | locA | locB | locC | BrandX | BrandY | BrandZ |
|---------|------|--------|-------|-----|---------|------|------|------|--------|--------|--------|
| Male | 496 | 0 | 197 | 259 | 40 | 371 | 72 | 53 | 303 | 96 | 97 |
| Female | 0 | 504 | 39 | 216 | 249 | 310 | 98 | 96 | 256 | 125 | 123 |
| nonHS | 197 | 39 | 236 | 0 | 0 | 212 | 12 | 12 | 171 | 34 | 31 |
| HS | 259 | 216 | 0 | 475 | 0 | 330 | 77 | 68 | 257 | 102 | 116 |
| College | 40 | 249 | 0 | 0 | 289 | 139 | 81 | 69 | 131 | 85 | 73 |
| locA | 371 | 310 | 212 | 330 | 139 | 681 | 0 | 0 | 515 | 82 | 84 |
| locB | 72 | 98 | 12 | 77 | 81 | 0 | 170 | 0 | 22 | 126 | 22 |
| locC | 53 | 96 | 12 | 68 | 69 | 0 | 0 | 149 | 22 | 13 | 114 |
| BrandX | 303 | 256 | 171 | 257 | 131 | 515 | 22 | 22 | 559 | 0 | 0 |
| BrandY | 96 | 125 | 34 | 102 | 85 | 82 | 126 | 13 | 0 | 221 | 0 |
| BrandZ | 97 | 123 | 31 | 116 | 73 | 84 | 22 | 114 | 0 | 0 | 220 |

Table 1: Simulated Burt Matrix

| | Male | Female | nonHS | HS | College | locA | locB | locC | BrandX | BrandY | BrandZ |
|---------|------|--------|-------|-----|---------|------|------|------|--------|--------|--------|
| Male | 272 | 234 | 197 | 259 | 40 | 371 | 72 | 53 | 303 | 96 | 97 |
| Female | 234 | 289 | 39 | 216 | 249 | 310 | 98 | 96 | 256 | 125 | 123 |
| nonHS | 197 | 39 | 68 | 114 | 59 | 212 | 12 | 12 | 171 | 34 | 31 |
| HS | 259 | 216 | 114 | 238 | 134 | 330 | 77 | 68 | 257 | 102 | 116 |
| College | 40 | 249 | 59 | 134 | 101 | 139 | 81 | 69 | 131 | 85 | 73 |
| locA | 371 | 310 | 212 | 330 | 139 | 345 | 117 | 115 | 515 | 82 | 84 |
| locB | 72 | 98 | 12 | 77 | 81 | 117 | 67 | 45 | 22 | 126 | 22 |
| locC | 53 | 96 | 12 | 68 | 69 | 115 | 45 | 66 | 22 | 13 | 114 |
| BrandX | 303 | 256 | 171 | 257 | 131 | 515 | 22 | 22 | 502 | 98 | 87 |
| BrandY | 96 | 125 | 34 | 102 | 85 | 82 | 126 | 13 | 98 | 57 | 21 |
| BrandZ | 97 | 123 | 31 | 116 | 73 | 84 | 22 | 114 | 87 | 21 | 46 |

Table 2: Simulated Modified Burt Matrix

| | Burt Matrix | | Modified Burt Matrix | |
|------|-------------------|----------|----------------------|-----------|
| | Principal inertia | Percent | Principal inertia | Percent |
| k=1 | 0.4843487 | 27.67707 | 0.2712651 | 29.59681 |
| k=2 | 0.3874350 | 22.13915 | 0.1691768 | 18.4583 |
| k=3 | 0.2988519 | 17.07725 | 0.1278038 | 13.94424 |
| k=4 | 0.2463057 | 14.07461 | 0.1044092 | 11.39173 |
| k=5 | 0.1253908 | 7.165191 | 0.09827268 | 10.7222 |
| k=6 | 0.0112663 | 6.437910 | 0.09176939 | 10.01265 |
| k=7 | 0.0095004 | 5.428824 | 0.02237811 | 2.4416 |
| k=8 | 0 | 0 | 0.02048708 | 2.235275 |
| k=9 | 0 | 0 | 0.006054817 | 0.6606204 |
| k=10 | 0 | 0 | 0.004917945 | 0.5365802 |

Table 3: Summary for simulated data

MCA Graphical display of Burt Matrix for Simulated Data

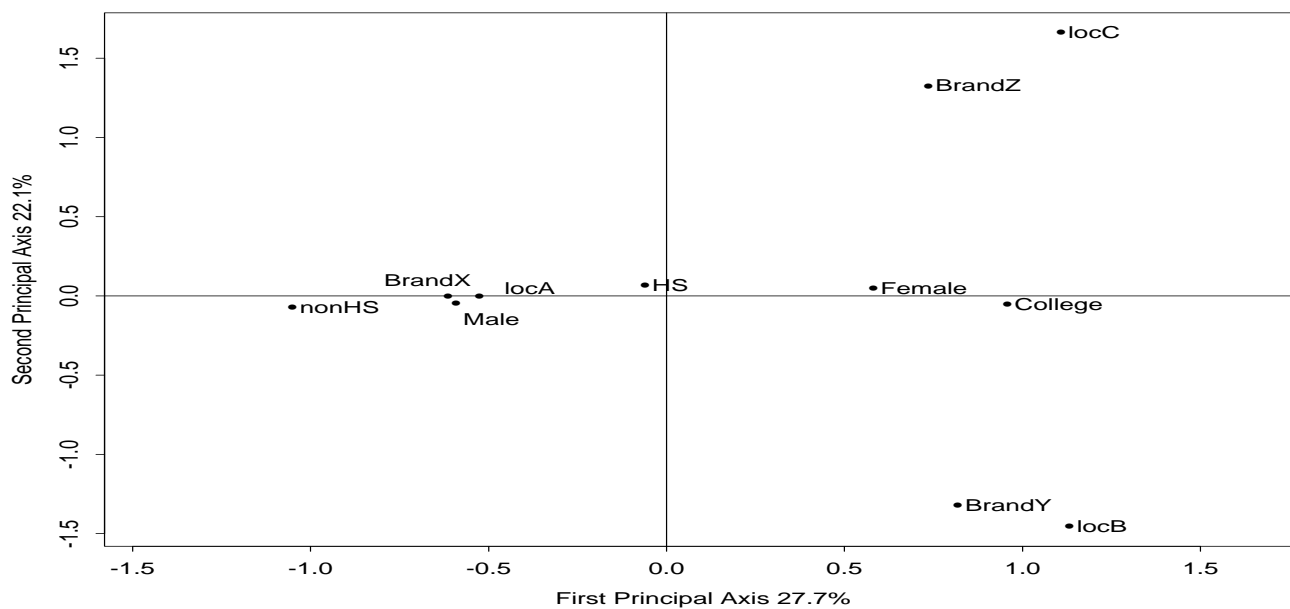


Figure 1: The two-dimensional graphical display obtained from Burt matrix.

MCA Graphical display of Modified Burt Matrix for simulated data

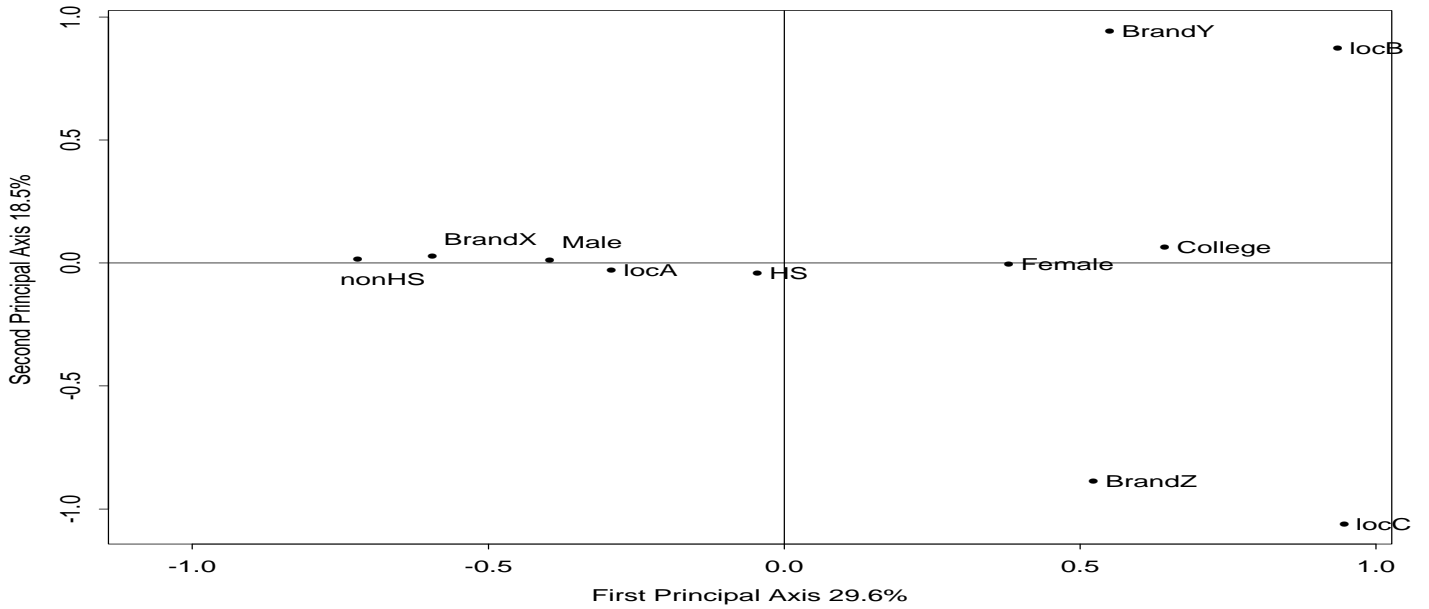


Figure 2: The two-dimensional graphical display obtained from modified Burt Matrix.

| | Brand1 | Brand2 | Male | Female | College | High School |
|-------------|--------|--------|------|--------|---------|-------------|
| Brand1 | 695 | 0 | 373 | 322 | 499 | 196 |
| Brand2 | 0 | 842 | 456 | 386 | 420 | 422 |
| Male | 373 | 456 | 829 | 0 | 457 | 372 |
| Female | 322 | 386 | 0 | 708 | 462 | 246 |
| College | 499 | 420 | 457 | 462 | 919 | 0 |
| High School | 196 | 422 | 372 | 246 | 0 | 618 |

Table 4: Burt Matrix

| | Brand1 | Brand2 | Male | Female | College | High School |
|-------------|--------|--------|------|--------|---------|-------------|
| Brand1 | 346 | 349 | 373 | 322 | 499 | 196 |
| Brand2 | 349 | 493 | 456 | 386 | 420 | 422 |
| Male | 373 | 456 | 468 | 361 | 457 | 372 |
| Female | 322 | 386 | 361 | 347 | 462 | 246 |
| College | 499 | 420 | 457 | 462 | 584 | 335 |
| High School | 196 | 422 | 372 | 246 | 335 | 283 |

Table 5: Modified Burt Matrix

| | | Singular Value | Principal Inertia | Chi-Square | Percent | Cum Percent |
|-------------|-------|----------------|-------------------|------------|---------|-------------|
| Burt Matrix | k=1 | 0.64473 | 0.41568 | 1993.5 | 41.57 | 41.57 |
| | k=2 | 0.57628 | 0.33210 | 1592.67 | 33.21 | 74.78 |
| | k=3 | 0.50222 | 0.25233 | 1209.64 | 25.22 | 100.00 |
| | total | | 1 | 4795.81 | 100 | |
| Modified | k=1 | 0.33318 | 0.11101 | 182.735 | 85.6 | 85.6 |
| Burt Matrix | k=2 | 0.13667 | 0.01868 | 30.749 | 14.4 | 100.00 |
| | total | | 0.12969 | 213.484 | 100 | |

Table 6: Summary for MCA of Burt and Modified Burt Matrix

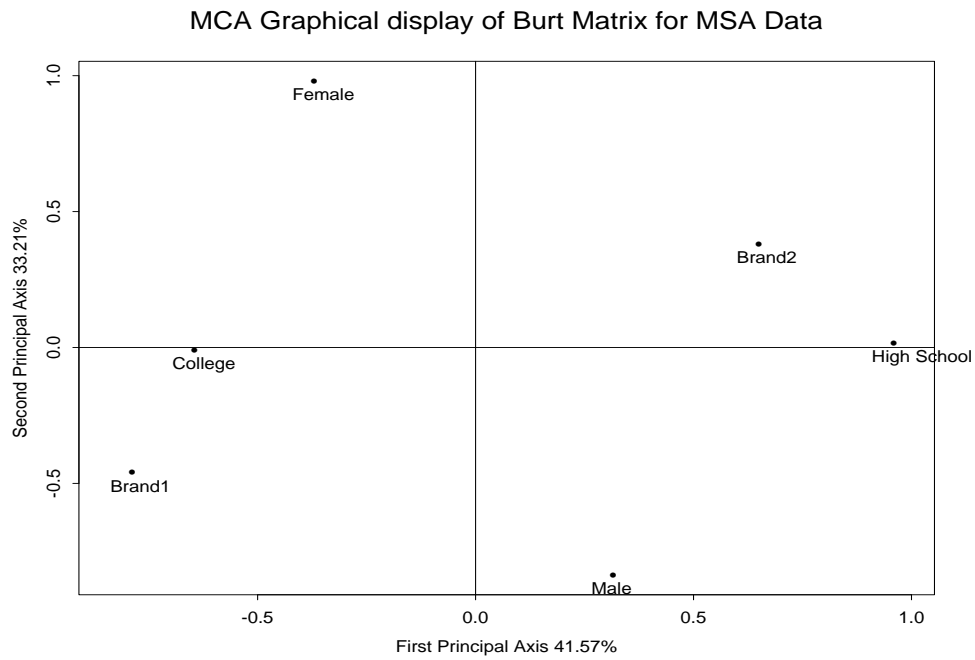


Figure 3: The two-dimensional graphical display obtained from Burt matrix.

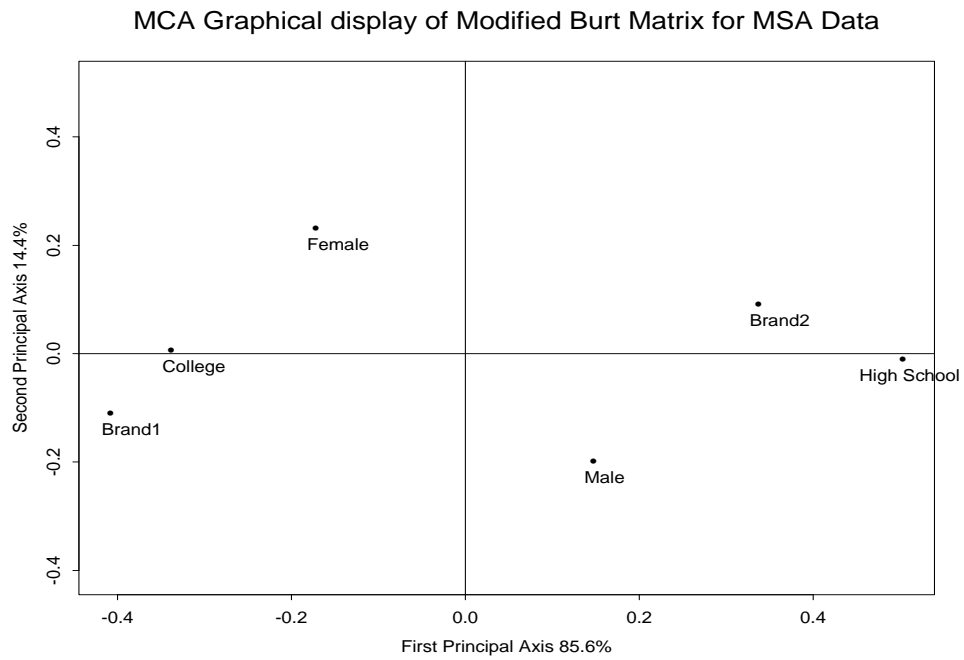


Figure 4: The two-dimensional graphical display obtained from modified Burt Matrix.