LUND UNIVERSITY
School of Economics and Management

Evaluation of scoring index with different normalization and distance measure with correspondence analysis

Anders Nilsson

Master Thesis in Statistics 15 ECTS
Spring Semester 2010

Supervisors: Professor Björn Holmquist and Antonio Marañon, Ph.D.

Contents

List of Figures

List of Tables

Abstract

The purpose of this thesis is to analyze data for a scoring system and evaluate different normalization procedures and distance measures for correspondence analysis. When bootstrapping 100 samples and evaluating coordinates for the row and column profiles the results show that the representation of the coordinates for the row and column profiles are similar when looking at the normalizations methods separately. The individual positioning of the attributes and brands does not change. However, the scaling is differently presented and when looking at biplots, combining row and column profiles, a different mapping of the rows and column profiles can be seen.

One has to be careful when choosing between the different normalization and distance measures. A guiding rule is to choose according to the underlying assumptions and according to the research objective. For this particular data, the relationship between the column profiles and row profiles are important hence symmetric normalization is preferred with euclidean distance.

Keywords: correspondence analysis, normalization, chi-squared distance, euclidean distance, Bootstrap

1 Introduction

1.1 Background

Particularly in market situations, it is important for business establishments or companies to find out how customers perceive their brands in terms of relevant product attributes. In forming market strategies, companies want to find out which attributes are associated with brand X, relative to the competitor's brand Y. One way to summarize the market positions is through a perceptual map, a graphical representation of brands and attributes in terms of coordinate systems where the positions of the different stimuli reveal how customers perceive the similarity of the objects. In terms of marketing, different strategies might be pursued to achieve differentitation.

In marketing research, the basis for a perceptual map is an $I \times J$ contingency table with, for example, the I attributes as the row variables and the different J brands as the column variables. The respondents evaluate the different brands based on the selected attributes. Likert scales of agreement or a simple dichotomy (agree/disagree or associated/not associated) are used to create a $I \times J$ table of brand-attribute scores. These values indicate some index of association between brands and attributes. This scoring index is used in other areas as well, such as the natural sciences, psychology and economics. In biology, the scoring system involves ratio scales which indicate the quantitative measure of some property for various species. In marketing science, a scoring structure is used to explain customer preference based on brand-attribute association scores. Respondents indicate which attributes are associated with each brand.

1.2 Problem

A common way to graphically summarize the scoring structure is to apply correspondence analysis on the derived 2-way contingency table. The likert scales are dichotomized as either (1) positive or negative (0), where the positive responses correspond to positive alternatives such as “good” and “very good”, while the negative responses correspond to indifference or the negative alternatives such as “bad” and “very bad”.

A problem using this kind of scoring index is that the number of answers is not equal to the number of respondents. Standard correspondence analysis is based on contingency tables where the scores are frequencies which sum up to the total number of observations. The number in each cell of a scoring index table is not unique, and one cannot tell which respondent has answered for the given brand-attribute combinations. Each respondent can evaluate multiple brand-attribute associations. One consequence of this scoring system is that some respondents with greater awareness of the brands can have greater influence on the outcome, while some of the less knowledgeable respondents contribute with minimal scores. Another obvious consequence is that leading brands tend to dominate the mapping. Popular brands tend to get more scores.

For the hypothetical example in the table below, we can see that several of the respondents associate brand D, probably indentified as the leading brand in the market, with all of the attributes.

	Brand A	Brand B	Brand C	Brand D	Sum
Attribute 1	55	40	10	150	255
Attribute 2	15	65	30	300	410
Attribute 3	20	10	49	95	174
Sum	90	115	89	545	839

Table 1: Point scoring index

One way to even out the influence would be to standardize the points column-wise; and derive the relative distributions. The resulting column percentages are as follows:

	Brand A	Brand B	Brand C	Brand D
Attribute 1	61.1	34.8	11.2	27.5
Attribute 2	16.7	56.5	33.7	55.1
Attribute 3	22.2	8.7	55.1	17.4
Sum %	100.0	100.0	100.0	100.0

Table 2: Point scoring index in terms of percentages

To map the brands and attributes based on a scoring index table, correspondence analysis is used. Correspondence analysis will determine the distances among brands and among attributes, using chi-squared measures that take into account the relative size of each of the columns and rows. The problem of mapping the scoring index in correspondence analysis is, however, a problem among market research practitioners. Standardization can compensate for brands with relatively lesser respondent awareness. There are available normalization techniques that can be used to convert the scoring index tables. However, which normalization technique to choose, why and what the differences among the techniques and the different results they yield are issues to look into.

1.3 Objectives

The aim of this thesis is to explore the results of the different normalization methods and distance measures and to tell which properties are to be considered when applying correspondence analysis to scoring index data.

1.4 Outline

- **Section 1** presents the marketing research problem and the objective of the study.
- **Section 2** presents the process of the computation of correspondence analysis, such as the chi-square measure and the singular value decomposition. The section ends with a presentation of the different normalization options.
- **Section 3** describes the data material, the software used and necessary pre-checks of the data set.
- **Section 4** presents results for correspondence analysis for the original data and for different normalization methods and distance measures.
- **Section 5** concludes and discusses the results, as well as makes further implications of the results.

2 Correspondence analysis

Correspondence analysis is a method for representing tabular data graphically, often involving nominal variables, Greenacre (1984), for example different brands and attributes. Correspondence analysis can handle both nonmetric data and nonlinear relationships. The categories are presented in a multidimensional space where the proximity indicates the level of association among row or column categories, Hair et al (2005).

The presentation of correspondence analysis follows the formulation of Greenacre (1984) but with adaptation to the notation used by SPSS. For a more general overview of the correspondence analysis methodology, readers are referred to the original authors Benzécri (1969) and Greenacre (1984). The correspondence algorithm in SPSS is composed of three major parts:

1. Singular value decomposition (SVD)
2. Centering and rescaling of the data and various rescaling of the results
3. Variance estimation by the delta method.

The emphasis of this section is to give a presentation of the general framework of correspondence analysis and how it is conducted. Hence, the general notation, singular value decomposition and the centering and rescaling is considered. The delta method is used to derive the approximation of the eigenvectors/eigenvalues.

2.1 Introductory notation

Let F be a matrix of frequencies with size $I \times J$ and with elements f_{ij}. F is a matrix composed of non-negative values with row and column sums which are non-zero. The correspondence matrix P, with size $I \times J$, is denoted as a matrix where all elements in F are divided with the grand total n, which is $P = \frac{1}{n}F$. Furthermore, let the vectors of row and column sums be denoted as $\mathbf{r}$ and $\mathbf{c}$, which can also be expressed as centroids of row and column clouds in their respective spaces $\mathbf{r}$ and $\mathbf{c}$, with row centroid defined as $\mathbf{c} = R^{'}\mathbf{r}$ and column centroid as $\mathbf{r} = C^{'}\mathbf{c}$.

The diagonal matrices of these sums are represented as $D_r = diag(r)$ and $D_c = diag(c)$. The respective row and column profiles of $\mathbf{P}$ are defined as the vectors of rows and columns of $\mathbf{P}$ divided by their respective sums. The row profile is defined as

$$R = D_r^{-1}\mathbf{P} = \begin{bmatrix} \tilde{\mathbf{r}}_1^{'} \\ \vdots \\ \tilde{\mathbf{r}}_I^{'} \end{bmatrix} \tag{1}$$

where $\tilde{\mathbf{r}}_I$ is each respective row profile (each attribute), and the column profile is defined as

$$C = D_c^{-1}\mathbf{P}^{'} = \begin{bmatrix} \tilde{\mathbf{c}}_1^{'} \\ \vdots \\ \tilde{\mathbf{c}}_J^{'} \end{bmatrix} \tag{2}$$

where $\tilde{\mathbf{c}}_J$is each respective column profile (each brand). A following relation can be seen for the marginal total of row i, $n\mathbf{r} = f_{i+}$ and for the marginal total of column j, $n\mathbf{c} = f_{j+}$.

The row and column profiles define two clouds of points in respective I and J dimensional weighted Euclidean space. The row clouds are defined by the I row profiles, with the points $\tilde{r_1}...\tilde{r_I}$, the mass being the I elements of $\mathbf{r}$, and the metric D_c^{-1}, weighted Euclidean with dimension defined by inverse of the elements of column $\mathbf{c}$. Likewise, the column cloud with J column profiles are defined as $\tilde{c_1}...\tilde{c_J}$ in the J dimensional space, with mass being the J elements of $\mathbf{c}$ and the metric D_r^{-1}, weighted Euclidean with dimension defined by inverse of $\mathbf{r}$.

2.2 The process of correspondence analysis

Correspondence analysis can be described as a process oriented method, a step by step computation process involving the distance of categorical variables in a $\mathbb{R}$ geometrical space based on the chi-square statistics. Chi-square can be described as a standardized measure of actual cell frequencies compared to expected frequencies, indicating some measure of association between row and column variables. The presentation of χ^2- statistic has the form of

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \tag{3}$$

which in matrix form can be expressed as

$$\chi^2 = (\mathbf{o} - \mathbf{e})' D_e^{-1} (\mathbf{o} - \mathbf{e}) \tag{4}$$

with $\mathbf{o}$ being the observed vector values

$$\mathbf{o}= \underbrace{\begin{bmatrix} \mathrm{o_1} \\ \mathrm{o_2} \\ \vdots \\ \mathrm{o_k} \end{bmatrix}}_{r\times 1}$$

and $\mathbf{e}$ being expected vector values

$$\mathbf{e} = \underbrace{\begin{bmatrix} \mathrm{e_1} \\ e_2 \\ \vdots \\ e_k \end{bmatrix}}_{r\times 1}$$

and De^{-1} be the diagonal inverse of the expected values

$$D_e^{-1} = \underbrace{\begin{bmatrix} \frac{1}{e_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{e_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{e_k} \end{bmatrix}}_{r\times c}$$

Correspondence analysis can also analyze and present the χ^2- statistic in terms of relative frequencies, as to express the results in terms of percentages. This representation of relative frequencies can be denoted as

$$\chi^2 = n^2 \left(\mathbf{p} - \overline{\mathbf{p}}\right)' D_e^{-1} \left(\mathbf{p} - \overline{\mathbf{p}}\right) \tag{5}$$

with observed relative frequencies expressed as $\mathbf{p}=\frac{1}{n}\mathbf{o}$ and the expected relative frequencies as $\overline{\mathbf{p}} = \frac{1}{n}\mathbf{e}$.

From the chi-square value we get the definition of inertia. Inertia can be defined as the integral of mass times the squared distance to the centroid, (Greenacre, 1984), or more simply, the chi-square value divided by the grand total. The total inertia is equal in both clouds (row and column), $in(I) = in(J)$, and is also equal to the mean square contingency coefficient calculated on n, that is

$$in(I) = in(J) = \sum_i \sum_j \frac{\left(f_{ij} - f_{+j} f_{i+}\right)^2}{f_{+j} f_{i+}} = \frac{\chi^2}{n}$$

2.2.1 Correspondence analysis by SPSS

In SPSS, the first step in order to perform correspondence analysis is to form the auxiliary matrix Z with the general element

$$z_{ij} = \frac{f_{ij}}{\sqrt{f_{i+} f_{+j}}} - \frac{\sqrt{f_{i+} f_{+j}}}{n}$$

$$z_{ij} = \frac{f_{ij} - f_{i+} f_{+j}/n}{\sqrt{f_{i+} f_{+j}}} \tag{6}$$

where $n = \mathbf{1}'F\mathbf{1}$

In matrix notation, Z can be expressed as

$$Z = (z_{ij}) = D_{rf}^{1/2} \left(F - \mathbf{r}_f \mathbf{c}_f'/n\right) D_{cf}^{-1/2} \tag{7}$$

where $D_{rf} = diag(\mathbf{r}_f)$ and $D_{cf} = diag\left(\mathbf{c}_f\right)$. This is the rcmean standardization, where both the row and column means are removed, with ehi-square measure. Note that the rcmean standardization is the only available option under chi-square distance while the euclidean distance has several other. It can also be written (with $\mathbf{P} = F/n$, $\mathbf{r} = \mathbf{P1} = \mathbf{r}_f/n$, $\mathbf{c}' = \mathbf{1}'\mathbf{P} = \mathbf{c}_f'/n$) as

$$Z = (z_{ij}) = \left(nD_r\right)^{-1/2} \left(nP - n\mathbf{r}\mathbf{c}'\right) \left(nD_c\right)^{-1/2}$$

$$Z = D_r^{-1/2} \left(P - \mathbf{r}\mathbf{c}'\right) D_c^{-1/2} \tag{8}$$

The defined auxiliary matrix Z is actually a form of Euclidean distance and can be represented as

$$\tilde{z}_{ij} = \tilde{f}_{ij} - \frac{\tilde{f}_{i+}\tilde{f}_{+j}}{n} \tag{9}$$

where $\tilde{f}_{ij} = f_{ij}$, $\tilde{f}_{i+} = f_{i+}$ and $\tilde{f}_{+j} = f_{+j}$. Then

$$z_{ij} = \frac{\tilde{z}_{ij}}{\sqrt{f_{i+}f_{+j}}}$$

$$z_{ij} = \frac{f_{ij} - f_{i+}f_{+j}/n}{\sqrt{f_{i+}f_{+j}}} \tag{10}$$

which is the same as z_{ij} for the Chi-square measure in equation (7).

2.3 Singular value decomposition

A central theme in correspondence analysis is singular value decomposition (SVD) which involves around the concept of dimension reduction of a data set. The aim of correspondence analysis is to find a low dimensional approximation of the data set to represent the row and column profiles that is the dimensions should be $min\{I, J\} - 1$. Another name for SVD is "Eckart-Young decomposition" after Eckart & Young (1936). SPSS makes a singular value decomposition of the auxiliary matrix Z into following three matrices

$$\underset{I\times J}{Z} = \underset{I\times M}{K}\,\underset{M\times M}{\Lambda}\,\underset{M\times J}{L}^{'} \tag{11}$$

Here K are the eigenvectors (left singular vectors) on the left-side of Z, being orthonormal basis for columns of Z. The matrix L being the eigenvectors (right singular vectors) on the right-side of Z, are the orthonormal basis for rows of Z. Finally Λ being the singular values of Z. M is the rank of A ($M \leq min\,\{I, J\}$).

A following relation is that $K'K = I$, $L'L = I$ and Λ is diagonal. Greenacre's approach is to perform the singular vale decomposition of the centered data,$P - \mathbf{rc}^{'}$, into $\mathbf{A}D_{\mu}\mathbf{B}^{'}$ where $\mathbf{A}^{'}D_r^{-1}\mathbf{A} = I$ and $\mathbf{B}^{'}D_c^{-1} = I$. A following relation is that $D_r^{-1/2}\mathbf{A} = K$ and $D_c^{-1/2}\mathbf{B} = L$.

2.4 Normalization

Normalization procedures in Correspondence analysis can be used to determine whether and how similarity of the row and column variables, as well as the relationship between them can be interpreted in terms of row and column coordinates and the origin of the plot. The different normalization options available in Stata (Internet) but which are also to be found in SPSS, are as follows:

Method	*Similarity row category*	*Similarity column category*	*Association row and column*
Symmetric	No	No	Yes
Principal	Yes	Yes	No
Row	Yes	No	Yes
Column	No	Yes	Yes

Table 3: Normalization methods

- **Symmetrical normalization**. According to SPSS Categories each dimension, rows are the weighted average of columns divided by the singular value. In similar way, columns are the weighted average of rows divided by the singular value, SPSS Categories. This normalization is useful when examining differences or similarities between rows and columns, between attributes and brands. This is the default normalization method by SPSS.

- **Principal normalization**. The distances between row points and column points are approximations of chi-square or euclidean distance. The distances represent the distance between the row or column and its corresponding average row or column profile. This normalization evaluates the differences between categories of the row variable and differences between categories of the column variable, SPSS Categories. According to SPSS differences between variables can not be interpreted.

- **Row principal normalization**. Distances between row points are approximations of chi-square distances (or of euclidean distances). This method yields row points that are weighted averages of the column points This normalization method is useful when examining differences or similarities between categories of the row variable, SPSS Categories.

- **Column principal normalization**. Distances between column points are approximations of chi-square distances (or of euclidean distances). This method yields column points that are weighted averages of the row points. This normalization is useful when examining differences or similarities between categories of the column variable, SPSS Categories.

In SPSS there is also a fith normalization option, which allows the user to specify any value in the range $[-1, 1]$ and spread the inertia over row/column scores to varying degress. A value of 1 is equal to the principal method, a value of 0 is equal to the Symmetrical method and a value of -1 is equal to the column principal method.

2.4.1 Normalization calculation

The normalization, described in SPSS user manual PASW Statistics 18 Algorithms, is calculated on on the row and column principal co-oordinates. The row principal co-oordinates of the row profiles can be expressed as

$$\mathbf{F} = \left(D_r^{-1}\mathbf{P} - \mathbf{1c}'\right) D_c^{-1}\mathbf{B}$$

$$\begin{aligned} \mathbf{F} &= D_r^{-1/2} Z L \\ \mathbf{F} &= D_r^{-1/2} K \Lambda^{\alpha} \end{aligned} \tag{12}$$

that is the row principal co-ordinates of the row profiles are dependent on the left side eigenvector K and eigenvalues Λ from the SVD, and with α being a constant. In a similar manner the principal co-oordinates of the column profiles can be defined as

$$\mathbf{G} = \left(D_c^{-1}\mathbf{P}^{'} - \mathbf{1r}^{'}\right) D_r^{-1}\mathbf{A}$$

$$\begin{aligned} \mathbf{G} &= D_c^{-1/2} Z^{'} L \\ \mathbf{G} &= D_c^{-1/2} L \Lambda^{\beta} \end{aligned} \tag{13}$$

with L being the right side eigenvector and eigenvalues Λ from the SVD and β being a constant. The adjustment to the row and column metric, the unstandardized principal row coordinates are calculated as

$$\tilde{r}_{is} = k_{is}/\sqrt{\tilde{f}_{i+}/N} \tag{14}$$

which is ${D_r^{-1/2}}$, and the principal column coordinates as

$$\tilde{c}_{js} = l_{js}/\sqrt{\tilde{f}_{+j}/N} \tag{15}$$

hence, the normalization implies that the row scores

$$r_{is} = \tilde{r}_{is}\lambda_s^{\alpha}$$

which implies that

$$r_{is} = D_r^{-1/2}K\Lambda^{\alpha} \tag{16}$$

and in similar manner for the column scores

$$c_{js} = \tilde{c}_{js}\lambda_s^{\beta}$$

which implies that

$$c_{js} = D_c^{-1/2}L\Lambda^{\beta} \tag{17}$$

Depending on the normalization option chosen in SPSS one can normalize the score. The normalization option q can be chosen to be any value in the interval $[-1, 1]$ or it can be specified according to the following designations:

$$q = \begin{cases} 0 & \alpha = 1/2, \beta = 1/2 \\ 1 & \alpha = 1, \beta = 0 \\ -1 & \alpha = 0, \beta = 1 \end{cases}$$

When q is equal to 0 the symmetric normalization method is choosen and it can be noted here that the square root of the eigenvalues are calculated. When q is equal to 1 we ge the row normalization and -1 gives column normalization. There is another possibility of choosing the designation principal normalization, which does not correspond to a q-value. When principal normalization is chosen, normalization parameters α for the rows and β for the columns are both set to 1. This normalization is the way how Greenacre (1984) presents the scaling of the principal row and column co-ordinates.

3 Data

The data material is obtained from the market research company GfK Sweden (Growth from Knowledge) and comprises of a study of the evaluation of the ratings of the attributes of different juice brands. In total 6 brands were selected and with 23 statements being evaluated with the option to answer yes or no, dicotomized as value 1 or 0. In total 503 participants were included in the study. The data is considered as non-metric. Since the data is considered to be confidential no further in-depth description of the material is presented. The programs used are SPSS PASW 18 and MATLAB R2007a.

Correspondence analysis is described as technique with relative freedom from assumptions, Hair et al (2005). The data material is considered as nonmetric, analysed as crosstabulated data. For the analysis of crosstabulated data the cases should be considered to be unique. However, each respondent may select to answer all of the questions in the survey, on the other hand the analysis of consumer preferences with correspondence analysis is using a data set that can be characterized as a scoring index. Except from this conceptual discrepancy of the underlying data set, no pre-assumptions of the data has been considered.

4 Results

4.1 Initial results from correspondence analysis

The input matrix for the correspondence analysis, the data from a GfK survey where respondents evaluated different brands and attributes. The data set with the row and column variables as well as the row and column marginal is presented in the following table:

	Column						
Row	Brand 1	Brand 2	Brand 3	Brand 4	Brand 5	Brand 6	Active margin
Attribute 1	383	270	15	129	139	81	1017
Attribute 2	14	96	355	220	1	133	819
Attribute 3	275	241	65	165	26	116	888
Attribute 4	54	127	308	152	12	65	718
Attribute 5	4	8	4	147	1	12	176
Attribute 6	63	158	311	193	23	181	929
Attribute 7	147	37	0	24	72	4	284
Attribute 8	125	135	182	131	18	145	736
Attribute 9	68	37	2	15	19	22	163
Attribute 10	104	127	101	95	37	92	556
Attribute 11	35	60	102	69	2	56	324
Attribute 12	3	5	74	28	3	33	146
Attribute 13	9	32	170	94	2	86	393
Attribute 14	186	136	12	48	62	7	451
Attribute 15	4	1	81	44	2	74	206
Attribute 16	5	35	168	92	1	84	385
Attribute 17	79	108	201	173	24	52	637
Attribute 18	24	30	90	51	5	22	222
Attribute 19	47	71	173	78	13	269	651
Attribute 20	1	3	32	11	1	89	137
Attribute 21	149	84	20	34	67	24	378
Attribute 22	7	9	23	2	3	8	52
Attribute 23	97	77	53	40	31	57	355
Active margin	1883	1887	2542	2035	564	1712	10623

Table 4: Input data for Correspondence analysis

The marginal row profiles (attributes) for the correspondence analysis is obtained by dividing each cell content with their respective row total. The column mass is obtained by dividing the row total with the total points, e.g. the mass for the first brand is $1883/10623 = 0.177$. The marginal row profiles and masses for the columns are presented in table 5:

	Column						
Row	Brand 1	Brand 2	Brand 3	Brand 4	Brand 5	Brand 6	Active margin
1	0.377	0.265	0.015	0.127	0.137	0.080	1.000
2	0.017	0.117	0.433	0.269	0.001	0.162	1.000
3	0.310	0.271	0.073	0.186	0.029	0.131	1.000
4	0.075	0.177	0.429	0.212	0.017	0.091	1.000
5	0.023	0.045	0.023	0.835	0.006	0.068	1.000
6	0.068	0.170	0.355	0.208	0.025	0.195	1.000
7	0.518	0.130	0.000	0.085	0.254	0.014	1.000
8	0.170	0.183	0.247	0.178	0.024	0.197	1.000
9	0.417	0.227	0.012	0.092	0.117	0.135	1.000
10	0.187	0.228	0.182	0.171	0.067	0.165	1.000
11	0.108	0.185	0.315	0.213	0.006	0.173	1.000
12	0.021	0.034	0.507	0.192	0.021	0.226	1.000
13	0.023	0.081	0.433	0.239	0.005	0.219	1.000
14	0.412	0.302	0.027	0.106	0.137	0.016	1.000
15	0.019	0.005	0.393	0.214	0.010	0.359	1.000
16	0.013	0.091	0.436	0.239	0.003	0.218	1.000
17	0.124	0.170	0.316	0.272	0.038	0.082	1.000
18	0.108	0.135	0.405	0.230	0.023	0.099	1.000
19	0.072	0.109	0.266	0.120	0.020	0.413	1.000
20	0.007	0.022	0.234	0.080	0.007	0.650	1.000
21	0.394	0.222	0.053	0.090	0.177	0.063	1.000
22	0.135	0.173	0.442	0.038	0.058	0.154	1.000
23	0.273	0.217	0.149	0.113	0.087	0.161	1.000
Mass	0.177	0.178	0.239	0.192	0.053	0.161	1.000

Table 5: Row profile

In a similar manner, the marginal column profile (brands) are obtained by dividing each cell content with the column total. The mass for each row profile is obtained in a similar manner by dividing the column total with the total number the total number of points. The column profiles and the row masses are presented in table 6:

	Column						
Row	Brand 1	Brand 2	Brand 3	Brand 4	Brand 5	Brand 6	Mass
1	0.203	0.143	0.006	0.063	0.246	0.047	0.096
2	0.007	0.051	0.140	0.108	0.002	0.078	0.077
3	0.146	0.128	0.026	0.081	0.046	0.068	0.084
4	0.029	0.067	0.121	0.075	0.021	0.038	0.068
5	0.002	0.004	0.002	0.072	0.002	0.007	0.017
6	0.033	0.084	0.122	0.095	0.041	0.106	0.087
7	0.078	0.020	0.000	0.012	0.128	0.002	0.027
8	0.066	0.072	0.072	0.064	0.032	0.085	0.069
9	0.036	0.020	0.001	0.007	0.034	0.013	0.015
10	0.055	0.067	0.040	0.047	0.066	0.054	0.052
11	0.019	0.032	0.040	0.034	0.004	0.033	0.030
12	0.002	0.003	0.029	0.014	0.005	0.019	0.014
13	0.005	0.017	0.067	0.046	0.004	0.050	0.037
14	0.099	0.072	0.005	0.024	0.110	0.004	0.042
15	0.002	0.001	0.032	0.022	0.004	0.043	0.019
16	0.003	0.019	0.066	0.045	0.002	0.049	0.036
17	0.042	0.057	0.079	0.085	0.043	0.030	0.060
18	0.013	0.016	0.035	0.025	0.009	0.013	0.021
19	0.025	0.038	0.068	0.038	0.023	0.157	0.061
20	0.001	0.002	0.013	0.005	0.002	0.052	0.013
21	0.079	0.045	0.008	0.017	0.119	0.014	0.036
22	0.004	0.005	0.009	0.001	0.005	0.005	0.005
23	0.052	0.041	0.021	0.020	0.005	0.033	0.003
Active Margin	1.000	1.000	1.000	1.000	1.000	1.000	

Table 6: Column profile

A summary table for the output from the correspondence analysis from SPSS is presented in table 7 below. A total of five dimensions are in the output, since $min\{16, 6\} - 1$. One usually would want to choose between one and three dimensions when representing the row and column profiles, as more dimensions are far too complex to interpret. The table below indicates both for the chi-square distance and with the euclidean distance that choosing 2 dimensions is suitable, as two dimensions represent roughly 90% of the cumulative inertia. From table 7, summarizing the results for the correspondence analysis with the chi-square distance and the euclidean distance, the observed chi-square value is significant, hence there is a relationship between the variables. No chi-square value is obtained for the euclidean distance.

For the chi-square distance, the two first dimensions have reported inertia (eigenvalue) of 0.321 and 0.070 respectively, accounting for 86.0% of the cumulative inertia. That is the two first dimensions accounts for explaining 86.0% of the total explained variance, which is 45.4%. The correlation between the two first dimension is also low, 0.066. For the euclidean distance the reported inertia for the two first dimensions are 0.435 and 0.065 respectively, which accounts for 92.5% of the accumulative inertia. That is the first dimension explains 92.5% of the total explained variance of 54.0% . The correlation between the two first dimensions is also here low, 0.090.

Next the row and column scores (coordinates), equation (16) and (17) for the two chosen dimension are calculated. Readers are referred to tables 12-15 for calculation of the row and column scores for the different normalization methods under Chi-square distance and Euclidean distance. These scores will be plotted in biplots in the following subsection.

4.1.1 Biplots for original data

Forming a biplot , a combined plot with both the column and row profiles, for the original data with each distance measure and normalization method, we get the following figures:

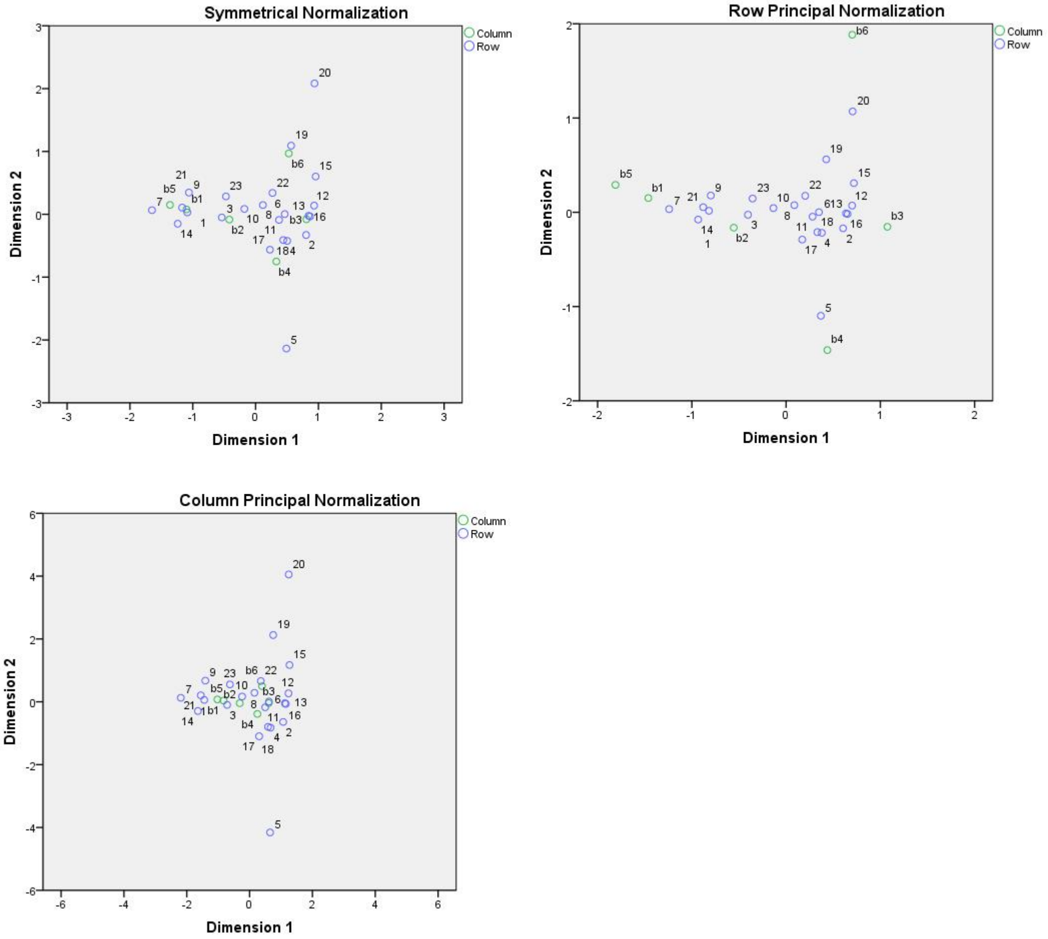

Figure 1: Biplot for different normalizations with Chi-square distance

One can note that the scaling are different for each biplot but there are some differences between the normalization methods. For example, brand 4 and brand 1 for row normalization is somewhat differently scaled. Principal normalization was not in the output by SPSS since the association between row and column profiles can not be interpreted.

						Proportion of inertia		Confidence Singular Value				
						Accounted	Cumulative	Standard	Correlation			
Distance	Dimension	Singular value	Inertia	Chi-square	Sig.	for		Deviation	2	3	4	5
Chi-square distance	1	0.567	0.321			0.707	0.707	0.007	0.066	0.057	0.160	-0.007
	2	0.264	0.070			0.153	0.860	0.001		0.349	0.030	0.003
	3	0.198	0.039			0.087	0.947	0.009			0.045	0.002
	4	0.146	0.021			0.047	0.993	0.011				-0.269
	5	0.055	0.003			0.007	1.000	0.011				
	Total		0.454	4827.508	0.000	1.000	1.000					
Euclidean distance	1	0.660	0.435			0.805	0.805	0.003	0.090	0.015	-0.022	-0.318
	2	0.254	0.065			0.120	0.925	0.007		0.118	0.005	-0.006
	3	0.158	0.025			0.046	0.971				0.128	-0.084
	4	0.113	0.013			0.024	0.995					-0.419
	5	0.053	0.003			0.005	1.000					
	Total		0.540			1.000	1.000					

Table 7: Summary table for Chi-square distance and Euclidean distance

The biplots for the different normalization methods under Euclidean distance can be seen in figure 2:

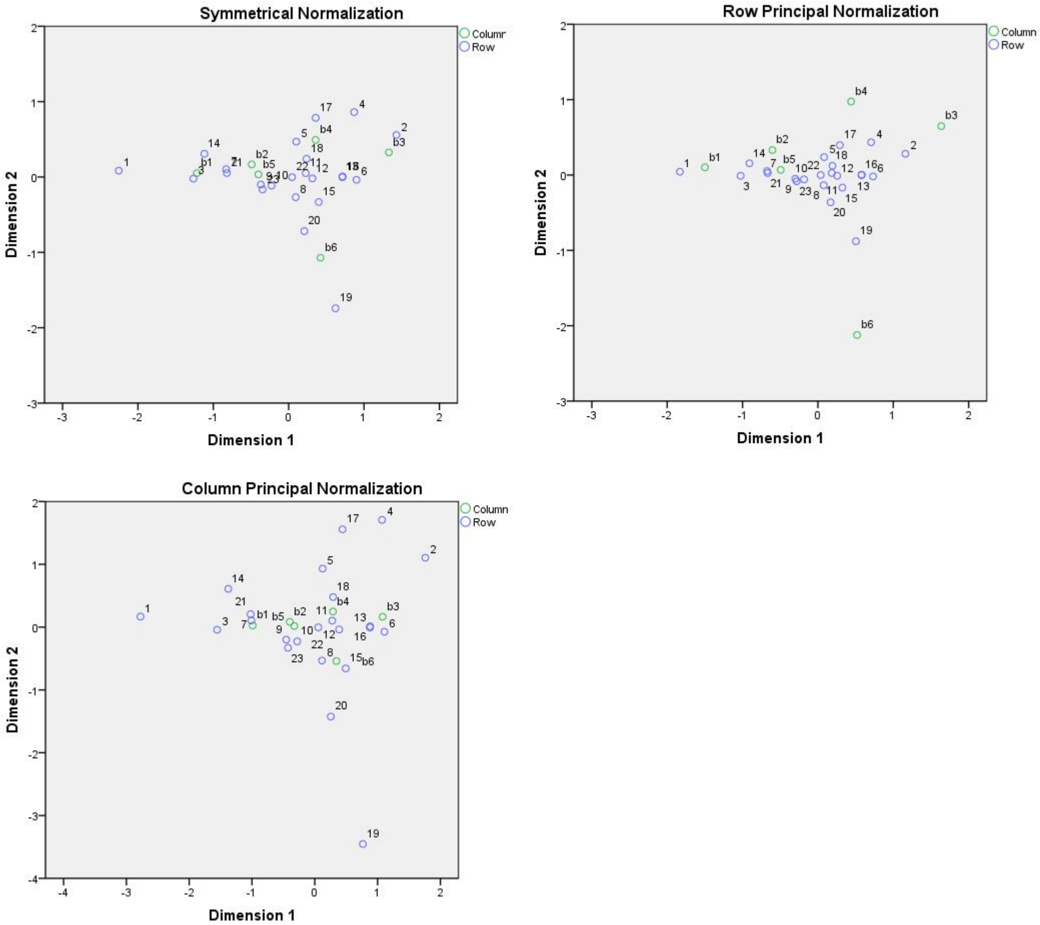

Figure 2: Biplot for different normalizations with Euclidean distance

The scaling is almost identical for biplots under Euclidean distance. Brand 6 can be seen as differently scaled for the row principal normalization method. As a final notice when comparing the biplots, the Chi-square distance and Euclidean distance tend give different representations of the row and column scores.

4.2 Validation of Correspondence analysis

The boostrap is a distribution free method by Efron (1979) of assessing sampling variability based on resampling from the empirical distribution. Boostrapping can be an approach when estimating the properties of an estimator, such as the mean or variance. In conjunction with Correspondence, bootstrap can be used to access the external validity, the stability of the results. A modified procedure of The Bootstrap MATLAB toolbox has been used, bootstrapping each individual and selecting 100 samples. Each sample has been used to rerun the correspondence analysis with different normalization methods (symmetric, row principal, column principal and principal) and distance measures (Euclidean or Chi-square).

The chi-square distance is used when the rows or columns are specified as equal. Similar, the Euclidean distance can be used when the rows or columns are not specified as equal. To study the dispersion of the different normalization methods we calculate the volume of the convex hull for each normalization method. For a set of points X in a real vector space V, the convex hull is the minimal convex set containing X.

4.2.1 Chi-square distance

For the different normalization methods with Chi-square distance we get the following volumes for the 100 bootstrapped samples for the row profiles (attributes):

	Normalization method			
Attribute	Symmetric	Row	Column	Principal
1	0.0148	0.0064	0.0394	0.0064
2	0.0236	0.0102	0.0643	0.0102
3	0.0309	0.0125	0.0765	0.0125
4	0.0439	0.0171	0.1144	0.0171
5	0.3913	0.1543	1.0053	0.1543
6	0.0174	0.0076	0.0411	0.0076
7	0.0796	0.0335	0.1969	0.0335
8	0.0299	0.0116	0.0771	0.0116
9	0.1827	0.0723	0.4693	0.0723
10	0.0544	0.0212	0.1402	0.0212
11	0.0781	0.0302	0.2023	0.0302
12	0.1262	0.0542	0.3114	0.0542
13	0.0379	0.0158	0.0953	0.0158
14	0.0322	0.0133	0.0853	0.0133
15	0.1163	0.0477	0.3084	0.0477
16	0.0483	0.0184	0.1305	0.0184
17	0.0279	0.0113	0.0717	0.0113
18	0.1329	0.0533	0.3347	0.0533
19	0.0612	0.0263	0.1536	0.0263
20	0.1872	0.0705	0.5104	0.0705
21	0.0410	0.0168	0.1054	0.0168
22	0.7951	0.3065	2.0637	0.3065
23	0.0783	0.0303	0.2020	0.0303
Average volume	0.1144	0.0453	0.2956	0.0453

Table 8: Volume for attributes under Chi-square distance

From the table above we can see that column normalization gives the largest volume. Attribute 5 and 22 under column normalization have a large calculated volume compared to the other normalization methods. Principal normalization gives the same results as row normalization, which has the lowest average volume.

A plotting of the bootstrapped coordinates for row profiles (in blue) as well as of the original data set (in red) can be seen in figure 3. The normalization methods under the Chi-square distance tend to give similar results. However, the scaling of the figures are somewhat different. It should be noted that the plotting of the row scores of the bootstrapped samples for attribute 5 and 22 are very scattered.

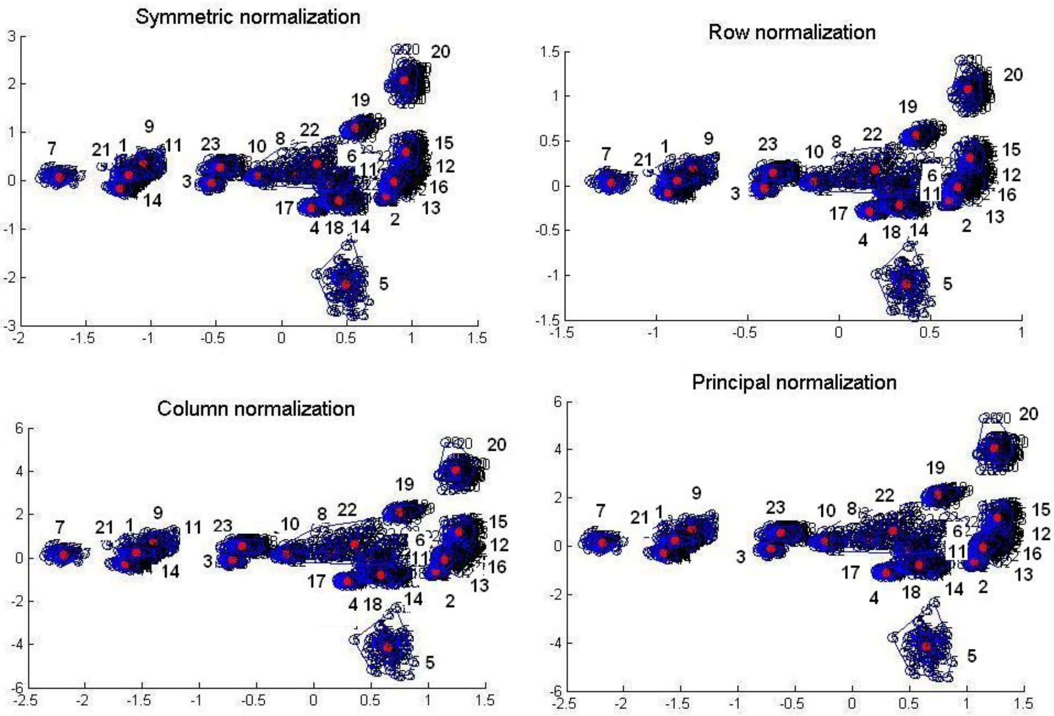

Figure 3: Normalization with Chi-square distance for attributes

In similar way we study the volume for the different column profiles (brands), these are given in table 9. From the table below we can see that the different normalization methods give a very low average volume. Row normalization will give the largest average volume and column normalization will give the lowest. Column and principal normalization give identical results.

	Normalization method			
Brand	Symmetric	Row	Column	Principal
1	0.0124	0.0299	0.0053	0.0053
2	0.0199	0.0510	0.0078	0.0078
3	0.0185	0.0563	0.0079	0.0079
4	0.0401	0.1041	0.0164	0.0164
5	0.0609	0.1458	0.0261	0.0261
6	0.0407	0.1024	0.0178	0.0178
Average volume	0.0321	0.0816	0.0135	0.0135

Table 9: Volume for brands under Chi-square distance

A plotting of the bootstrapped coordinates for column profiles (in green) as well as of the original data set (in red) can be seen in figure 4. From the figure below we can see that the plotting of the column scores are almost identical, only some minor differences in terms of scaling.

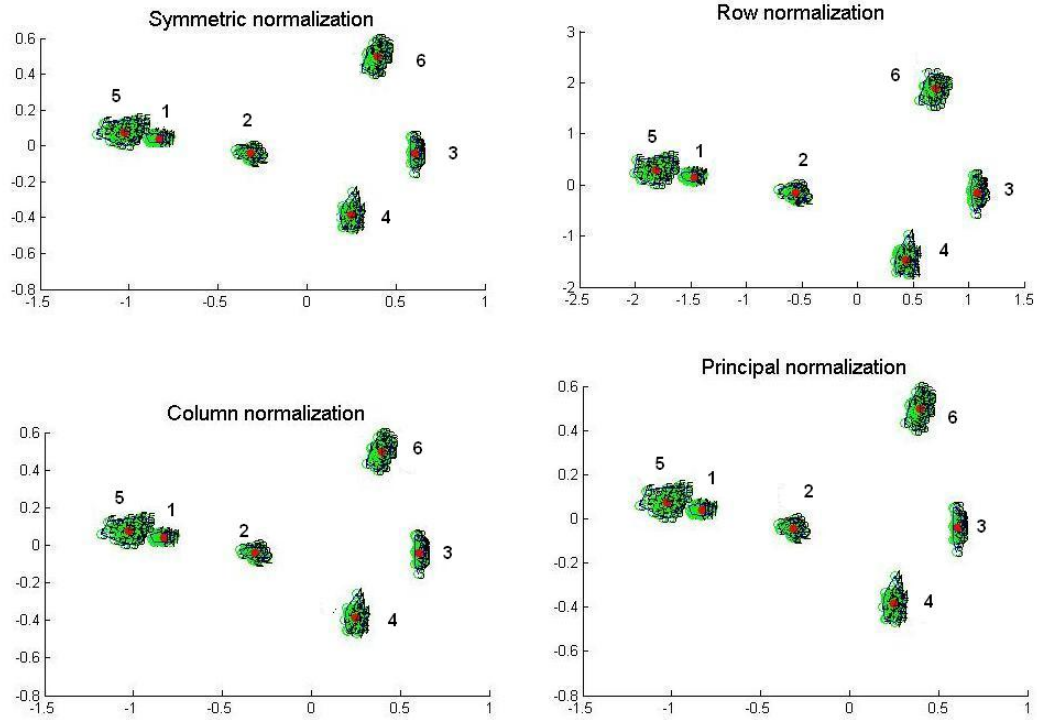

Figure 4: Normalization with Chi-square distance for brands

4.2.2 Euclidean distance

For the different normalization methods with Euclidean distance we get the following volumes for the 100 bootstrapped samples for the row profiles (attributes)

	Normalization method			
Attribute	Symmetric	Row	Column	Principal
1	0.0581	0.0311	0.1995	0.0311
2	0.2585	0.1230	0.5570	0.1230
3	0.0728	0.0348	0.1615	0.0348
4	0.3242	0.1368	0.7804	0.1368
5	0.1259	0.0497	0.3192	0.0497
6	0.0707	0.0335	0.1503	0.0335
7	0.0323	0.0133	0.0813	0.0133
8	0.1412	0.0578	0.3448	0.0578
9	0.0485	0.0193	0.1237	0.0193
10	0.1027	0.0426	0.2479	0.0426
11	0.0519	0.0211	0.1287	0.0211
12	0.0271	0.0115	0.0646	0.0115
13	0.0386	0.0167	0.0907	0.0167
14	0.1404	0.0547	0.3643	0.0547
15	0.1166	0.0476	0.2962	0.0476
16	0.0444	0.0182	0.1099	0.0182
17	0.3882	0.1598	0.9515	0.1598
18	0.0947	0.0394	0.2290	0.0394
19	0.8029	0.3354	1.9404	0.3354
20	0.1407	0.0598	0.3422	0.0598
21	0.0361	0.01476	0.0910	0.0148
22	0.0124	0.0051	0.0309	0.0051
23	0.1125	0.0461	0.2762	0.0461
Average volume	0.1409	0.0600	0.3392	0.0600

Table 10: Volume for attributes under Euclidean distance

.

From table 10 column normalization has the largest calculated average volume and row normalization the lowest. Row and principal normalization have the same calculated average volume. Attribute 19, 4 and 2 for column normalization have larger calculated average volumes compared to the other normalization methods.

A plotting of the bootstrapped coordinates for row profiles (in blue) as well as of the original data set (in red) can be seen in figure 5.

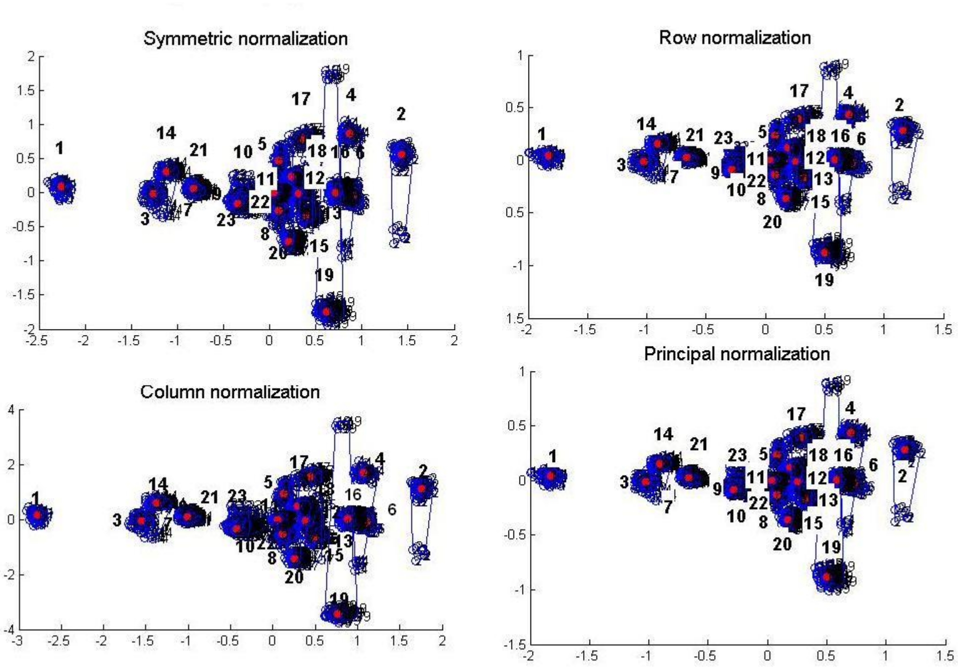

Figure 5: Normalization with Euclidean distance for attributes

From the figure below we can see that the plotting of the row scores are almost identical, only some minor differences in terms of scaling. We can see that attribute 19, 4 and 2 tend to have large variations.

In similar way we study the volume for the different column profiles (brands), these are presented in table 11 below. From table 11 Symmetric normalization has very large average volume, especially for brand 1, 3 and 6. Column normalization has the lowest average volume. Column and principal normalization show identical results.

	Normalization method			
Brand	Symmetric	Row	Column	Principal
1	0.9074	0.0248	0.0056	0.0056
2	0.2920	0.1733	0.0351	0.0351
3	0.8952	0.1960	0.0433	0.0433
4	0.1800	0.4212	0.0678	0.0678
5	0.1185	0.0295	0.0047	0.0047
6	1.8354	0.6730	0.1232	0.1232
Average volume	0.70473	0.2529	0.0466	0.0466

Table 11: Volume for brands under Euclidean distance for brands

A plotting of the bootstrapped coordinates for column profiles (in green) as well as of the original data set (in red) can be seen in figure 6. From figure 4 below we can see that symmetric normalization have bootstrappped column scores that tend to give very scattered results. Overall brand 4 tend to vary for all the normalization methods.

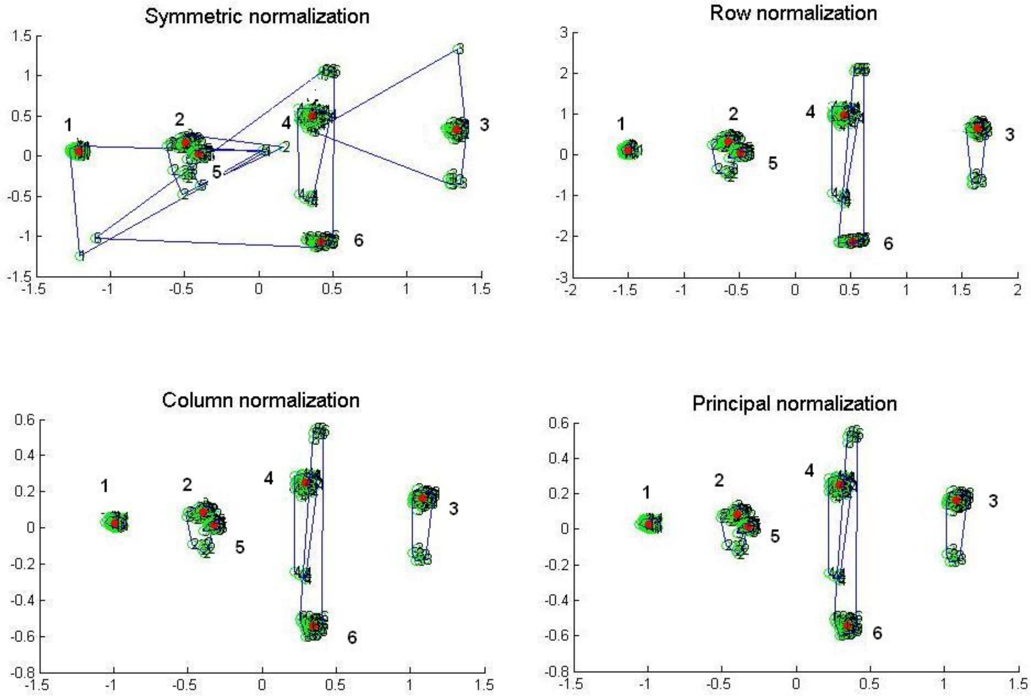

Figure 6: Normalization with Euclidean distance for brands

5 Discussion and conclusions

By inspection of the previous figures, figure 3-6, presented in the previous section, there are no general shifts of the individual brands, however the figures of the different normalization methods are differently scaled. There are also certain changes in the dispersion of the brands and attributes for the various normalization options and most noteworthy between the different distance measure.

The calculated volume for the different normalization options was greater for the Euclidean measure compared to the Chi-square measure. The normalization options under chi-square measure does not differ in great amount in terms of position and in terms of scaling for the different figures. The positioning of the brands and attributes for the different normalization methods with the euclidean distance measure, the scaling tends to differ much more. A comparison between the brands is brand 4, 5 and 6 notably differs between the distance measures.

Row normalization was most efficient when presenting the row profiles (attributes) and column normalization was most efficient when presenting the column profiles (brands). Symmetric normalization was least effective presenting the brands under euclidean normalization, which is due to greater dispersion for some bootstrap sample. Hence the symmetric normalization procedure is not regarded as a very stable normalization procedure. Even though symmetric normalization is not regarded as the most stable normalization procedure under euclidean distance, it would be the recommended choice when looking at a final biplot presenting the brands and attributes in a same graphical representation.

A final note is that even though the Euclidean distance and Chi-square measure are similarly computed, they tend to represent the row and column scores different when under normalization.

An alternative approach for this study would have been to bootstrap coordinates for the row and column profiles. This suggestion has been proposed by Greenacre (1984) since bootstrapping each individual sample and rerunning the analysis would have been too time consuming and all too computer intensive. This suggestion was initially performed, however, since SPSS performs a different way of rescaling and normalization than Greenacre (1984), this approach was abandoned.

For this study, one limitation was that the sample size could have been increased. However, it is hard to simplify the process of rerunning 100 analysis for different distance measures as well as different normalization methods and selecting the row and column profiles coordinates along the selected dimensions.

This study shows the dilemma positioning of the coordinates under different distance measures and normalizations, and one can ask how you evaluate the actual positioning of the row and column profiles. The results from a correspondence analysis has been criticized due to its difficulty to interpret the distance measure, Hair et al (2005). One way to evaluate the results from a correspondence analysis would be to let the company who have conducted or asked for the correspondence analysis to follow up and evaluate the positioning of the row and attributes. For instance, there could be an evaluation of the conducted marketing strategy in relation to profit. Did the planned strategy achieve the desired positioning we wanted?

A Row and column coordinates for original data

	Symmetric		Row principal		Column principal		Principal	
Attribute	Scoring in dimension		Scoring in dimension		Scoring in dimension		Scoring in dimension	
	1	2	1	2	1	2	1	2
1	-1.087	0.030	-0.818	0.016	-1.444	0.059	-0.818	0.016
2	0.803	-0.330	0.605	-0.170	1.067	-0.642	0.605	-0.170
3	-0.538	-0.050	-0.405	-0.026	-0.715	-0.098	-0.405	-0.026
4	0.503	-0.423	0.379	-0.218	0.668	-0.824	0.379	-0.218
5	0.490	-2.137	0.369	-1.098	0.651	-4.158	0.369	-1.098
6	0.463	0.003	0.349	0.002	0.616	0.006	0.349	0.002
7	-1.649	0.065	-1.241	0.034	-2.190	0.127	-1.241	0.034
8	0.116	0.147	0.087	0.076	0.154	0.287	0.087	0.076
9	-1.061	0.346	-0.799	0.178	-1.410	0.673	-0.799	0.178
10	-0.179	0.086	-0.135	0.044	-0.238	0.168	-0.135	0.044
11	0.373	-0.090	0.281	-0.046	0.495	-0.175	0.281	-0.046
12	0.931	0.138	0.701	0.071	1.237	0.269	0.701	0.071
13	0.844	-0.024	0.635	-0.012	1.121	-0.046	0.635	-0.012
14	-1.240	-0.151	-0.933	-0.077	-1.647	-0.293	-0.933	-0.077
15	0.955	0.601	0.719	0.309	1.269	1.170	0.719	0.309
16	0.866	-0.034	0.652	-0.017	1.151	-0.066	0.652	-0.017
17	0.228	-0.564	0.171	-0.290	0.302	-1.097	0.171	-0.290
18	0.441	-0.410	0.332	-0.211	0.585	-0.797	0.332	-0.211
19	0.565	1.092	0.426	0.561	0.751	2.125	0.426	0.561
20	0.938	2.082	0.706	1.070	1.245	4.052	0.706	1.070
21	-1.169	0.107	-0.880	0.055	-1.553	0.207	-0.880	0.055
22	0.269	0.339	0.203	0.174	0.357	0.660	0.203	0.174
23	-0.473	0.284	-0.356	0.146	-0.628	0.553	-0.356	0.146

Table 12: Coordinates for row profiles for normalizations under Chi-square distance

	Symmetric		Row principal		Column principal		Principal	
Brand	Scoring in dimension		Scoring in dimension		Scoring in dimension		Scoring in dimension	
	1	2	1	2	1	2	1	2
1	-1.101	0.078	-1.462	0.151	-0.829	0.040	-0.829	0.040
2	-0.418	-0.084	-0.555	-0.164	-0.315	-0.043	-0.315	-0.043
3	0.809	-0.079	1.074	-0.154	0.609	-0.041	0.609	-0.041
4	0.329	-0.751	0.437	-1.461	0.248	-0.389	0.248	-0.386
5	-1.363	0.149	-1.810	0.291	-1.026	0.077	-1.026	0.077
6	0.528	0.968	0.702	1.884	0.398	0.497	0.398	0.497

Table 13: Coordinates for column profiles for normalizations under Chi-square distance

	Symmetric		Row principal		Column principal		Principal	
Attribute	Scoring in dimension		Scoring in dimension		Scoring in dimension		Scoring in dimension	
	1	2	1	2	1	2	1	2
1	2.255	0.085	-1.831	0.043	-2.776	0.168	-1.831	0.043
2	1.429	0.557	1.161	0.281	1.1760	1.105	1.161	0.281
3	-1.263	-0.021	-1.026	-0.011	-1.555	-0.042	-1.026	-0.011
4	0.869	0.861	0.706	0.434	1.070	1.707	0.706	0.434
5	0.101	0.470	0.082	0.237	0.124	0.931	0.082	0.237
6	0.900	-0.037	0.731	-0.019	1.109	-0.074	0.731	-0.019
7	-0.830	0.104	-0.674	0.052	-1.022	0.206	0.052	0.025
8	0.093	-0.268	0.076	-0.135	0.115	-0.532	-0.135	0.002
9	-0.369	-0.100	-0.300	-0.051	-0.455	-0.199	-0.051	0.004
10	-0.227	-0.114	-0.184	-0.058	-0.280	-0.227	-0.058	0.003
11	0.226	0.052	0.183	0.026	0.278	0.103	0.026	0.002
12	0.315	-0.019	0.256	-0.009	0.388	-0.037	-0.009	0.003
13	0.713	-0.003	0.579	-0.001	0.878	-0.005	-0.001	0.015
14	-1.119	0.308	-0.909	0.155	-1.377	0.610	0.155	0.038
15	0.399	-0.332	0.324	-0.167	0.492	-0.658	-0.167	0.006
16	0.715	0.007	0.581	0.003	0.880	0.013	0.003	0.015
17	0.358	0.785	0.291	0.396	0.441	1.556	0.396	0.011
18	0.237	0.785	0.193	0.122	0.292	0.479	0.122	0.002
19	0.621	-1.742	0.505	-0.879	0.765	-3.454	-0.879	0.045
20	0.207	-0.718	0.168	-0.362	0.255	-1.424	-0.362	0.007
21	-0.820	0.053	-0.666	0.027	-1.010	0.105	0.027	0.021
22	0.045	-0.002	0.037	-0.001	0.056	-0.003	-0.001	0.000
23	-0.347	-0.166	-0.282	-0.084	-0.427	-0.329	-0.084	0.004

Table 14: Coordinates for row profiles for normalizations under Euclidean distance

	Symmetric		Row principal		Column principal		Principal	
Brand	Scoring in dimension		Scoring in dimension		Scoring in dimension		Scoring in dimension	
	1	2	1	2	1	2	1	2
1	-1.218	0.051	-1.499	0.102	-0.989	0.026	-0.989	0.026
2	-0.490	0.167	-0.604	0.331	-0.398	0.084	-0.398	0.084
3	1.330	0.327	1.637	0.648	1.080	0.165	1.080	0.165
4	0.357	0.492	0.440	0.975	0.290	0.248	0.290	0.248
5	-0.401	0.034	-0.494	0.067	-0.326	0.017	-0.326	0.017
6	0.422	-1.071	0.520	-2.123	0.343	-0.540	0.343	-0.540

Table 15: Coordinates for column profiles for normalizations under Euclidean distance

B Matlab code for plotting row coordinates from excel

```
* Plotting of row coordinates from excel
* Starting position from the second column
* Selecting the two following columns for the first sample
* and the next next two columns from the second sample until the 101 sample

* The 101 sample is the original data which has been used for bootstrap.
* The samples are plotted, the bootstrap data are marked blue
* and the original data is marked red.

NUMERIC=xlsread('L:\Data.xls');

hold on
for i=1:23. Y=[];
for k=1:101.
  y=NUMERIC(:.[2+(k-1)*2.3+(k-1)*2]);
  Y=[Y;y(i.:)];
end

for k=1:101.
plot(Y(k.1).Y(k.2).'bo')
if k==101,
  pp=plot(Y(k,1),Y(k,2),'r.');
  set(pp,'MarkerSize',20)
end
  text(Y(k.1).Y(k.2).num2str(i))
end

[K.A] = convhull(Y(:.1).Y(:.2));
Ks=size(K);
for j=2:Ks
  plot([Y(K(j-1).1).Y(K(j).1)]'.[Y(K(j-1).2).Y(K(j).2)]'.'-')
end

disp(['Arean: ' num2str(A)])

end
```

C Matlab code for plotting column coordinates from excel

```
* Plotting of column coordinates from excel
* Starting position from the second column
* Selecting the two following columns for the first sample
* and the next next two columns from the second sample until the 101 sample

* The 101 sample is the original data which has been used for bootstrap.
* The samples are plotted, the bootstrap data are marked green
* and the original data is marked red.

NUMERIC2=xlsread(’L:\CA\Data4\Data\Euclid\Principal_brands.xls’);

hold on
for i=1:6. Y=[];
for k=1:101.
  y=NUMERIC2(:.[2+(k-1)*2.3+(k-1)*2]);
  Y=[Y;y(i.:)];
end

for k=1:101.
plot(Y(k.1).Y(k.2).’go’)
if k==101
  plot(Y(k,1),Y(k,2),
  set(pp,’MarkerSize’,20,’ro’)
end
text(Y(k.1).Y(k.2).num2str(i))
end

[K.A] = convhull(Y(:.1).Y(:.2));

Ks=size(K);
for j=2:Ks
  plot([Y(K(j-1).1).Y(K(j).1)]’.[Y(K(j-1).2).Y(K(j).2)]’,’-’)
end

disp([’Arean: ’ num2str(A)])

end
```

D Correspondence analysis for SPSS

```
CORRESPONDENCE

/TABLE=ALL(23,6)
/MEASURE=EUCLID
/DIMENSION=5
/NORMALIZATION=SYMMETRIC
/PRINT = TABLE RPOINTS CPOINTS RPROFILES CPROFILES RCONF CCONF
/PLOT = NDIM(1,MAX) BIPLOT(20) RPOINTS(20) CPOINTS(20) TRROWS(20)
TRCOLUMNS (20).


*The different distance measures are
* EUCLID = Euclidean
* CHISQ = Chi square.


*The different normalization options are
*SYMMETRIC = Symmetric normalization
*RPRINCIPAL = Row principal normalization
*CPRINCIPAL = Column principal normalization
*PRINCIPAL = Principal normalization.
```

References

[1] Abdelhak M.Z. and Iskander D.R., (1998). *Bootstrap Matlab Toolbox*, Communications and Information Processing Group Cooperative Research Centre for Satellite Systems School of Electrical & Electronic System Engineering Queensland University of Technology, Brisbane, Australia.

[2] Benzécri J.F. (1969). Statistical analysis as a tool to make patterns emerge from data. *In Methodologies of Pattern Recognition*, S Watanabe, ed. New York, Academic press.

[3] Eckart C.H., and Young G. (1936). The approximation of one matrix by another one of lower rank. *Psychometrica*. 211-218

[4] Efron B. (1979). Bootstrap methods - another look at the jacknife. *Ann. Statist.* 1-26

[5] Greenacre M.J. (1984). *Theory and Applications of Correspondence Analysis,* Academic Pr.

[6] Hair J.F., Black B., Babin B., Anderson R.E. and Tatham R.L. (2005). *Multivariate Data Analysis*. 6th edition. Prentice Hall. New Jersey.

[7] SPSS. *PASW Statistics 18 Algorithms*. SPSS Inc

[8] SPSS. *SPSS Categories 18*. SPSS Inc

[9] Stata, *Stata 11 help for ca*, http://www.stata.com/help.cgi?ca