BOOTSTRAPPING CORRESPONDENCE ANALYSIS

Meral Yay-Elif Özge Özdamar

Abstract

Correspondence analysis, also known as homogeneity analysis is a statistical visualization method for graphically representing of two way contingency tables. Correspondence analysis is a method for exploring associations between sets of categorical variables and is one of the ordination techniques aimed at reducing data sets into a manageable number of variables. Relationships between categorical data can be analyzed with correspondence analysis and the results are obtained graphically in a two-dimensional space. Conversely to advantages of correspondence analysis, impact of any changes in the data cannot be determined by this analysis. At this point, bootstrapping correspondence analysis gives additional information by exhibiting ellipses of row and column profiles of the contingency table. This study applies bootstrap ellipses for graphical visualization of correspondence analysis on market research data of a baby food brand. The results show that with the bootstrapped ellipses, it is convenient to merge categories of the contingency table, resulting in dimension reduction. This result cannot be obtained with regular correspondence analysis.

Key words: Correspondence Analysis, Bootstrapping, Confidence Ellipses

JEL Code: C15, C38

Introduction

Today, most of scientific researches rely on consultancy of statistical methods. Data obtained during scientific researches should be analyzed with the appropriate statistical methods. Examination of relationships between variables is one of the frequent problematic in researches. Structures of examined variables lead researchers to select the convenient methods for the analysis processes. Nowadays, it is essential to work with large number of observations and variables in most research areas, and with the help of the computer power, it can be claimed that yesterday’s large dataset problem has turned into an advantage. However, distribution characteristics and measurement types of the variables make it obligatory to work with complex datasets, whereas restricting the applicable methods. Thus, structure and measurement type of the variables used for the analysis play an important role in selection of the method.

Complex data sets are examined with multivariate analysis. However, visualization aspects of these analyses are inadequate. On the other hand, correspondence analysis (CA) examines multivariate relationships with graphically tools. Introduced by Benzecri in 1973 and followed by Greenacre in 1984, CA is relatively free from assumptions about the nature of the data. CA can work with counts (frequencies) and does not require data that appropriate to a normal distribution (Greenacre, 1984). The main assumption of correspondence analysis is that all of the relevant variables are included in the analysis (Hair et al., 1995).

1 Correspondence Analysis

CA is a statistical technique preferred in studies and researches on categorical data. It combines cross-tabular data in the form of frequencies called contingency tables with mathematical and graphical techniques in order to explicit an apparent understanding of the categories of observation units.

CA explains the relationship between two categorical variables via a two dimensional contingency table constructed with frequency values of subgroups of the categories in the cells. It aims to visualize and interpret the relationship between variables in two-dimensional space, enabling to observe the categories of the units in the cells in an easily understandable way, with no significant loss of information. Visualization of the relationship between row and column categories of the contingency table determinates the factors allowing to plot the table and enables dimension reduction. Contingency tables contain the frequencies of the variable categories of the dataset.

Many statistical methods have been developed in order to analyze relationships between variables designed as contingency tables, such as Fisher's exact test, Chi-square analysis, G statistics, and Log-linear models. Among those, Chi-square analysis has a frequent usage.

Analyses on cross-tabular data obligate the analysts to merge cell frequencies in the situations where there is no observed frequency or very small amount of frequency in the cells. Merging the categories causes loss of information in the process of data summarizing, therefore while visually summarizing the relationships between variables, alternative methods have been developed in order to eliminate this cost. An example of a two-dimensional rxc contingency table is given below, where r refers to rows and, c to columns.

Tab. 1: Two way contingency table

	Variable 2					
Variable 1	Cat 1	Cat 2	Cat 3	...	Cat c	Row ToTal
Cat 1	x_{11}	x_{12}	x_{13}	...	x_{1c}	$x_{1.}$
Cat 2	x_{21}	x_{22}	x_{23}	...	x_{2c}	$x_{2.}$
Cat 3	x_{31}	x_{32}	x_{33}	...	x_{3c}	$x_{3.}$
…	...	...	...	…	...	...
Cat r	x_{r1}	x_{r2}	x_{r3}	…	x_{rc}	$x_{r.}$
Column Total	$x_{.1}$	$x_{.2}$	$x_{.3}$	…	$x_{.c}$	n

In the rxc contingency table, frequencies belonging to the categories of two variables are shown. Here; $i\ (i=1,2,3,...,r)$ stands for row indices, $j\ (j=1,2,3,.....,c)$ for the columns, $x_{r.}$ for sum of rows, $x_{.c}$ for sum of columns, and x_{ij} for the frequency value in the cell of i .th row and j .th column. In order to interpret reliable and understandable results from contingency tables, both rows and columns of the table should be considered at the same time. For this purpose, the frequency values in rows and columns are proportioned to the total of row and column frequencies respectively, in order to add up to 1. These relative frequencies are named as profiles. Column and row profiles of the contingency table ensure fundamental information for CA (3). For example row profile of the first cell of row “i” is calculated as $x_{i1}/x_{i.}$ and, column profile of the first cell of column “j” is as $x_{1j}/x_{.j}$. In this way, profiles for each cell of the contingency table are achieved. Hence, average row and column profiles can be obtained. Average row profile is the division of sum of each column frequency to grand total. Average row profile for c columns is calculated as $\left(\frac{x_{.1}}{n},\frac{x_{.2}}{n},.....,\frac{x_{.c}}{n}\right)$ and likewise average row profiles for r rows as $\left(\frac{x_{1.}}{n},\frac{x_{2.}}{n},.....,\frac{x_{r.}}{n}\right)$. Average row and column profiles are thought as center points, and the location of each profile in the contingency table is valued to average row and column profiles. The distance of the each profile points to average profiles is evaluated by proximity or length. For instance, if the distance between a profile and the average profile is larger than the other distances, the point is said to be distant from the origin, and vice versa.

As an exploratory statistical analysis, the purpose of CA is to identify a sample space with points determining row and column profiles. At this point, equalizing the distance between the points become crucial. Hence, each distance is defined as ratios called "mass" calculated by weighing each point. Mass is a measure of the effect of a frequency in the contingency table to its marginal frequency, and obtained by dividing row and column sums to overall sum. The matrix of mass values is also called correspondence matrix (P). Expected values of the correspondence matrix are expressed as row and column probabilities and shown as rc'. The difference between each value in the matrix P and rc' are calculated and estimation residuals are obtained. Residuals are divided by square root of rc', and standardized values are obtained as follows:

$$s_{ij} = \left(p_{ij} - r_j c_j\right) / \sqrt{r_j c_j} \tag{1}$$

CA is based on the chi-squared distances. The chi-squared distance, also named as weighted Euclidean distance, is the weighted distance between the row and column profiles. Based on these distances, the information exhibited by row and column profiles are perceptible. The Euclidean distance between the i.th row profile and the average row profile is obtained follows: (Clausen, 1998, p.11):

$$s(i,i') = \sqrt{\sum_{j=1}^{c}\left(\frac{x_{ij}}{x_{r.}} - \frac{x_{.c}}{n}\right)^2} \tag{2}$$

Chi-square distance uses inverse of average profile values as weights. Chi-square distance, between i.th row profile and the average row profile, taking into account the frequencies, is calculated as follows:

$$d(i,i') = \sqrt{\sum_{j=1}^{c}\frac{\left(\frac{x_{ij}}{x_{r.}} - \frac{x_{.c}}{n}\right)^2}{\frac{x_{.c}}{n}}} \tag{3}$$

While calculating the chi-square distance, only one variable's profiles are considered. By means of the achieved chi-square distance, inertia value is obtained as follows.

$$\Lambda^2 = \frac{\chi^2}{n} \tag{4}$$

The inertia is an indicator of how much of the variation in the original data is "preserved" in the dimensional solution (Bendixen, 1996). When the inertia is low, the row profiles are not scattered very much and lie close to their average profile. The higher the total inertia, the greater is the association between the rows and columns, displayed by the higher dispersion of the profile points in the profile space. If the inertia value is close to zero, there is no association between the row and column average profiles (Greenacre and Blasıus, 1994). To make an interpretation of the inertia value on the graph, it can be said that if the points are close to each other, there is no association, and distant points indicate strong association.

2 Bootstrapping Correspondence Anaylsis

CA, also known as a homogeny analysis, is commonly used as a virtualization tool for cross-tabular data. On the other hand, it is possible to determine the location of points from the graph of CA, but conversely stability of the points cannot be obtained. In other words, CA does not help to identify the impact of any changes that may occur in the data for each point. At this point, using bootstrapping, stability and bias effects can easily be observed.

Let us name the rxc contingency table given above as matrix X and get ‘ B ’ bootstrap samples, with each bootstrapped matrices with the same dimension as the contingency table. As bootstrap samples are generated with replacement, it is possible to say that any cell value of the contingency table may occur in more than one bootstrapped sample, or in none of them. Following, CA is applied B times, on every bootstrap sample, and visualizations on bootstrapped values are gathered. Thus, the plot constructed with the bootstrapped row and column profiles, stability of the dataset can be directly observed. Stability now can be acquired as the difference between the original CA coordinates and the bootstrapped points, with the help of mean square errors:

$$MSE_i = \frac{1}{B}\sum_{b=1}^{B}\left(f_{ib} - \hat{f}_i\right)'\left(f_{ib} - \hat{f}_i\right) \tag{5}$$

f_{ib}, is the principal row of CA on the $b.$th bootstrap sample and, $\hat{f}_i$ is the principal row coordinate of CA on original contingency table. Same calculations are accomplished for the columns and,

$$MSE_i = \frac{1}{B}\sum_{b=1}^{B}\left(g_{jb} - \hat{g}_j\right)'\left(g_{jb} - \hat{g}_j\right) \tag{6}$$

equation is achieved. Here, g_{jb}, is the principal column coordinate of CA on the $b.$th bootstrap sample, and ; $\hat{g}_j$ is the principal column coordinate of CA on original contingency table. In order to compare the stability of different solutions, instead of mean square error, relative mean square would be more accurate to use. For this, total sum of squares is needed and obtained for rows and columns as follows:

$$TSS_{rows} = \sum_{i=1}^{n} \hat{f}_i' \hat{f}_i \quad and \quad TSS_{columns} = \sum_{j=1}^{n} \hat{g}_j' \hat{g}_j \qquad (7), (8)$$

Relative mean square errors for rows and columns are obtained via total sum of squares as follows respectively:

$$RMSE_{rows} = \frac{\sum_{i=1}^{n} MSE_i}{TSS_{rows}} \quad and \quad RMSE_{columns} = \frac{\sum_{i=1}^{n} MSE_j}{TSS_{columns}} \qquad (9),(10)$$

Total sum of squares for all observations is the average of total sum of squares of rows and columns:

$$RMSE = \frac{1}{2}\left(RMSE_{rows} + RMSE_{columns}\right) \qquad (11)$$

Hence, applying CA on B set of bootstrap samples, as the same size and obtained from the X dataset matrix, bootstrap ellipses are constructed.

It is important to clearly indicate which bootstrap ellipses belong to which category in the contingency table. Hence, plotting all bootstrap points is not a feasible choice and it is more insightful to display (1-α) % confidence ellipses around the bootstrap means. These ellipses are constructed in such a way that for each category, the ellipse contains exactly (1-α) % of the corresponding bootstrap points. Using confidence ellipses, the relative positions of the categories are clearly represented, and the sizes and shapes of the ellipses nicely visualize stability and dependencies among the categories (Van de Velden et al.,2012).

3 Application

The aim of CA is to explain the association between row and column categories of the contingency table with fewer categories by visualization. This study examines the dataset of 742 observations belonging to a new product of baby food in terms of “consumer” and “purchase” types. Frequencies are obtained by purchase decision of the baby food in respect to its color, smell, consistency and smoothness. The consumer variable has five categories; working mother (1), non-working mother (2), care-taker (3), grandmother (4), and pregnant (5). The purchase variable has also five categories; “I don’t purchase absolutely” (1), “I don’t purchase" (2), "Unstable" (3), "I purchase" (4) and "I purchase absolutely" (5) respectively. 5x5 contingency table of the category frequencies is given in Table 2.

Tab. 2: Contingency table of baby food

Consumer	Purchase					
	I don't purchase absolutely	I don't purchase	unstable	I purchase	I purchase absolutely	Active Margin
working mother	50	23	15	1	1	90
non-working mother	20	29	18	4	5	76
caretaker	15	10	49	26	13	113
grandmother	18	24	40	74	81	237
pregnant	23	15	45	60	83	226
Active Margin	126	101	167	165	183	742

In order to analyze purchase decisions about baby food of consumers simple CA is applied. The profiles of rows and columns required for the CA are given in Table 3. and Table 4. respectively.

Tab. 3: Row profiles

Consumer	Purchase					
	I don't purchase absolutely	I don't purchase	unstable	I purchase	I purchase absolutely	Active Margin
working mother	,556	,256	,167	,011	,011	1,000
non-working mother	,263	,382	,237	,053	,066	1,000
caretaker	,133	,088	,434	,230	,115	1,000
grandmother	,076	,101	,169	,312	,342	1,000
pregnant	,102	,066	,199	,265	,367	1,000
Active Margin	,170	,136	,225	,222	,247	

As mentioned before, if the distance between a profile to the average profile is larger than the other distances, the point is said to be distant from the origin. When the row profiles are examined, we can claim that the "pregnant" category of the consumer variable is close to the average row profile.

Tab. 4: Column profiles

Consumer	Purchase					
	I don't purchase absolutely	I don't purchase	unstable	I purchase	I purchase absolutely	Active Margin
working mother	,397	,228	,090	,006	,005	,121
non-working mother	,159	,287	,108	,024	,027	,102
caretaker	,119	,099	,293	,158	,071	,152
grandmother	,143	,238	,240	,448	,443	,319
pregnant	,183	,149	,269	,364	,454	,305
Active Margin	1,000	1,000	1,000	1,000	1,000	

When the column profiles are examined, we can claim that “I purchase” category of the purchase variable is close to the average column profile. When the row and column profiles are examined at the same time, we can assert that “pregnant” and “I purchase” categories are close to origin as shown in Figure 1. Graphical inference of CA is interpreted according to location of the points to the origin. In the case of a point is close to another, we can assert that the points are positively correlated. Otherwise, they are negatively correlated. The farthest the point is, the strongest the relationship.

Fig. 1: Biplot of correspondence analysis

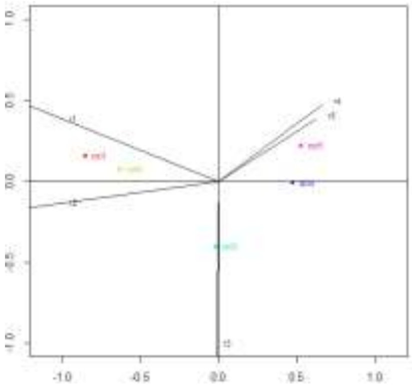


Figure 1 shows that rows and columns are clustered in three different regions. Taking account of the location of the categories, we can interpret that 5 categories of the variables in the contingency table can be represented as 3 categories.

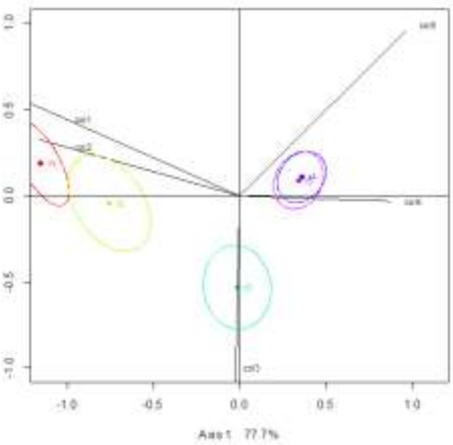
As described previously, inertia, calculated using the chi-square, is a measure of stability. The chi-square value is calculated as $\chi^2 = 283.266$ based on the contingency table given in Table1. Inertia value is computed as $\Lambda = \frac{\chi^2}{n} = \frac{283.266}{742} = 0.382$. The inertia value indicates that the row and column profiles are not far from the average profiles. Inertia values of categories are given in Table 5.

Tab. 5: Inertia values of the variables

Purchase	Inertia	Consumer	Inertia
I don’t purchase absolutely	0,138	Working Mother	0,172
I don’t purchase	0,074	Non-working Mother	0,078
Unstable	0,037	Caretaker	0,044
Purchase	0,052	Grandmother	0,047
Purchase Absolutely	0,081	Pregnant	0,041
Total	0,382	Total	0,382

After application of CA, 1000 bootstrap sample matrixes are simulated as the same size of contingency table. These bootstrap values are used in calculation of points in the graph, where named as bootstrap ellipses. The bootstrap ellipses are constituted for row and column profiles of the contingency table and shown in Figure 2 and Figure 3 respectively.

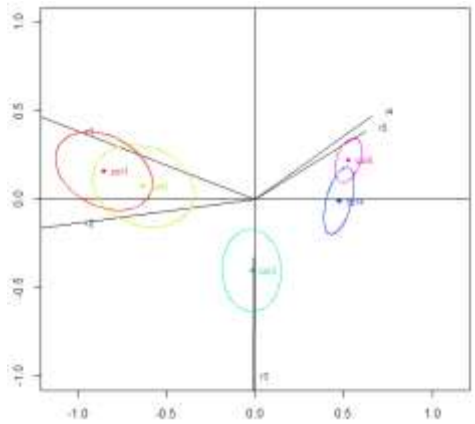
Fig. 2: Confidence regions for rows, 1000 resamples



The graph of bootstrapped CA exhibits that; categories of the variables can be visualized as ellipses, not as points. Consequently, we can claim that areas of the ellipses gain importance instead of points in bootstrapped CA. Bootstrapped CA can generate more successful ellipses with bigger data sets (Ringrose, T., 2011).

There is a very large overlap between ellipses of consumer variables category four and five, and rather less between category one and two. Category three has a relatively large severity from the other categories. Another conclusion that can be made from the Figure 2 is that, the overlapping ellipses, indicating row profiles four and five, can be represented as one category. Similar interpretation can be made for the column profiles as shown in Figure 3.

Fig. 3: Confidence regions for columns, 1000 resamples



There is a very large overlap between ellipses of purchase variables category one and two, and rather less between category four and five. Column profiles one and two can be represented as one category.

4 Conclusion

CA, an analysis to visualize data in contingency tables with a lot of categories, is applied on a new released baby food in terms of marketing research. The contingency tabbe is composed of two variables, both with five categories, named as “consumer” and “purchase” respectively. The biplot generated by CA suggests that the contingency table may be a 3x3 matrix, in the meaning of dimension reduction. The bootstrapped CA also supports this inference, also adding important informance to the analysis. With the help of bootstrap ellipses, it comes to light that row 4 and 5, “pregnant” and “grandmother” categories, in the contingency table may be added together. The same situation is also applicable for the column profiles. Column one and two, “I don’t purchase absolutely” and “I’don’t purchase” categories may be added for dimension reduction.

References

Andersen, E. B. (1990). *The statistical analysis of categorical data*. Berlin: Springer-Verlag.

Bendixen, M. (1996). A practical guide to use of correspondence analysis in marketing research. Marketing Research On-line, 16-38. Retreiwed from http://mro.massey.ac.nz/ca.html

Clausen, S. (1998). *Applied correspondence analysis an introduction*. Thousand Oaks, CA: Sage Publications.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Greenacre, M. J.,& Blasius, J. (1994). *Correspondence analysis in the social sciences*. London: Academic Press.

Hair, J. F., Anderson, R.E., Tatham, R.L., & Black,W.C. (1995). *Multivariate data analysis* (4th ed.). Upper Saddle River, N.J.: Prentice Hall.

Murtagh, F. (2005). *Correspondence analysis and data coding with Java and R*. Boca Raton, FL: Chapman & Hall/CRC.

Ringrose, T. J. Bootstrap confidence regions for correspondence analysis. *Journal of Statistical Computation and Simulation*, 1-17.

Van de Velden, M., Beuckelaer, A., Groenen, P.J.F,& Busing, F.M.T.A. (2012). Nonmetric unfolding of marketing data: degeneracy and stability

Contact

Meral Yay
Mimar Sinan Fine Arts University
Cumhuriyet Mah. Silahşör Cad. No:89 Bomonti, Istanbul, Turkey
Meral.yay@msgsu.edu.tr
Elif Özge Özdamar
Mimar Sinan Fine Arts University
Cumhuriyet Mah. Silahşör Cad. No:89 Bomonti, Istanbul, Turkey
Ozge.ozdamar@msgsu.edu.tr