The Use of Bertin Graphs in Product Research

Assist. Prof. Dr. Serkan GÜNEŞ
Department of Industrial Design,
Gazi University
Kirim Cad. 6 Sok. Emek Ankara
TURKEY

Abstract

Many design research uses several comparative methods to compare and contrast products or brand images among user attitudes, purchasing decisions, ergonomics, functionality and other design related issues. With the development of computer technology, design researchers became capable to several statistical software packages. These software packages provide statistical graphs and charts varying from simple pie charts through advanced plots. However, each type of graph has its advantages and disadvantages according their design. This paper introduces the use of Bertin Graphs as a tool to present more legible outputs in comparison with conventional Correspondence Analysis plots in product research. For this, the paper reviews the relevant use of Correspondence Analysis in product research and compares two output alternatives with examples.

Keywords: Design Research, Correspondence Analysis, Bertin Graphs.

1. Introduction

Well-designed graphs make it easier to identify trends and relationships among variables (Tufte, 1997) and well-chosen graphics can effectively communicate a large amount of information efficiently (Larkin & Simon, 1987). Recent developments in computer technologies offer sophisticated tools for producing statistical graphics in an easier and simpler manner. However, many choices on variety of forms, style and alternatives can influence final interpretation (Zacks et al., 1998). On this account, graphs should be drawn in a manner that leads viewers to reach conclusions consistent with those that they would reach when analyzing the underlying numeric data upon which the graphs are based (Arunachalam, 2002). Many researches mostly concentrate on data collection and analyze method, but limited numbers of them pay attention for the final outputs of their valuable data or remain blind to their importance. On the contrary, yet the researches communicate with their graphs, a research should be planned and built around the representation.

2. Product research

The relationship between emotions and products became an important sphere of interest after rational choice framework enclosed non-rational elements (Hirschman & Hollbrook 1982) such as traditional or habitual action, emotional or affectual action and various forms of value-oriented action. Nevermore, there is still wide divergence in the content of emotions studied in consumer research (Laros & Steenkamp 2003). Proposed emotion sets in literature are mostly in bipolar characteristics such as pleasant vs. unpleasant (Desmet et. al 2005), negative vs. positive (Richins 1997 ; Laros & Steenkamp 2003) or in linear combination of two independent systems such as valence dimension (unpleasant-pleasant) and arousal dimension (activation – deactivation (Posner et. al 2005). However, many studies eliminate the number of emotions by cluster analyzes whether hierarchical or not according similarity or the relevancy to the case or for the sake of calculability in analysis or legibility in the final Correspondence Analysis outputs. Despite their theoretical and practical implications, studies are still weak in representation of their valuable plots.

CA in product research is used to perform portfolio analyses. It mostly aid to monitor overall competitive market structure or niches to fill, defining and prioritizing design concepts and keeping track of design decisions according to targeted strategies. With the help of emotion sets, researchers observe consumers' needs and desirable product features or growth/niche opportunities in market. These analyses also help designers and marketers to optimize product strategies; to define target markets; to design consumer-focused products to maximize profits. Most of the product researches depend on the variety of attributes of respondents. In this methods, the brands are arranged in rows and respondents are asked to rate each brand according to their attributes, perceptions or the emotions that are triggered.

Rating scales are one of the most widely used tools in consumer research and product research (Dawes, 2008). When responding to a survey item, respondents specify their level of agreement to a statement by Likert scales, semantic differential scales or other visual scales that best fit to their attitudes. In this way, a set of categorized data is formed. The obtained data is interpreted in the context of conceptual space. In this space or perceptual maps, similarities and differences between among brands are observed with respect to the attitudes. Simple relationships among brands and attributes can be observed by contingency tables however more complex and advanced relationships among items need visual methods. To reduce the complexity, numbers of attributes are limited at first or data are reduced as in Correspondence Analysis. To study every aspect of the Correspondence Analysis in depth, it should be better to mention about contingency tables to lay the foundations.

2.1 Contingency tables

The cross-tabulation of categorical data is one of the most common forms of analysis in product research (Hoffman & Franke 1986). In brief, a default contingency table shows relationship between categorical variables located at columns and rows as well as with marginal totals in the right-hand column and the bottom row. The degree of association between variables can be interpreted easily in simple tables however larger ones need analytical procedures for assessment like X^2 (Chi-square). The chi-squared (X^2) statistic measures the discrepancy between the observed frequencies in a contingency table and the expected frequencies calculated under a hypothesis of homogeneity of the row profiles (or of column profiles) (Greenacre, 1993). In graphical representation of contingency tables X^2 distance is used to show the distances between profiles (row and column) in an Euclidian space. The presentation is highly related to the number of columns and rows. If there are n columns (or rows), then perfect representation can be achieved in n-1 dimensions (Bendixen, 1996). However, most contingency tables have many more rows and columns and the profiles lies in a space of much higher dimensionality that are difficult to interpret and even visualize (Greenacre, 1993). In this situation, to obtain low-dimensional subspace, reduction of dimensionality is applied with a certain amount of loss of information as in Correspondence Analysis.

2.2 Correspondence analysis

Correspondence analysis is an exploratory data analysis technique for the graphical display of contingency tables and multivariate categorical data (Hoffman & Franke 1986). Problem of correspondence analysis (CA) is to find an optimal plot of cross tabulation in a lower dimensional space to locate columns and rows are on the same scale. It is chiefly a graphical method of data analysis (Greenacre, 1993). The purpose of correspondence analysis is to reproduce the distances between the row and/or column points in a two-way table in a lower-dimensional display. So, it is better to focus on its plots and their interpretation rather than mathematical procedures and technical details. In order to analyze the subject, an example of a contingency table with 5 columns and 8 rows are given in Table 1. This table has 5 product groups cross-tabulated by 8 variables of emotions with the 600 (n=600) of sample.

Table 1: Example Product Research Table to Demonstrate the CA

	VAR1	VAR2	VAR3	VAR4	VAR5	VAR6	VAR7	VAR8	TOTAL
PRODUCT A	11	8	13	19	8	1	9	35	104
PRODUCT B	21	7	16	18	16	1	22	37	138
PRODUCT C	16	14	18	23	12	3	12	21	119
PRODUCT D	5	22	26	11	13	3	12	44	136
PRODUCT E	27	5	5	18	5	5	19	19	103
TOTAL	80	56	78	89	54	13	74	156	600

Here, the number of emotion are not three-dimensional but in 8 and the perfect presentation can be achieved in n-1 dimension, in other words 7. To create two-dimensional presentation, lower-dimensional projections are applied by identifying the closest to all profile points to project on to such a subspace; however there occur data losses which may be expressed as a percentage of the total inertia.

As a rule, the lower the loss, the higher the quality and the higher the loss the lower the quality (Greenacre, 1993).

In Table 2, the SPSS based comparison of dimension and data loss of the example is introduced. The section labeled "Cumulative Inertia" points out that over 76.5% of the X^2 for association is accounted for by two dimensions with %23.5 data loss. Note that 5 dimensional plots represent 100% of data with no data loss. However due to the impossibility of realization or even imagine of 5 dimensions, through accepting a specific amount of data loss, 2 dimensional plots are preferred.

Table 2: Dimensions and data loss of the example

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation 2
1	,251	,063			,536	,536	,005	,064
2	,164	,027			,229	,765	,005	
3	,129	,017			,141	,906		
4	,079	,006			,054	,960		
5	,069	,005			,040	1,000		
Total	,117	4588,010	,000(a)		1,000	1,000		
a 395 degrees of freedom								

The correspondence plot of the example is shown at below (Fig. 1). This 2 dimensional representation is made up of 71.5% of data from dimension 1 and 19.3% from dimension 2. To have a good grip of the dimension problem, it is better to imagine dimensions as lines which cut through the point cloud as closely as possible to lessen data loss in a multidimensional profile space. This means that, the points are projected perpendicular on this lines or in other words dimensions. However, these projected points are not their true positions. The difference among original and projected position creates data loss. To interpret the plot, first, it is better to summarize the logic behind these types of plots. To put the matter in hand, Figure 2 is introduced for extensive conception.

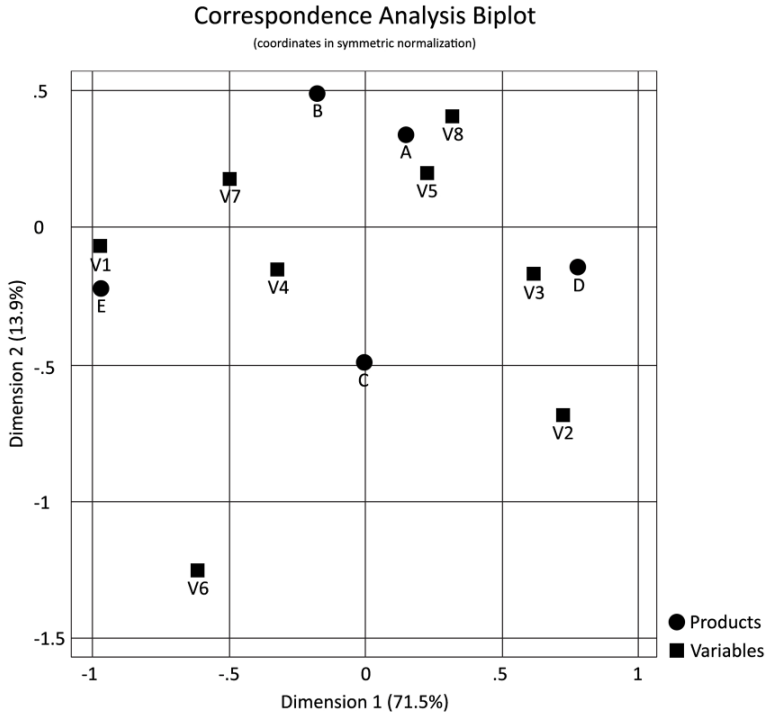

Figure 1: CA Plot of the example

To interpret the plot, first, it is better to summarize the logic behind these types of plots. To put the matter in hand, Figure 2 is introduced for extensive conception.

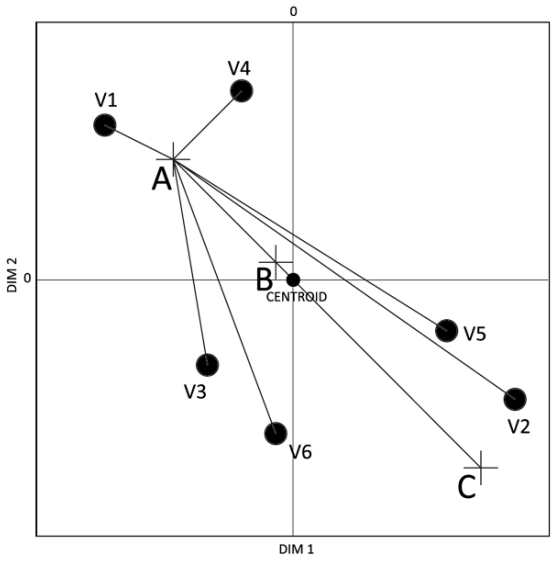

Figure 2: Sample CA plot to demonstrate associations

In Figure 2, the locations of elements are determined by the mass and the shape of distribution. The masses of elements are fading in according their relativity to overall frequency. The map is anchored by the centroid (0, 0). The centroid is the position of the 'middle' of the cloud of points or a center of gravity (but not always at middle in CA plots) where average row profile and the average column profile are located. In interpreting the sample map, column profiles similar to the average profile (showing no strong association to any row) are plotted close to the centroid of the map (as B) and vice versa. Profiles highly dissimilar to the average profile are plotted near the margins of the plot. Points which have anti-correlated profiles are located in opposite directions from the centroid (as A and C).

According to Bertin, resemblance, order and proportional are the three sign fields in graphics (Bertin, 1983). So, A, B, C are distinguishable and B is between A and C. BC is twice as long as AB. Small distances between profiles (as A and V1) suggest high association, while large distances (C and V4) indicate low association.

After this brief explanation, in Figure 1, it can be observed that, to form the plot, the bulk amount of data (71.5 %) is gained from Dimension 1. Thus, the main axis for the interpretation is Dimension 1. C, A and B products are close to one another and the centroid (0, 0). These points will have a relatively similar profile. In other words, the respondents have answered to the variables in similar proportions for C, A and B. However, product C is exactly located on Dimension 1 axis of the centroid that means product C is the closest to the average among others. Despite B and A is located around the centroid, they are at opposite side of Dimension 1 axis of the centroid. So the differentiation classification of these three products will be as B, C, A. Product E and D are far from centroid and located at the margins of plot. Then the overall differentiation classification of all products will be interpreted as E, B, C, A and D. Similarly, when centroid and dominant Dimension 1 axis considered, differentiation classification of variables will be as V1, V6, V7, V4, V5, V8, V3 and V2. The overall interpretation of plot is:

- Most differentiated products are E and D,
- E is differentiated from D mostly by variables V1 and V6 and D is differentiated from E mostly by variables V2 and V3,
- Variables V7, V4, V5 and V8 are relatively common for all products.

3. Bertin graphics

According to Chauchat and Risson, in some CA plots, inaccurate conclusions are possible, yet these CA plots does not depicts the raw data after rows and columns have been permuted on their order on the first CA axis.

To improve the legibility of the CA plots Bertin developed a graphic method. The purpose of the method was permuting the rows and columns of a matrix for revealing hidden structure in data matrix. Bertin's graphics can be seen as a type of scatter plot: coordinates from CA become ranks, and the area of each rectangle is proportional to the number of observations/cases with those ranks (Chauchat and Risson, 1998). In 1977 J. Bertin introduced a display and an analysis strategy for multivariate data with low or medium sample size. He tries to make the information in a data set understandable. He does not fit models: he tries to provide simple tools to interrogate data. The tools operate simultaneously on cases and variables, combining aspects otherwise separately encountered in cluster analysis (on cases) and principal component analysis or factor analysis (on variables) (Bertin, 1983).

In abstract terms, a Bertin matrix is a matrix of displays. Bertin matrices allow rearrangements to transform an initial matrix to a more homogeneous structure. The rearrangements are row or column permutations, and groupings of rows or columns. To fix ideas, think of a data matrix, variable by case, with real valued variables. For each variable, draw a bar chart of variable value by case. Highlight all bars representing a value above some sample threshold for that variable (De Falguerolles, 1996) (Figure 3).

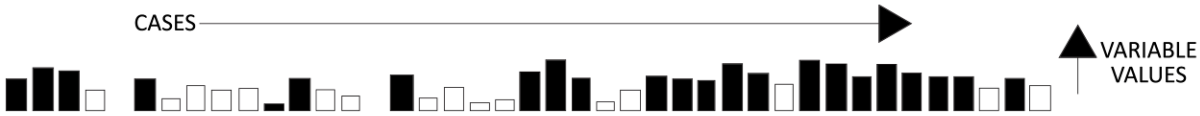

Figure 3: Univariate simple Bertin plot of one quantitative variable

First, the logic of the system is based on simple reordering. Bertin first reordered the matrix tables depending on visual re-classing. By this way reordered matrices become readable by defined characteristic groups with particular situations. Thence, the columns or rows are either rearranged or inversed (Figure 4).

In the re-orderable matrix the elementary areas are equal. In the weighted matrix, x and y vary in a certain quantity. The areas become meaningful; the rows and/or columns are unequal. The weighted matrix must therefore be drawn and can only be applied to tables of limited dimensions (Bertin, 1977). To produce plots Bertin uses four steps for the construction (Figure 5):

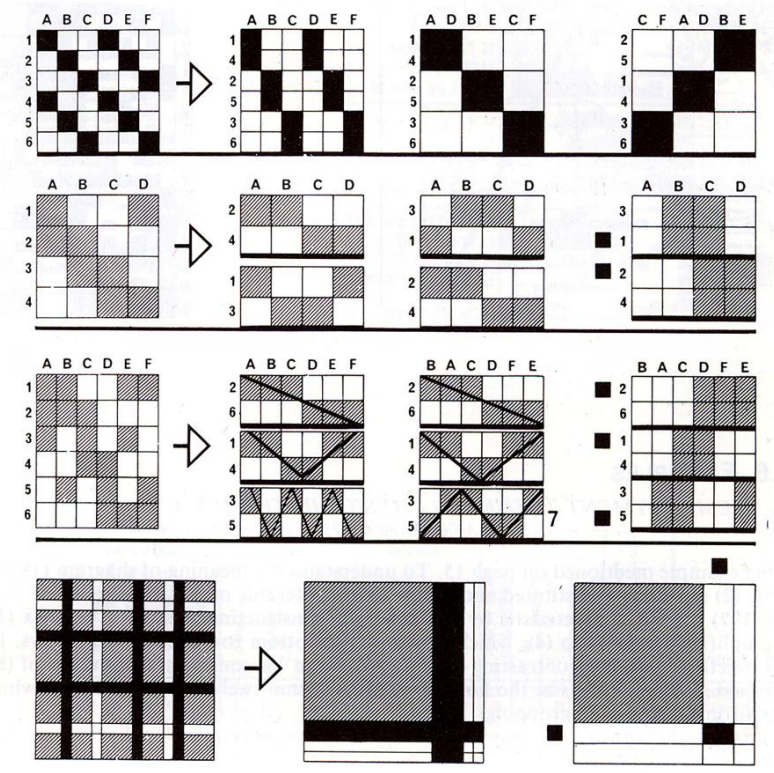

Figure 4: Reordering the matrixes

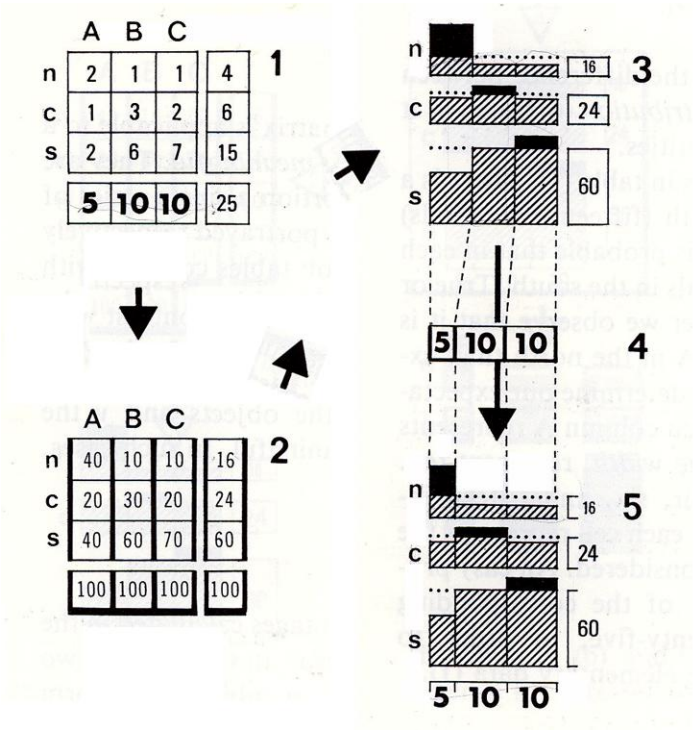

Figure 5: Construction of the Bertin Graph

1- Calculating the vertical percentages of the table,
2- Construction of drawing directly according to these percentages. Darken whatever exceeds the mean per row. Re-classification of rows and columns.
3- Giving the columns a width proportional to the totals obtained from table.
4- In the final drawing; writing the totals per column.

The weighted matrix shows:
- The totals by column profile along X,
- The percentage of each column profile in each row profiles along Y,
- The partial quantities by area,
- And whatever exceeds the mean in black (Y'), that is, whatever characterizes each row data and each column data (Figure 6).

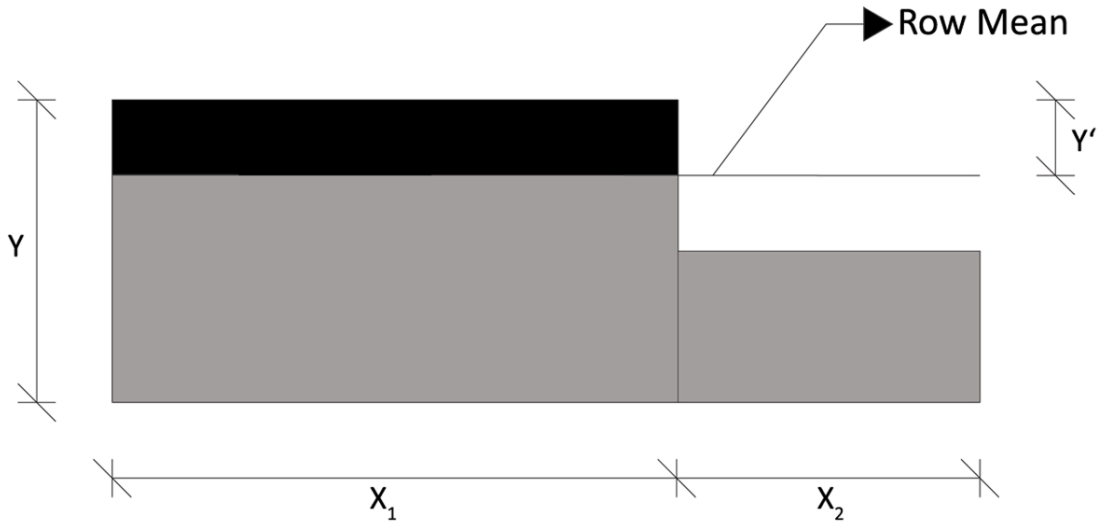

Figure 6: Drawing of Bertin Graph according to percentages and totals

In the light of these explanations the Bertin type graphic steps of the previous example will be as follows:
Step 1: Calculating the vertical percentages of the table (Table 3)

Table 3: Cross tabulation of percentages

	V1	V2	V3	V4	V5	V6	V7	V8	Active Margin
Product A	13,8	14,3	16,7	21,3	14,8	7,7	12,2	22,4	17,3
Product B	26,3	12,5	20,5	20,2	29,6	7,7	29,7	23,7	23,0
Product C	20,0	25,0	23,1	25,8	22,2	23,1	16,2	13,5	19,8
Product D	6,3	39,3	33,3	12,4	24,1	23,1	16,2	28,2	22,7
Product E	33,8	8,9	6,4	20,2	9,3	38,5	25,7	12,2	17,2
Profile Total	100%	100%	100%	100%	100%	100%	100%	100%	100%
Total	80	56	78	89	54	156	78	56	600

Bertin Graph of the example will as follows before re-classification rows and columns (Fig.7).

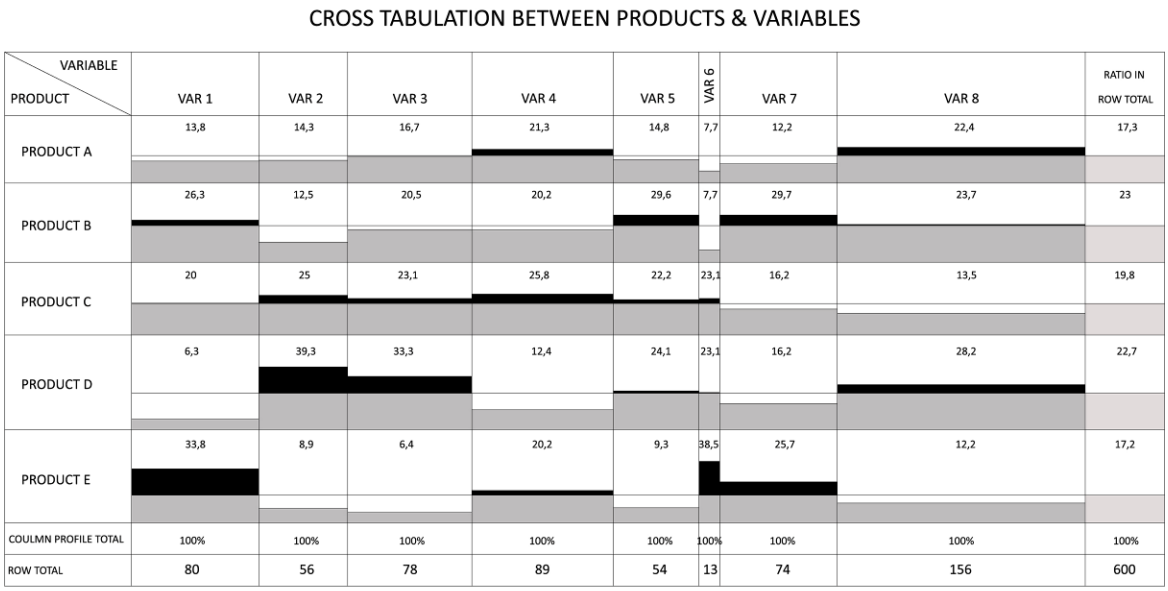

Figure 7: Bertin Graph of the example before reordering

Step 2: Re-classification of rows and columns according to their percentages (Table 4).

Table 4: Cross tabulation of percentages after reordering

	V1	V6	V7	V4	V5	V8	V3	V2	Active Margin
Product E	33,8	38,5	25,7	20,2	9,3	12,2	6,4	8,9	17,2
Product B	26,3	7,7	29,7	20,2	29,6	23,7	20,5	12,5	23,0
Product C	20,0	23,1	16,2	25,8	22,2	13,5	23,1	25,0	19,8
Product A	13,8	7,7	12,2	21,3	14,8	22,4	16,7	14,3	17,3
Product D	6,3	23,1	16,2	12,4	24,1	28,2	33,3	39,3	22,7
Profile Total	100%	100%	100%	100%	100%	100%	100%	100%	100%
Total	80	13	74	89	54	156	78	56	600

The final Bertin Graph of the example is introduced in Figure 8. Note the percentages that exceed the mean per row are darkened.

CROSS TABULATION BETWEEN PRODUCTS & VARIABLES : BERTIN GRAPHICS

VARIABLE / PRODUCT	VAR 1	VAR 6	VAR 7	VAR 4	VAR 5	VAR 8	VAR 3	VAR 2	RATIO IN ROW TOTAL
PRODUCT E	33,8	38,5	25,7	20,2	9,3	12,2	6,4	8,9	17,2
PRODUCT B	26,3	7,7	29,7	20,2	29,6	23,7	20,5	12,5	23,0
PRODUCT C	20,0	23,1	16,2	25,8	22,2	13,5	23,1	25,0	19,8
PRODUCT A	13,8	7,7	12,2	21,3	14,8	22,4	16,7	14,3	17,3
PRODUCT D	6,3	23,1	16,2	12,4	24,1	28,2	33,3	39,3	22,7
COLUMN PROFILE TOTAL	100%	100%	100%	100%	100%	100%	100%	100%	100%
ROW TOTAL	80	13	74	89	54	156	78	56	600

Figure 8: Final Bertin Graph after reordering of rows and columns depending on inter relations

For the graphical legibility the Bertin graphics mostly organizes permutations in diagonal equivalences (from Product E and V1 to Product D and V2). In this way, the graph becomes clearer by the cluster groups. For this, columns are reorganized for simplifying the image. Note that Figure x has similar force of expression according to Figure X without data loss. One advantage of the Bertin graphics is the graphic communication. Graphic communication involves transcribing and telling others what you have discovered. Its aim is to simplify rapid perception and, potentially, memorization of the overall information. Graphic communication poses problems on the level of simplification and selectivity (Bertin, 1977).

There are several superiorities of Bertin Graphs compared to CA plots. First, it frames more legible and homogeneous cluster groups. Second, differentiation in the dimensions of the rectangles gives more information about row and column scores. Third, the darker areas mark more clear and tangible differentiation percentages that exceed the mean per row. Fourth, no advanced calculations are required to find Euclidian distances. And last, Bertin graphs are more legible compared to CA plots when there are more variables in calculation as it introduced in second example. When the variables increase the CA plots become complex for recognition and interpretation. In CA, when the number of variables increases, the legibility of the plot decreases. The points on plots lose their singularity and realized as clusters or point clouds. In this complexity, interpretation gets difficult. Therefore, to define relationships, there is a tendency to cluster the points into imaginary drawn sets. This means that, the relationships in the plot is not points dependent, but cluster dependent. To prevent complexity in final output, many researchers tend to use hierarchical cluster algorithms as Ward at analyze level. In this situation, the plot becomes inadequate to reflect overall picture. Because, the plot will need extra hierarchical trees (Dendograms) and tables to specify the contents of each cluster.

To analyze the complexity of interpretation CA plot, an example of a contingency table with 56 columns (28 bipolar variables as X+ and X-) and 6 rows (Products as A,B,C,D,E,F) are introduced. The STATA plot of the CA is shown in Figure 9.

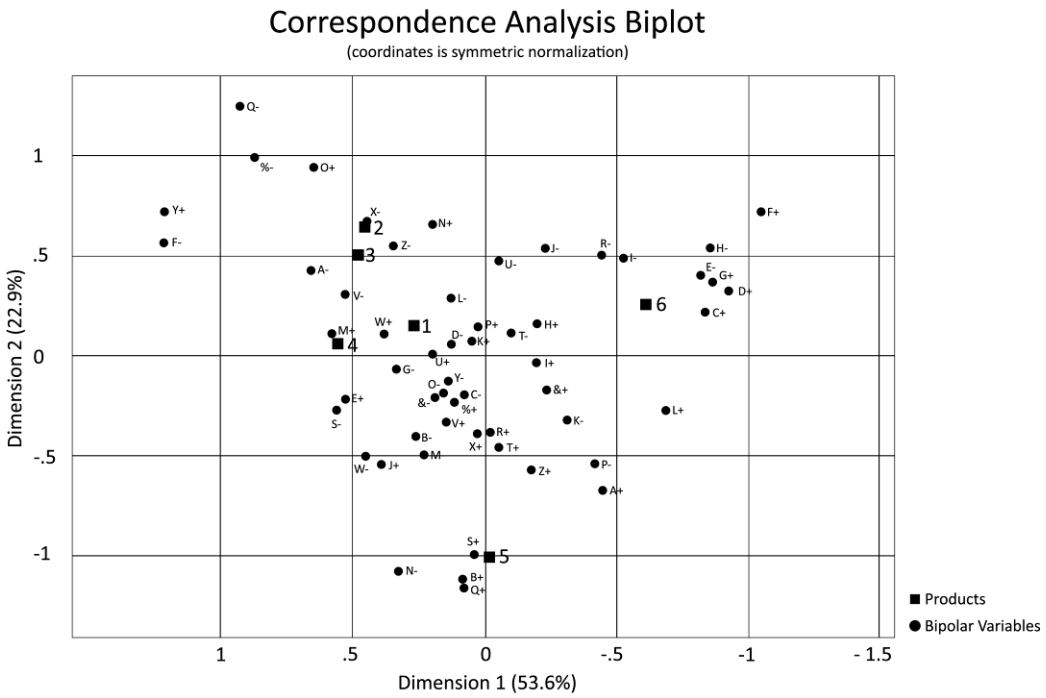

Figure 9: CA Plot of the second example

Above the CA plot (Figure 9) and below Bertin Graph (Figure 10) of the example is shown. Both graphs are produced from same contingency table. Bertin Graph is directly produced from the contingency table so there is no data loss for the occasion.

When two graphs are compared, Bertin Graphs are built in advantages. By rearranging rows and columns, the graph produces associations and clusters naturally. It also clearly defines differentiations among products by diagonally lying exceeded percentages mean per row. On the other hand, CA plot is much more chaotic in comparison with Bertin. With a respectable amount of data loss, the observer should spend a notable time to find out associated points while simultaneously acting in a confused manner in the point cloud.

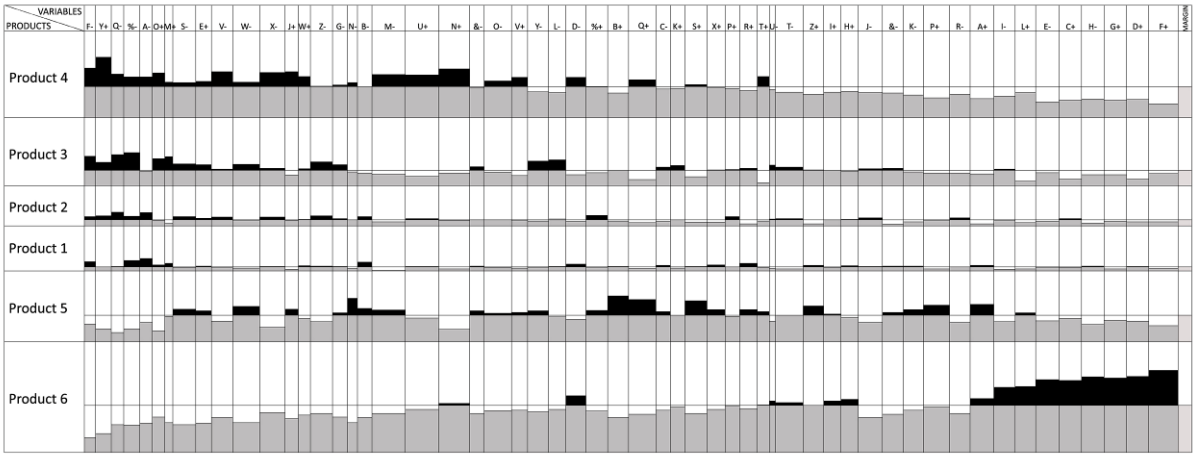

Figure 10: Final Bertin graph of the second example

4. Discussion and conclusion

Although graphical plots of CA ease product positioning interpretation rather than numerical data set, they represent approximate values due to data reduction. Moreover, increased variable numbers causes plots more complicated and crabbed to interpret associations.

CA plots should not be used with data source to verify its findings due to internal procedures of analysis as data reduction. On the other hand, Bertin Graphs acquire data directly from contingency table and naturally create discriminative cluster groups for further interpretations. Data loss is minimized without complicated calculations that CA required.

Nevermore, Bertin Graphs has limits in practice. First, despite the rectangle lengths in X-axis gives valuable information about column profile totals, notable differences among totals create difficulties in legibility because of their proportional length differences among the shortest and the longest as in VAR 6 in Figure 8. Occasionally, they noticed as if a single line where there are huge differences between columns totals. Next, high variable numbers in columns and rows causes graphs to overflow the paper limits. Third statistical software's do not extensively offer Bertin Graphs as a plot alternative and academic interest on subject is still limited.

5. References

Arunachalam, V., Pei, B. & Steinbart, P. J. (2002). Impression management with graphs: Effects on choices, Journal of Information Systems, 16(2), 183-202.

Bendixen, M. (1996). A practical guide to the use of correspondence analysis in marketing research, Marketing Research On-Line, 1, 16-38.

Bertin, J. (1983). Semiology of graphics (William J Berg, Trans.), Wisconsin: University of Wisconsin Press Madison.

Chauchat, J., Risson, A. (1998). Visialization of catagorical data, California: Academic Press.

Dawes, J. (2008). Do data characteristics change according to the number of scale points used?, International Journal of Market Research, 50(1), 61-77.

De Falguerolles, A. (1996), A tribute to J. Bertin's graphical data analysis, [Online] Available: http://www.statlab.uni- heidelberg.de/reports/by.series/beitrag.34.pdf (1996).

Desmet, P. M. A. (2003). Measuring emotions: Development of an instrument to measure emotional responses to products. In M. Blythe, K. Overbeeke, A. Monk & P. Wright (Eds.), Funology: from usability to enjoyment report. Dordrecht: Kluwer Academic.

Greenacre, M. J. (1993). Correspondence analysis in practice, London: Academic Press.

Hirschman, E. C., Holbrook, M. B. (1982). Hedonic consumption: Emerging concepts, methods and propositions, Journal of Marketing, 46, 92–101.

Hoffman, D. L., Franke, G. R. (1986). Correspondence analysis: Graphical representation of categorical data in marketing research, Journal of Marketing Research, 23, 213-227.

Larkin, J. H., Simon, H. (1987). Why a diagram is (sometimes) worth ten thousand words, Cognitive Science, 11, 65-100.

Laros, F. J. M., Steenkamp, J. E. M. (2005). Emotions in consumer behavior: a hierarchical approach, Journal of Business Research, 58, 1437– 1445.

Posner, J., Russell, J.A. & Peterson, B.S. (2005). The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology, Development and Psychopathology, (17), 715–734.

Richins, M. L. (1997). Measuring emotions in the consumption experience, Journal of Consumer Research, 24 (2), 127– 46.

Tufte, E.R. (1983). Visual explanations, Cheshire CT: Graphics Press.

Zacks, J., Levy, E., Tversky, B., & Schiano, D. (1998). Reading bar graphs: Effects of extraneous depth cues and graphical context, Journal of Experimental Psychology: Applied, 4, 119-138.