A Practical Guide to the Use of Correspondence Analysis in Marketing Research

Mike Bendixen

This paper illustrates the application of correspondence analysis in marketing research.

Keywords: Correspondence Analysis

Introduction

The emphasis is on the interpretation of results rather than the technical and mathematical details of the procedure.

Contingency Tables in Research

The cross-tabulation of categorical data is perhaps the most commonly encountered and simple form of analysis in research (Hoffman & Franke1986). Consider, for example, the following contingency table showing the frequency of usage of four common brands of toothpaste in three geographic regions among a random sample of 120 users:

Table 1. Brand by Region Contingency Table

	Region 1	Region 2	Region 3	Total
Brand A	5	5	30	40
Brand B	5	25	5	35
Brand C	15	5	5	25
Brand D	15	5	0	20
Total	40	40	40	120

Interpreting this contingency table is a relative easy task in this simple example. Visual inspection indicates that Brand A is dominant in Region 3, Brand B is dominant in Region 2. Region 1 prefers Brands C and D and Brand D has no support in Region 3 -perhaps it has no distribution network in that region.

Larger contingency tables, i.e. those with more rows and/or columns, can become very complex to interpret and several aids are brought to bear to assist in this process. Examination of row and column profiles allows the researcher to examine the relative position of the columns and rows to each other and thus establish distinguishing characteristics. The row and column profiles of the contingency table in the example are presented in Tables 2 and 3 respectively. A brief examination of these tables confirms that the brand usage pattern established by visual inspection is correct.

Another analytical procedure that can be applied to a contingency table is the Chi-square test of independence. This statistical test is used to determine whether the rows and columns are independent of one another, or phrased differently, whether there is a statistically significant dependence between the rows and columns. In the present example, this would be tantamount to establishing whether brand usage is influenced by region.

It is important to note that this test should only be applied when the expected frequency of any cell is at least 5. Also, while this test may be used to establish dependence, little information is provided as to the nature of the dependence.

Table 2. Row Profile

	Region 1 (%)	Region 2 (%)	Region 3 (%)	Total (%)
Brand A	12.5	12.5	75.0	100.0
Brand B	14.3	71.4	14.3	100.0
Brand C	60.0	20.0	20.0	100.0
Brand D	75.0	25.0	0.0	100.0
Total	33.3	33.3	33.3	100.0

Table 3. Column Profile

Region 1 (%)	Region 2 (%)	Region 3 (%)	Total (%)
12.5	12.5	75.0	33.3
12.5	62.5	12.5	29.2
37.5	12.5	12.5	20.8
37.5	12.5	0.0	16.7
100.0	100.0	100.0	100.0

The Chi-square statistic is calculated as follows:

$$c^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}}$$

where there are r rows and c columns in the contingency table and the observed and expected frequencies of the cell in row i and column j are denoted by fo_{ij} and fe_{ij} respectively. This calculated statistic is compared to the critical value (obtained from statistical tables) with (r-1)(c-1) degrees of freedom.

Applying this formula to the contingency table in the example yields a Chi-square value of 79.07. At 5% significance level with 6 degrees of freedom, the critical Chi-square value is 12.592. The calculated statistic is greater than this value and thus we must conclude that region and brand usage are not independent of each other. The contribution of each cell to the total Chi-square score is presented in Table 4.

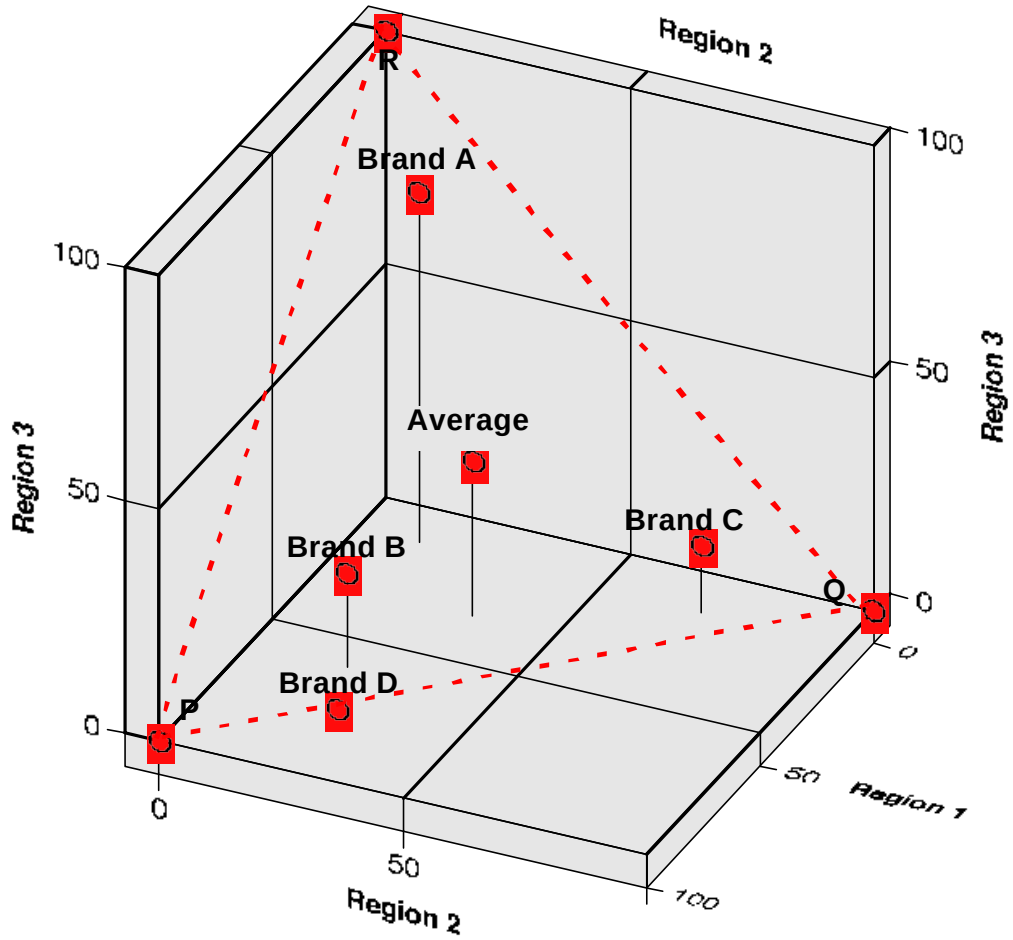
Table 4. Calculation of Chi-square Statistic

	Region 1	Region 2	Region 3	Total
Brand A	5.208	5.208	20.833	31.250
Brand B	3.810	15.238	3.810	22.857
Brand C	5.333	1.333	1.333	8.000
Brand D	10.417	0.417	6.667	17.500
Total	24.768	22.196	32.643	79.607

The relative contributions of each cell to the total Chi-square statistic give some indication of the nature of the dependency between region and brand usage. From Table 3, it can be seen that the cells representing Brand A in Region 3, Brand B in Region 2 and Brand D in Region 1 contribute about 58% to the total Chi-square score and thus account for most of the difference between expected and observed values. This confirms the earlier visual interpretation of the data. As stated earlier, visual interpretation may not be clear in larger contingency tables and the contribution of one cell to the total Chi-square score becomes a useful way of establishing the nature of dependency.

Graphical Representation of a Contingency Table

An alternative means of extracting the nature of the dependency between the rows and columns of the contingency table is to represent the row or column profiles graphically. To illustrate, the row profiles presented in Table 2 may be plotted in three-dimensional space with each of the dimensions representing a different region. Each brand is positioned in this space according to its profile. This is illustrated, together with the plot of the average brand profile in Figure 1.



(N. B. All points lie on the plane of triangle PQR)

Figure 1. 3-Dimensional Representation of Row Profiles

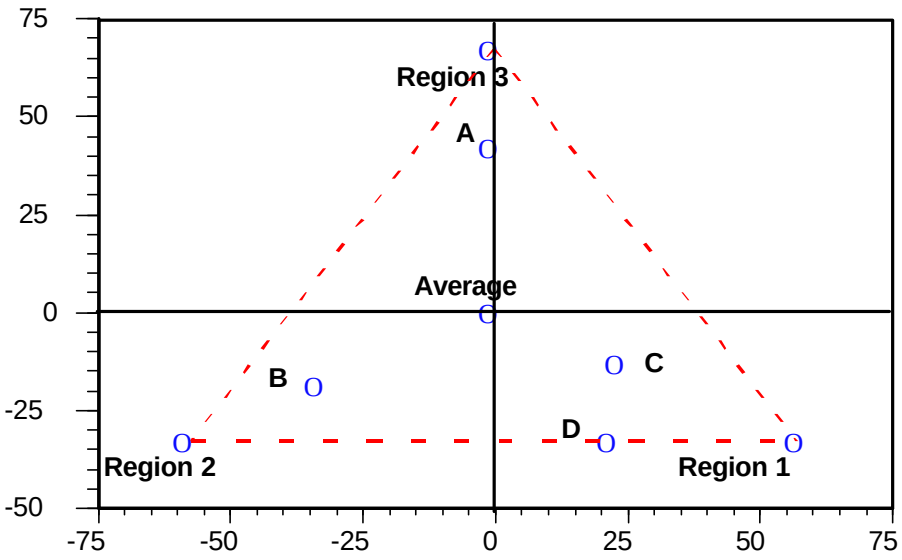


Figure 2. 2-Dimensional Representation of Row Profiles

The complexity of interpreting the position of the brands in three dimensions is immediately apparent. However, as the sum of each row profile is 100%, all of the points plotted lie on a plane in the three-dimensional space. (This plane is bounded by the triangle PQR illustrated in Figure 1.) Hence it is possible, and considerably more convenient, to represent this plot in only two-dimensional space. This simpler illustration, with the average brand profile arbitrarily taken as the zero point on each axis, is presented in Figure 2.

From Figure 2, the proximity of Brand A to the apex representing Region 3 indicates that Brand A is strongly "associated" with Region 3 which is clearly the case from the profile presented in Table 2, i.e. 75% of Brand A users reside in Region 3. Likewise, the proximity of Brand B to the apex representing Region 2 and Brands C and D to the apex representing Region 1 indicates the higher frequency of usage of those brands in those regions. Also, the fact that Brands C and D are positioned relatively closely, indicates a similarity in their regional usage profiles. The fact that Brand A is positioned relatively far away from Brands C and D indicates that Brand A has very different regional usage profile from Brands C and D.

As a matter of terminology, when row profiles are plotted simultaneously with apexes representing the columns, the plot is termed ***asymmetric***.

The Idea of Correspondence Analysis

In the previous section, it was demonstrated that the rows of contingency table can be represented graphically in column space. In the example used, there were 3 columns and perfect representation could be achieved in two dimensions. In can be shown in general that if there are n columns (or rows), then perfect representation can be achieved in n-1 dimensions. It can be seen that perfect graphical representation becomes problematic when there are more than three or four columns (or rows) involved. For instance, if in the previous example we wished to represent columns in row space, we would have to use three dimensions for perfect representation. While this can be done, it is far more difficult to interpret (and even visualise) three-dimensional plots compared to two-dimensional plots. Thus, the graphical procedure demonstrated is only useful for contingency tables that have a maximum of 3 rows or columns.

There is considerable appeal in representing contingency tables graphically in low-dimensional space for easy interpretation of any dependency between rows and columns. This is the idea behind correspondence analysis which analysis allows the optimal representation of a contingency table in low-dimensional space. For instance, it would be necessary to resort to 15-dimensional space for ***perfect*** graphical representation of a 16x16 contingency table, perhaps 75% of the subtlety of the table could be retained in just two dimensions. This represents an enormous gain in simplicity (2 versus 15 dimensions) for an acceptable trade-off in accuracy of representation (75% versus 100%).

The mathematical procedures involved in correspondence analysis are complex - unless you are very familiar with matrix algebra. Interested students are referred to Greenacre (1984) or Hoffman and Franke (1986) for technical details. What is of concern in this paper is the practical application and interpretation of correspondence analysis rather than the mathematical and statistical details.

The asymmetric plot that is produced using correspondence analysis for the previous example is illustrated in Figure 3. Note that besides the change in orientation, this plot is essentially the same as that illustrated in Figure 2.

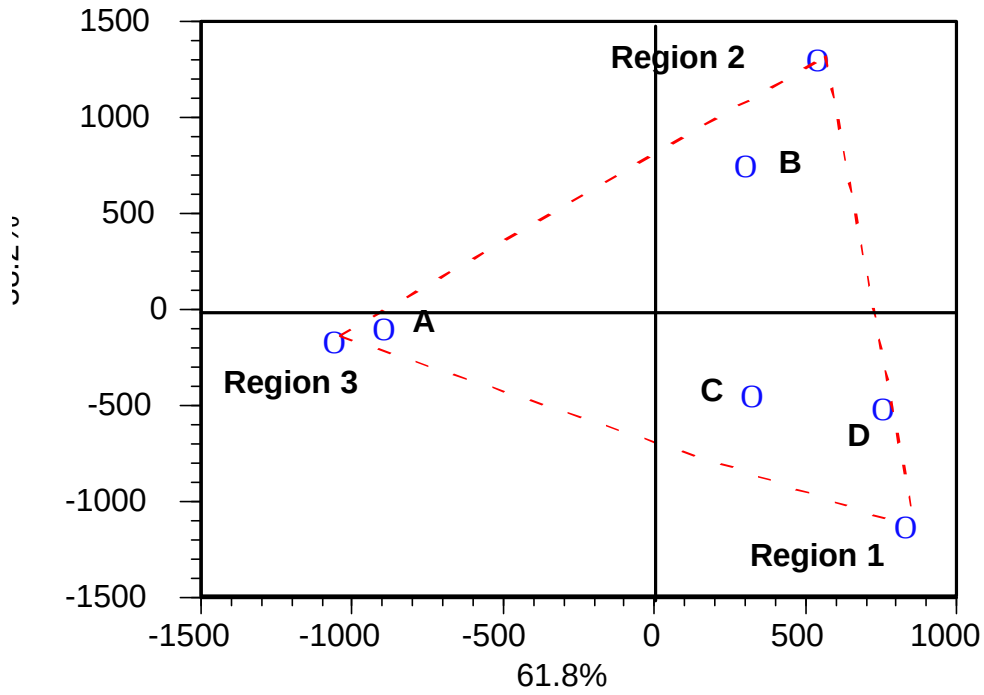


Figure 3. Asymmetric Plot of Row Profiles and Column Vertices

Detailed Worked Example

In order to illustrate the interpretation of output from correspondence analysis, the following example is worked through in detail.

A sample of 100 housewives were asked which of the 14 statements listed below they associated with any of 8 breakfast foods. Note that multiple responses were allowed.

The key to the statements and foods is presented in Table 5 and the frequency of responses is presented in Table 6.

Table 5. Key to Statements and Foods

Statement		Breakfast Foods	
A	Healthy	Cereals	CER
B	Nutritious	Muesli	MUE
C	Good in summer	Porridge	POR
D	Good in winter	Bacon and eggs	B&E
E	Expensive	Toast and tea	T&T
F	Quick and easy	Fresh fruit	FRF
G	Tasty	Stewed fruit	STF
H	Economical	Yoghurt	YOG
I	For a treat		
J	For weekdays		
K	For weekends		
L	Tasteless		
M	Takes too long to prepare		
N	Family's favourite		

Table 6. Frequency of Response

	CER	MUE	POR	B&E	T&T	FRF	STF	YOG	TOTAL
A	14	38	25	18	8	31	28	34	196
B	14	28	25	25	7	32	26	31	188
C	42	22	11	13	7	37	16	35	183
D	10	10	32	26	6	11	19	8	122
E	6	33	5	27	3	9	18	10	111
F	54	33	8	2	15	26	8	20	166
G	24	21	16	34	11	33	26	26	191
H	24	3	20	3	16	7	3	7	83
I	5	3	3	31	4	4	16	17	83
J	47	24	15	9	13	11	6	10	135
K	12	5	8	56	16	10	23	18	148
L	8	6	2	2	0	0	2	1	21
M	0	0	9	35	1	0	10	0	55
N	14	4	10	31	5	7	2	5	78
TOTAL	274	230	189	312	112	218	203	222	1760

The output of a 2-axis correspondence analysis solution to the above problem is presented in Appendix 1. A step-by-step interpretation of the problem follows.

Significance of Dependencies

The first step in the interpretation of correspondence analysis is to establish whether there is a significant dependency between the rows and columns. There are two approaches to establish significance. Firstly, the ***trace*** is examined. This appears in the eigenvalue report. The square root of the trace may be interpreted as a correlation co-efficient between the rows and columns.

As a rule of thumb, any value of this correlation co-efficient in excess of 0,2 indicates significant dependency.

In this example,

$$\sqrt{Trace} = \sqrt{0.3678} = 0.6065$$

thus indicating a strong dependency between the statements and the breakfast foods.

This is a rough and ready approximation and a more thorough approach is to calculate the chi-square statistic from the following formula:

$$c^2 = Trace * \sum_{i=1}^{r} \sum_{j=1}^{c} fo_{ij}$$

In this example,

$$c^2 = 0.3678 * 1760 = 647.3$$

There are (8-1)*(14-1) = 91 degrees of freedom in this problem. At a confidence level of 5%, the critical chi-square value is 70,0 thus indicating a significant dependency between rows and columns.

(Question: Why is the application of the chi-square test of independence not strictly applicable in this example?)

Dimensionality of the Solution

The second step in interpretation is to determine the appropriate number of dimensions to use in the solution. This is achieved by examining the eigenvalue report in more detail. The sum of the eigenvalues is equal to the trace. The ratio of the eigenvalue of any axis to the trace represents the proportion of the total "inertia" (or chi-square value) explained by that axis.

In this example, there are 8 columns, thus, if the data were purely random with no significant dependencies, the average axis should account for 100/(8-1) = 14.3% of the inertia. Likewise, the average axis should account for 100/(14-1) = 7.7% in terms of the 14 rows. Thus, any axis contributing more than the maximum of these two percentages should be regarded as significant and included in the solution. Thus, as the third axis in this example accounts for only 11.92% of the inertia, only a 2-dimensional solution should be used. Note that a higher number of dimensions may be used but the additional dimensions are unlikely to contribute significantly to the interpretation of nature of the dependency between the rows and columns.

The first and second axes account for 52.50% and 21.13% of the inertia respectively, i.e. a cumulative total of 73.64%. This latter figure is often referred to as the ***retention*** of the solution. Obviously, the higher the retention, the more subtlety in the original data is retained in the low-dimensional solution.

Interpreting the Axes

It is common practice to simply plot the co-ordinates presented in the correspondence analysis output. This is termed the ***French plot*** or ***symmetric plot***. While this plot may be useful, it may also lead to misinterpretation if examined in isolation or only visually. The reason for this is that ***principal co-ordinates*** are presented for both rows and columns. These co-ordinates represent the row and column profiles and not the apexes for which the ***standard co-ordinates*** are required. This means that while the distances between any row items and the distance between column items is meaningful and may be interpreted, the distance between any row and column items is not! In order to interpret any inter-point distances, the columns (profiles) must be presented in row space (vertices) or vice-versa. The French plot represents the row and column profiles simultaneously in a common space.

This problem may be overcome in one of two ways. The simplest way is to present only asymmetric plots. The apices of either the rows (or the columns) are plotted from the standard co-ordinates and the profiles of the columns (or the rows) are plotted from the principle co-ordinates. The standard and principle co-ordinates for any axis are related as follows:

$$P_{ij} = \sqrt{l_j}\, S_{ij}$$

where P_{ij} and S_{ij} are the principal and standard co-ordinates of row (or column) i on axis j and λ_j in the eigenvalue of axis j.

The asymmetric plot representing breakfast food profiles and statement apexes is presented in Figure 4. (Note that the plot in Figure 3 for the previous example was handled in this way).

A more complex, but very much more satisfying way of overcoming this problem in terms of richness in meaning, is to interpret the axes in terms of the rows (or the columns) and plot only the column points (or row points) in the space of the labeled axes.

The first step in this procedure is to decide whether to interpret the axes in terms of rows or columns. In this example, this requires a decision as to whether to interpret breakfast foods in statement space or statements in breakfast food space. After a little consideration, it seems likely that the former choice would be most appropriate.

The axes are interpreted by way of the contribution that each element (in this case each statement) makes towards the total inertia accounted for by the axis. In this example there are 14 statements, thus, any contribution greater than 100/14 = 7,1% would represent significance greater than what would be expected in the case of a purely random distribution of statements over the axes. (Please note that the figures quoted in Appendix 1 are multiplied by 1000).

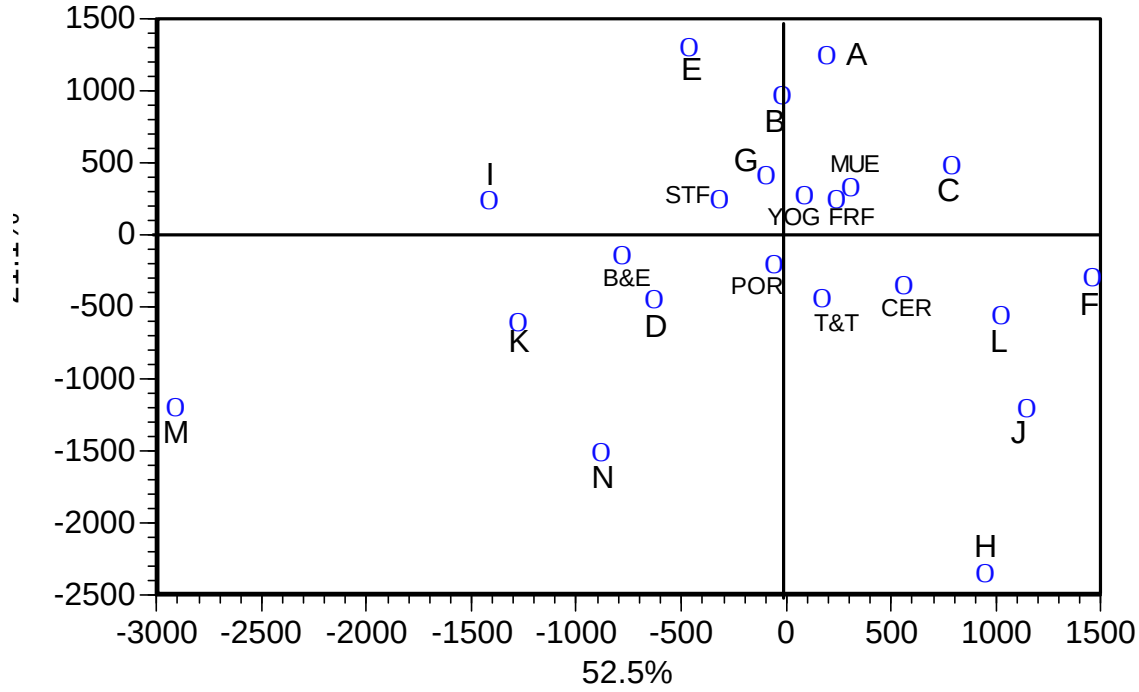


Figure 4. Asymmetric Plot of Breakfast Food Data with Statements in Standard Co-ordinates

Examining the detailed report for the rows in Appendix 1, statements M, F, K, J and I meet this criterion and "determine" the first axis. However, while statements F and J have positive co-ordinates, statements M, K and I have negative co-ordinates. The opposite poles of this axis are interpreted differently and as follows:

-ve		+ve	
M	Takes too long to prepare	Quick and easy	F
K	For weekends	For weekdays	J
I	For a treat		

It is reasonably clear from the loading of the statements that the first axis represents "convenience". While the positive side represents breakfasts that are everyday and easy to prepare, the negative side represents breakfasts that are special and take a long time to prepare.

The second axis is interpreted in the same fashion:

-ve		+ve	
H	Economical	Healthy	A
J	For weekdays	Expensive	E
N	Family's favourite	Nutritious	B

Notice that, unlike the interpretation of the first axis, the opposite poles of this axis are not entirely logical opposites. While economical is opposite to expensive, healthy and nutritious

have no polar opposites. It is also interesting to note that healthy and nutritious product are associated with being expensive.

The principle co-ordinates can now be plotted with the axes labeled as above. This plot is illustrated in Figure 5.

The relative positioning of the various breakfast foods is now not only correct but the statements determining their positioning is also clear.

The Quality of Representation

So far it is apparent that correspondence analysis has been successful in representing the contingency table in low dimensional space. An overall retention of 76,34% has been achieved in two dimensions. However, not all of the statements or products are equally well represented. Determining the ***quality*** of representation of a particular row or column provides additional richness to the interpretation of the relationships in the contingency table.

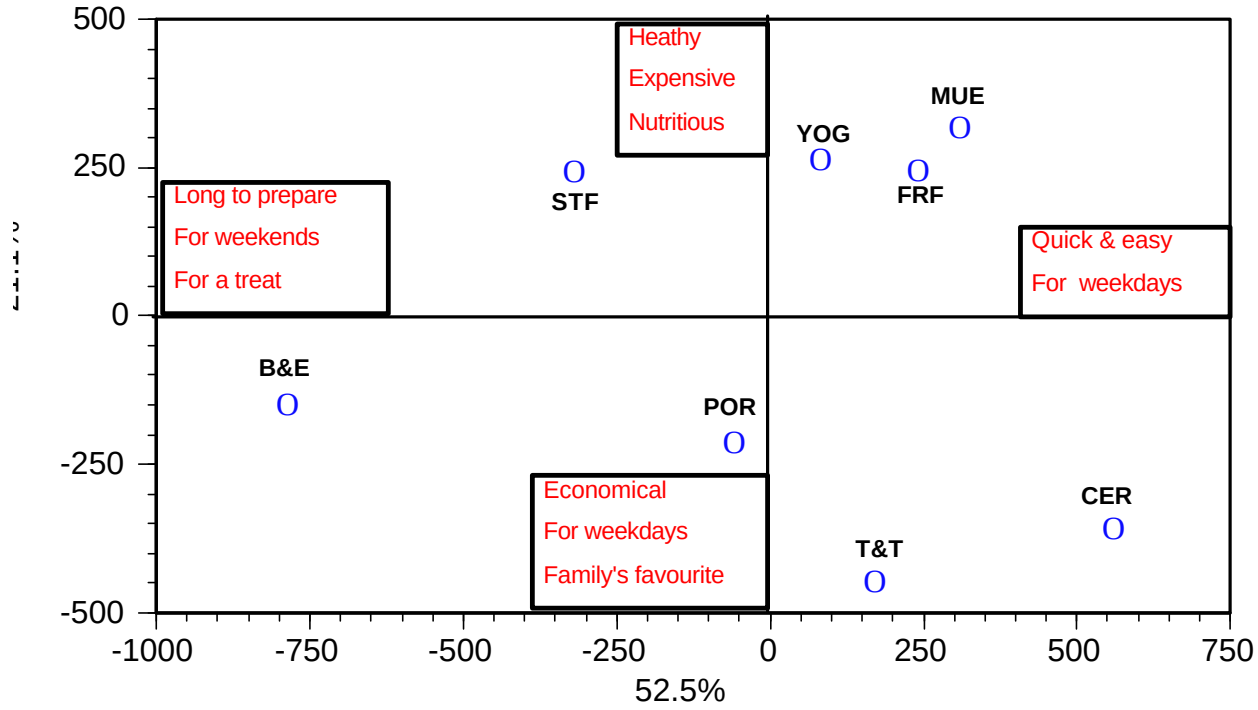


Figure 5. Plot of Breakfast foods in Statement Labeled Space

Some software packages provide details of the quality automatically. However, should this information not be available, the quality of representation is easily calculated from the ***correlations*** or ***squared correlations*** given in the output. Consider the detailed report presented for the columns presented in Appendix 1. The squared correlation presented for any column measures the degree of association between that column and a particular axis. So, for instance, the squared correlation between cereals (CER) and the first and second axes is 0.647 and 0.262 respectively. This implies that cereals are strongly associated with the first axis but only weakly associated with the second axis. This may be further interpreted by taking the sign of the co-ordinates of cereals on the two axis into account (i.e. examining the position of cereals on the plot presented in Figure 5): cereals are strongly associated with being quick &

easy to prepare and for weekdays (first axis); cereals are weakly associated with being economical, for weekdays and the family's favourite (second axis).

The quality of representation of a row or column in n dimensions is simply the sum of the squared correlations of that row or column over the n dimensions. In this example, the quality of representation of the breakfast foods in two dimensions is calculated as follows:

CER	0.647 + 0.262 = 0.909
MUE	0.289 + 0.299 = 0.588
POR	0.008 + 0.128 = 0.136
B&E	0.917 + 0.033 = 0.950
T&T	0.081 + 0.529 = 0.610
FRF	0.308 + 0.305 = 0.613
STF	0.520 + 0.308 = 0.828
YOG	0.047 + 0.439 = 0.486

Note that all products except porridge (POR) are well represented in the two dimensions. This implies that some caution is needed when interpreting porridge in this space and a higher dimensional solution is probably necessary to understand the relationship between porridge and the statements.

Supplementary Points

One of the most flexible aspects of correspondence analysis is the possibility of representing supplementary data points in the same low-dimensional space. All that is required is that the supplementary data must have either the rows or the columns in common with the original data. For this worked example. suppose that the following frequency of usage data was collected simultaneously with the attribute associations:

Table 7. Frequency of Product Usage

	CER	MUE	POR	B&E	T&T	FRF	STF	YOG
I	24	3	4	8	18	2	9	11
II	58	15	8	13	16	10	10	29
III	6	10	12	46	8	14	15	8
IV	2	4	28	9	4	47	4	2
V	10	68	48	24	54	27	62	50

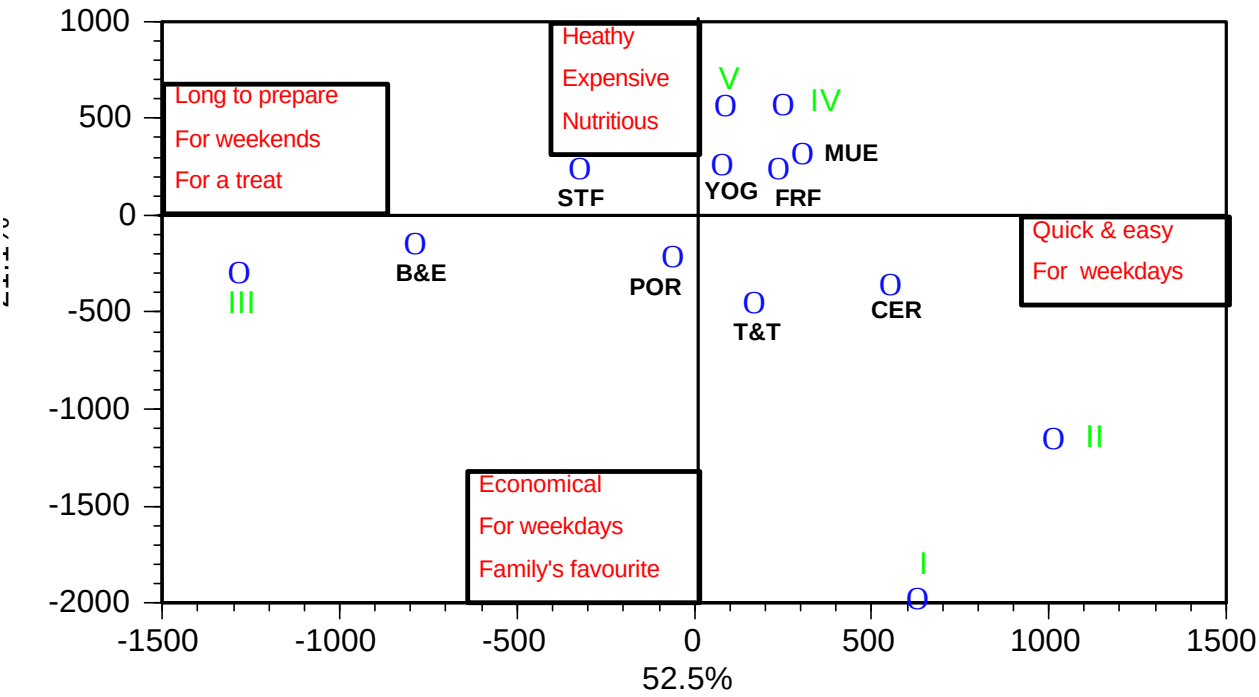
I = Daily, II = Several times per week, III = Several times per month, IV = Every few months, V = Never.

This data has the breakfast food products in common with the original data and can therefore be represented as supplementary rows on the correspondence analysis plot. Note that supplementary points are merely represented in this space and have no influence in determining either its nature or orientation. The detail report for the supplementary rows is presented in Table 8. Notice that in this report both the contributions and weights are zero confirming the fact that these supplementary points had no influence in determining the axes.

Table 8. Detail Report - Supplementary Rows

Label	Weight	Coordinate		Contribution		Sq. Correl.	
		F1	F2	CTR1	CTR2	COR1	COR2
I	0	280	-551	0	0	101	393
II	0	448	-321	0	0	425	218
III	0	-562	-82	0	0	866	19
IV	0	114	161	0	0	8	16
V	0	42	157	0	0	4	61

These supplementary rows are plotted in the correspondence analysis as illustrated in Figure 6 enriching the interpretation of the data even further.



(N.B. Supplementary rows plotted in standard co-ordinates)

Figure 6. Plot of Breakfast Foods with Supplementary Rows

Outliers

From time to time the data contained in a contingency table may contain one or more "outliers" in the rows and/or columns. The effect of such outliers is to dominate the interpretation of one or more of the axes. In doing so. the remaining row and/or column points tend to be tightly clustered in the resulting plots and thus difficult to interpret.

Potential outliers may be detected by seeking rows or columns that ***have both high absolute co-ordinate values and high contributions.*** The co-ordinates reported in correspondence analysis output represent the number of standard deviations the row (or column) is away from

the barycentre. Outliers are typified by being at least one standard deviation away from the barycentre as well as contributing significantly to the interpretation to one pole of an axis.

In the worked example. statement M is potentially an outlier as it is 1.278 standard deviations below the barycentre on the first axis and has the highest contribution (26.4%) in determining this axis. There are no other apparent outliers in this data.

Outliers may be treated as supplementary points and the correspondence analysis re-run without these points being allowed to determine the nature or orientation of the principle axes. In the case of the worked example. besides the exclusion of statement M. the interpretation of the axes remains virtually identical and the relative positioning of the rows and columns in the two-dimensional space remains unchanged - including the outlier viz. statement M which is now positioned as a supplementary point. This is unusual and probably indicates that statement M was not an outlier in the first place. The fact that the plotted points are not strongly clustered save for statement M is additional evidence for this conclusion.

Note that when an outlier is detected. the resulting interpretation of the axes must be seen in the light of the suppression of the outlier. The resulting interpretation must be qualified in that it is only relevant after the dominating influence of the outlier has been suppressed. Also, when an outlier is identified in a row, there are usually associated columns that are also outliers and *vice-versa*.

Conclusion

The above discussion represents a basic introduction to the use of correspondence analysis for the analysis of contingency tables in marketing research, specifically the construction of perceptual maps. A more advanced treatment of such data, including the segmentation of markets into perceptual groups, is discussed by Bendixen (1995).

The use of itemised rating scales is common in research, e.g. a 5-point Likert scale. As a matter of convenience, such data is usually assumed to behave in an interval fashion whereas it is strictly only ordinal. Correspondence analysis may be used to rescale such data. Dual or optimal scaling involves considering the co-ordinates only on the first principal axis (Greenacre, 1984). More recently, Bendixen and Sandler (1995) have proposed that the first two axes be used for this purpose so as to accommodate non-linear effects.

Hoffman and Franke (1986. p225-226). conclude that in the context of marketing research:

> *"Correspondence analysis is very flexible. Not only is it flexible in terms of data requirements. but also allows for incorporation of marketing knowledge.*
>
> *"Categorical data are common products of marketing research. However. the analysis of such data often is hindered by the requirements and limitations of many familiar research tools. Correspondence analysis is a versatile and easily implemented analytical method that can do much to assist researchers in detecting and explaining relationships among complex marketing phenomena."*

References

Bendixen, MT (1995). Compositional perceptual mapping using chi-squared trees analysis and correspondence analysis. *Journal of Marketing Management*, 11 (6), 571-581.

Bendixen MT & Sandler M(1995). Converting verbal scales to interval scales using correspondence analysis. *Management Dynamics: Contemporary Research*, 4 (1), 31-49.

Greenacre M J (1984). *Theory and Applications of Correspondence Analysis*, Academic Press. London.

Hoffman DL & Franke GR (1986). Correspondence analysis: Graphical representation of categorical data in marketing research. *Journal of MarketingResearch*, XXIII, (August), 213-227.

Lebart L; Morineau A & Warwick KM (1982). *Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*, John Wiley & Sons. Inc.. New York.

Mike Bendixon is in the Graduate School of Business Administration, University of Witwatersrand, South Africa.

Appendix 1

Eigenvalue Report

Axis	Eigenvalue	Percent	
		Individual	Cumulative
1	0.19309454	52.50	52.50
2	0.07773081	21.13	73.64
3	0.04385414	11.92	85.56
4	0.03280421	8.92	94.48
5	0.01225680	3.33	97.81
6	0.00568740	1.55	99.36
7	0.00236309	0.64	100.00
Trace	0.36779100		

Detail Report – Rows

		Coordinate		Contribution		Sq. Correl.	
Label	Wght	F1	F2	CTR1	CTR2	COR1	COR2
A	111	87	346	4	171	46	734
B	107	-9	269	0	99	1	676
C	104	349	133	66	24	552	80
D	69	-277	-127	28	14	227	48
E	63	-201	361	13	106	100	322
F	94	644	-85	202	9	889	15
G	109	-40	113	1	18	41	328
H	47	417	-657	42	262	216	537
I	47	-619	65	94	3	707	8
J	77	507	-338	102	113	611	272
K	84	-560	-171	137	32	724	68
L	12	451	-158	13	4	258	32
M	31	-1278	-334	264	45	894	61
N	44	-388	-422	35	101	343	405

Note: All numbers were multiplied by 1000.

Detail Report – Columns

		Coordinate		Contribution		Sq. Correl.	
Label	Wght	G2	G2	CTR1	CTR2	COR1	COR2
CER	156	563	-358	255	257	647	262
MUE	131	313	319	66	171	289	299
POR	107	-54	-213	2	63	8	128
B&E	177	-783	-148	563	50	917	33
T&T	64	175	-447	10	164	81	529
FRF	124	246	245	39	96	308	305
STF	115	-315	243	59	87	520	308
YOG	126	86	264	5	113	47	439

Note: All numbers were multiplied by 1000.