THE USE OF CORRESPONDENCE ANALYSIS

IN BUILDING LOG-LINEAR MODELS

BY

Charles David Heber Parry

THESIS

Submitted in fulfilment of the

Requirements for the degree of

MASTER OF SCIENCE

IN THE DEPARTMENT OF

MATHEMATICAL STATISTICS

UNIVERSITY OF CAPE TOWN

Supervisor : Prof. J. M. Juritz

September 1983

To my parents

ACKNOWLEDGEMENTS

I wish to express my gratitude to the following :

i) To my supervisor, PROF. JUNE JURITZ for her encouragement and help throughout this study.

ii) To the MEDICAL RESEARCH COUNCIL for their support and sponsorship of this degree.

iii) To my COLLEAGUES AT THE INSTITUTE FOR BIOSTATISTICS for their assistance, and in particular Mrs Marie Kotzé and Miss Antoinette van Zyl for their typing of this thesis and Mr René Gonin for his editing and proofreading of the manuscript.

iv) To Mrs Lena Green for proofreading the manuscript.

ABSTRACT

Data collected in the biomedical and social sciences by means of questionnaires is in most instances qualitative in nature. Such data, typically set out in the form of (multi-dimensional) contingency tables, is usually subjected to hypothesis testing in order to assess the inter-relationships between the questions. Prior to undertaking confirmatory procedures, we argue that exploratory techniques should be used to gain a "feel" for the data. Correspondence Analysis (an exploratory data analysis procedure) and Log-linear Model building (a confirmatory data analysis procedure) are discussed before an investigation is undertaken to ascertain whether they can be used in conjunction. We found that correspondence analysis : (i) detects questions that are "strictly" independent/unrelated, (ii) detects pairwise relationships between questions (2-factor interactions) and thus can be used to suggest a splitting of large data sets into two or more subsets of questions that are independent, each of which can be analysed separately, and (iii) cannot be used to select log-linear models in general because it does not detect higher order interactions.

TABLE OF CONTENTS

LIST OF TABLES

* - Appendix A1

* - Appendix A1

** - Appendix A2

LIST OF FIGURES

CHAPTER 1

1. INTRODUCTION

1.1 Questionnaires

Data in the biomedical and social sciences are often collected by means of a questionnaire. Questionnaires are defined by the Concise Oxford Dictionary (6-th edition, 1976) to be a "formulated series of questions especially for statistical study..." and according to Babbi (1973), their purpose is to aid the researcher in describing, exploring and explaining the phenomena he is interested in. In this thesis we shall neither be concerned with the design of the questionnaire nor with the sampling scheme used to collect the data, although the statistician can perform a useful role in both these matters. We shall rather focus on some steps in the analysis of data which arises from questionnaires. Such data is essentially multivariate in nature and has to be analysed accordingly.

But what sort of multivariate analysis? The answer to this question rests essentially on two issues :

(i) the nature of the data, and

(ii) the problems we are seeking to solve.

We shall not address the latter issue at present except to make the comment that with questionnaire data the problem is often one of assessing the inter-relationships between variables and possibly setting up prediction

equations. The nature of the data, on the other hand, defines the statistical techniques which are appropriate to achieving these aims.

With questionnaire data most of the variables are invariably qualitative and only a few are quantitative. Usually the responses to a given question have been formulated in advance and the respondent chooses the response which most closely coincides with his position. Thus the data generated by such a procedure is largely categorical in nature and some method of discrete multivariate analysis will be appropriate. Quantitative data from questionnaires can be handled either by discretizing the data into defined categories or by using methods which can handle both qualitative and quantitative data (e.g. GLIM, see Nelder and Wedderburn, 1972). In this thesis the focus will be on qualitative data though we shall also refer to methods for handling mixed data.

However, before any sophisticated analysis of the data is undertaken, care should be taken to gain some understanding of it.

1.2 Exploratory and Confirmatory Data Analysis

'Exploratory data analysis needs restoring to its rightful place alongside confirmatory data analysis', this is the view of Tukey (1977) in his refreshing book entitled "Exploratory Data Analysis". He recognizes the importance of confirmatory techniques in the modern statistician's arsenal, but points out that along with their development has come a downgrading of those techniques which are not explicitly linked to confirmatory procedures and that despite their usefulness they have come to be labeled "mere

descriptive statistics". He feels that statisticians have lost much by limiting themselves to understanding those things which can be confirmed, often under a set of very restrictive assumptions which are almost impossible to check in practice. Both Tukey and Everitt (1978) stress that exploratory and confirmatory data analyses are complementary.

Turning our attention to the former, we note that many authors such as Solomon (1977), Friedman & Rubin (1967), Tukey (1977) and Benzécri (see Greenacre, in press) make the same point, namely that we should 'let the data speak to us'. In other words that we should, at least at the initial stages of the statistical analysis, be less involved with confirmatory procedures and more concerned with the data itself. They use phrases such as : "...we wish the data to tell us what is going on...", "...the investigator... wants the data to 'suggest' natural categories...to lend insight into the structure of the data so as to suggest more formal models for further analysis... ", "... looking at data to see what it seems to say... " and "... the model should follow the data, not the inverse". They are all making the point that exploratory data analysis is a necessary part of any statistical analysis. But what sort of exploratory techniques should we use?

Clearly we need procedures which are able to describe the data in such a way that the relevant information is easily assimilated by the investigator. Pictures and graphs are an obvious starting place. Tukey and Tukey (1981:189) points out that :

> "Pictures are particularly valuable in an exploratory setting because not only can they confirm or contradict what we thought we knew in advance about the data, but they can also reveal in a dramatic way things that we did not even suspect."

The usefulness of pictures and graphs is perhaps best summed up in a quote by Greenacre (in press, chapter 1) : " -a picture is worth a thousand numbers ". In Chapter 2 we discuss a number of graphical/pictorial methods of exploratory data analysis.

1.3 Preliminary steps in the analysis of questionnaire data

In this section we discuss a number of methods to simplify the data matrix which results from compiling the information on a questionnaire. These methods should be considered before one goes on to more formal exploratory and confirmatory data analytic procedures. They involve (1) procedures for data elimination that can be done without the prior use of statistical tests and (2) procedures for summarizing the data, both of which help one from the outset to gain a better understanding of the data.

The data is typically in the form of a matrix of N (i = 1,2,..,N) rows (subjects/observations) and M (j = 1,2,..,M) columns (variables/ questions) where the ij-th entry is the i-th subject's response to the j-th question. The matrix might be data collected from a survey with a large majority of the variables being discrete (e.g. sex : male/female), while a few of the variables might be continuous (e.g. age, income). In addition, there are a number of "book-keeping" type variables comprising case numbers, identity numbers, card numbers, serial/reference numbers and possibly dates of collection e.g. 080382 . Variables of this type would be omitted from the statistical analysis, but are important when seeking the source or explanation of unusual or outlying data points.

1.3.1 Initial screening of the data

After the data has been coded and punched it is helpful to run computer programs such as BMDP P1D, P2D, or SAS PROC FREQ which give, amongst other things, 1-way frequency distributions of all the variables. From these, one is not only able to get a good idea of the distributions of the responses to the various questions asked, but one is often also able to detect errors in the data due to incorrect coding or punching, for example, a code of 5 when there were only 4 possible responses to a particular question. In such a situation it would be advisible to go back to the questionnaire in order to ascertain from where the inappropriate code originated. If the incorrect code appeared on the questionnaire then, if at all possible one should try to trace the respondent and fill in the appropriate response. The output from these programs is also useful in the data simplification/reduction techniques mentioned in Sections 1.3.2 and 1.3.3 below.

Given that the sample size is large enough, if one finds that there are variables on which all the subjects give the same response then these should be noted down and thereafter deleted from the analysis as they are unable to help in making distinctions between subjects.

1.3.2 Combining questions

In many cases, when questionnaires are drawn up, little thought is given to the actual data matrix which will result from encoding the responses of the subjects to the various questions. This often leads to wasteful usage of a

number of questions of the "yes/no" variety when in fact one question with several mutually exclusive responses could be used. The problem is one of being able to distinguish between variables and levels of a single variable. Such a condensation is not possible if the various variables (which one is recoding into a single new variable) are not mutually exclusive.

1.3.3 Combining response categories of questions

From the output from the BMDP and SAS procedures mentioned in section 1.3.1 it is easy to ascertain which categories of the variables have no or only a few responses. Faced with such a situation, many statisticians would consider dropping the categories having no observations and combining the categories having only a few responses, of course taking practical considerations into account. A word of caution is needed. In the case of the former it is important for the statistician to be aware that zero observations may have arisen due to problems in his sampling scheme. Care must be taken when combining categories, as it has been shown (Reynolds, 1977) that significant interactions can be "created" simply as a result of the particular way in which the levels of the variables involved have been collapsed. Rules for collapsing over categories of variables are given in Bishop et al. (1975) and Whittmore (1978).

1.4 Concluding remarks

The focus of this thesis is on the use of exploratory data analysis in aiding the researcher in gaining an understanding of the data before moving on to hypothesis testing, model building and other confirmatory procedures. We suggest that a preliminary step be added to the exploratory data analysis stage - a step which involves the procedures outlined above. We propose that the steps represented schematically below should be followed when working with data from a questionnaire.

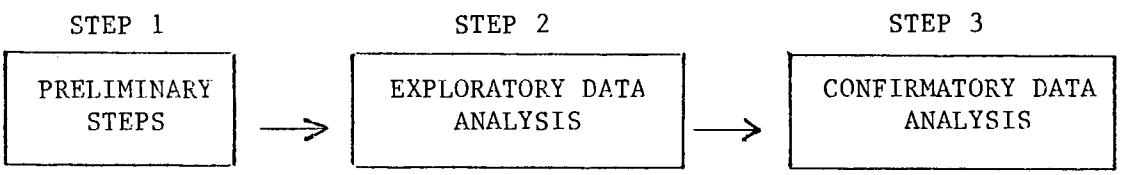


Figure 1 : Schematic representation indicating the movement from the preliminary steps through exploratory to confirmatory data analysis

CHAPTER 2

2. GRAPHICAL/PICTORIAL METHODS OF EXPLORATORY DATA ANALYSIS

...a brief overview.

In Table 1 are listed a number of graphical and pictorial techniques for displaying multivariate data. Each of which is dependent both on what it is we wish to display and also on the nature of the data with which we are working. This list is not exhaustive , but it should be complete enough to situate Correspondence Analysis, a method which we shall subsequently show to be suited to performing exploratory data analysis on questionnaire data. Information on the type of data matrix and the nature of the data required for each technique, as well as a number of references for each, are given in Table 1. We shall not refer to each technique individually but shall briefly comment on each group of techniques, singling out only a few for specific comment.

Suppose N subjects have answered a questionnaire with M questions, then the resulting data can be set out in the form of an NxM profile matrix. In addition, it is possible to transform the profile data matrix into two symmetric matrices: the one, an NxN matrix of inter-subject distances or similarities and the other, an MxM matrix of inter-variable (question) distances or similarities. One may then perform the exploratory data analysis procedures on either the profile matrix or the symmetric inter-subject or inter-variable matrices.

TABLE 1 GRAPHICAL AND PICTORIAL TECHNIQUES FOR MULTIVARIATE DATA

TECHNIQUE	DATA MATRIX	DISPLAY OF	REFERENCES*
I. ORDINATION/DIMENSION REDUCING METHODS			
(a) The biplot			
(i) Principal Components Biplot	Operates on a profile matrix of continuous quantitative data.	both rows and columns	1 2 3
(ii) Generalized Principal Components analysis			1
(iii) Principal Components Analysis of Standardized Data			1
(iv) Covariance Biplot			1 2 3 4
(v) Correlation Biplot			1
(vi) Symmetric Biplot			1 4 5
(vii) Canonical Variate Analysis			1
(b) Correspondence Analysis	Operates on a profile matrix. Essentially on qualitative but can be extended to quantitative data.	both rows and columns	1 6 7 8 9 10
(c) Discriminant Analysis	Operates on a profile matrix of continuous (quantitative) data. It can be extended to qualitative data.	rows (observations)	11 26

* - see list of codes at end of table

TABLE 1 (Continued)

(d) Multi-dimensional Unfolding	Operates on a profile matrix. Both quantitative and qualitative data - preferably of ordinal or higher level of measurement but can handle nominal data.	both rows and columns	8 11
(e) Methods operating on a triangular matrix of distances/similarities			
(i) Classical Scaling	Operates on a triangular matrix of both qualitative or quantitative data of all levels of measurement.	either rows or columns or both (separately)	1
(ii) Non-linear Mapping			1 3
(iii) Least Squares Scaling			1
(iv) Principal Co-ordinates Analysis			1 3 12
(v) Non-metric Multi-dimensional Scaling			1 3 13 14 15 16
II. GROUPING METHODS FOR MULTIVARIATE DATA			
(a) Hierarchical Clustering Methods	Operates on a similarity/distance matrix. On both qualitative* and quantitative data. * - classify into "like" and "unlike" and work with binary matrix.	both rows (observations) and columns (variables)	3 17 18
(b) Hierarchical Segmentation Methods			

TABLE 1 (Continued)

III. TECHNIQUES FOR MULTIVARIATE DATA BY REPRESENTING FURTHER DIMENSIONS ON A 2-DIMENSIONAL PLOT			
(a) Characters and Glyphs			
(i) -simple-			
- qualitatively distinct characters, crosses or polygons			
- whisker directions whisker length			
- heaviness of line using some fixed character			
- size of character for characters of some fixed shape	Quantitative data. Best with continuous data, but could use discrete (ordinal) data.	rows (observations)	3 19 20 21
- colour			
(ii) -compound-			
- weathervane symbols			
- stars or polygons			
- glyphs			
- trees, castles and related symbols			
- schematic cells			
- multi-aspect characters			
- Chernoff faces			
(b) Andrew's Plots	as for III (a)	rows (observations)	3 22
IV. MISCELLANEOUS TECHNIQUES			
(a) A.I.D. (Automatic Interaction Detection)	Either quantitative or qualitative data of all levels of measurement.	columns (variables)	23 24 25

TABLE 1 - CODES

Code	Reference
1	Greenacre and Underhill (1982)
2	Gabriel (1971)
3	Everitt (1978)
4	Gabriel (1981)
5	Bradu and Gabriel (1978)
6	Benzécri (1969)
7	Hill (1974)
8	Greenacre (1978a)
9	Greenacre (1978b)
10	Greenacre (1981)
11	Greenacre (in press)
12	Gower (1966)
13	Kruskal (1964a)
14	Kruskal (1964b)
15	Shepard (1962a)
16	Shepard (1962b)
17	Everitt (1974)
18	Hawkins & ten Krooden (1982)
19	Tukey & Tukey (1981)
20	Chernoff (1973)
21	Solomon (1977)
22	Andrews (1972)
23	Kass (1975)
24	Hawkins & Kass (1982)
25	Bishop et al. (1975)
26	Hand (1981)

2.1 Ordination/dimension reducing methods (I)

Ordination (or scaling) methods are involved with reducing the dimensionality of the data while still maintaining as far as possible the structure of the original data space, and thereafter displaying the 'lower' dimensioned data graphically. A number of techniques ((a)(i)- (vi),(b),(c) and (d)) operate directly on an NxM (profile) data matrix. With the exception of Multidimensional Unfolding (d), these scaling methods are termed by Greenacre and Underhill (1982) 'basic structure displays'. They refer to the remaining ordination techniques ((e)(i)-(v)) as 'multidimensional scaling techniques'. These operate on an NxN (MxM) symmetric matrix of inter-individual (inter-column) distances/ dissimilarities or similarities.

Given a matrix P, Principal Components Analysis is, according to Everitt (1978), a procedure for transforming a set of points $\{P_{ij}\}$ (i=1,2,...,N and j=1,2,..,M) into a second set of points $\{Q_{ij}\}$ such that the points in the second set are related to a different set of orthogonal axes to those in the first. Typically only the first few axes are used and thus the relative (Euclidean) distances between points in the second set are approximations of those in the first. The biplot is a way of displaying the relationships between the rows and columns of matrix P (in a Euclidean space of lower dimension) on the same set of axes. Bradu and Gabriel (1978) put forward the Symmetric Biplot as a method for diagnosing models of 2-way tables. Their method essentially involves the displaying of the rank two approximation of a data matrix by plotting what they term row and column 'markers'. Collinearity of markers and the angle between a line drawn

through certain row and column markers can be used to make a diagnosis of the type of model underlying the 2-way table. Besides this technique necessitating that the rank two fit of the data be good, there is an obvious difficulty in applying this to questionnaire data, namely that it requires that the data be quantitative.

Correspondence Analysis, like the biplot, is an ordination method for displaying the rows and columns of an NxM profile matrix. It was especially designed for a table of counts (though it can accommodate other types of data) and does not have such strict requirements about the nature of the data as does the biplot. In addition, Greenacre (1978a) points out that these two approaches differ in terms of their interpretations of within-set (row-to-row/column-to-column) and between- set (column-to-row) distances on the graphical display/biplot. A detailed discussion of the theory behind Correspondence Analysis is given in section 3.2.

Multidimensional Unfolding is a method for representing a profile matrix of 'pseudo distances' (simmilarities/dissimilarities) and, according to Greenacre and Underhill (1982:260), is aimed at scaling "the set of rows and the set of columns as points in a joint low-dimensional Euclidean space so as to fit the pseudo distances".

2.2 Grouping methods for multivariate data (II)

Clustering techniques aim to group observations that are similar on the basis of some or other criterion. In Table 1 we refer to two related

cluster analysis procedures which yield a graphical display in the form of a dendrogram. Hierarchical techniques operate on a symmetric matrix of inter-subject/inter-variable distances or similarities and yield a solution in which clusters are nested within each other. Hierarchical segmentation techniques are essentially the same as the 'ordinary' hierarchical methods, except that certain restrictions are placed on cluster membership.

The techniques under sections I(e) and II in Table 1 can be applied to questionnaire data, but not without considerable effort. For example, to set up a matrix of inter-variable distances or similarities one would first have to decide upon the criterion for inter-variable distance or similarity (say the log-likelihood ratio chi-square statistic to measure the relationship between pairs of questions) and then calculate each entry in the matrix before using these techniques.

2.3 Techniques for multivariate data by representing further dimensions on a two dimensional plot (III)

Tukey and Tukey (1981) discuss a number of methods for displaying multivariate data which utilize symbols or faces on a two-dimensional plot to give some indication of further dimensions. Two problems with these techniques are firstly, that all the dimensions are not weighted equally and secondly, that there is a limit to the number of dimensions that can be represented in this way. Their usefulness (with the exception of Chernoff faces) lies in their application together with one or other of the ordination methods (I).

Andrews (1972) seeks to overcome one of the limitations of the above methods, the inability to go beyond three or four dimensions, by proposing a method for plotting high dimensional data which is based on a Fourier expansion.

2.4 Miscellaneous techniques (IV)

Automatic Interaction Detection (A.I.D.), although primarily a confirmatory technique, can be used in an exploratory manner to help the investigator to decide which type of statistical analysis to proceed with or which model to fit to the data (Kass, 1975; Hawkins and Kass, 1982). It operates on a data matrix consisting of one dependent variable and one or more predictor (explanatory) variables. This technique uses the predictor variables to split the data by successive binary divisions into a number of subgroups. The output is a tree-like structure (similar to the dendrogram in cluster analysis) which provides an explanation of the variability in the dependent variable.

Some of the techniques set out in Table 1 lean towards the confirmatory (e.g. A.I.D.), nevertheless they can all profitably be used for exploratory purposes. Of these techniques, the biplot methods, Correspondence Analysis, Multidimensional Unfolding, A.I.D. and the scaling and clustering methods for an MxM symmetric matrix of inter- variable similarities/dissimilarities (I(e) and II) all deserve consideration when we are faced with performing exploratory data analysis on an NxM matrix of questionnaire data. We

dismiss both the biplot and Multidimensional Unfolding, the former as it requires that the data be quantitative and the latter as it requires that the data matrix be one of "pseudo distances" . Neither of these requirements are usually met in questionnaire data. A.I.D. will only be used when we have certain questions which we know in advance are dependent on other questions. Of the two remaining techniques, we favour Correspondence Analysis as it provides information on the observations and variables simultaneously and plots both the observation and variable points on a joint graphical display.

CHAPTER 3

3. LOG-LINEAR MODEL BUILDING AND CORRESPONDENCE ANALYSIS

In Figure 1 (section 1.4) we illustrated the progressive steps that could be followed in the statistical analysis of questionnaire data. In this chapter we concentrate on one technique for exploratory data analysis (correspondence analysis) and one confirmatory data analysis procedure (log-linear models) - the former, a technique for displaying data graphically and the latter, a method for building models with multivariate categorical data. We now discuss each in turn, making no attempt to link the two (that is reserved for chapters 4 and 5), and wherever possible we shall direct our comments to the analysis of questionnaire data.

3.1 Log-linear models

Much of confirmatory data analysis involves the fitting of models in order to :

(i) help one understand the data,

(ii) help in assessing the size of the interaction/correlation between variables, and

(iii) obtain predicted or smoothed values of population parameters.

We shall contain our interest to the fitting of log-linear models, which are extensively used by researchers seeking to describe multivariate categorical data.

3.1.1 Definition of the log-linear model

Consider a simple example of a questionnaire with four questions A,B,C and D with I,J,K, and L response categories respectively. Suppose a random sample of N subjects answers the questionnaire then the responses can be coded in an Nx4 profile data matrix. Let $f_{ijk\ell}$ be the number of subjects whose responses fell into the i-th category of question A, the j-th of B, the k-th of C and the ℓ-th of D ($i=1,2,..,I$, $j=1,2,..,J$, $k=1,2,..,K$, $\ell=1,2,..,L$). The joint probability distibution of the $f_{ijk\ell}$ is an IxJxKxL multinomial distribution with joint probability mass function :

$$\frac{N!}{\prod_{ijk\ell} f_{ijk\ell}!} \prod_{ijk\ell} p_{ijk\ell}^{f_{ijk\ell}} \qquad \text{where } \sum_{ijk\ell} f_{ijk\ell} = N$$

where $p_{ijk\ell}$ is the probability of a subject falling into category i of A, j of B, k of C and ℓ of D (Cochran,1952 and Bishop et al., 1975). The $p_{ijk\ell}$'s of this multinomial distribution have a natural structure because of the possible relationships between A,B,C and D.

The $f_{ijk\ell}$'s can be regarded as IxJxKxL independent Poisson variables subject to the condition that $\sum_{ijk\ell} f_{ijk\ell} = N$. If $m_{ijk\ell}$ is the expected

value of $f_{ijk\ell}$, then

$$E(f_{ijk\ell}) = Np_{ijk\ell} = m_{ijk\ell} .$$

It can be shown (Nelder and Wedderburn, 1972) that the structure can be expressed as the log-linear model :

$$\log_e m_{ijk\ell} = \theta \qquad \text{..constant}$$

$$+ \lambda^A_i + \lambda^B_j + \lambda^C_k + \lambda^D_\ell \qquad \text{..main effects}$$

$$+ \lambda^{AB}_{ij} + \lambda^{AC}_{ik} + \lambda^{AD}_{i\ell} + \lambda^{BC}_{jk} + \lambda^{BD}_{j\ell} + \lambda^{CD}_{k\ell} \qquad \text{..2-way interactions}$$

$$+ \lambda^{ABC}_{ijk} + \lambda^{ABD}_{ij\ell} + \lambda^{ACD}_{ik\ell} + \lambda^{BCD}_{jk\ell} \qquad \text{..3-way interactions}$$

$$+ \lambda^{ABCD}_{ijk\ell} \qquad \text{..4-way interaction}$$

where

$$\sum_i \lambda^A_i = 0 = \ldots = \sum_\ell \lambda^D_\ell = 0$$

$$\sum_i \lambda^{AB}_{ij} = \sum_j \lambda^{AB}_{ij} = 0 = \ldots = \sum_k \lambda^{CD}_{k\ell} = \sum_\ell \lambda^{CD}_{k\ell} = 0$$

$$\sum_i \lambda^{ABC}_{ijk} = \ldots = \sum_k \lambda^{ABC}_{ijk} = \ldots = \sum_j \lambda^{BCD}_{jk\ell} = \ldots = \sum_\ell \lambda^{BCD}_{jk\ell} = 0$$

$$\sum_i \lambda^{ABCD}_{ijk\ell} = \ldots = \sum_\ell \lambda^{ABCD}_{ijk\ell} = 0 .$$

If fitted, this (saturated) model will reproduce the observed values. Unsaturated models contain only some of the terms above. We restrict ourselves to hierarchical log-linear models. In these models if an

interaction term, λ^{ABC}_{ijk} for example, is included in the model, then all lower order 2-factor interaction terms λ^{AB}_{ij}, λ^{AC}_{ik} and λ^{BC}_{jk} and main effects λ^{A}_{i}, λ^{B}_{j} and λ^{C}_{k} are also included.

3.1.2 Estimation of parameters

The parameters of the model can be estimated by maximum likelihood estimation (M.L.E.). Since the multinomial distribution belongs to the class of exponential distributions (Hogg and Craig, 1970), these estimates are functions of sufficient statistics. The M.L.E.'s are identical to the minimum discrimination information (M.D.I.) estimates of Gokhale and Kullback (1978) in the case of hierarchical models. In some unsaturated models the estimates of the cell frequencies can be written as explicit functions of the marginals (direct models). This is not generally true and usually the estimated cell frequencies have to be calculated by an iterative scaling procedure (Bishop et al., 1975).

3.1.3 The relationship between contingency tables and log-linear models

The data generated by the model $\log_e m_{ijk\ell}$ can be set out in a 4-dimensional contingency table with IxJxKxL cells. There is a one-to-one correspondence between the marginals of the multidimensional contingency table and the λ's included in the model.

If for example, the unsaturated model

$$\log_e m_{ijk\ell} = \theta + \lambda^{A}_{i} + \lambda^{B}_{j} + \lambda^{C}_{k} + \lambda^{AB}_{ij}$$

fits the data, then the data can be described to a reasonable degree by two contingency tables: a two dimensional contingency table containing the 2-way A and B marginals

$$f_{ij..} = \sum_{k\ell} f_{ijk\ell}$$

and a 1-way table of C marginals

$$f_{..k.} = \sum_{ij\ell} f_{ijk\ell}$$

Such an unsaturated model will be denoted by AB,C.

3.1.4 Goodness-of-fit statistics

How well a particular model describes the data can be assessed by goodness-of-fit statistics. Let $\hat{m}_{ijk\ell}$ be the estimated cell frequency of cell $ijk\ell$ under some model. Two commonly used goodness-of-fit statistics are :

(i) the likelihood ratio chi-square statistic,

$$G^2 = -2 \sum_{ijk\ell} f_{ijk\ell} \log_e \frac{\hat{m}_{ijk\ell}}{f_{ijk\ell}}$$

which has a chi-square distribution with degrees of freedom appropriate for $\hat{m}_{ijk\ell}$ and

(ii) the traditional Pearson's chi-square statistic,

$$\chi^2 = \sum_{ijk\ell} \frac{(f_{ijk\ell} - \hat{m}_{ijk\ell})^2}{\hat{m}_{ijk\ell}}$$

which also has a chi-square distribution with degrees of freedom appropriate for $\hat{m}_{ijk\ell}$, but which does not have the additive property of (i). Kotze (1982) argued that the less conservative likelihood ratio chi-square statistic is to be prefered in general.

3.1.5 Model selection techniques

A questionnaire with N subjects and M questions could in principle be described by an M-way multidimensional contingency table, but in practice a number of problems may be encountered :

(i) The sample size would have to be extremely large otherwise the table would have a very large number of empty cells. This would cause problems in fitting models and with the applicability of goodness-of-fit tests.

(ii) Currently no computer programmes are available to handle the data. BMDP P4f, for example, cannot deal with tables having more than 2999 cells.

(iii) The resulting model would be unwieldly and possibly very difficult to interpret.

Usually the sample size is not large enough to adequately fit log-linear models to more than 4 or 5 variables. Even with 4 or 5 variables there are still very many unsaturated hierarchical models that could be fitted. Bishop et al. (1975) point out that with only 4 variables there are 113 hierarchical models involving all 4 variables to choose from.

In reviewing various strategies for selecting log-linear models, Kotze (1982) asserted that a good indication of the relative importance of each interaction in a multi-dimensional contingency table can be found by conducting tests for marginal (Brown, 1976) and partial (Birch, 1972) association. Two widely used model selection strategies reviewed by Kotze are the forward and backward stepping procedures (Goodman, 1971), which are in some ways analogous to the stepwise procedures in regression analysis, and Aitkin's (1978, 1979, 1980) simultaneous test procedure which is based on a similar technique to that used to select models for unbalanced cross -classifications in the ANOVA.

3.2 Correspondence Analysis

3.2.1 Historical background

"Correspondence analysis" is a translation of the French "analyses factorielle des correspondances", a term which was first coined by a group of French statisticians led by the linguist J-P Benzécri to refer to this method of metric scaling. Since the mid 1930's, researchers working independently in various statistical and applied fields have put forward a number of theoretically equivalent techniques such as simultaneous linear regression, dual scaling and reciprocal averaging. However, it is this group of Frenchmen who from the early 1960's have used and refined this technique in the context of graphical data displays.

3.2.2 Correspondence Analysis of contingency tables and 'indicator matrices'

Correspondence analysis is, by definition, applicable to the analysis of a 2-way contingency table. The theory behind and application of correspondence analysis on 2-way contingency tables have been clearly outlined by Greenacre in a number of publications (1978a, 1978b, 1981, 1982). In his book, (in press), Greenacre (chapter 5) showed that this was the same as the canonical correlation analysis of the contingency table data in the form of an 'indicator matrix' which is defined below :

> "The rows of the indicator matrix correspond to the observational units (individuals, cases, subjects,..) of the study, while the columns correspond to the categories of the two discrete variables defining the rows and columns of the contingency table. Each row has two non-zero elements (usually ones) which indicate the categories into which the observational units fall."

For example, the indicator matrix corresponding to the following 2-way contingency table

		B	
		b1	b2
A	a1	1	3
	a2	4	2

is as follows :

a1	a2	b1	b2	cell
1	0	1	0	a1b1
1	0	0	1	a1b2
1	0	0	1	
1	0	0	1	
0	1	1	0	a2b1
0	1	1	0	
0	1	1	0	
0	1	1	0	
0	1	0	1	a2b2
0	1	0	1	

Greenacre extended this to the case where there are more than two (discrete) variables and introduced the concept of multivariate indicator matrices. Multiple correspondence analysis is the term used for correspondence analysis on such matrices. We shall demonstrate that the data obtained from a typical questionnaire can easily be coded into the form of a multivariate indicator matrix and multiple correspondence analysis performed on it. Our focus will be on this rather than on the correspondence analysis of a 2-way contingency table.

3.2.3 Multiple Correspondence Analysis with particular reference to questionnaire data

Suppose N subjects responded to a questionnaire containing M questions where the q-th question has J_q possible response categories. From this data it is hypothetically possible to set up an M-way $(J_1 x J_2 x \ldots x J_M)$ multidimensional contingency table. The information in this contingency table can be used to construct an NxJ indicator matrix Z where $J = J_1 + J_2 + .. + J_M$ (the sum of the number of response categories to each individual question). The elements of the matrix comprise a series of "1's" and "0's" - "1" indicates a subjects affirmation of a particular response category and "0" that that particular response category is not applicable. See Figure 2.

	Question 1	Question 2		Question q		Question M
	1 2...J_1	1 2...J_2	. . .	1 2...J_q	. . .	1 2...J_M
1	1 0...0	1 0...0	. . .	1 0...0	. . .	1 0...0
.						
.						
.						
N	0 0...1	0 0...1	. . .	0 0...1	. . .	0 0...1

Figure 2 : Indicator matrix Z

Greenacre terms the rows and columns of Z, subjects and objects respectively. A number of features of Z may be noted :

(i) It contains only "1's" and "0's".

(ii) Since the information on each subject is contained in one particular row of Z, there will be no rows comprising of only "0's" as each subject will have responded to at least one question.

(iii) The information in each column of Z pertains to a particular response category of a certain question. We assume that there are no columns consisting purely of "0's" since we will have deleted such categories in the preliminary statistical analysis.

If each subject answers each question:

(iv) There will be NxM "1's" scattered throughout Z. The rest of the entries will be "0".

(v) Each row of Z will sum to M, and the column sums will indicate the 1-way marginals of all the response categories for all the variables /questions .

The information in Z can be condensed by adding similar rows together and forming a new matrix, W (which we have termed the "weighted" multivariate indicator matrix) with dimension IxJ, where I is the number of cells in the hypothetical multidimensional contingency table and J is the total number

of response categories to all the questions. We assume that $I < N$ otherwise we will not be condensing Z. The "1's" in each row of the indicator matrix are replaced by the frequency of a particular cell in the contingency table. This procedure reduces one dimension of the input matrix for the multiple correspondence analysis and thus reduces the storage space and the computer time required to run the program. Justification for this move comes from the Principle of Distributional Equivalence quoted in Greenacre (in press, chapter 4) which states :

> "If two row profiles (say) are identical then the corresponding rows of the original data matrix may be replaced by their summation (a single row) without affecting the geometry of the column profiles."

Comments on the effect of performing correspondence analysis on W instead of Z will be made after we have defined a number of concepts such as mass and inertia.

We shall now investigate the multiple correspondence analysis of W. The material for this section comes from Greenacre (1978a, 1978b, 1981, 1982, in press).

Correspondence matrix

It is convenient to rescale W so that the sum of its elements is 1:

Let

$$P = [\, P_{ij} \,] = W/NM$$

where N and M are scalars.

Row and column sums of P

$$\underset{(IX1)}{\underset{\sim}{r}} = P\underset{\sim}{1}_{(J)} \quad \text{where } r_i > 0 \quad (i=1,2,..,I)$$

- is the vector of row masses corresponding not to individual subjects, but to groups of subjects (Principle of Distributional Equivalence).
- From now on when we refer to subjects we shall in fact mean 'groups of subjects'.
- $\underset{\sim}{1}$ is the appropriate vector of "1's".

$$\underset{(JX1)}{\underset{\sim}{c}} = P^T\underset{\sim}{1}_{(I)} \quad \text{where } c_j > 0 \quad (j=1,2,..,J)$$

- is the vector of column masses. It is proportional to the 1-way marginals of all the response categories of all the questions/variables.

$$D_r = \text{diag}(r_1, r_2, \ldots, r_I)$$

$$D_c = \text{diag}(c_1, c_2, \ldots, c_J)$$

Row and column profiles

The row and column profiles refer to the rows and columns of P divided by their respective totals :

$$R = D_c^{-1} P \qquad \text{....row profile}$$

$$C = D_c^{-1} P^T \qquad \text{....column profile}$$

The row and column profiles define the I and J points in J- and I-dimensional vector spaces respectively. These profiles are assigned their respective (chi-square) metrics defined by D_c^{-1} and D_r^{-1} in their respective spaces. (See Greenacre and Underhill, 1982). The problem, according to Greenacre (1982:8) is

> "..posed of finding the principal axes of the points in each of those spaces, that is finding the co-ordinates F and G of the points with respect to the p-dimensional subspaces which are closest to the profile points in terms of minimum weighted sum of squared distance, where the points are weighted by their masses."

The co-ordinate matrices F and G which satisfy this criterion are derived from the p right and left (generalized) singular vectors respectively of $P - \underset{\sim\sim}{r}c^T$, in the metrics D_r^{-1} and D_c^{-1}, corresponding to the largest p singular values.

The multiple correspondence analysis in p dimensions of W

Given that

$$\tilde{A} \equiv D_r^{-\frac{1}{2}} (P - \underset{\sim}{r}\underset{\sim}{c}^T) D_c^{-\frac{1}{2}}$$

(IxJ)

Then the singular-value decomposition of A is

$$\tilde{U}\, \tilde{D}_\mu\, \tilde{V}^T$$

where

$\tilde{U}$ is an IxI matrix

$\tilde{D}_\mu$ is an IxJ matrix

$\tilde{V}$ is an JxJ matrix.

The multiple correspondence analysis in p dimensions of W is obtained from the rank p basic structure of :

$\tilde{A} \simeq \hat{A}_{(p)} = U\, D_\mu V^T$ - for ease of notation we denote $\hat{A}_{(p)}$ by A

where

(i) U, V and D_μ are components of $\tilde{U}$, $\tilde{V}$ and $\tilde{D}_\mu$

(ii) D_μ is a diagonal (pxp) matrix with elements

$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_p > 0$

(iii) U (Ixp) and V (Jxp) are the left and right basic vectors of $\hat{A}_{(p)}$ and U and V contain information about the co-ordinates of the subjects and objects respectively.

The co-ordinates of the grouped-subjects (F) and the objects (G) in p-dimensional space

These are found from the rows of $D_r^{-\frac{1}{2}} U D_\mu$ and $D_c^{-\frac{1}{2}} V D_\mu$ respectively. A detailed step-by-step example of correspondence analysis is given in Chapter 5.

Inertia

A further concept discussed by Greenacre is that of inertia. The total inertia (In(total)) is a measure of the total variability of the elements of the data matrix. It is equal to (J/M)-1 (where J is the number of objects and M is the number of questions) and is decomposed along various principal axes (k=1,2,...,(J-M)). The moment of inertia for each axis is equal to the squared diagonal element (μ_k) of the matrix D_μ i.e. the square of the singular-values resulting from the decomposition of A.

The number of non-trivial dimensions/axes having positive inertia is equal to J-M. For example, in a situation where a questionnaire has 5 questions each with 3 possible response categories, the number of dimensions with positive inertia will be 10. Usually we will only be interested in examining the first few (say 3 or 4) principal axes and the plot positions of the subjects and objects on these. The percentage of inertia accounted for by looking at the first p principal axes is defined to be :

$$\tau_p = \frac{\sum_{k=1}^{p} \mu_k^2}{\text{in (total)}} \times 100$$

where in (total) = $\sum_{k=1}^{J-M} \mu_k^2$

and where J-M is the number of non-trivial singular-values of A. According to Greenacre (1978a:87-88), the total inertia may be broken up in the following ways:

$$\text{in(total)} = \sum_i \sum_j \left\{ \frac{(P_{ij} - r_i c_j)^2}{r_i c_j} \right\}$$

- the terms in curly brackets refer to the contribution of the ij-th data point to the total inertia.

$$= \sum_i \left\{ r_i \left[\sum_j \frac{1}{c_j} \left(\frac{P_{ij}}{r_i} - c_j \right)^2 \right] \right\}$$

- the terms in curly brackets refer to the contribution of the i-th subject to the total inertia. The quantity in square brackets is the squared chi-square distance of the subject profile to the centre of gravity c in the subject space $\mathbb{R}^J$ and r_i is the mass of the i-th subject (grouped subject).

$$= \sum_j \left\{ c_j \left[\sum_i \frac{1}{r_i} \left(\frac{P_{ij}}{c_j} - r_i \right)^2 \right] \right\}$$

- the terms in curly brackets refer to the contribution of the j-th object to the total inertia. The quantity in square brackets is the squared chi-square distance of the object profile to the centre of gravity r in the subject space $\mathbb{R}^I$ and c_j is the mass of the j-th object.

$$= \sum_k \{ \mu_k^2 \}$$

where $1 \geq \mu_1^2 \geq \mu_2^2 \geq \ldots \geq \mu_{J-M}^2 \geq 0$

-the term in curly brackets is the contribution of the k-th axis to to the total inertia.

$= \sum_k [\sum_i \{r_i f_{ik}^2\}]$ -the term in curly brackets is the contribution of the i-th subject (grouped) to the inertia along the k-th axis.

$= \sum_k [\sum_j \{c_j g_{jk}^2\}]$ -the term in curly brackets is the contribution of the j-th object to the inertia along the k-th axis.

The terms above in curly brackets are known as 'absolute' contributions. There is another group of contributions termed 'relative' contributions. This consists of the contributions of the axes to the inertia of the subject or object points :

$\frac{f_{ik}^2}{\sum_k f_{jk}^2}$ is the relative contribution of the k-th axis to the inertia of the i-th subject (row)

$\frac{g_{jk}^2}{\sum_k g_{jk}^2}$ is the relative contribution of the k-th axis to the inertia of the j-th object (column)

Transition formulae

Furthermore, Greenacre shows that the co-ordinates of the subjects can be calculated from those of the objects and visa versa. The two (transition)

formulae whereby this can be achieved are given below:

$$F = D_r^{-1} PGD_\mu$$

$$G = D_c^{-1} P^T FD_\mu$$

These formulae highlight the interrelationships between subjects and objects. For example, from the equation for calculating G from F, we see that the co-ordinate of the j-th object along axis k is the weighted centre of gravity of all subject points on that axis and as a result an object is attracted to those subjects for which the column (object) profile is large. From this, and the fact that the inertias and their decompositions along the k-th axis are the same, it seems justifiable that both subjects and objects be displayed on the same set of axes.

The effect of performing correspondence analysis on W instead of Z (i.e. of invoking the Principal of Distributional Equivalence)

Since the column (object) profiles are unaffected by performing correspondence analysis on W as opposed to Z (see definition of this principle),

(i) the masses, inertia, absolute and relative contributions of the objects remain the same,

(ii) exactly the same graphical display of the objects is yielded, and in addition,

(iii) the inertia of the various principal axes remains unchanged.

Thus no information on the questions is lost by performing correspondence analysis on the weighted form of Z. All that is lost is information on individual subjects. This has been replaced by information on groups of subjects falling into a particular cell of a multidimensional contingency table. In terms of our goal, the fitting of log-linear models, the information on groups of subjects may be more meaningful than the information on individual subjects which has been replaced.

3.2.4 Output from the computer program

In chapter 4, in our efforts to show how correspondence analysis can provide an indication of the structure of data sets in terms of the interrelationships between variables, numerous references will be made to the output from a computer program for correspondence analysis written by Tabet (1973). We now give a brief indication of the output from performing correspondence analysis on a typical data matrix using this program. See Greenacre (1978a) for more detail.

The output can essentially be divided into 5 sections :

(i) A printout of the data matrix, along with row and column sums.

(ii) Information relating to the decomposition of the total inertia (see for example Table 12, section 4.3.1). A table is printed

which provides amongst other things, information on the moments of inertia (under the heading 'EIGENVALUE'..these have been squared), each moment of inertia expressed as a percentage of the total inertia (under 'PERCENT'), the cumulated percentages of inertia (under 'CUMUL') and a histogram of inertia relative to the first moment of inertia (i.e. eigenvalue number 2).

(iii) Information on the singular-value decomposition of A and in in particular the elements of V from which we can determine the co-ordinates of the objects.

(iv) The decomposition of the largest moments of inertia for each of the sets of subjects and objects (see for example Table 14, section 4.3.1) in separate tables. The column marked QLT gives the cosine squared of the angle the point makes with the subspace defined by the first p (5 in this case) principal axes. For each point, the mass and inertia are given in the columns headed 'MASS' and 'INR' respectively. The co-ordinate on the axis, the relative contribution (COR), and absolute contribution (CTR) for each point (subject or object) is given for each principal axis. Some of the values in these tables are multiplied by 1000 to make printing and visual inspection easier.

(v) A plot of the positions of the subjects and objects in the plane of any two principal axes (e.g. axes 1 and 2 in Figure 5, section 4.3.1). The subjects (in fact the 'group' of subjects falling into different cells of a multi-dimensional contingency table) are designated by an "O" (by choice) and the objects by an "X" (also

by choice). It is possible to suppress the printing of the display positions of either the subjects or objects if we are only interested in the one set and do not want the graphical display to get too cluttered. This is what we have done with the subjects in the other graphical displays (Figures 6-10).

3.3 The connection between Correspondence Analysis and log-linear model building

Having discussed these two procedures, the one being essentially exploratory and the other confirmatory, with special emphasis on questionnaire data, we proceed in Chapters 4 and 5 to ascertain whether they can be used in conjunction.

CHAPTER 4

4. USING CORRESPONDENCE ANALYSIS TO BUILD LOG-LINEAR MODELS

4.1 Introduction

Data collected in the bio-medical and social sciences by means of questionnaires are in most instances qualitative in nature. Examples include demographic information such as sex, race and marital status; indications of the presence or absence of particular medical or social conditions; clinical presentation of illnesses or manifestations of social conditions; information relating to etiological factors and variables relating to the effects of intervention strategies. Even variables which are quantitative by nature, such as age, are often categorized.

In principle an entire questionnaire with M questions could be summarized by forming an M-dimensional contingency table and fitting a log-linear model with all main effects and appropriate interactions to it. However, in practice sample sizes are usually too small to attempt this and the information is summarized by a number of tables of lower order, probably with dimension at most 4 or 5. From a questionnaire with M questions there are 2^r-1 ways of forming tables of dimension $r=1,2,\ldots,M$ and one is faced with the problem of deciding which of these will most usefully summarize the information in the data. Naturally, tables which would allow the testing of á priori hypotheses would always be constructed, but many others remain.

The question we ask is: can correspondence analysis help us in

(i) selecting variables which can be meaningfully used for forming r-way tables, and

(ii) deciding on the particular log-linear model that should be fitted to these tables?

4.2 Method of investigation

4.2.1 Outline

To investigate the two questions asked above, we constructed contingency tables with a known hierarchical structure and formed data sets from these. We then used the data sets in a correspondence analysis program (Tabet, 1973) and determined the appropriate log-linear model by examining the tabular output and graphical displays.

4.2.2 Construction of the tables

Initially we considered only four variables (A,B,C and D) and later extended our investigation to include eight (A to H). The variables, their indices, number of levels and degrees of freedom are given in Table 2.

Table 2 : Summary of the variables (A to H)

variable	index	number of levels	df
A	i	3	2
B	j	2	1
C	k	2	1
D	l	2	1
E	m	2	1
F	n	2	1
G	o	2	1
H	p	2	1

The notation to be used will be as follows :

$f_{ijk\ell}$ = observed occurrence in cell $ijk\ell$ of the 4-way contingency table

= the number of counts simultaneously falling into category i of A, j of B, k of C and ℓ of D.

$f_{ijk\ell mnop}$ = observed occurrence in cell $ijk\ell mnop$ of the 8-way contingency table

= the number of counts simultaneously falling into category i of A, j of B, k of C, ℓ of D, m of E, n of F, o of G and p of H.

We define the 1-, 2- and 3-way marginals in the following manner:

$f_{i...}$	= the sum over j,k and l of $f_{ijk\ell}$	1-way marginal
$f_{ij..}$	= the sum over k and l of $f_{ijk\ell}$	2-way marginal
$f_{ijk.}$	= the sum over l of $f_{ijk\ell}$	3-way marginal

Some hierarchical models are known as "direct" models. In these the cell entries can be calculated directly from the lower order marginal subtables. In addition, such models often have a simple probabilistic interpretation. In "indirect" models the cell entries cannot be expressed as a function of the marginals, but have to be calculated using an iterative procedure. No models of this type were used in our study.

Bishop et al. (1975) list the direct models for a 4-way table. To construct the 4-way contingency tables we used the cell estimates corresponding to these direct models. We constructed the 8-way tables by a slight adjustment of the direct cell estimates corresponding to pairs of 4-way models. The direct 4-way models are given in Table 3 and their inter-relationships are shown diagramatically in Figure 3. Some of the possible direct 8-way models are given in Table 4 and their inter-relationships are shown in Figure 4.

TABLE 3 : SOME 4-WAY MODELS WITH DIRECT CELL ESTIMATES

Model	Direct cell estimate
1. A,B,C,D	$f_{i...} f_{.j..} f_{..k.} f_{...\ell} / N^3$
2. BC,A,D	$f_{.jk.} f_{i...} f_{...\ell} / N^2$
3. AD,B,C	$f_{i..\ell} f_{.j..} f_{..k.} / N^2$
4. AC,AD,B	$f_{i.k.} f_{i..\ell} f_{.j..} / f_{i...} N$
5. BC,AD	$f_{.jk.} f_{i..\ell} / N$
6. AC,AD,BD	$f_{i.k.} f_{i..\ell} f_{.j.\ell} / f_{i...} f_{...\ell}$
7. AB,AC,AD	$f_{ij..} f_{i.k.} f_{i..\ell} / f_{i...}^2$
8. ABC,D	$f_{ijk.} f_{...\ell} / N$
9. ABC,AD	$f_{ijk.} f_{i..\ell} / f_{i...}$
10. ABC,ABD	$f_{ijk.} f_{ij.\ell} / f_{ij..}$

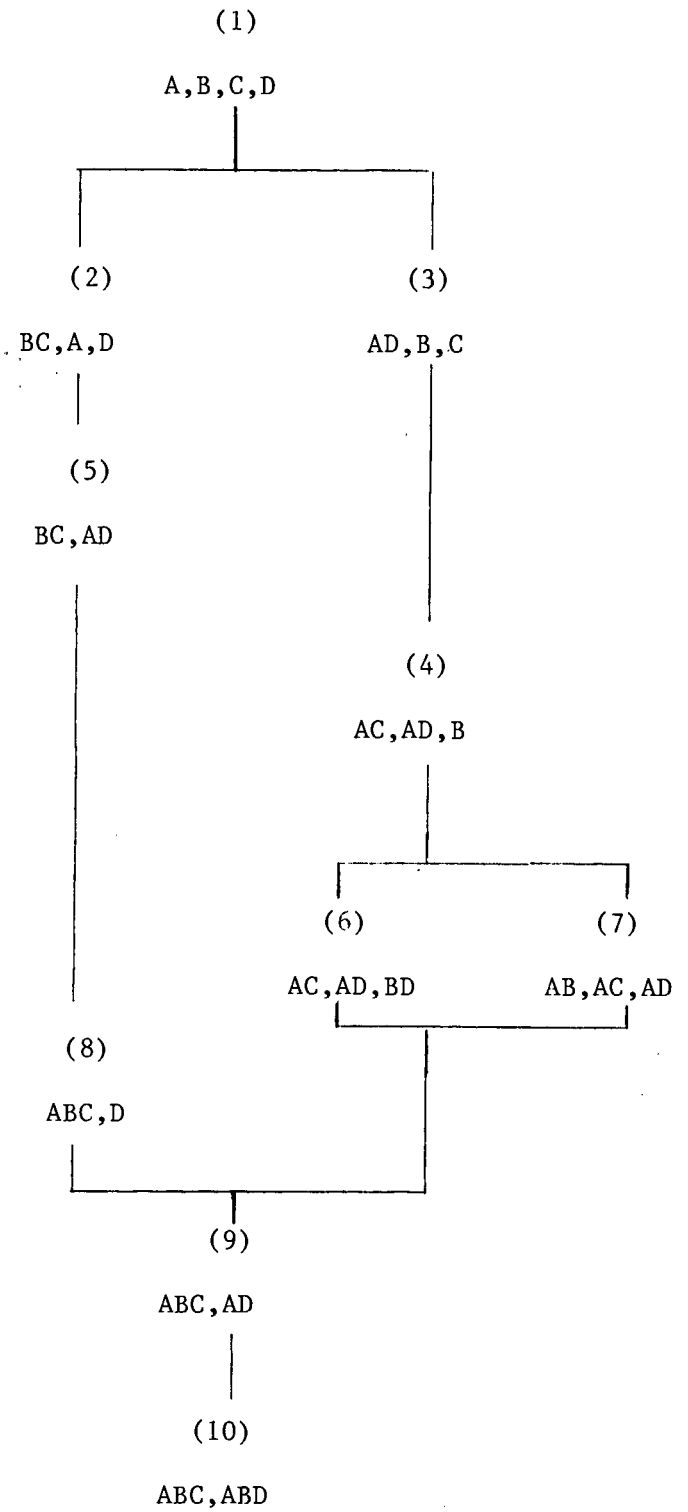


Figure 3 : Diagramatical representation of the inter-relationships between the 4-way models in Table 3 in terms of increasing complexity

TABLE 4 : SOME 8-WAY MODELS WITH DIRECT CELL ESTIMATES

Model	Direct cell estimate
11. A,B,C,D,E,F,G,H	$f_{i.....} f_{.j....} f_{..k...} f_{...\ell..} f_{....m..}$ $f_{.....n.} f_{......o} f_{.......p} / N^7$
12. BC,AD,EF,EG,H	$f_{.jk....} f_{i..\ell..} f_{....mn.} f_{....m.o}$ $f_{.......p} / f_{....m...} N^3$
13. ABC,D,EF,EG,FH	$f_{ijk....} f_{...\ell..} f_{....mn.} f_{....m.o}$ $f_{.....n.p} / f_{....m...} f_{.....n..} N^2$
14. ABC,AD,EF,EG,EH	$f_{ijk....} f_{i..\ell..} f_{....mn.} f_{....m.o}$ $f_{....m..p} / f_{i.....} f^2_{....m...} N$
15. ABC,ABD,EFG,EH	$f_{ijk....} f_{ij.\ell..} f_{....mno} f_{....m..p} / N$ $f_{ij.....} f_{....m...}$

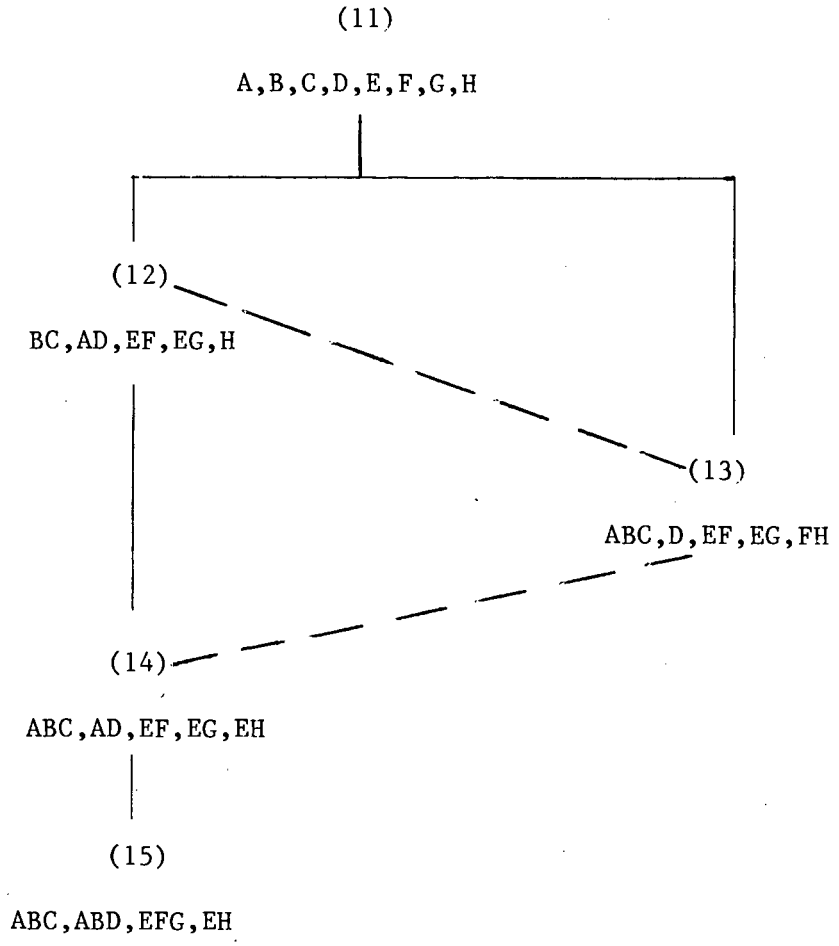


Figure 4 : Diagramatical representation of the inter-relationships between the 8-way models in Table 4 in terms of increasing complexity

1-way marginals were decided upon for each of the variables (see Table 5) and were used throughout. In choosing the marginals we tried to cover a selection of 1-way probability structures, ranging from the equally likely (variable C) to the more extreme case (variable F) where level 1 had a probability of 0,1 and level 2 a probability of 0,9.

Table 5 : Table of 1-way marginals for variables A to H

variable	level 1	level 2	level 3	total
A	3000	4000	3000	10000
B	6000	4000		10000
C	5000	5000		10000
D	7000	3000		10000
E	2000	8000		10000
F	1000	9000		10000
G	3000	7000		10000
H	4000	6000		10000

Cell entries for a table in which all variables are independent can be constructed from the product of the 1-way marginals. In addition, since the inclusion of an interaction term, λ_{ij}^{AB} say (see Bishop et al.,1975), in the log-linear model, implies that the AB marginals of the fitted model are equal to the observed AB marginals, a multidimensional table with an AB interaction can be constructed by defining the AB marginals suitably. We illustrate the method used for setting up the two-way marginals by refering to the 2-way AB marginal table. For A and B independent, the second-order AB marginals are given in the following table :

Table 6 : Second-order AB marginal table (A and B independent)

	A1	A2	A3	1st order marginals
B1	1800	2400	1800	6000
B2	1200	1600	1200	4000
1st order marginals	3000	4000	3000	10000

where each entry $f_{ij..} = f_{i...} f_{.j..} /N$

However, in cases where we need the AB marginals to calculate direct cell estimates, such as in model (2), A and B are in fact dependent and the 2-way marginals should differ from those found in the independence model given above. Since Table 7 has 2 degrees of freedom, two of the cell entries can be assigned arbitrarily (marked with an *) and the remaining entries are defined by the given marginal constraints.

Table 7 : Second-order AB marginal table (A and B dependent)

	A1	A2	A3	1st order marginals
B1	800*	3000*	2200	6000
B2	2200	1000	800	4000
1st order marginals	3000	4000	3000	10000

Other 2-way interactions were constructed in the same way.

In order to construct 3-way interaction tables keeping the 1- and 2-way marginals consistent throughout, the procedure discussed below was developed. We illustrate the procedure used by refering to the 3-way ABC marginal table (see Table 8).

Table 8 : Third-order ABC marginal table (A,B and C dependent)

8(a)

		C1	C2	1-st order marginals
A1	B1	$f_{111.}$	$f_{112.}$	
	B2	$f_{121.}$	$f_{122.}$	3000
A2	B1	$f_{211.}$	$f_{212.}$	
	B2	$f_{221.}$	$f_{222.}$	4000
A3	B1	$f_{311.}$	$f_{312.}$	
	B2	$f_{321.}$	$f_{322.}$	3000
1st order marginals		5000	5000	10000

8(b)

		C1	C2	1-st order marginals
A1	B1	300*	500	
	B2	700	1500	3000
A2	B1	2400*	600	
	B2	600	400	4000
A3	B1	400	1800	
	B2	600	200	3000
		5000	5000	10000

The cell frequencies in Table 8(b) were obtained by solving the 24 linear equations given below and using the fact that the third order ABC marginals have two degrees of freedom. We thus selected to define the two marginal totals marked with an *, bearing in mind that the lower order marginals still had to be satisfied : namely,

for the AB marginals

$$\sum_{k\ell} f_{ijk\ell} = f_{ij..} \qquad i=1,2,3 \qquad j=1,2$$

for the AC marginals

$$\sum_{j\ell} f_{ijk\ell} = f_{i.k.} \qquad i=1,2,3 \qquad k=1,2$$

for the BC marginals

$$\sum_{i\ell} f_{ijk\ell} = f_{.jk.} \qquad j=1,2 \qquad k=1,2$$

and the 1-way marginals

$$\sum_{jk\ell} f_{ijk\ell} = f_{i..} \qquad i=1,2,3$$

$$\sum_{ik\ell} f_{ijk\ell} = f_{.j..} \qquad j=1,2$$

$$\sum_{ij\ell} f_{ijk\ell} = f_{..k.} \qquad k=1,2$$

$$\sum_{ijk} f_{ijk\ell} = f_{...\ell} \qquad \ell=1,2$$

and the grand total

$$\sum_{ijk\ell} f_{ijk\ell} = f_{....}$$

Once all the cell frequencies had been obtained, the significance of the ABC interaction was tested using a BMDP P4f computer program. This criterion was satisfied at the 5% level of significance. Other 3-way interactions were constructed in a similar manner.

Listings of the marginal tables are given in Tables 22 to 33 (see Appendix A1). Using these marginals and the procedure outlined above, the direct cell estimates for the models given in Tables 3 and 4 were calculated.

In some of the tables minor adjustments were made to round off the cell frequencies to whole numbers. If we recall that the Poisson variation of the cell counts is proportional to the square root of the count, we realize that the round-off errors introduced will be negligible. In order to check the effect this rounding would have on our procedure, we fitted the correct

log-linear model in each case and calculated the log-likelihood ratio chi-square goodness-of-fit statistic using the BMDP P4f computer program. In the 4-dimensional models the likelihood ratio chi-square statistics were close to zero while in the 8-dimensional cases the rounding errors were extremely small relative to the degrees of freedom of the chi-square statistic. The results are given in Table 9 below :

Table 9 : Likelihood ratio chi-square statistics for the 15 models

Model	df	L.R. χ^2 statistic
1	18	0.0
2	17	-0.04
3	16	0.0
4	14	-0.02
5	15	0.0
6	13	0.01
7	12	0.01
8	11	0.0
9	9	0.01
10	6	0.01
11	374	11.32
12	369	20.33
13	364	26.66
14	362	27.63
15	357	29.06

Apart from this, no other random variation of the cell frequencies was introduced since it was felt at this stage that random variation would only blur the structure suggested by the correspondence analysis.

4.2.3 Construction of data matrix W

For ease of exposition we shall consider only the 4-dimensional table, 3x2x2x2 in our case. Extension to the 8-dimensional table follows a similar pattern - only the number of subscripts needs to be increased.

The entry $f_{ijk\ell}$ in cell ijkℓ of the table is the number of subjects, who fall into category i of A, j of B, k of C, and ℓ of D. Thus we can construct an NxJ indicator matrix Z (where N = the $\sum_{ijk\ell} f_{ijk\ell}$ = the number of subjects and J = the total number of response categories to all the questions) whose columns are the levels of the variables A to D (one column per level) and whose rows correspond to the subjects. The first f_{1111}, the second f_{1112}, ..., the last f_{IJKL} rows are the same. The information in Z can be condensed by adding the identical rows together and forming a new matrix with dimension IxJxKxL (the number of cells in the 4-way contingency table - 3x2x2x2 = 24 in the example) by J (defined above), where the non-zero row entries are $f_{ijk\ell}$ - by invoking the Principal of Distributional Equivalence (see section 3.2.3). We call this new matrix the "weighted" multivariate indicator matrix W.

4.3 Results

4.3.1 Main findings

Correspondence analysis was performed on each of the 15 data sets using a program developed by Tabet (1973) (refer to section 3.2.4 for comments on the output of a correspondence analysis program). The following criteria were used for selecting models using the output from the correspondence analysis:

(1) In data sets where each of the principal axes of inertia accounted for the same or nearly the same percentage of the total inertia, this was taken as evidence that there was very little structure in the data and that the variables in the data set were in fact independent.

(2) The contribution made by a variable/question to the inertia of a particular axis was found by summing the contribution made by each of its levels, (e.g. the contribution made by variable A was the sum of the contributions of the 3 levels (objects) A1, A2, and A3 to that axis).

(3) It was decided that a variable should be included in the model if its total contribution to the inertia of a particular axis was more than $1/(J-M) \times 100$ % of the total inertia (where J = the number of objects, M is the number of questions and J-M is the maximum number of non-trivial dimensions with positive inertia (Greenacre, op. cit.)).

Only variables which contributed more than 20% to the inertia of an axis in the case of 4-way models, and 11%, in the case of 8-way models, were therefore included in the proposed model.

(4) In instances where several variables satisfied the criterion given in (3) for a particular axis, we assumed that these variables interacted (i.e. the correspondence analysis indicates a model which includes these interactions). In terms of the graphical display, if variables are dependent, i.e. they interacted, then they fall along the same axes on the graphical display.

(5) If, however, variables or sets of variables, fall on different axes, they are considered to be independent (a variable falling on to a particular axis implies that all the levels of the variable fall on the axis). Two sets of variables, e.g. A & B and C & D are therefore considered independent if A & B and C & D contribute significantly to the inertia of different axes.

Using these criteria, the output from the 15 correspondence analyses was examined without prior knowledge of the true model. The results are tabulated in Tables 10 and 11. Table 10 contains information on the decomposition of the total inertia for each of the 15 data sets with known structure (underlying model). For each model we are particularly concerned with those axes which account for approximately 100/J-M % of the total inertia. These axes have been marked with a "+" in Table 10. In addition,

TABLE 10 : DECOMPOSITION OF THE TOTAL INERTIA ALONG THE PRINCIPAL AXES FOR MODELS 1-15

MODEL	DESCRIPTION	% OF THE TOTAL INERTIA EXPLAINED BY AXIS NUMBER								
		1	2	3	4	5	6	7	8	9
1	A,B,C,D	20,0	20,0	20,0	20,0	20,0				*
2	BC,A,D	20,8+	20,0+	20,0+	20,0+	19,2				
3	AD,B,C	30,5+	20,0+	20,0+	20,0+	9,5				
4	AC,AD,B	33,3+	24,7+	20,0+	13,0	9,0				
5	BC,AD	30,5+	20,8+	20,0+	19,2	9,5				
6	AC,AD,BD	33,5+	24,8+	19,7+	13,0	9,0				
7	AB,AC,AD	33,3+	31,5+	16,0	10,3	9,0				
8	ABC,D	31,1+	25,2+	20,0+	14,0	9,7				
9	ABC,AD	33,3+	31,0+	17,0	9,7	9,0				
10	ABC,ABD	33,8+	30,9+	16,3	9,7	9,3				
11	A,B,C,D,E,F,G,H	11,2	11,2	11,2	11,1	11,1	11,1	11,1	11,1	11,0 *
12	BC,AD,EF,EG,H	18,7+	16,9+	11,6+	11,1+	11,1+	10,7	9,5	5,3	5,2
13	ABC,D,EF,EG,FH	17,3+	14,7+	14,1+	14,0+	11,1+	8,1	7,8	7,5	5,4
14	ABC,AD,EF,EG,EH	20,3+	18,5+	17,2+	10,0	9,4	9,3	5,4	5,0	4,9
15	ABC,ABD,EFG,EH	19,6+	18,7+	17,2+	11,0+	9,6	9,1	5,3	5,2	4,3

* - indicates independence

TABLE 11 : PRINCIPLE CONTRIBUTORS TO THE INERTIA OF THE SIGNIFICANT AXES AND FITTING OF THE MODELS SUGGESTED BY CORRESPONDENCE ANALYSIS

MODEL	PRINCIPLE CONTRIBUTORS TO THE INERTIA OF AXIS					MODEL SUGGESTED BY CORRESPON-DENCE ANALYSIS	LR. CHI SQ STATISTIC OF SUGGESTED MODEL		DF	PROB
	1	2	3	4	5					
1:A,B,C,D ≠						A,B,C,D	0,0	α	18	1,0000
2:BC,A,D	BC	A	D	A		BC,A,D	-0,4	α	17	1,0000
3:AD,B,C	AD	A	C	B		AD,B,C	0,0	α	16	1,0000
4:AC,AD,B	ACD	AC	B			ACD,B	-0,02	**	11	1,0000
5:BC,AD	AD	BC	A			BC,AD	0,0	α	15	1,0000
6:AC,AD,BD	AD	AC	B			AD,AC,B	80,15	*	14	0,0000
7:AB,AC,AD	ACD	AB	(BCD)			ACD,AB,BCD	0,01	**	6	1,0000
8:ABC,D	ABC	ABC	D			ABC,D	0,0	α	11	1,0000
9:ABC,AD	AD	AB	(BC)			AB,BC,AD	2701,22	*	13	0,0000
10:ABC,ABD	AD	AB	(BC)			AB,BC,AD	3207,01	*	13	0,0000
11:A,B,C,D,E,F,G,H ≠						A,B,C,D,E,F,G,H	11,32	α	374	1,0000
12:BC,AD,EF,EG,H	EFG	AD	BC	AH	AH	BC,AD,EFG,AH	20,44	**	365	1,0000
13:ABC,D,EF,EG,FH	ABC	EG	FH	ABC	D	ABC,D,EG,FH	2660,10	*	366	0,0000
14:ABC,AD,EF,EG,EH	EFGH	ACD	ABC	(GH)		ABC,ACD,EFGH	27,56	**	351	1,0000
15:ABC,ABD,EFG,EH	EFGH	ACD	ABC	FG		ABC,ACD,EFGH	481,85	**	351	1,0000

* The suggested model has too few terms (it is under-fitted)
** The suggested model has too many terms (it is over-fitted)
α The suggested model is correct
≠ Independence model

Table 11 contains information relating to the fit of the models 'suggested' by correspondence analysis using the criteria outlined above. Tables 34 and 35 (see Appendix A2) give the % contribution of each variable to the inertia of the first five principal axes of inertia. Table 11 includes a summary of this information.

Comparing the results suggested by the correspondence analysis and the true models, we found that:

(i) where each of the principal axes of inertia accounted for the same percentage of the total inertia (criterion (1)), that the variables in the data set were indeed independent. This was found in the case of the data sets corresponding to models (1) and (11) (see Tables 12 and 13).

(ii) if two variables or groups of variables were independent then they contributed significantly to the inertia of different axes. This was found for the following models: (2),(3),(4),(5),(8),(12), (13),(14) and (15).

Consider for example model 5 : BC,AD. From the output given in Table 14 and Figure 5, we observe that:

a) Variables A and D each account for 50% of the inertia of axes 1 and 5 (see columns headed 'CTR') while variables B and C each account for 50% of the inertia of axis 2 and 4. Variable A accounts for all the inertia in axis 3. See also Table 34, Appendix A2.

Table 12 : Model (1) : A, B, C, D
— the moments of inertia and their percentage of the total inertia

THE EIGENVALUES VAL (1) = 0.99999517

NUM	ITER	EIGENVALUE	PERCENT	CUMUL	*	HISTOGRAM OF THE EIGENVALUES (MOMENTS OF INERITIA)
2	1	0.25000054	20.000	20.000	*	••••••••••••••• / ••••••••••••••• / ••••••••••••••• / •••••••••••••••
3	1	0.25000012	20.000	40.000	*	••••••••••••••• / ••••••••••••••• / ••••••••••••••• / •••••••••••••••
4	4	0.24999988	20.000	60.000	*	••••••••••••••• / ••••••••••••••• / ••••••••••••••• / •••••••••••••••
5	1	0.24999940	20.000	80.000	*	••••••••••••••• / ••••••••••••••• / ••••••••••••••• / •••••••••••••••
6	1	0.24999857	20.000	100.000	*	••••••••••••••• / ••••••••••••••• / ••••••••••••••• / •••••••••••••••
7	0	0.00000024	0.000	100.000	*	
8	2	0.00000003	0.000	100.000	*	
9	3	-.00000005	-0.000	100.000	*	

We note that each axis accounts for 20% of the inertia

Table 13 : Model (11) : A, B, C, D, E, F, G, H
— the moments of inertia and their percentage of the total inertia

THE EIGENVALUES VAL (1) = 0.99999672

NUM	ITER	EIGENVALUE	PERCENT	CUMUL	*	HISTOGRAM OF THE EIGENVALUES (MOMENTS OF INERITA)
2	2	0.12583649	11.186	11.186	*	••••••••••••••• / ••••••••••••••• / ••••••••••••••• / •••••••••••••••
3	2	0.12572253	11.176	22.361	*	••••••••••••••• / ••••••••••••••• / ••••••••••••••• / •••••••••••••••
4	2	0.12544096	11.151	33.512	*	••••••••••••••• / ••••••••••••••• / ••••••••••••••• / •••••••••••••••
5	2	0.12522262	11.131	44.643	*	••••••••••••••• / ••••••••••••••• / ••••••••••••••• / •••••••••••••••
6	2	0.12513393	11.123	55.766	*	••••••••••••••• / ••••••••••••••• / ••••••••••••••• / •••••••••••••••
7	1	0.12505525	11.116	66.882	*	••••••••••••••• / ••••••••••••••• / ••••••••••••••• / •••••••••••••••
8	2	0.12498218	11.110	77.992	*	••••••••••••••• / ••••••••••••••• / ••••••••••••••• / •••••••••••••••
9	1	0.12456810	11.073	89.065	*	••••••••••••••• / ••••••••••••••• / ••••••••••••••• / ••••••••••••••
10	0	0.12301737	10.935	100.000	*	••••••••••••••• / ••••••••••••••• / ••••••••••••••• / ••••••••••••••
11	0	0.00000029	0.000	100.000	*	
12	4	0.00000001	0.000	100.000	*	
13	1	-.00000003	-0.000	100.000	*	
14	4	-.00000007	-0.000	100.000	*	
15	2	-.00000022	-0.000	100.000	*	
16	2	-.00000032	-0.000	100.000	*	
17	1	-.00000052	-0.000	100.000	*	

We note that each axis accounts for approximately 11% of the total inertia

- refer to section 3.2.4 for details on the output of the correspondence analysis program

FIGURE 5 : MODEL(5)
BC,AD

PROJECTION OF SUBJECTS & OBJECTS ON AXES 1 AND 2

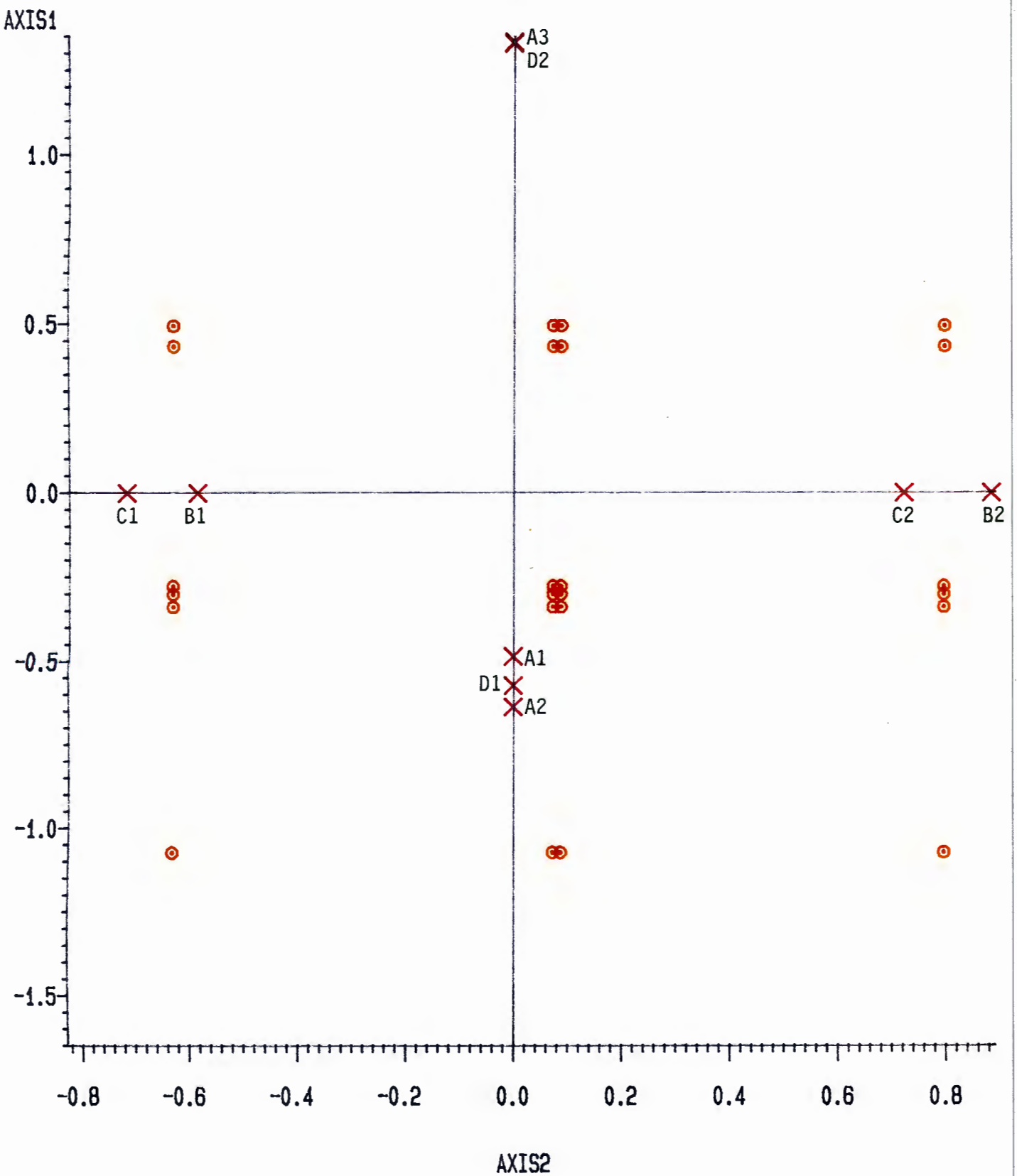


Table 14 : Model (5) : BC, AD
— Decomposition of the first 5 moments of inertia in terms of the objects

JI	NAME	QLT	MASS	INR	1#F	COR	CTR	2#F	COR	CTR	3#F	COR	CTR	4#F	COR	CTR	5#F	COR	CTR
1	A1	999	75	140	484	100	46	0	0	0	-1423	868	608	0	0	0	270	31	46
2	A2	1000	100	120	635	269	106	0	0	0	986	647	388	0	0	0	354	84	106
3	A3	1000	75	140	-1329	759	348	0	0	0	110	5	4	0	0	0	-741	236	348
4	B1	1000	150	80	0	0	0	589	520	200	0	0	0	565	480	200	0	0	0
5	B2	1000	100	120	0	0	0	-883	520	300	0	0	0	-847	480	300	0	0	0
6	C1	1000	125	100	0	0	0	721	520	250	0	0	0	-692	480	250	0	0	0
7	C2	1000	125	100	0	0	0	-720	520	250	0	0	0	693	480	250	0	0	0
8	D1	1000	175	60	572	763	150	0	0	0	0	0	0	0	0	0	-318	237	150
9	D2	1000	75	140	-1333	763	350	0	0	0	0	0	0	0	0	0	744	237	350
K =			•4000E + 05	1000			1000			1000			1000			1000			1000

Table 15 : Model (14) : ABC, AD, EF, EG, EH
— Decomposition of the first 5 moments of inertia in terms of the objects

JI	NAME	QLT	MASS	INR	1#F	COR	CTR	2#F	COR	CTR	3#F	COR	CTR	4#F	COR	CTR	5#F	COR	CTR
1	A1	749	38	78	1	0	0	207	18	8	1304	729	330	0	0	0	70	2	2
2	A2	752	50	67	5	0	0	794	420	151	-700	328	127	1	0	0	77	4	3
3	A3	757	38	78	-6	0	0	-1264	686	288	-368	58	26	0	0	0	-171	13	10
4	B1	844	75	44	0	0	0	-106	17	4	-617	572	148	3	0	0	412	255	120
5	B2	844	50	67	2	0	0	161	17	6	927	572	222	-3	0	0	-617	255	180
6	C1	899	63	56	3	0	0	539	290	87	-441	195	63	-7	0	0	-642	414	244
7	C2	899	63	56	-2	0	0	-538	290	87	442	195	63	8	0	0	643	414	244
8	D1	812	88	33	4	0	0	513	614	111	117	32	6	1	0	0	266	166	59
9	D2	812	38	78	-8	0	0	-1196	614	258	-273	32	15	-2	0	0	-621	166	137
10	E1	729	25	89	1707	729	320	-11	0	0	0	0	0	-30	0	0	-8	0	0
11	E2	729	100	22	-426	729	80	3	0	0	0	0	0	8	0	0	2	0	0
12	F1	530	13	100	2177	528	260	-12	0	0	-3	0	0	-141	2	2	-36	0	0
13	F2	530	112	11	-241	528	29	1	0	0	0	0	0	16	2	0	4	0	0
14	G1	665	38	78	835	299	115	-5	0	0	0	0	0	-922	366	284	29	0	0
15	G2	665	87	33	-357	299	49	3	0	0	0	0	0	396	366	122	-12	0	0
16	H1	801	50	67	635	268	98	0	0	0	0	0	0	895	533	355	5	0	0
17	H2	801	75	44	-421	268	59	0	0	0	0	0	0	-595	533	236	-3	0	0
K =			• 8000E + 05	1000			1000			1000			1000			1000			1000

- refer to section 3.2.4 for details on the output of the correspondence analysis program

b) From the graphical display of the 1st and 2nd principal axes we observe that variables B & C and variables A & D fall on different axes.

These results are compatible with the findings of Gabriel (1981: 160-161). He was interested in using the biplot (see comments chapter 2) as a diagnostic tool for models of 2-way tables. He too attempted to use the graphical display as a means of inferring the mathematical model underlying a data matrix. Referring to row and column markers (the former relating to the one variable and the latter to the other) he noted that :

"if the row markers are seen to be collinear, and the column markers are also noted to be collinear, and two lines are at right angles to each other, one may infer that an additive model will fit the data closely i.e. $y_{ij} = \alpha_i + \beta_j$ (no interaction) for one set of the α_i's and β_j's ..... also, if one observes that all markers for, both rows and columns, are on one and the same line, it is obvious that the matrix is of rank one and so the model is $y_{ij} = \alpha_i\beta_j$ (i.e. interaction)." ..(author's comments in brackets)

This view was corroborated by the findings of our study (although (i) our data was qualitative and (ii) we looked at situations involving more than two variables), since firstly, when variables lie on different axes, the structure of the data set supports the view that the variables are in fact independent of each

other and secondly that when the variables lie along the same axis they are in fact dependent.

This feature is very useful as it can suggest a split of the variables into two or more independent sets. This was found to be the case in the 8-variable models when correspondence analysis was performed on the data sets relating to models (13), (14) and (15). We illustrate this using model (14):

ABC,AD,EF,EG,EH

From Table 15 we note that the two groups of variables (A,B,C,D and E,F,G,H) and subsets of these two groups never contribute significantly to the inertia of the same axes (see the columns headed 'CTR' and also Table 35, Appendix A2). From the graphical display of the 1st and 2nd principal axes in Figure 6, we observe that these two groups of variables fall on different axes. This was consistent in the graphical displays of the other axes.

Thus in a practical situation with 8 variables, we would split the variables into the two independent sets, A,B,C,D, and E,F,G,H, and fit models to each independently.

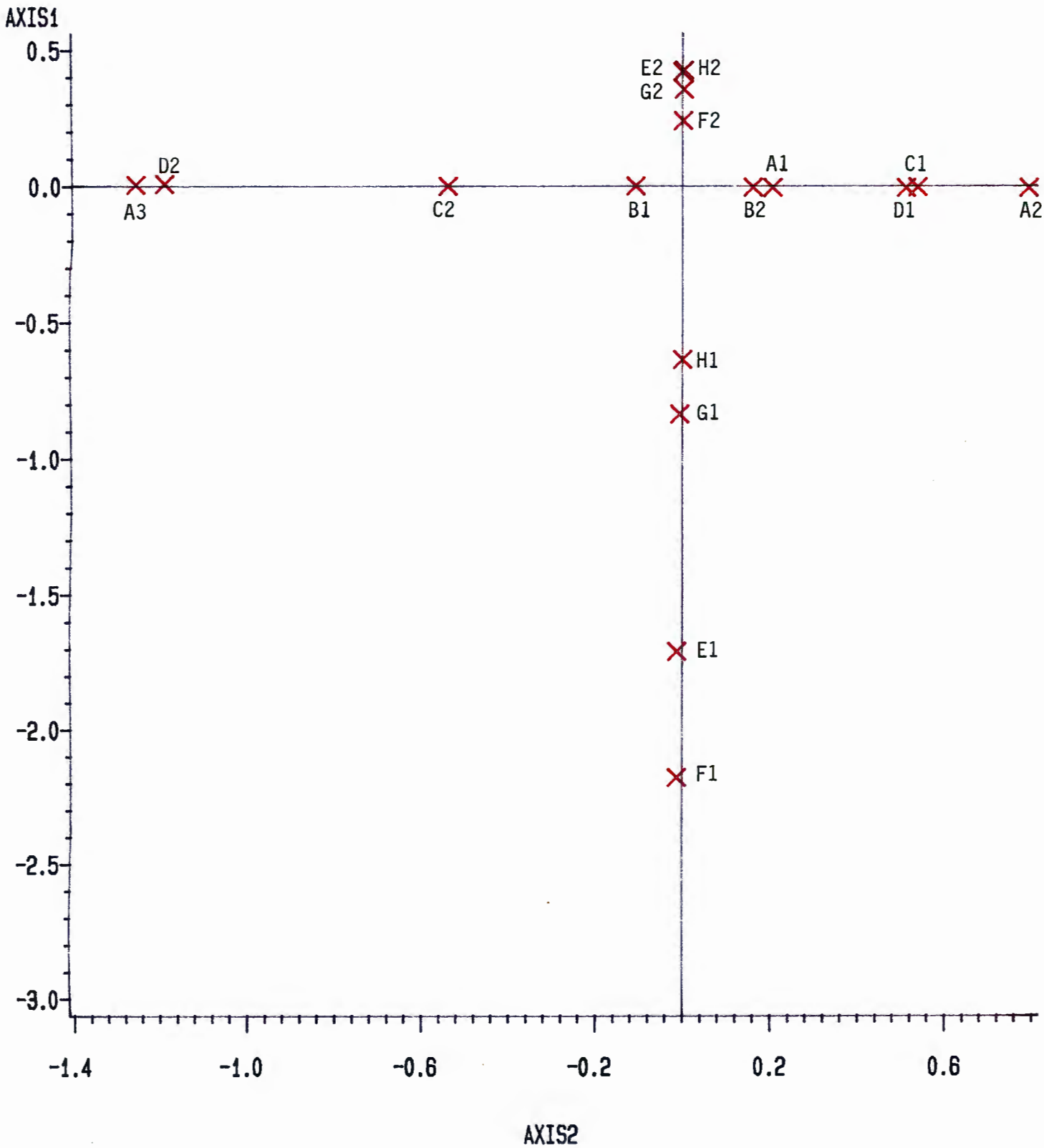
(iii) Correspondence analysis seems to detect 2-way interaction better than 3-way. This was evident most clearly in the case of the data sets corresponding to models (9) and (10) :

model (9) : ABC,AD

model (10) : ABC,ABD

in which the output from the correspondence analyses gave no clear indication of the 3-way interactions amongst variables (see Table 11). The view that correspondence analysis is not as

FIGURE 6 : MODEL(14)
ABC,AD,EF,EG,EH



sensitive to 3-way interactions as it is to 2-way, is also apparent when we compare the models suggested by the output from the correspondence analysis with the actual models underlying the data for models (14) and (15) (see Table 11).

However it is of interest to observe that in model (8) :

ABC,D

and model (13) :

ABC,D,EF,EG,FH,

the correspondence analysis detected the independence of ABC from D. In model (13) for example, we note (Table 16) that variables A,B & C account for all of the inertia of axes 1 and 4 while variables E & G, F & H and D account for all of the inertia of axes 2, 3 & 5 respectively. We also observe (Figures 7 and 8) that the three groups of variables A,B & C, E,F,G & H and D fall on different axes.

(iv) The correspondence analysis program sometimes detected indirect relationships between variables which were not directly related. For example in model (4) : AC,AD,B the output indicated that the model ACD,B be fitted to the data i.e. it detected a non-existant CD interaction.

(v) Using the criteria listed above,we found that in 40% of the cases (6 out of 15) the correct model (i.e. the model from which the data was generated) was suggested by the correspondence analysis. The correspondence analysis suggested models with too many interactions in 33,3% of the cases (5 out of 15), and under-fitted models in 26,7% of the cases (4 out of 15).

Table 16 : Model (13) : ABC, D, EF, EG, FH
— Decomposition of the first 5 moments of inertia in terms of the objects

JI	NAME	QLT	MASS	INR	1≠F	COR	CTR	2≠F	COR	CTR	3≠F	COR	CTR	4≠F	COR	CTR	5≠F	COR	CTR
1	A1	737	38	78	1183	599	270	-7	0	0	9	0	0	-567	138	77	0	0	0
2	A2	725	50	67	-905	547	211	4	0	0	-1	0	0	-515	178	84	0	0	0
3	A3	676	38	78	26	0	0	3	0	0	-5	0	0	1256	676	375	0	0	0
4	B1	707	75	44	-550	455	117	3	0	0	-8	0	0	410	252	80	0	0	0
5	B2	707	50	67	826	455	175	-4	0	0	13	0	0	-614	252	120	0	0	0
6	C1	686	63	56	-593	352	113	2	0	0	0	0	0	-577	334	132	0	0	0
7	C2	686	63	56	594	352	113	-1	0	0	0	0	0	578	334	132	0	0	0
8	D1	1000	88	33	0	0	0	-3	0	0	1	0	0	0	0	0	-654	1000	300
9	D2	1000	38	78	0	0	0	9	0	0	-1	0	0	0	0	0	1527	1000	700
10	E1	663	25	89	-9	0	0	-1630	663	400	-14	0	0	1	0	0	4	0	0
11	E2	663	100	22	2	0	0	406	663	100	4	0	0	0	0	0	0	0	0
12	F1	636	12	100	-21	0	0	-13	0	0	2396	636	450	16	0	0	-14	0	0
13	F2	636	113	11	2	0	0	2	0	0	-265	636	50	-1	0	0	2	0	0
14	G1	663	37	78	-7	0	0	-1243	663	350	-12	0	0	4	0	0	6	0	0
15	G2	663	86	33	4	0	0	533	663	150	6	0	0	-1	0	0	-2	0	0
16	H1	636	50	67	-8	0	0	-13	0	0	976	636	300	7	0	0	6	0	0
17	H2	636	75	44	6	0	0	9	0	0	-651	636	200	-4	0	0	-4	0	0
K =		●8000E + 05		1000			1000			1000			1000			1000			1000

- refer to section 3.2.4 for details on the output of the correspondence analysis program

FIGURE 7 : MODEL(13)
ABC,D,EF,EG,FH

PROJECTION OF OBJECTS ON AXES 1 AND 2

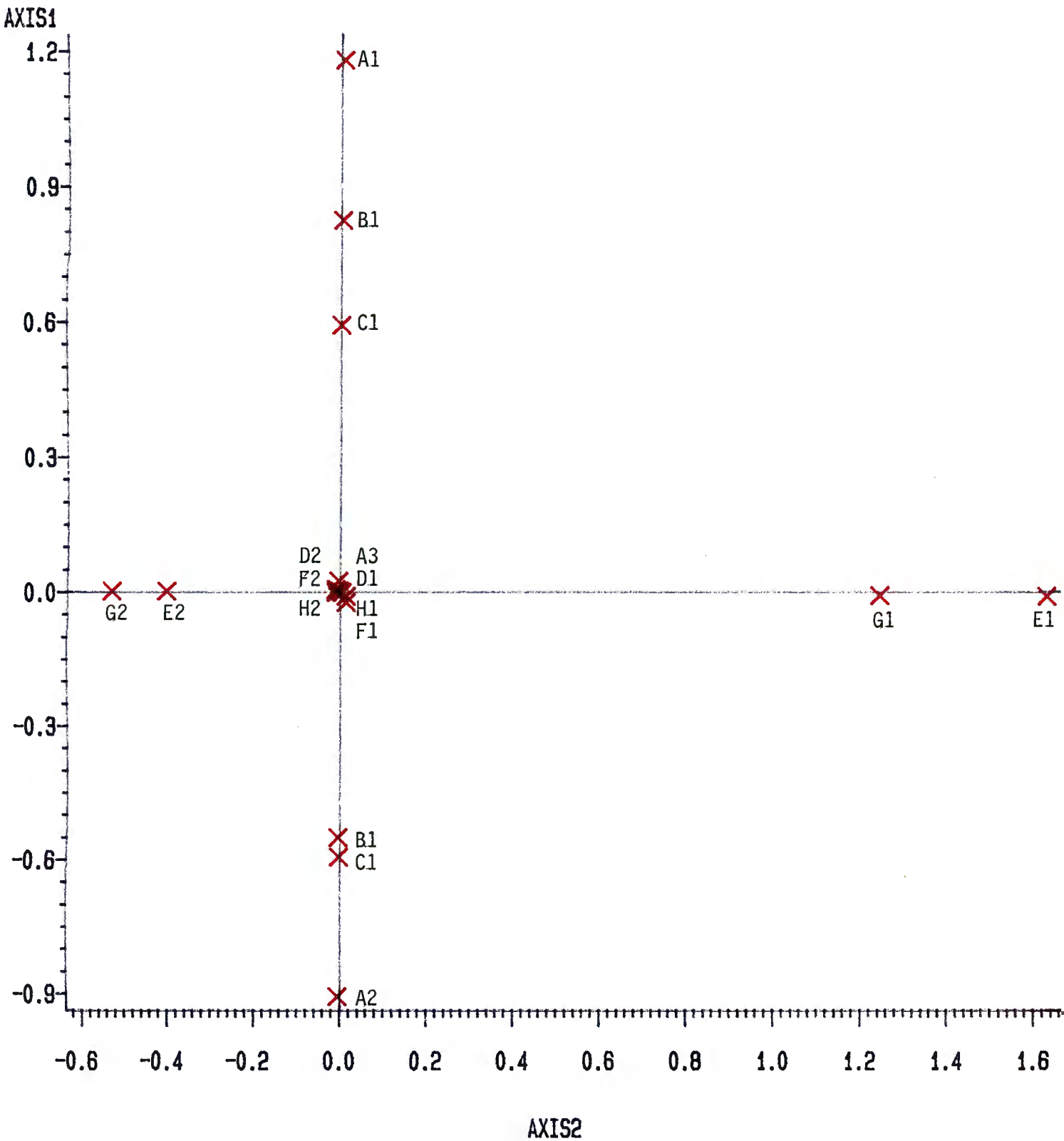
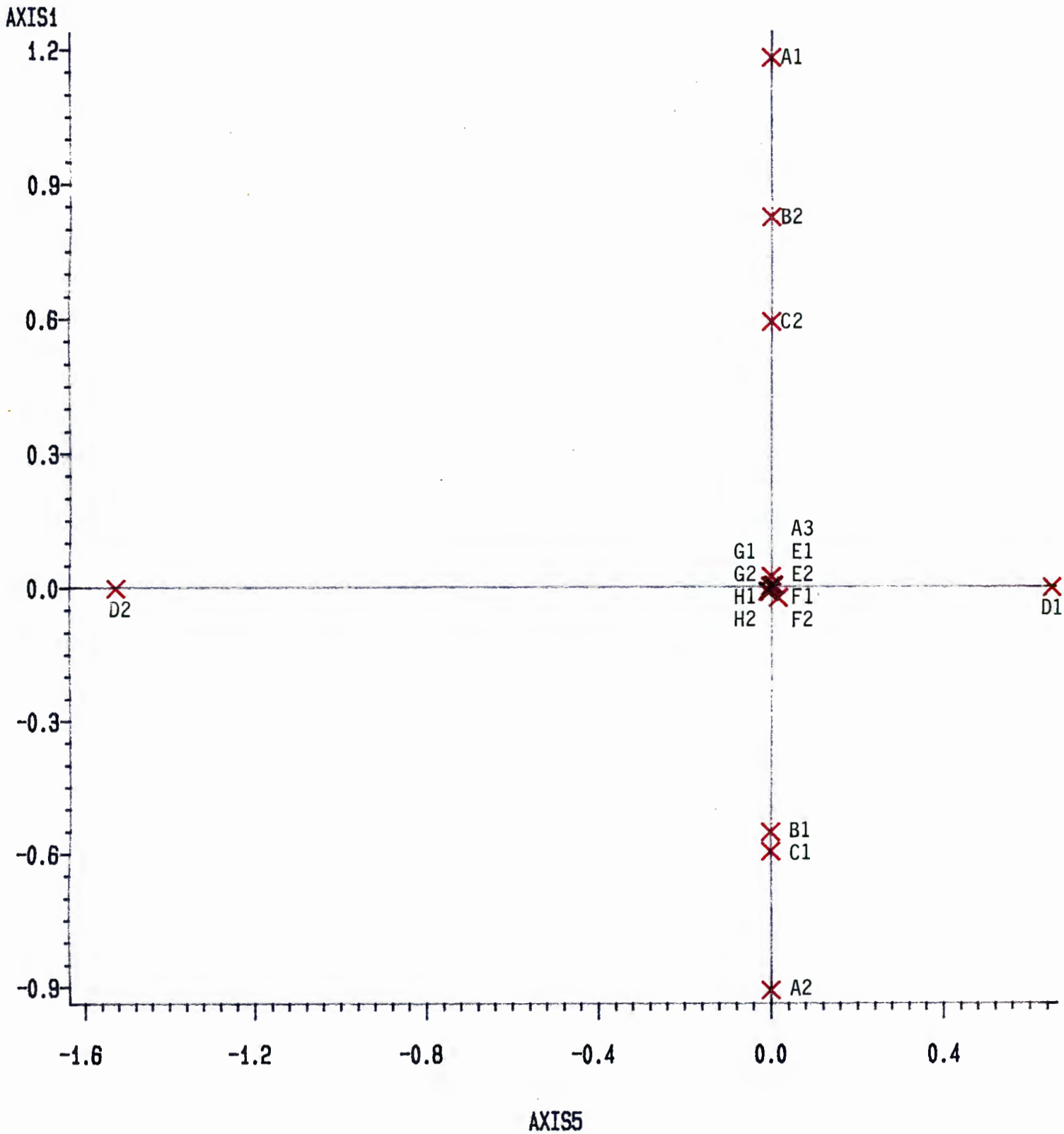


FIGURE 8 : MODEL(13)
ABC,D,EF,EG,FH

PROJECTION OF OBJECTS ON AXES 1 AND 5



4.3.2 Conclusions based on the above investigation

At the outset of this study we asked two questions, can correspondence analysis help us in :

(i) Selecting variables which can be meaningfully used for forming r-way tables?

(ii) Deciding on the particular log-linear model that should be fitted to these tables?

The answer to the first question is "yes". We have seen how correspondence analysis is able to split the variables into two or more sets which are independent of each other. These sets could then be analysed separately and log-linear models fitted to each using one of the usual methods such as that offered by Aitkin (1978, 1979, 1980).

The answer to the second question is "only in special cases", namely when the model either contains 2-way interactions or else when the variables are independent. Correspondence analysis may not pick up higher order interactions. We shall show later, in the cases where we decided that a 3-way interaction should be included, that what in fact we were detecting was the absence of certain 2-factor interactions.

In addition, the investigation highlighted the patterns of zero's in the contributions of the objects to the inertia of the various principal axes (see, for example the columns headed 'CTR' in Tables 14-16). These patterns imply that certain variables do not contribute to the inertia of certain axes. An explanation of this feature will be given in Chapter 5.

We note, however, that the observed relationships are obeyed exactly in the models we constructed because there is no random variation. With real data sets, the observed frequencies, $f_{ijk\ell}$ would not obey exact relationships such as,

$$f_{ijk\ell} = f_{ij..}f_{..k\ell}/N \quad \text{when AB and CD are independent,}$$

and so the patterns of zero and non-zero contributions by the objects to the inertia of the principal axes would similarly not be as clearcut. We discuss the effect of random variation on the data in the next section.

4.4 Correspondence analysis on data sets with known structure but where random variation of the cell frequencies is introduced

4.4.1 Method of investigation

In constructing the contingency tables in the previous section negligible random variation of the cell frequencies was permitted, as it was felt at that stage that it would only blur the structure suggested by the

correspondence analysis. Now that we have gained some insight into the value of correspondence analysis in fitting log-linear models, we need to investigate the effect of sampling variation. This will be achieved by performing correspondence analysis on data sets with known structure, but where random variation of the cell frequencies has been introduced.

Six data sets corresponding to 6 of the 15 models above were constructed using the following procedure :

> The structure of a multidimensional contingency table is determined by the marginals. Such a table also has a number of degrees of freedom which determine, within a particular set of marginal constraints, how many of the cell frequencies are not uniquely determined. Using a given set of marginal probabilities, those cell frequencies which were free to vary, were determined by selecting values from a binomial distribution with (i) mean equal to the cell frequency in the situation of no random error and (ii) with fixed sample size (10000 in five of the data sets and 500 in the other one). The other values in the contingency table are uniquely determined as a result of the marginal constraints. This gives us contingency tables with specified structures, but where the cell frequencies have a random binomial error. The one data set was constructed having only 500 observations in order to see how the introduction of both random error and reduced sample size would effect the correspondence analysis.

4.4.2 Results

Correspondence analysis was then performed on these data sets to predict their underlying structure. The underlying structure of the data sets as well as indications of the goodness-of-fit of the models are given in Table 17.

The results of performing correspondence analysis on these 6 data sets are tabulated in Tables 18 and 19. For each model we are particularly concerned with those axes which account for approximately 100/J-M % of the total inertia. These axes have been marked with a "+" in Table 18. Table 36, Appendix A2 gives the % contribution of each variable to the inertia of the first five principal axes. Table 19 includes a summary of the information given in this appendix. In Table 20 we set out the models suggested by correspondence analysis (a) when the data contains negligible random error (condition I) and (b) after random error has been introduced (condition II).

In comparing the output obtained from performing correspondence analysis on the data sets under these two conditions, we note the following :

(1) There are very small differences when we compare their moments of inertia and the decomposition of the total inertia along the various principal axes. (See Tables 10 and 18).

TABLE 17 : THE UNDERLYING STRUCTURE OF THE 6 "RANDOM" DATA SETS AND THE GOODNESS OF FIT OF THE STRUCTURE TO THESE DATA SETS

DATA SET	MODEL AND UNDERLYING STRUCTURE	DF	L.R. CHI-SQUARE STATISTIC	PROB
1	BC,AD	15	13,02	0,6006
2	ABC,AD	9	7,44	0,5919
3	AC,AD,BD	13	9,64	0,7228
4	ABC,AD,EF,EG,EH	362	389,95	0,1496
5	ABC,D,EF,EG,FH	364	342,55	0,7841
6	A,B,C,D,E,F,G,H	374	380,05	0,4034

TABLE 18 : DECOMPOSITION OF THE TOTAL INERTIA ALONG THE PRINCIPAL AXES FOR DATA SETS 1-6

DATA SET	% OF THE TOTAL INERTIA EXPLAINED BY AXIS NUMBER								
	1	2	3	4	5	6	7	8	9
1	30,4+	20,9+	19,9+	19,2	9,6	-	-	-	-
2	33,1+	30,7+	17,2	9,8	9,1	-	-	-	-
3	32,7+	25,2+	19,6+	13,6	8,8	-	-	-	-
4	20,1+	18,3+	16,9+	10,0	9,9	9,1	5,4	5,2	5,0
5	17,3+	14,7+	14,4+	14,1+	11,1+	8,1	7,7	7,4	5,2
6	11,5	11,5	11,3	11,2	11,1	11,0	10,9	10,8	10,8 *

* - indicates independence

TABLE 19 : PRINCIPLE CONTRIBUTORS TO THE INERTIA OF THE SIGNIFICANT AXES FOR DATA SETS 1-6

DATA SET	PRINCIPLE CONTRIBUTORS TO THE INERTIA OF AXES				
	1	2	3	4	5
1	AD	BC	A		
2	AD	AB			
3	AD	AC	B		
4	EFGH	ACD	ABC	(GH)	
5	ABC	EG	ABC	FH	D
6*					

*Independence model

TABLE 20 : THE MODELS SUGGESTED BY CORRESPONDENCE ANALYSIS FOR THE DATA SETS WITH AND WITHOUT RANDOM ERROR

DATA SET	TRUE MODEL	SUGGESTED MODEL FOR THE OUTPUT OF THE CORRESPONDENCE ANALYSIS	
		NO RANDOM ERROR	RANDOM ERROR
1	5:BC,AD	BC,AD	BC,AD
2	9:ABC,AD	AB,BC,AD	AD,AB
3	6:AC,AD,BD	AD,AC,B	AD,AC,B
4	14:ABC,AD,EF,EG,EH	ABC,ACD,EFGH	ABC,ACD,EFGH
5	13:ABC,D,EF,EG,EH	ABC,D,EG,FH	ABC,D,EF,FH
6	11:A,B,C,D,E,F,G,H	A,B,C,D,E,F,G,H	A,B,C,D,E,F,G,H

(2) Similarly there are only very small differences when we compare the percentage contributions of each variable to the inertia of the first 5 axes of inertia (see Tables 34-36, Appendix A2). We do however note that for data set (5)/Model (13) that the correspondence analysis has decomposed the total inertia differently for the two conditions, namely that variables A,B and C in condition II are the major contributors to the inertia of the 3rd axis and not the 4th axis as they were in condition I. Likewise F and H are the major contributors to the inertia of the 4th axis and not the 3th as was the case in condition I.

(3) The graphical displays show similar patterns, though in the output from the correspondence analysis on the data sets with random error, the plot positions of the objects vary more about the principal axes.

(4) We now make a few specific comments on the output obtained from performing correspondence analysis on each of the 5 data sets.

Data set (1): true model : BC,AD

Using the criteria set out in Section 4.3 above, both the information relating to the contributions of the variables to the inertia of the various axes as well as the graphical displays supported the fitting of the model : BC,AD. From Table 20, we note that this is the same model as was suggested when the data contained no random error.

Data set (2): true model : ABC,AD

Correspondence analysis proposes that the model : AD,AB be fitted to the data. This model has too few interactions.

Data set (3): true model : AC,AD,BD

Correspondence analysis proposes that the model : AC,AD,B be fit to the data. As in the case with data set (2), the model has too few interactions.

Data set (4): true model : ABC,AD,EF,EG,EH

The correspondence analysis suggests a split of the variables into two groups of 4 variables - the one part comprising A,B,C and D and the other, E,F,G and H. With regard to the variables A,B,C and D, as was the case in condition I, it detects the ABC interaction, but also suggests ACD interaction. In both conditions I and II, it seems to be detecting an indirect relationship between C and D because of the direct relationship between A & C and A & D. Furthermore in both conditions no clear indication was given by the correspondence analysis as to the interactions among the other group of variables.

Data set (5): true model : ABC,D,EF,EG,FH

A similar splitting of the data set into two parts was found

when correspondence analysis was performed on this data set. Likewise no indication was given as to the interactions among the group of variables E,F,G and H. With regard to the variables A,B,C and D, the interaction between the variables A, B and C and their independence from D was clearly detected. These results were also found when the data set with negligible random error was used.

Data set (6): true model : A,B,C,D,E,F,G,H

Correspondence analysis on this data set yielded a decomposition of the total inertia into equal parts along each of the principal axes under both conditions I and II.

In this study the conclusions given in section 4.3.2 remain consistent despite the introduction of random variation into the data sets (and in one case a data set with 1/200th the number of observations). In principle, it would be possible to repeat the procedure above a very large number of times and obtain information about the sampling distribution of the contributions to the inertias and the display positions. This has not been done in this study.

In Chapter 5 we shall investigate the above results from a theoretical point of view.

CHAPTER 5

5. THEORETICAL INTERPRETATION OF THE CONNECTION BETWEEN CORRESPONDENCE ANALYSIS AND LOG-LINEAR MODEL BUILDING

In chapter 4 we saw that the plot positions of the objects (levels of the variables) are in some way indicative of the relationships between the variables and in particular the independence between variables or groups of variables. In terms of the vector/matrix notation used in section 3.2 and defined again in section 5.1 below, we note that the correspondence analysis of W, the data matrix in weighted form, in p dimensions is obtained from the rank p basic structure of :

$$\tilde{A} \equiv D_r^{-\frac{1}{2}} (P - \underset{\sim\sim}{r}\underset{\sim}{c}^T) D_c^{-\frac{1}{2}} \approx \hat{A}_{(p)} = U D_\mu V^T$$

where

$\hat{A}_{(p)}$ (=A) is an IxJ matrix,

D_μ is a pxp matrix,

U is an Ixp matrix,

V is a Jxp matrix

and $U^T U = V^T V = I_p$.

If I = J then $A = U D_\mu U^T$ and U = V. But, as is usually the case I ≠ J and we are concerned with the singular-value decomposition of A.

If $A = UD_{\mu}V^{T}$ where U and V are orthogonal matrices (I>J)

then (i) V consists of the orthogonal eigenvectors of $A^{T}A$,

(ii) U consists of the orthogonal eigenvectors of AA^{T},

and (iii) the diagonal elements of D_{μ} are the non-negative square roots of the eigenvalues of $A^{T}A$

(Rosen et al., 1970).

The plot positions of the objects are determined from their principal co-ordinates on various axes (as given in matrix G) and these in turn are functions of the basic vectors in V :

$$G = D_{c}^{-\frac{1}{2}} VD_{\mu}$$

- where V is the matrix of p basic vectors of the matrix $A^{T}A = VD_{\mu}^{2}V^{T}$.

Similarly, the plot positions of the subjects are determined from their principal co-ordinates on various axes (as given in matrix F) and these in turn are functions of the basic vectors in U :

$$F = D_{r}^{-\frac{1}{2}} UD_{\mu}$$

- where U is the matrix of p basic vectors of the matrix $AA^{T} = UD_{\mu}^{2}U^{T}$.

The structure of the data set, in so far as the objects are concerned, is thus contained in the matrix $A^{T}A$. We shall now explain the connection between correspondence analysis and log-linear model building by :

(i) explaining how the matrix A^TA contains information pertaining to the structure of the data set (in terms of the relationships between the variables) and

(ii) showing how the eigenvalue decomposition of this matrix leads to a matrix G of principal co-ordinates which, when groups of variables are independent (e.g. model (5) : BC,AD), contains a structure of 'zero' and 'non-zero' elements. This results in some of the objects falling on specific axes and not on others.

In providing the explanation we shall refer to the data set corresponding to the following 3-way table involving the variables A,B and C, each with two levels.

Table 21 : 3-way table involving the variables A,B and C

		C1	C2	1-st order marginals
A1	B1	10	40	
	B2	4	16	70
A2	B1	2	8	
	B2	4	16	30
1-st order marginals		20	80	100

The data set corresponding to this contingency table has the structure AB,C (with no sampling variation) and was constructed in the manner described in the previous chapter.

5.1 Composition of the matrix A^TA in terms of the relationships between the variables

If W (IxJ) is the matrix (in weighted form) on which we are performing the correspondence analysis, then the multiple correspondence analysis in p dimensions of W is obtained from the rank p basic structure of :

$$\underset{\sim}{A} \equiv D_r^{-\frac{1}{2}} (P - \underset{\sim\sim}{r}\underset{\sim}{c}^T) D_c^{-\frac{1}{2}}$$

where P (IxJ) = W/NM

where N is the number of subjects, and

M is the number of questions/variables

$\underset{\sim}{r}$ (Ix1) = $P\underset{\sim}{1}_{(J)}$

$\underset{\sim}{c}$ (Jx1) = $P^T\underset{\sim}{1}_{(I)}$

where J is the number of objects/response categories

D_r (IxI) = a square matrix with the elements of $\underset{\sim}{r}$ along the diagonal

D_c (JxJ) = a square matrix with the elements of $\underset{\sim}{c}$ along the diagonal.

With reference to the data set with structure AB,C :

$$W = \begin{bmatrix}
f_{111} & 0 & f_{111} & 0 & f_{111} & 0 \\
f_{112} & 0 & f_{112} & 0 & 0 & f_{112} \\
f_{121} & 0 & 0 & f_{121} & f_{121} & 0 \\
f_{122} & 0 & 0 & f_{122} & 0 & f_{122} \\
0 & f_{211} & f_{211} & 0 & f_{211} & 0 \\
0 & f_{212} & f_{212} & 0 & 0 & f_{212} \\
0 & f_{221} & 0 & f_{221} & f_{221} & 0 \\
0 & f_{222} & 0 & f_{222} & 0 & f_{222}
\end{bmatrix}$$

where f_{ijk} = the observed occurrence (frequency) in cell ijk of the 3-way table involving the variables A,B and C (the index of A is i, of B is j and of C is k).

$P = W/NM$ where W is a matrix and N and M are scalars

$$\underset{\sim}{r} = \frac{1}{N} \begin{bmatrix} f_{111} \\ f_{112} \\ f_{121} \\ f_{122} \\ f_{211} \\ f_{212} \\ f_{221} \\ f_{222} \end{bmatrix}$$

$$\underset{\sim}{c}^T = \frac{1}{NM} (f_{1..}\ f_{2..}\ f_{.1.}\ f_{.2.}\ f_{..1}\ f_{..2})$$

where $f_{i..}$ is the one way marginal for variable A, i=1,2

$f_{.j.}$ is the one way marginal for variable B, j=1,2

$f_{..k}$ is the one way marginal for variable C, k=1,2.

$$rc^T = \frac{1}{MN^2}\begin{bmatrix} f_{111}f_{1..} & f_{111}f_{2..} & f_{111}f_{.1.} & f_{111}f_{.2.} & f_{111}f_{..1} & f_{111}f_{..2} \\ f_{112}f_{1..} & f_{112}f_{2..} & f_{112}f_{.1.} & f_{112}f_{.2.} & f_{112}f_{..1} & f_{112}f_{..2} \\ f_{121}f_{1..} & f_{121}f_{2..} & f_{121}f_{.1.} & f_{121}f_{.2.} & f_{121}f_{..1} & f_{121}f_{..2} \\ f_{122}f_{1..} & f_{122}f_{2..} & f_{122}f_{.1.} & f_{122}f_{.2.} & f_{122}f_{..1} & f_{122}f_{..2} \\ f_{211}f_{1..} & f_{211}f_{2..} & f_{211}f_{.1.} & f_{211}f_{.2.} & f_{211}f_{..1} & f_{211}f_{..2} \\ f_{212}f_{1..} & f_{212}f_{2..} & f_{212}f_{.1.} & f_{212}f_{.2.} & f_{212}f_{..1} & f_{212}f_{..2} \\ f_{221}f_{1..} & f_{221}f_{2..} & f_{221}f_{.1.} & f_{221}f_{.2.} & f_{221}f_{..1} & f_{221}f_{..2} \\ f_{222}f_{1..} & f_{222}f_{2..} & f_{222}f_{.1.} & f_{222}f_{.2.} & f_{222}f_{..1} & f_{222}f_{..2} \end{bmatrix}$$

$$D_r^{-\frac{1}{2}} = \sqrt{N}\begin{bmatrix} f_{111} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & f_{112} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & f_{121} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & f_{122} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & f_{211} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & f_{212} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & f_{221} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & f_{222} \end{bmatrix}^{-1/2}$$

$$D_c^{-\frac{1}{2}} = \sqrt{MN}\begin{bmatrix} f_{1..} & 0 & 0 & 0 & 0 & 0 \\ 0 & f_{2..} & 0 & 0 & 0 & 0 \\ 0 & 0 & f_{.1.} & 0 & 0 & 0 \\ 0 & 0 & 0 & f_{.2.} & 0 & 0 \\ 0 & 0 & 0 & 0 & f_{..1} & 0 \\ 0 & 0 & 0 & 0 & 0 & f_{..2} \end{bmatrix}^{-1/2}$$

$$\tilde{A} \equiv D_r^{-\frac{1}{2}} (P - \underset{\sim}{r}\underset{\sim}{c}^T) D_c^{-\frac{1}{2}}$$

(8x6)

Let $D = A^T A$. If $d_{k\ell}$ is the $k\ell$-th element of D then

$$d_{k\ell} = \sum_{i=1}^{I} a^T_{ki} a_{i\ell}$$

where a^T_{ki} is the ki-th element of A^T and $a_{i\ell}$ is the $i\ell$-th element of A. But $A^T_{ki} = a_{ik}$ so

$$d_{k\ell} = \sum_{i=1}^{I} a_{ik} a_{i\ell}$$

Thus $d_{k\ell}$ is the sum of the corresponding products of the elements of the k- and ℓ-th columns of A.

$$\text{and } A^T A = \begin{bmatrix} d_{11} & d_{12} & d_{13} & d_{14} & d_{15} & d_{16} \\ d_{21} & d_{22} & d_{23} & d_{24} & d_{25} & d_{26} \\ d_{31} & d_{32} & d_{33} & d_{34} & d_{35} & d_{36} \\ d_{41} & d_{42} & d_{43} & d_{44} & d_{45} & d_{46} \\ d_{51} & d_{52} & d_{53} & d_{54} & d_{55} & d_{56} \\ d_{61} & d_{62} & d_{63} & d_{64} & d_{65} & d_{66} \end{bmatrix}$$

With multiway data such as was used in this study, independence between variables implies that the cell frequencies, $f_{ijk\ell}$, are suitably adjusted products of the marginal frequencies. We found that the relationhips between $f_{ijk\ell}$ and its marginals causes certain elements in D to be zero.

Proof of this for the general case appears to be difficult because of the intricate relationhips that may exist between variables. However, we can

demonstrate this in detail for the simple case for the data set with the structure AB,C.

For this data set the following matrix A (given below to 3 decimal places) was obtained using the SAS program given in Appendix A3.

$$A = \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \\ A1 & A2 & B1 & B2 & C1 & C2 \end{array} \\ \left[\begin{array}{cccccc} 0,065 & -0,100 & 0,094 & -0,115 & 0,327 & -0,163 \\ 0,131 & -0,200 & 0,189 & -0,231 & -0,163 & 0,082 \\ 0,041 & -0,063 & -0,089 & 0,110 & 0,207 & -0,103 \\ 0,083 & -0,126 & -0,179 & 0,219 & -0,103 & 0,052 \\ -0,068 & 0,104 & 0,042 & -0,052 & 0,146 & -0,073 \\ -0,137 & 0,209 & 0,084 & -0,103 & -0,073 & 0,037 \\ -0,097 & 0,148 & -0,089 & 0,110 & 0,207 & -0,103 \\ -0,193 & 0,295 & -0,179 & 0,219 & -0,103 & 0,052 \end{array}\right] \end{array} \begin{array}{c} \text{obj/subj} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array}$$

and $A^TA = D$

$$= \begin{array}{c} \begin{array}{cccccc} A1 & A2 & B1 & B2 & C1 & C2 \end{array} \\ \left[\begin{array}{cccc|cc} 0,100 & -0,153 & 0,041 & -0,050 & \underline{0,000} & 0,000 \\ -0,153 & 0,233 & -0,063 & 0,077 & 0,000 & 0,000 \\ 0,041 & -0,063 & 0,133 & -0,163 & 0,000 & 0,000 \\ -0,050 & 0,077 & -0,163 & 0,200 & 0,000 & 0,000 \\ \hline 0,000 & 0,000 & 0,000 & 0,000 & 0,267 & -0,133 \\ 0,000 & 0,000 & 0,000 & 0,000 & -0,133 & 0,067 \end{array}\right] \end{array} \begin{array}{c} \text{object} \\ A1 \\ A2 \\ B1 \\ B2 \\ C1 \\ C2 \end{array}$$

Such a patterning is not suprising since we know that the data set has the structure AB,C. i.e. A & C and B & C are independent. However, it is not obvious why this patterning occurs, i.e. why certain of the $d_{k\ell}$ equal zero while others do not.

Above we have stated that $d_{k\ell} = \sum_{i=1}^{I} a_{ik} a_{i\ell}$. To obtain the entry d_{15} one multiplies (by row) each element of columns 1 and 5 of A and then sums the products :

$$
\begin{aligned}
d_{15} &= \sum_{i=1}^{8} a_{i1} a_{i5} \\
&= a_{11}a_{15} + a_{21}a_{25} + a_{31}a_{35} + a_{41}a_{45} + \\
&\quad a_{51}a_{55} + a_{61}a_{65} + a_{71}a_{75} + a_{81}a_{85} \\
&= 0{,}065\text{x}0{,}327 + 0{,}131\text{x}-0{,}163 + 0{,}041\text{x}0{,}207 + 0{,}083\text{x}-0{,}103 + \\
&\quad -0{,}068\text{x}0{,}146 + -0{,}137\text{x}-0{,}073 + -0{,}097\text{x}0{,}207 + -0{,}193\text{x}-0{,}103 \\
&= 0{,}000
\end{aligned}
$$

In terms of the notation given earlier in this section, the 1-st and 5-th columns of A may be denoted as :

Column 1

$$D_r^{-\frac{1}{2}}(P-\underset{\sim}{r}\underset{\sim}{c}^T)D_c^{-\frac{1}{2}}$$

$$\left(\frac{f_{111}}{MN} - \frac{f_{111}f_{1..}}{MN^2}\right)\times\frac{1}{\sqrt{f_{111}/N}}\times\frac{1}{\sqrt{f_{1..}/MN}}$$

$$\left(\frac{f_{112}}{MN} - \frac{f_{112}f_{1..}}{MN^2}\right)\times\frac{1}{\sqrt{f_{112}/N}}\times\frac{1}{\sqrt{f_{1..}/MN}}$$

$$\left(\frac{f_{121}}{MN} - \frac{f_{121}f_{1..}}{MN^2}\right)\times\frac{1}{\sqrt{f_{121}/N}}\times\frac{1}{\sqrt{f_{1..}/MN}}$$

$$\left(\frac{f_{122}}{MN} - \frac{f_{122}f_{1..}}{MN^2}\right)\times\frac{1}{\sqrt{f_{122}/N}}\times\frac{1}{\sqrt{f_{1..}/MN}}$$

$$\left(0 - \frac{f_{211}f_{1..}}{MN^2}\right)\times\frac{1}{\sqrt{f_{211}/N}}\times\frac{1}{\sqrt{f_{1..}/MN}}$$

$$\left(0 - \frac{f_{212}f_{1..}}{MN^2}\right)\times\frac{1}{\sqrt{f_{212}/N}}\times\frac{1}{\sqrt{f_{1..}/MN}}$$

$$\left(0 - \frac{f_{221}f_{1..}}{MN^2}\right)\times\frac{1}{\sqrt{f_{221}/N}}\times\frac{1}{\sqrt{f_{1..}/MN}}$$

$$\left(0 - \frac{f_{222}f_{1..}}{MN^2}\right)\times\frac{1}{\sqrt{f_{222}/N}}\times\frac{1}{\sqrt{f_{1..}/MN}}$$

Column 5

$$D_r^{-\frac{1}{2}}(P-\underset{\sim}{r}\underset{\sim}{c}^T)D_c^{-\frac{1}{2}}$$

$$\left(\frac{f_{111}}{MN} - \frac{f_{111}f_{..1}}{MN^2}\right)\times\frac{1}{\sqrt{f_{111}/N}}\times\frac{1}{\sqrt{f_{..1}/MN}}$$

$$\left(0 - \frac{f_{112}f_{..1}}{MN^2}\right)\times\frac{1}{\sqrt{f_{112}/N}}\times\frac{1}{\sqrt{f_{..1}/MN}}$$

$$\left(\frac{f_{121}}{MN} - \frac{f_{121}f_{..1}}{MN^2}\right)\times\frac{1}{\sqrt{f_{121}/N}}\times\frac{1}{\sqrt{f_{..1}/MN}}$$

$$\left(0 - \frac{f_{122}f_{..1}}{MN^2}\right)\times\frac{1}{\sqrt{f_{122}/N}}\times\frac{1}{\sqrt{f_{..1}/MN}}$$

$$\left(\frac{f_{211}}{MN} - \frac{f_{211}f_{..1}}{MN^2}\right)\times\frac{1}{\sqrt{f_{211}/N}}\times\frac{1}{\sqrt{f_{..1}/MN}}$$

$$\left(0 - \frac{f_{212}f_{..1}}{MN^2}\right)\times\frac{1}{\sqrt{f_{212}/N}}\times\frac{1}{\sqrt{f_{..1}/MN}}$$

$$\left(\frac{f_{221}}{MN} - \frac{f_{221}f_{..1}}{MN^2}\right)\times\frac{1}{\sqrt{f_{221}/N}}\times\frac{1}{\sqrt{f_{..1}/MN}}$$

$$\left(0 - \frac{f_{222}f_{..1}}{MN^2}\right)\times\frac{1}{\sqrt{f_{222}/N}}\times\frac{1}{\sqrt{f_{..1}/MN}}$$

It was found that

$$a_{11}a_{15} + a_{21}a_{25} = 0,$$
$$a_{31}a_{35} + a_{41}a_{45} = 0,$$
$$a_{51}a_{55} + a_{61}a_{65} = 0,$$
and $$a_{71}a_{75} + a_{81}a_{85} = 0.$$

But why for example does $a_{11}a_{15} + a_{21}a_{25} = 0$?

We shall now show why this is so.

$$a_{11}a_{15} + a_{21}a_{25} =$$

$$\left\{\left(\frac{f_{111}}{NM} - \frac{f_{111}f_{1..}}{MN^2}\right)x\frac{1}{\sqrt{f_{111}/N}}\ x\ \frac{1}{\sqrt{f_{1..}/NM}}\right\}x\left\{\left(\frac{f_{111}}{NM} - \frac{f_{111}f_{..1}}{NM^2}\right)x\frac{1}{\sqrt{f_{111}/N}}\ x\ \frac{1}{\sqrt{f_{..1}/NM}}\right\}$$

$$+\left\{\left(\frac{f_{112}}{NM} - \frac{f_{112}f_{1..}}{MN^2}\right)x\frac{1}{\sqrt{f_{112}/N}}\ x\ \frac{1}{\sqrt{f_{1..}/NM}}\right\}x\left\{-\left(\frac{f_{112}f_{..1}}{MN^2}\right)x\frac{1}{\sqrt{f_{112}/N}}\ x\ \frac{1}{\sqrt{f_{..1}/NM}}\right\}$$

$$=\left\{\frac{(Nf_{111}-f_{111}f_{1..})}{N\sqrt{M}\sqrt{f_{111}}\sqrt{f_{1..}}}\right\}x\left\{\frac{(Nf_{111}-f_{111}f_{..1})}{N\sqrt{M}\sqrt{f_{111}}\sqrt{f_{..1}}}\right\}-\left\{\frac{(Nf_{112}-f_{112}f_{1..})}{N\sqrt{M}\sqrt{f_{112}}\sqrt{f_{1..}}}\right\}x\left\{\frac{f_{112}\ f_{..1}}{N\sqrt{M}}\right\}$$

$$=\left\{\frac{(Nf_{111}-f_{111}f_{1..})(Nf_{111}-f_{111}f_{..1})}{MN^2f_{111}\sqrt{f_{1..}}\sqrt{f_{..1}}}\right\}-\left\{\frac{\sqrt{f_{112}}(Nf_{112}-f_{112}f_{1..})}{MN^2\sqrt{f_{1..}}}\right\}$$

$$=\left\{\frac{f_{111}(N - f_{1..})(N - f_{..1})}{MN^2\sqrt{f_{1..}}\sqrt{f_{..1}}}\right\} -\left\{\frac{f_{112}\sqrt{f_{..1}}\ (N - f_{1..})}{MN^2\sqrt{f_{1..}}}\right\}$$

$$=\left\{\frac{f_{111}f_{2..}f_{..2}}{MN^2\sqrt{f_{1..}}\sqrt{f_{..1}}}\right\}-\left\{\frac{f_{112}\sqrt{f_{..1}}\ f_{2..}}{MN^2\sqrt{f_{1..}}}\right\}$$

$$=\left\{\frac{f_{111}f_{2..}f_{..2}}{MN^2\sqrt{f_{1..}}\sqrt{f_{..1}}}\right\} - \left\{\frac{f_{112}\ f_{..1}\ f_{2..}}{MN^2\sqrt{f_{1..}}\sqrt{f_{..1}}}\right\}$$

$$= \frac{f_{11.}f_{..1}f_{2..}f_{..2} - f_{11.}f_{..2}f_{2..}f_{..1}}{MN^3\sqrt{f_{1..}}\sqrt{f_{..1}}}$$

since $f_{ijk} = f_{ij.}\ f_{..k}\ /N$ due to the structure of the data set (AB,C).

$$= 0$$

Similarly, it can be shown how $a_{31}a_{35} + a_{41}a_{45} = 0$,

$a_{51}a_{55} + a_{61}a_{65} = 0$, and

$$a_{71}a_{75} + a_{81}a_{85} = 0.$$

and thus how $d_{15} = 0$.

The matrix A can, because of the structure of the data set (AB,C), be divided into two subsets, the one corresponding to the first four columns (A1, A2, B1, B2), and the other to the last two columns (C1, C2). On the basis of what we have shown above and from investigating the composition of the matrix A^TA for the each of the 15 data sets whose structure was known, we conclude that :

$d_{k\ell} = 0$ if k and ℓ are in different sets
(e.g. $k \in \{A1,A2,B1,B2\}$
and $\ell \in \{C1,C2\}$).
This is because of the marginal relationships between independent sets, which causes the expression A^TA to equal zero.

$d_{k\ell} \neq 0$ if k and ℓ are in the same set
(e.g. k and $\ell \in \{A1,A2,B1,B2\}$
or k and $\ell \in \{C1,C2\}$).

5.2 The eigenvalue decomposition of the matrix A^TA and the matrix G of principal co-ordinates of the objects

We have seen how the matrix A^TA contains information pertaining to the relationships between the variables and in particular we have observed that,

in cases where there is independence between variables or groups of variables, this matrix contains a specific structuring of zero and non-zero elements. We now examine the eigenvalue decomposition of A^TA.

$$A^TA = VD_\mu^2V^T$$

where D_μ contains the eigenvalues and

V is a matrix of basic vectors.

Using the notation defined above, the matrix A^TA for the data set with structure AB,C could be rewritten as $A^TA = D$

$$= \left[\begin{array}{cccc|cc} d_{11} & d_{12} & d_{13} & d_{14} & 0 & 0 \\ d_{21} & d_{22} & d_{23} & d_{24} & 0 & 0 \\ d_{31} & d_{32} & d_{33} & d_{34} & 0 & 0 \\ d_{41} & d_{42} & d_{43} & d_{44} & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & d_{55} & d_{56} \\ 0 & 0 & 0 & 0 & d_{65} & d_{66} \end{array}\right]$$

Because of the zeros we can rewrite matrix D as

$$D = \begin{bmatrix} D_{11} & 0 \\ 0 & D_{22} \end{bmatrix}$$

where D_{11} is a 4x4 matrix and

D_{22} is a 2x2 matrix.

The basic values are the values of λ that satisfy

$$\left| \begin{bmatrix} D_{11} & 0 \\ 0 & D_{22} \end{bmatrix} - \lambda \begin{bmatrix} I_{11} & 0 \\ 0 & I_{22} \end{bmatrix} \right| = 0$$

$$\begin{vmatrix} D_{11} - \lambda I_{11} & 0 \\ 0 & D_{22} - \lambda I_{22} \end{vmatrix} = 0$$

where I_{11} and I_{22} are identity matrices with dimensions equal to those of D_{11} and D_{22} respectively,

$$= |D_{11} - \lambda I_{11}||D_{22} - \lambda I_{22}| = 0$$

There will be J-M (where J=6 is the sum of the levels of the variables and M=3 is the number of questions) non-trivial singular values for the matrix $A^T A$.

In our example we found that all the non-zero eigenvalues were different :

$$\lambda_1 = 0{,}672 \qquad \text{i.e. } D_\mu = \begin{bmatrix} 0{,}672 & 0 & 0 \\ 0 & 0{,}577 & 0 \\ 0 & 0 & 0{,}464 \end{bmatrix}$$
$$\lambda_2 = 0{,}577$$
$$\lambda_3 = 0{,}464$$

The expression :

$$|D_{11} - \lambda I_{11}||D_{22} - \lambda I_{22}| = 0$$

if either $|D_{11} - \lambda I_{11}| = 0$

or $|D_{22} - \lambda I_{22}| = 0$

or both $|D_{11} - \lambda I_{11}|$ and $|D_{22} - \lambda I_{22}| = 0$

In our example both λ_1 and λ_3 make

$$|D_{11} - \lambda I_{11}| = 0 \quad \text{and} \quad |D_{22} - \lambda I_{22}| \neq 0$$

while λ_2 makes

$$|D_{11} - \lambda I_{11}| \neq 0 \quad \text{and} \quad |D_{22} - \lambda I_{22}| = 0 .$$

Given $\underset{\sim}{V}_k^T = (\underset{\sim}{V}_k^{11}, \underset{\sim}{V}_k^{22})$ where k = 1,2,...,J-M.

Let $\underset{\sim}{V}_1$ be the characteristic vector that corresponds to λ_1, then $\underset{\sim}{V}_1$ is the solution of :

$$(D - \lambda_1 I)\underset{\sim}{V}_1 = \underset{\sim}{0}$$

where I is a matrix with the same dimension as D

$$= \begin{bmatrix} D_{11} - \lambda_1 I_{11} & 0 \\ 0 & D_{22} - \lambda_1 I_{22} \end{bmatrix} \begin{bmatrix} \underset{\sim}{V}_1^{11} \\ \underset{\sim}{V}_1^{22} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{0}^{11} \\ \underset{\sim}{0}^{22} \end{bmatrix}$$

Then $(D_{11} - \lambda_1 I_{11})\underset{\sim}{V}_1^{11} = \underset{\sim}{0}^{11}$ and,

$(D_{22} - \lambda_1 I_{22})\underset{\sim}{V}_1^{22} \neq \underset{\sim}{0}^{22}$

For λ_1, the expression $|D_{11} - \lambda I_{11}| = 0$ and the system of equations $(D_{11} - \lambda_1 I_{11})\underset{\sim}{V}_1^{11} = \underset{\sim}{0}$ has a non-trivial solution, i.e. $\underset{\sim}{V}_1^{11} \neq \underset{\sim}{0}$. In addition for λ_1, the expression $|D_{22} - \lambda I_{22}| \neq 0$. This implies that the only solution to $(D_{22} - \lambda_1 I_{22})\underset{\sim}{V}_1^{22} = \underset{\sim}{0}$ is $\underset{\sim}{V}_1^{22} = \underset{\sim}{0}$. Hence $\underset{\sim}{V}_1^T = (\underset{\sim}{V}_1^{11}, \underset{\sim}{0})$. Following the same argument for λ_2 and λ_3 we find that $\underset{\sim}{V}_2^T = (\underset{\sim}{0}, \underset{\sim}{V}_2^{22})$ and $\underset{\sim}{V}_3^T = (\underset{\sim}{V}_3^{11}, \underset{\sim}{0})$.

In the example we found that

$$V = \begin{bmatrix} \underset{\sim}{V}_1^{11} & \underset{\sim}{V}_2^{11} & \underset{\sim}{V}_3^{11} \\ \underset{\sim}{V}_1^{22} & \underset{\sim}{V}_2^{22} & \underset{\sim}{V}_3^{22} \end{bmatrix} =$$

	AXES			
	1	2	3	Objects
	-0,387	0	0,387	A1
	0,592	0	-0,592	A2
	-0,447	0	-0,447	B1
	0,548	0	0,547	B2
	0	-0,894	0	C1
	0	0,447	0	C2

We observe that the column co-ordinates of the objects in the 1st and 3rd axes are either the same or differ merely by a +ve or -ve sign. If we examine the square of the singular-values: $\lambda_1^2 = 0,452$, $\lambda_2^2 = 0,333$ and $\lambda_3^2 = 0,215$, it is clear that the 1st principal axis explains 45% of the total inertia, the 2nd, 33% and the 3rd, 22%. Greenacre (op. cit.) argues that in general only those axes which account for more than 1/M x 100% of the total inertia are "interesting". The other axes being "artefacts" of the analysis, will not add much, if anything, to our study of the inter-relationships between variables. In the example, as M is 3, we are therefore only interested in axes 1 and 2 which each explain a third or more of the total inertia. From inspecting the column co-ordinates of the objects on the 3rd axis it is clear that they will not add to the information already given in the 1st axis.

Let $\underset{\sim}{V}_k$, an Jx1 vector, be the k-th basic vector (k=1,2,..,T where T = J-M,

the number of non-trivial singular-values). We have shown how, depending on the structure of independence between variables or groups of variables, the elements of the vector $\underset{\sim}{V}_k$ will be 'zero' or 'non-zero'. This structure of 'zero' and 'non-zero' elements in the matrix V is passed on to the matrix G, which contains the principal co-ordinates of the objects on the various axes since the elements of G are simply scalar products of the elements of V :

$$G = D_c^{-\frac{1}{2}} V D_\mu \quad \text{where both } D_c^{-\frac{1}{2}} \text{ and } D_\mu \text{ are diagonal matrices.}$$

In the example we found

$$G = \begin{array}{c} \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \text{AXES} \\ \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \left[\begin{array}{ccc} -0,539 & 0 & 0,371 \\ 1,258 & 0 & -0,867 \\ 0,672 & 0 & -0,463 \\ 1,009 & 0 & 0,695 \\ \hline 0 & -2,000 & 0 \\ 0 & 0,500 & 0 \end{array}\right] \end{array} \begin{array}{l} \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{l} \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{l} \\ \text{A1} \\ \text{A2} \\ \text{B1} \\ \text{B2} \\ \text{C1} \\ \text{C2} \end{array}$$

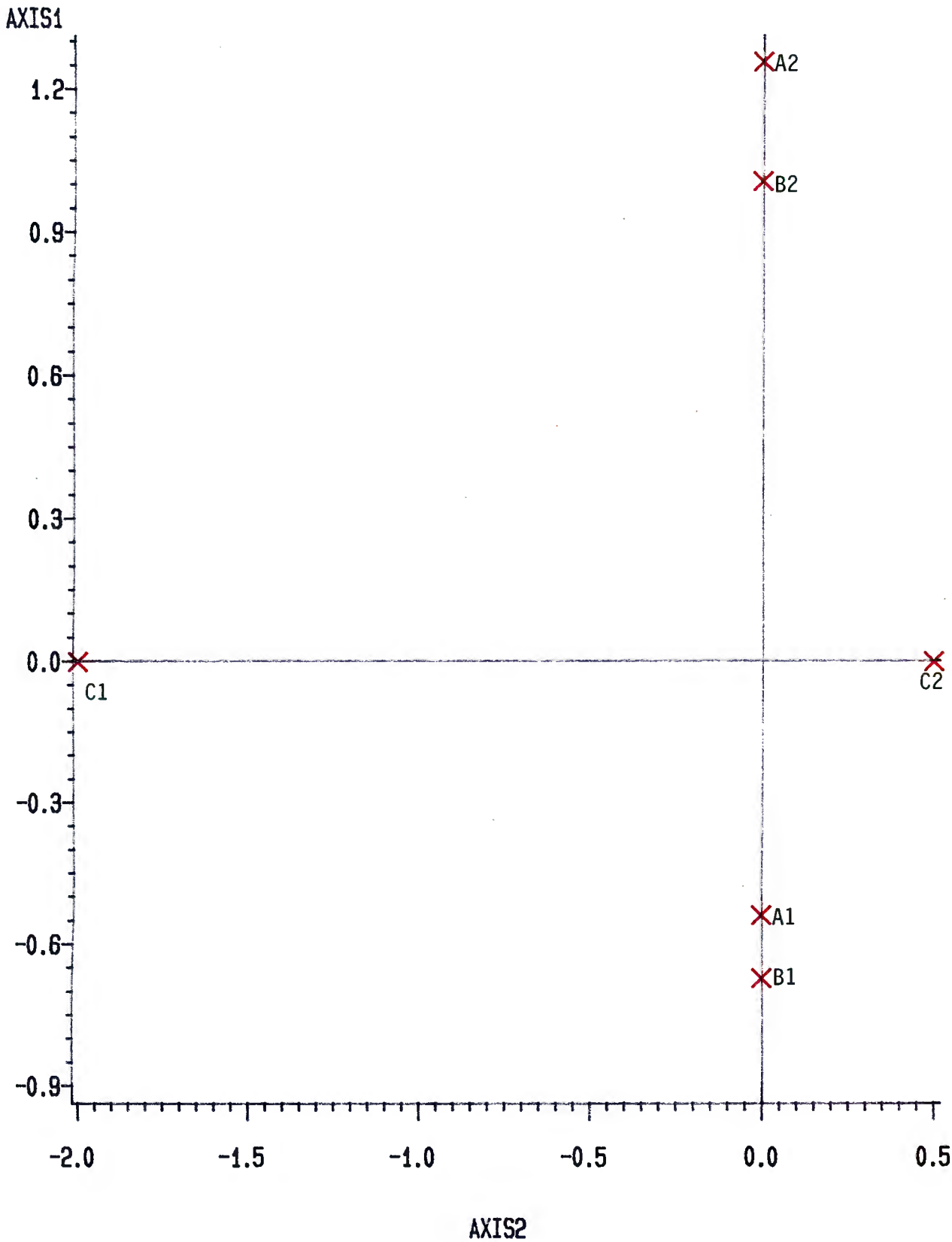
Plotting these standardized/principal co-ordinates on the first two principal axes, we obtain the graphical display presented in Figure 9. When an object falls on a particular axis (i.e. its co-ordinate on that axis is zero), it follows that its contribution to the inertia of that axis is also zero, since :

inertia = mass of object x squared distance

(see formulae in section 3.2)

= mass x 0.

FIGURE 9 : DATA SET WITH STRUCTURE AB,C



5.3 General Comments

5.3.1 Generalization

The above comments hold for other data sets where there is independence between variables or groups of variables. The marginal relationships which exist when variables or sets of variables are independent result in patterns of zeros in A^TA. This matrix can always be rearranged by interchanging columns and rows to give a block diagonal structure as shown above for the model with structure AB,C. This rearrangement will not affect the eigenvalue decomposition of A^TA.

Suppose we have S sets of independent variables, each set containing one or more variables. $D = A^TA$ can be set out in the form :

$$D = \begin{bmatrix} D_{11} & & 0 & & 0 \\ & \ddots & & & \\ 0 & & D_{ss} & & 0 \\ & & & \ddots & \\ 0 & & 0 & & D_{SS} \end{bmatrix}$$

The basic structure will be the solution of

$$\prod_{s=1}^{S} | D_{ss} - \lambda_{k(s)} I_{ss} | = 0.$$

where $\lambda_{k(s)}$ refers to the k-th eigenvalue which is associated with the s-th subset,

and I_{ss} is an identity matrix with the same dimension as D_{ss}.

D has T basic vectors (where T = J - M, the number of response categories minus the number of questions):

$$\underset{\sim}{V}_1, \underset{\sim}{V}_2, \ldots, \underset{\sim}{V}_T$$

These basic vectors can each be partitioned into S subvectors:

$$\begin{bmatrix} \underset{\sim}{V}_1^{11} \\ \underset{\sim}{V}_1^{22} \\ \cdot \\ \cdot \\ \underset{\sim}{V}_1^{SS} \end{bmatrix} \begin{bmatrix} \underset{\sim}{V}_2^{11} \\ \underset{\sim}{V}_2^{22} \\ \cdot \\ \cdot \\ \underset{\sim}{V}_2^{SS} \end{bmatrix} \cdots \begin{bmatrix} \underset{\sim}{V}_T^{11} \\ \underset{\sim}{V}_T^{22} \\ \\ \cdot \\ \underset{\sim}{V}_T^{SS} \end{bmatrix}$$

where the subvector $\underset{\sim}{V}_k^{ss}$ relates to the basic vector associated with the s-th independent set of objects on the k-th axis.

We deal with two situations:

(1) The roots (λ_k) are all different :

$|D_{ss} - \lambda_{k(s)} I_{ss}| = 0$ for a specific k and s
(k = 1,2,..,T and s = 1,2,..S) and

$|D_{ss'} - \lambda_{k(s')} I_{ss'}| \neq 0$ otherwise (s' refers to the complementary sets to s).

Then V_k will have the following structure :

$$\begin{bmatrix} 0 \\ 0 \\ \cdot \\ \underset{\sim}{V}_k^{ss} \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

where $\underset{\sim}{V}_k^{ss}$ is not a vector of zeros.

This is passed on to the matrix G of principal co-ordinates. It accounts for the patterns of zeros that we found in models (2),(3),(5),(8),(12), (13),(14),(15) and hence the finding that certain variables or groups of variables lay on specific axes in the graphical displays.

(2) The roots (λ_k) are all equal :

In the cases where there is "strict" independence between all the variables (as was the case with models (1) and (11)) :

model (1) : A,B,C,D

model (11) : A,B,C,D,E,F,G,H

it was found that the non-zero eigenvalues were all equal.

Using the data set corresponding to model (1), where A has 3 levels and B,C and D have 2, it was found that $D = A^TA$ has the block diagonal structure:

$$D_{9\times 9} = \begin{bmatrix} D_{11} & 0 & 0 & 0 \\ 0 & D_{22} & 0 & 0 \\ 0 & 0 & D_{33} & 0 \\ 0 & 0 & 0 & D_{44} \end{bmatrix}$$

where D has a rank of 5 and

D_{11} is a 3x3 matrix with rank 2 and

D_{22}, D_{33} and D_{44} are 2x2 matrices with rank 1.

This occurs because of the marginal relationships between the frequencies:

$$f_{ijk\ell} = f_{i...} f_{.j..} f_{..k.} f_{...\ell} / N^3.$$

The eigenstructure of D is found by solving

$$|D - \lambda_{k(s)} I| = 0$$

$$= \prod_{s=1}^{4} |D_{ss} - \lambda_{k(s)} I_{ss}| = 0$$

and we find that the 5 non-zero eigenvalues are all equal,

$$\lambda_{1(1)} = \lambda_{2(1)} = \lambda_{3(2)} = \lambda_{4(3)} = \lambda_{5(4)} = \lambda = 0{,}5$$

The 5 basic vectors $\underset{\sim}{V}_1$, $\underset{\sim}{V}_2$, $\underset{\sim}{V}_3$, $\underset{\sim}{V}_4$, $\underset{\sim}{V}_5$ are linearly independent and are the solution to:

$$(D - \lambda I)V = \underset{\sim}{0}$$

But $\underset{\sim}{V}_1$, $\underset{\sim}{V}_2$,..., $\underset{\sim}{V}_5$ can be partitioned as

$$\underset{\sim}{V}_k = \begin{bmatrix} \underset{\sim}{V}_k^{11} \\ \underset{\sim}{V}_k^{22} \\ \underset{\sim}{V}_k^{33} \\ \underset{\sim}{V}_k^{44} \end{bmatrix} \qquad k=1,2,...,5$$

where $\underset{\sim}{V}_k^{11}$ is a 3x1 vector

and $\underset{\sim}{V}_k^{22}$, $\underset{\sim}{V}_k^{33}$, $\underset{\sim}{V}_k^{44}$ are 2x1 vectors.

We can write $(D - \lambda I)V = \underset{\sim}{0}$ as

$$\begin{bmatrix} D_{11} - \lambda I_{11} & 0 & 0 & 0 \\ 0 & D_{22} - \lambda I_{22} & 0 & 0 \\ 0 & 0 & D_{33} - \lambda I_{33} & 0 \\ 0 & 0 & 0 & D_{44} - \lambda I_{44} \end{bmatrix} \begin{bmatrix} \underset{\sim}{V}_k^{11} \\ \underset{\sim}{V}_k^{22} \\ \underset{\sim}{V}_k^{33} \\ \underset{\sim}{V}_k^{44} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{0} \\ \underset{\sim}{0} \\ \underset{\sim}{0} \\ \underset{\sim}{0} \end{bmatrix}$$

which can be written as 4 separate sets of linear equations:

$(D_{11} - \lambda I_{11})\underset{\sim}{V}_k^{11} = \underset{\sim}{0}$ where D_{11} has rank 2

$(D_{22} - \lambda I_{22})\underset{\sim}{V}_k^{22} = \underset{\sim}{0}$ D_{22} has rank 1

$(D_{33} - \lambda I_{33})\underset{\sim}{V}_k^{33} = \underset{\sim}{0}$ D_{33} has rank 1

$(D_{44} - \lambda I_{44})\underset{\sim}{V}_k^{44} = \underset{\sim}{0}$ D_{44} has rank 1

Since the λ's are all equal,

$$| D_{ss} - \lambda I_{ss} | = 0 \quad \text{for all } s=1,2,3,4.$$

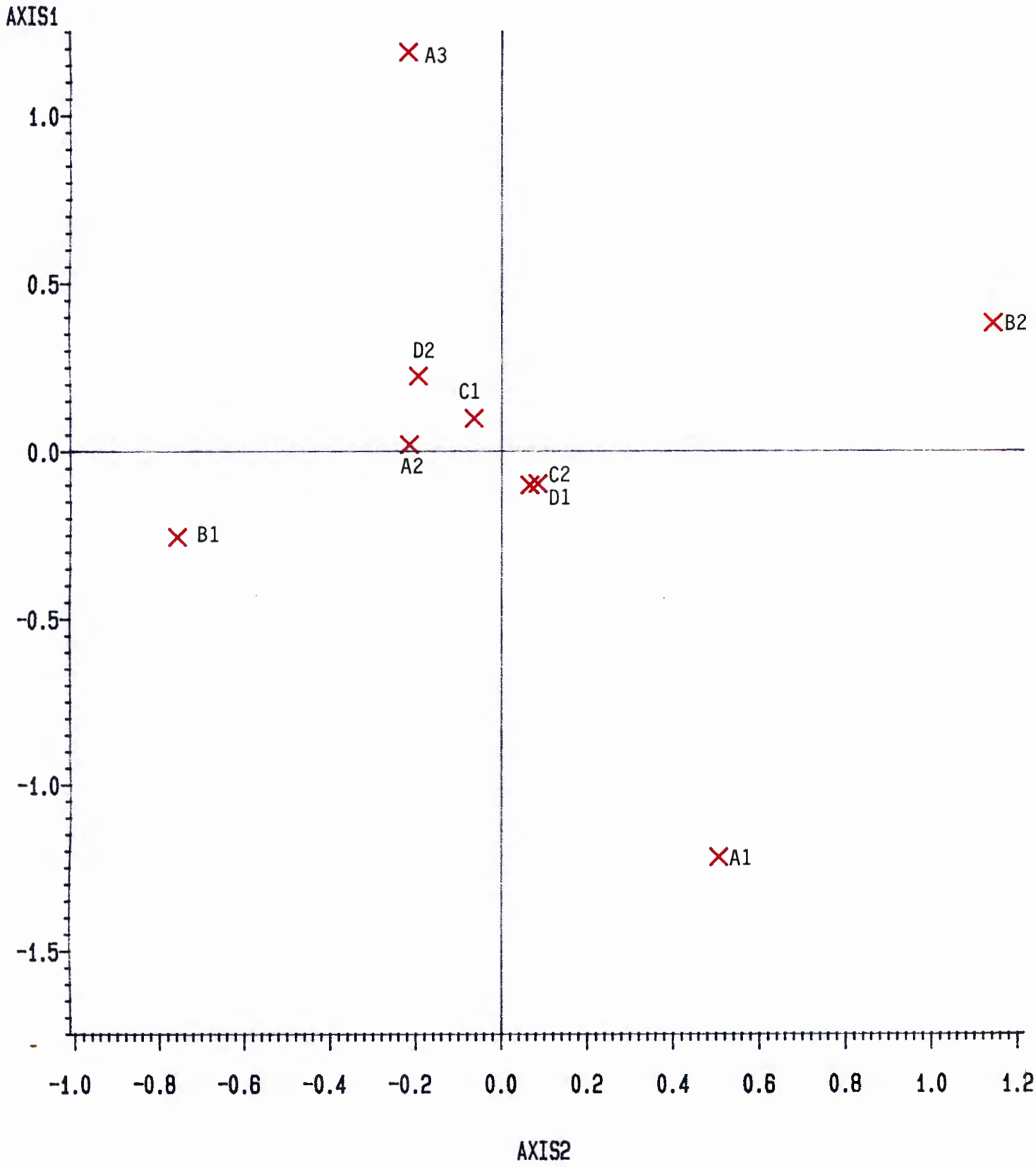
implying that

$$(D_{ss} - \lambda I_{ss})\underset{\sim}{V}_k^{ss} = 0$$

has a non-trivial solution and therefore $\underset{\sim}{V}_k^{ss} \neq 0$ for all k and s. As a result none of the objects will lie on the principal axes in the graphical displays (see Figure 10). This solution for the $\underset{\sim}{V}_k^{ss}$ is, however, not unique. Suppose that we use the block diagonal structure of $D = A^T A$ and solve $(D_{ss} - \lambda I_{ss})\ \underset{\sim}{V}_k^{ss} = \underset{\sim}{0}$ <u>separately</u> for each set (s = 1,2,3,4) and obtain the eigenvectors E_1^{11}, E_2^{11}, E_3^{22}, E_4^{33} and E_5^{44}. We can augment these solutions by

FIGURE 10 : *MODEL*(1)
A,B,C,D

PROJECTION OF OBJECTS ON AXES 1 AND 2



appending zero vectors of appropriate order to obtain the set of 5 basic vectors with the form:

$$\begin{bmatrix} \underset{\sim}{E}_1^{11} \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} \underset{\sim}{E}_2^{11} \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ \underset{\sim}{E}_3^{22} \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \underset{\sim}{E}_4^{33} \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ \underset{\sim}{E}_5^{44} \end{bmatrix}$$

which satisfy $(D - I)V = \underset{\sim}{0}$ and moreover lie in 5 mutually orthogonal subspaces. With this choice of basic vectors, the independent variables A,B,C and D would lie on different axes in the graphical displays.

Thus, even though the output from Tabet's (1973) correspondence analysis programme on data sets where there is complete independence between variables does not result in the variables falling on different axes, it would be possible to 'engineer' this using the procedure outlined above.

5.3.2 Detection of 3-way interaction

In chapter 4 we noted that correspondence analysis appears to detect 2-way better than 3-way interactions. We found in the case of the data set corresponding to model (10) : ABC,ABD , that the output from the correspondence analysis gave no clear indication of the 3-way interactions amongst variables. However, we also remarked that in both

models (8) : ABC,D

and (13) : ABC,D,EF,EG,FH

the correspondence analysis detected the independence of ABC from D. From analysing the decomposition of the matrix A^TA in the cases of models (8) and (10), it seems that correspondence analysis is unable to provide a clear indication of the relationships between the variables when there is not "strict" independence i.e. when the independent variables/groups of variables are not mutually exclusive e.g. AB,BC. Moreover, when it does detect 3-way interaction (e.g. the ABC interaction and independence of A, B and C from D in models (8) and (13)), it is because there is "strict" independence between the variables in the one interaction set (A,B and C) and the variable(s) in the other set(s) (D). In other words, correspondence analysis does not pick up 3-way interactions, but rather pinpoints the absence of 2-way interactions (i.e. independence between A & D, B & D, and C & D).

Greenacre (op. cit., chapter 5) when discussing the correspondence analysis of a multivariate indicator matrix Z in terms of the correspondence analysis of the Burt matrix Z^TZ (which he shows to be equivalent), illustrates that these analyses should be considered as "joint bivariate" rather than multivariate, and thereby supports the argument presented above.

5.3.3 Correspondence Analysis on real data sets

Above we have shown a special property of the characteristic roots of the matrix A^TA in cases of independence between variables/groups of variables. We realize, however, that in working with real as opposed to contrived data sets the observed frequencies f_{ijk} (for example) will not obey the multiplicative relationships exactly such as $f_{ijk} = f_{ij.}\ f_{..k}/N$ for a data set where A and B

are related but independent of C). Hence A^TA will not have an exact block diagonal structure and as a result the matrices V and G will never have zero values, though some of their values will be close to zero. This was found when performing correspondence analysis on the six data sets which had random error added (see chapter 4). One would like to develop a statistical test to see if any of the roots are zero. This would involve finding the properties of the distribution of the roots which would be a complicated procedure. However, such a test is not really necessary as we can test for 2-way independence using the ordinary Chi-square test.

CHAPTER 6

6. CONCLUSION

This chapter has been set out in three sections. Firstly, we reiterate the main research findings and thereafter point out certain topics for future research. In the third section we outline steps that could be followed in using correspondence analysis to fit log-linear models to questionnaire data.

6.1 The findings of this study

6.1.1 Correspondence analysis can detect variables (questions) that are "strictly" independent.

6.1.2 Correspondence analysis can be used to detect pairwise relationships between variables (2-way interactions).

6.1.3 Correspondence analysis does not pick up higher order interactions between variables, but rather detects the absence of pairwise interactions. We therefore support the contention of Greenacre (op. cit., chapter 5) that correspondence analysis should be

considered as a "joint bi-variate" rather than as a multivariate technique. This does not invalidate its usage in analysing multivariate data, but it does suggest an area for future research.

6.1.4 As a result, correspondence analysis cannot be used to select log-linear models in general because such models often have 3- or 4-factor interactions. However, this technique could be used in the initial stages of the model selection process to detect 2-factor interactions.

6.1.5 Correspondence analysis can also be used to suggest a splitting of large data sets into two or more subsets of variables which are independent. Thereafter other model selection techniques could be used to determine the particular log-linear model to be fit to the subsets.

6.2 Topics for future research

This thesis has answered a number of questions, but as is inevitably the case with most research, it has raised many more than were initially asked. We now outline a number of topics for future research.

6.2.1 Altering the number of response categories (levels) of the questions (variables)

Questionnaire data usually comprise information on a number of subjects on each of several variables, each having a number of levels. Such data are

invariably set out in multidimensional contingency tables, the number of dimensions being determined by both the number of variables and the number of response categories to each. Whittmore (1978) highlights two problems of tables of high dimension, namely, that the relationships between variables are difficult to detect and that the probability of empty cells increases as the number of dimensions increases. She suggests that one way to overcome this problem is to collapse such a table by summing the frequencies of certain levels of some of the variables.

In our study the number of levels of each variable was arbitrarily decided upon (3 for A and 2 for the other variables) and remained fixed throughout. Further research is needed to assess the sensitivity of correspondence analysis in detecting the independence between variables while varying the number of levels of each variable. This is particularly important in the light of the finding by Reynolds (1977), that significant interactions between variables could be created simply as a result of the particular way in which the levels of the variables involved had been collapsed.

Another way to reduce the number of dimensions, a way which has become apparent in our study, is to allow correspondence analysis to suggest a splitting of the variables into two or more subsets and to analyse these separately i.e. fit multidimensional contingency tables to each subset of the variables.

6.2.2 Increasing the number of questions

Questionnaires typically involve the asking of a large number of questions. Further research is therefore needed to investigate the analysis of data sets

involving a large number of variables (more than eight anyway). However, as the number of variables increases, the number of dimensions of the contingency table to which we are seeking to fit a log-linear model also increases. Together with this comes the problem mentioned in section 6.2.1 of the expected frequency in each cell decreasing. In addition, with regard to correspondence analysis, the number of non-trivial axes needed to explain the total inertia increases as we include more variables and this in turn complicates the analysis.

Greenacre (op. cit., chapter 8) proposes a solution to this problem by suggesting firstly, that a method of reciprocal averaging be used because this method only permits the evaluation of the co-ordinates of the subjects and objects on the first few axes and secondly, that only the addresses of the non-zero elements in the multivariate indicator matrix be stored - the remaining elements are zero by default.

6.2.3 Other issues

Greenacre (op. cit.) mentions a number of issues which are not considered in this thesis, but which deserve attention in the future.

a) From the outset, we admitted that we were containing our interest to data sets where all the variables were discrete or had been made discrete. Reynolds (1977) showed experimentally that categorizing continuous data in different ways can lead to different conclusions and that as a result researchers must think carefully before they do this. As mentioned, Greenacre (chapter 5) discusses a way of performing correspondence analysis on mixed (i.e. both continuous and

discrete) data without discretizing the continuous variables. This suggests an interesting topic for future research : that of the use of correspondence analysis (on mixed data) in the fitting of generalized linear models (Nelder and Wedderburn, 1972).

b) Other topics discussed by Greenacre (chapters 5 and 8) which deserve further attention, concern the treatment of non-responses and missing values in the analysis of questionnaire data, in particular in relation to the fitting of log-linear models.

c) In chapter 4, we looked at the effect of the introduction of random error into the data sets with pure structure on the ability of correspondence analysis to detect independence between variables. Further research is needed to obtain information about the sampling distributions of the contributions of the objects (response categories of the variables) to the inertias of the axes. Such research would be related to the topic of the sampling distribution of the co-ordinates of the objects on the various axes. Some work in this regard is discussed by Greenacre (Chapter 8).

d) Greenacre (Chapter 5) discusses a way to attach different weights to some of the variables so that their inertias will be proportional to certain pre-assigned values. This might be useful if we regard certain questions as more important than others. Another way to weight questions is to assign those we desire to have no weight at all as supplementary points. (Chapter 3). This issue of the weighting of the questions in the statistical analysis of questionnaire data deserves attention in the future.

6.2.4 Additional topics for future research

a) The size of the sample :

Although the sample size is of importance in the fitting of log-linear models to multidimensional contingency tables, it is not of importance in performing correspondence analysis on the multivariate indicator matrix of questionnaire data. What is of importance are the relative proportions in Z or W (see section 3.2). These will clearly be more stable if the sample size is large. The stability of these proportions is thus another topic for further research.

b) The subjects :

In order to contain this study within certain limits, we have not concentrated on the information provided by the correspondence analysis on the subjects. Further studies are needed which utilize this information, and in particular the plot positions of the subjects on the various axes in relation to the positions of the objects.

6.3 Proposed steps to be followed in using correspondence analysis to fit log-linear models

One of the most expedient ways of explaining/understanding multivariate categorical data is to fit log-linear models. In this thesis we have suggested that correspondence analysis be used as an exploratory technique prior to the fitting of log-linear models to provide insight into the data.

In chapter 4 we showed how this might be achieved. We recommend that the following steps be followed:

STEP 1 : Preliminary steps

[1]. Run a computer program (e.g. SAS or BMDP program) to look at the frequency distributions of the variables. On the basis of this :

a) detect and correct errors,

b) omit book-keeping variables,

c) drop variables on which all the subjects give the same response,

d) regroup variables which are levels of a single variable,

e) possibly collapse categories,

f) possibly discretize continuous variables,

g) rerun the computer program on the new data set to detect further errors.

[2]. (Optional) Set up a multidimensional contingency table with all the variables or a number of tables with subsets of the variables. Look at the patterns in the cells of the contingency table(s). On the basis of this one may decide to drop certain variables from the analysis or to collect more data to fill up those cells having (non-structural) zeroes.

STEP 2 : Exploratory data analysis

[1]. Code the data into the weighted multivariate indicator matrix form, W (A simple program can be written to do this)

[2]. Run a correspondence analysis program on this indicator matrix, suppressing the plotting of the 'grouped' subjects on the graphical display if there are many of them.

[3]. Examine the output :

a) Examine the moments of inertia.

b) Examine the decomposition of the largest moments of inertia. In particular examine the contribution of each variable (sum over its objects) to the inertia of each axis in conjunction with the graphical display of the objects on the various planes of the principal axes.

c) In terms of the guidelines in section 4.3, answer the following questions :

(1) Are variables or groups of variables independent of each other?

(2) Does the analysis suggest a splitting of the data into two or more sets of variables whose interrelationships can be analysed separately.

STEP 3 : Confirmatory data analysis

Fit models to the data sets (or reduced sets involving subsets of the variables) using the information gained in Step 2 as a starting point for one of the model selection procedures mentioned in section 3.1.

6.4 Concluding remark

R.A. Fisher (1925), quoted in Everitt (1978:1), perhaps provides the best summary of the main argument of this thesis :

> "The prelimary examination of most data is facilitated by the use of diagrams. Diagrams prove nothing, but bring outstanding features readily to the eye; they are therefore no substitute for such critical tests as may be applied to data, but are valuable in suggesting such tests and explaining conclusions founded upon them."

REFERENCES

Aitkin, M. (1978). The Analysis of Unbalanced Cross-classifications. J. Royal. Statist. Soc., A, 141(2), 195-223.

Aitkin, M. (1979). A Simultaneous Test Procedure for Contingency Table Models. Applied Statistics, 28(3), 233-242.

Aitkin, M. (1980). A Note on the Selection of Log-Linear Models. Biometrics, 36, 173-178.

Andrews, D.F. (1972). Plots of high dimensional data. Biometrics, 28, 125-136.

Babbi, E.R. (1973). Survey Research Methods. Wadsworth Publishing Company Inc., Belmont, California.

Benedetti, J.K. and Brown, M.B. (1978). Strategies for the Selection of Log-Linear Models. Biometrics, 34, 680-686.

Benzécri, J.P. (1969). Statistical analysis as a tool to make patterns emerge from data. In Watanabe, S (ed.), Methodologies of Pattern Recognition, 35-74. Academic Press, New York.

Birch, M.W. (1972). Maximum likelihood in three-way contingency tables. J. Royal. Statist. Soc., B, 25, 220-233.

Bishop, Y., Fienberg, S., and Holland, P.W. (1975). Discrete Multivariate Analysis : Theory and Practice. M.I.T. Press, Cambridge, Massachusetts.

Bishop, Y.M.M. (1971). Effects of collapsing multidimensional contingency tables. Biometrics, 27, 545-562.

Bradu, D. and Gabriel, K.R. (1978). The Biplot as a Diagnostic Tool for Models of Two-Way Tables. Technometrics, 20, 47-68.

Brown, M.B. (1976). Screening effects in multidimensional contingency tables. Applied Statistics, 25, 37-46.

Chernoff, H. (1973). The Use of Faces to Represent Points in k-Dimensional Space Graphically. J.A.S.A., 68(342), 361-368.

Cochran, W. (1952). The χ^2 test of goodness-of-fit. Ann. Math. Statist., 23, 315-345.

Craig, C.C. (1953). Combination of neighbouring cells in contingency tables. J.A.S.A., 48, 104-112.

Dixon, W.J. et al. (1982). BMDP Statistical Software 1982. University of California Press, Berkeley, California.

Dahlquist, G. and Björck, A. (1974). Numerical Methods. (trans. Anderson, N.). Prentice Hall Inc., New Jersey.

Everitt, B.S. (1974). Cluster Analysis. Heinemann, London.

Everitt, B.S. (1978). Graphical Techniques for Multivariate Data. Heinemann, London.

Fielding, A. and O'Muircheartaigh, C.A. (1977). Binary segmentation in Survey Analysis with Particular Reference to AID. The Statistician, 26(1), 17-28.

Friedman, H.P. and Rubin, J. (1967). On some invariant criteria for grouping data. J.A.S.A., 62, 1159-1178.

Gabriel, K.R. (1971). The biplot graphic display of matrices with application to principal component analysis. Biometrika, 58, 453-467.

Gabriel, K.R. (1981). Biplot display of multivariate matrices for inspection of data and diagnosis. In Barnett, V. (ed.), Interpreting Multivariate Data, 147-174. Wiley, Chichester, U.K.

Gokhale, D.V. and Kullback, S. (1978). The Information in Contingency Tables. Marcel Dekker Inc., New York and Basel.

Goodman, L.A. (1971). The analysis of multidimensional contingency tables; stepwise procedures and direct estimation methods for building models for multiple classifications. Technometrics, 13(1), 33-61.

Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53, 325-338.

Greenacre, M.J. (1978a). Some Objective Methods of Graphical Display of a Data Matrix. Special Report, Department of Statistics and Operations Research, University of South Africa (November).

Greenacre, M.J. (1978b). Graphical display of data matrices. UNISA Module STA444 (supplementary notes), University of South Africa.

Greenacre, M.J. (1981). Practical Correspondence Analysis. In Barnett, V. (ed.), Interpreting Multivariate Data, 119-146. Wiley, Chichester, U.K.

Greenacre, M.J. (1982). Correspondence Analysis of Ratings and Marks. (Research Report), University of South Africa.

Greenacre, M.J. (in press). Theory and Application of Correspondence Analysis. Academic Press, London.

Greenacre, M.J. and Underhill, L.G. (1982). Scaling a data matrix in a low-dimensional Euclidean space. In Hawkins, D.M. (ed.), Topics in Applied Multivariate Analysis, 183-268. Cambridge University Press, Cambridge, U.K.

Hand, D.J. (1981). Discrimination and Classification. John Wiley & Sons, New York.

Hartigan, J.A. (1972). Direct Clustering of a Data Matrix. J.A.S.A., 67(337), 123-129.

Hawkins, D.M. and Kass, G.V. (1982). Automatic Interaction Detection. In Hawkins, D.M. (ed.), Topics in Applied Multivariate Analysis, 267-300. Cambridge University Press, Cambridge, U.K.

Hawkins, D.M., Muller, M.W. and ten Krooden, J.A. (1982). Cluster Analysis. In Hawkins, D.M. (ed.), Topics in Applied Multivariate Analysis, 301-351. Cambridge University Press, Cambridge, U.K.

Hill, M.O. (1974). Correspondence Analysis : A Neglected Multivariate Method. Applied Statistics, 23, 340-354.

Hogg, R.V. and Craig, A.T. (1968). Introduction to Mathematical Statistics. Macmillan, New York.

Jones, B. (1979). Cluster analysis of some social survey data. Bulletin in Applied Statistics (BIAS), 6/1, 25-56.

Kass, G.V. (1975). Significance Testing in Automatic Interaction Detection (A.I.D.). Applied Statistics, 24, 178-189.

Kotze, T.J.v.W. (1982). The log-linear model and its application to multiway contingency tables. In Hawkins, D.M. (ed.), Topics in Applied Multivariate Analysis, 142-182. Cambridge University Press, Cambridge, U.K.

Kruskal, J.B. (1964a). Multidimensional scaling by optimizing goodness-of-fit to a non-metric hypothesis. Psychometrika, 29, 1-27.

Kruskal, J.B. (1964b). Non-metric multidimensional scaling: a numerical method. Psychometrika, 29, 115-129.

Kullback, S. (1973). Program Manual. Dept. of Statistics. The George Washington University, Washington, D.C. 20052.

Lachenbruch, P.A. (1975). Discriminant analysis. Hafner Press, London.

Lanczos, C. (1961). Linear Differential Operators. Van Nostrand, London.

Mandel, J. (1961). Non-additivity in two-way analysis of variance. J.A.S.A., 56, 878-888.

Morgan, J.N. and Sonquist, J.A. (1963). Problems in the analysis of survey data, and a proposal. J.A.S.A., 58, 415-434.

Nelder, J.A. (1974). Log Linear Models for Contingency Tables : A Generalization of Classical Least Squares. Applied Statistics, 23(3), 323-329.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized Linear Models. J. Royal. Statist. Soc., A, 3, 370-384.

Plackett, R.L. (1974). The Analysis of Categorical Data. Charles Griffin & Company Ltd., London.

Ray, A.A. (ed.) (1982). SAS User's Guide : Statistics. SAS Institute Inc., Cary, U.S.A.

Reynolds, H.T. (1977). Some Comments on the Causal Analysis of Surveys with log-linear models. American Journal of Sociology, 83(1), 127-143.

Rosen, J.B., Mangasarian, O.L. and Ritter, K. (1970). Non Linear Programming. Academic Press, New York.

Shepard, R.N. (1962a). The analysis of proximities : multidimensional scaling with an unknown distance function I. Psychometrika, 27, 125-140.

Shepard, R.N. (1962b). The analysis of proximities : multidimensional scaling with an unknown distance function II. Psychometrika, 27, 219-246.

Solomon, H. (1977). Data Dependent Clustering Techniques. Classification and Clustering, 155-173.

Sonquist, J.N. and Morgan, J.A. (1964). The detection of interaction effects. Monograph No. 35, Survey Research Centre, Institute for Social Research, University of Michigan.

Sykes, J.B. (ed.) (1976). The Concise Oxford Dictionary of Current English. Sixth Edition.

Tabet, N. (1973). Programme d'analyse des correspondances. Part of doctoral thesis, 3e cycle, Université de Paris VI, Paris.

Tukey, J.W. (1977). Exploratory Data Analysis. Addison-Wesley, Reading, Massachussetts.

Tukey, P.A. and Tukey, J.W. (1981). Graphical display of data sets in three or more dimensions. 1. Preparation; prechosen sequence of views. 2. Data-driven selection; agglomeration and sharpening. 3. Summarization; smoothing; supplemented views. In Barnett, V. (ed.), Interpreting Multivariate Data, 189-278. Wiley, Chichester, U.K.

Whittmore, A.S. (1978). Collapsibility of Multidimensional Contingency Tables. J. Royal. Statist. Soc., B, 40(3), 328-340.

APPENDIX A

Appendix A1 : Tables of 2- and 3-way marginals for the 15 data sets with known structure

Table 22 : AB, 2-way marginal table

	A1	A2	A3	1-st order marginal
B1	800	3000	2200	6000
B2	2200	1000	800	4000
1-st order marginal	3000	4000	3000	10000

Table 23 : AC, 2-way marginal table

	A1	A2	A3	1-st order marginal
C1	1000	3000	1000	5000
C2	2000	1000	2000	5000
1-st order marginal	3000	4000	3000	10000

Table 24 : AD, 2-way marginal table

	A1	A2	A3	1-st order marginal
D1	2500	3500	1000	7000
D2	500	500	2000	3000
1-st order marginal	3000	4000	3000	10000

Table 25 : BC, 2-way marginal table

	C1	C2	1-st order marginal
B1	3100	2900	6000
B2	1900	2100	4000
1-st order marginal	5000	5000	10000

Table 26 : BD, 2-way marginal table

	D1	D2	1-st order marginal
B1	4000	2000	6000
B2	3000	1000	4000
1-st order marginal	7000	3000	10000

Table 27 : EF, 2-way marginal table

	F1	F2	1-st order marginal
E1	800	1200	2000
E2	200	7800	8000
1-st order marginal	1000	9000	10000

Table 28 : EG, 2-way marginal table

	G1	G2	1-st order marginal
E1	1200	800	2000
E2	1800	6200	8000
1-st order marginal	3000	7000	10000

Table 29 : EH, 2-way marginal table

	H1	H2	1-st order marginal
E1	1400	600	2000
E2	2600	5400	8000
1-st order marginal	4000	6000	10000

Table 30 : FH, 2-way marginal table

	H1	H2	1-st order marginal
F1	800	200	1000
F2	3200	5800	9000
1-st order marginal	4000	6000	10000

Table 31 : ABC, 3-way marginal table

		C1	C2	1-st order marginal
A1	B1	300	500	
	B2	700	1500	3000
A2	B1	2400	600	
	B2	600	400	4000
A3	B1	400	1800	
	B2	600	200	3000
1-st order marginal		5000	5000	10000

Table 32 : ABD, 3-way marginal table

		D1	D2	1-st order marginal
A1	B1	600	200	
	B2	1900	300	3000
A2	B1	2800	200	
	B2	700	300	4000
A3	B1	600	1600	
	B2	400	400	3000
1-st order marginal		7000	3000	10000

Table 33 : EFG, 3-way marginal table

		G1	G2	1-st order marginal
E1	F1	200	600	
	F2	1000	200	2000
E2	F1	120	80	
	F2	1680	6120	8000
1-st order marginal		3000	7000	10000

The data sets for the 15 models with known structure, as well as for the the 6 which had random error added, will be made available on request. They are not be given in this thesis due to limitations on space.

Appendix A2 : Tables giving the contribution of each variable to the first 5 axes of inertia for the 15 data sets with known structure as well as for the 6 which had random errors added.

See Tables 34-36.

TABLE 34 : THE CONTRIBUTION OF EACH VARIABLE TO THE FIRST 5 AXES OF INERTIA (MODELS (1)-(10))

MODEL	Contribution to the inertia of axis																			
	1				2				3				4				5			
	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
1. A,B,C,D*																				
2. BC,A,D	0,0	50,0	50,0	0,0	88,4	0,0	0,0	11,7	12,3	0,0	0,0	87,7	99,3	0,0	0,0	0,7	0,0	50,0	50,0	0,0
3. AD,B,C	50,0	0,0	0,0	50,0	100,0	0,0	0,0	0,0	0,0	0,0	100,0	0,0	0,0	100,0	0,0	0,0	50,0	0,0	0,0	50,0
4. AC,AD,B	48,0	0,0	20,6	34,4	60,1	0,0	26,4	13,4	0,0	100,0	0,0	0,0	0,0	100,0	0,0	0,0	53,9	0,0	7,2	38,9
5. BC,AD	50,0	0,0	0,0	50,0	0,0	50,0	50,0	0,0	100,0	0,0	0,0	0,0	0,0	50,0	50,0	0,0	50,0	0,0	0,0	50,0
6. AC,AD,BD	43,9	1,7	19,6	34,7	58,1	2,4	27,6	12,1	4,0	95,4	0,0	0,6	40,6	0,4	45,8	13,2	53,4	0,2	7,0	39,4
7. AB,AC,AD	44,9	0,0	21,0	34,2	46,2	38,1	10,0	5,8	0,7	29,2	50,0	20,0	54,3	31,5	13,8	0,3	53,9	1,2	5,2	39,7
8. ABC,D	48,1	29,2	22,6	0,0	53,6	20,0	26,4	0,0	0,0	0,0	0,0	100,0	46,5	22,1	31,4	0,0	51,9	28,7	19,4	0,0
9. ABC,AD	44,7	1,0	17,4	36,9	48,3	37,0	12,6	2,1	1,5	30,0	48,8	19,6	51,6	31,9	14,2	2,3	53,9	0,0	7,0	39,2
10.ABC,ABD	42,7	1,0	19,6	36,8	48,5	38,0	11,4	2,2	3,3	31,0	48,6	17,3	52,0	28,1	19,8	0,0	53,4	2,0	0,8	43,7

* Independence model

TABLE 35 : THE CONTRIBUTION OF EACH VARIABLE TO THE FIRST 5 AXES OF INERTIA (MODELS (11) - (15))

MODEL	Contribution to the inertia of axis 1								Contribution to the inertia of axis 2							
	A	B	C	D	E	F	G	H	A	B	C	D	E	F	G	H
11. A,B,C,D,E,F,G,H *																
12. BC,AD,EF,EG,H	0,0	0,0	0,0	0,0	43,4	35,4	21,1	0,0	50,0	0,0	0,0	50,0	0,0	0,0	0,0	0,0
13. ABC,D,EF,EG,FH	48,1	29,2	22,6	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	50,0	0,0	50,0	0,0
14. ABC,AD,EF,EG,EH	0,0	0,0	0,0	0,0	40,0	28,9	16,4	14,7	44,7	1,0	17,4	36,9	0,0	0,0	0,0	0,0
15. ABC,ABD,EFG,EH	0,6	0,0	0,2	0,6	42,9	27,1	11,7	16,7	42,2	2,0	17,5	36,9	0,6	0,4	0,1	0,2

MODEL	Contribution to the inertia of axis 3								Contribution to the inertia of axis 4							
	A	B	C	D	E	F	G	H	A	B	C	D	E	F	G	H
11. A,B,C,D,E,F,G,H *																
12. BC,AD,EF,EG,H	0,0	50,0	50,0	0,0	0,0	0,0	0,0	0,0	46,5	0,0	0,0	0,0	0,0	0,0	0,0	53,4
13. ABC,D,EF,EG,FH	0,0	0,0	0,0	0,0	0,0	50,0	0,0	50,0	53,6	20,0	26,4	0,0	0,0	0,0	0,0	0,0
14. ABC,AD,EF,EG,EH	48,3	37,0	12,6	2,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,2	40,6	59,1
15. ABC,ABD,EFG,EH	48,6	36,9	13,3	1,3	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	28,8	71,1	0,0

* Independence model

MODEL	Contribution to the inertia of axis 5							
	A	B	C	D	E	F	G	H
11. A,B,C,D,E,F,G,H *								
12. BC,AD,EF,EG,H	53,2	0,0	0,0	0,0	0,0	0,0	0,0	46,7
13. ABC,D,EF,EG,FH	0,0	0,0	0,0	100,0	0,0	0,0	0,0	0,0
14. ABC,AD,EF,EG,EH	1,5	30,0	48,8	19,6	0,0	0,0	0,0	0,0
15. ABC,ABD,EFG,EH	0,0	0,0	0,0	0,0	2,6	12,8	4,4	80,0

* Independence model

TABLE 36 : THE CONTRIBUTION OF EACH VARIABLE TO THE INERTIA OF THE FIRST 5 AXES OF INERTIA (DATA SETS 1-6)

DATA SET	Contribution to the inertia of axis 1								Contribution to the inertia of axis 2							
	A	B	C	D	E	F	G	H	A	B	C	D	E	F	G	H
(1)	50,1	0,0	0,0	50,0	–	–	–	–	9,1	45,7	45,1	0,0	–	–	–	–
(2)	44,9	1,7	16,6	36,8	–	–	–	–	48,7	36,4	13,5	1,4	–	–	–	–
(3)	46,2	1,7	15,2	36,9	–	–	–	–	49,6	6,7	33,3	10,4	–	–	–	–
(4)	0,4	0,5	0,0	0,0	39,3	27,8	16,9	15,0	45,2	0,6	17,9	36,1	0,0	0,1	0,0	0,0
(5)	48,3	30,5	19,4	0,0	0,1	0,9	0,2	0,5	5,0	2,9	2,8	0,0	44,0	0,3	44,7	0,2
(6)*																
	Contribution to the inertia of axis 3								Contribution to the inertia of axis 4							
	A	B	C	D	E	F	G	H	A	B	C	D	E	F	G	H
(1)	90,9	3,9	5,2	0,0	–	–	–	–	0,0	50,3	49,6	0,0	–	–	–	–
(2)	1,2	29,2	48,3	21,2	–	–	–	–	51,3	32,6	12,7	3,4	–	–	–	–
(3)	11,3	88,0	0,0	0,7	–	–	–	–	40,8	3,4	43,0	12,8	–	–	–	–
(4)	48,8	36,5	11,0	2,7	0,3	0,1	0,4	0,2	0,0	0,0	0,0	0,0	0,0	0,8	37,1	62,1
(5)	46,1	16,1	26,6	0,0	5,4	0,8	4,6	0,5	1,2	0,2	0,8	0,4	0,5	47,9	0,6	48,4
(6)*																

* Independence model

	Contribution to the inertia of axis 5							
	A	B	C	D	E	F	G	H
(1)	50,1	0,0	0,0	50,0	-	-	-	-
(2)	53,9	0,0	8,9	37,1	-	-	-	-
(3)	52,1	0,1	8,4	39,3	-	-	-	-
(4)	1,4	29,7	47,6	18,2	1,0	0,9	1,3	0,9
(5)	0,0	0,0	0,0	99,2	0,0	0,0	0,1	0,5
(6) *								

* Independence model

Appendix A3 : SAS program to perform Correspondence Analysis on the data set with structure AB,C (see Chapter 5)

```
*                                                                    ;
*THE FOLLOWING STATEMENTS DETERMINE THE PRINCIPAL CO-ORDINATES OF THE ;
*SUBJECTS AND OBJECTS IN THE 3-DIMENSIONAL SPACE :                   ;
*                                                                    ;
DATA;
INPUT A1  A2  B1  B2  C1  C2;
*                                                                    ;
*DATA SET IN WEIGHTED FORM                                           ;
* (FREE FORMAT)                                                      ;
*                                                                    ;
CARDS;
10  0 10  0 10  0
40  0 40  0  0 40
 4  0  0  4  4  0
16  0  0 16  0 16
 0  2  2  0  2  0
 0  8  8  0  0  8
 0  4  0  4  4  0
 0 16  0 16  0 16
;
RUN;
PROC MATRIX PRINT;
FETCH W;
P = W #/SUM(W);
R = P(,+);
C = P(+,);
SDR = INV(DIAG(SQRT(R)));
SDC = INV(DIAG(SQRT(C)));
Q = P - (R*C);
A = SDR*Q*SDC;
*                                                                    ;
*WHERE W IS THE DATA SET IN THE WEIGHTED FORM                        ;
*      P IS THE MATRIX OF PROPORTIONS                                ;
*      R AND C ARE THE VECTORS OF ROW AND COLUMN SUMS OF P AND       ;
*      SDR,SDC,Q & A ARE FUNCTIONS OF P,R & C                        ;
*      THE NEXT 14 LINES OF THE PROGRAM CALCULATE THE EIGENVALUE     ;
*      DECOMPOSITION OF ATA & AAT AND THE PRINCIPAL CO-ORDINATES OF  ;
*      THE SUBJECTS AND OBJECTS ON THE FIRST 3 AXES OF INERTIA       ;
*                                                                    ;
ATA = A'*A;
AAT = A*A';
EIGEN DA V ATA;
EIGEN DD U AAT;
GS V T L V;
GS U T L U;
DAP = DA>10E-8;
DAP =SUM(DAP);
```

```
U = U(,1:DAP);
V = V(,1:DAP);
DA=DA(1:DAP,);
DA = SQRT(DA);
F = SDR*U*DIAG(DA);
G = SDC*V*DIAG(DA);
*                                                                              ;
*      F IS THE MATRIX OF ROW (SUBJECT) CO-ORDINATES                           ;
*      G IS THE MATRIX OF COLUMN (OBJECT) CO-ORDINATES                         ;
*      THE NEXT FEW LINES OF THE PROGRAM ORGANIZE THE OUTPUTTING OF            ;
*      THE OBJECT CO-ORDINATES SO THAT THEY CAN BE PLOTTED USING THE           ;
*      SAS PLOTTER                                                             ;
*                                                                              ;
R ='A1' 'A2' 'B1' 'B2' 'C1' 'C2';
C = 'AXIS1' 'AXIS2' 'AXIS3';
OUTPUT G ROWNAME=R COLNAME=C OUT=COUT;
RUN;
*                                                                              ;
*      THE FOLLOWING STATEMENTS FACILITATE THE PLOTTING OF THE                 ;
*      OBJECTS ON THE FIRST 2 AXES OF INERTIA                                  ;
*                                                                              ;
PROC GPLOT;
GOPTIONS VSIZE=10 HSIZE=8;
TITLE1 .C=B .H=1 .F= ITALIC FIGURE 9 : DATA SET WITH STRUCTURE;
TITLE2 .C=B .H=1 .F= ITALIC AB,C;
TITLE3 ;
TITLE4 .C=G .H=1 .F= SIMPLEX PROJECTION OF OBJECTS ON AXES 1 AND 2;
TITLE5 ;
PLOT AXIS1*AXIS2=ROW/OVERLAY HREF=0 VREF=0 NOLEGEND;
SYMBOL1 V=X C=RED;
SYMBOL2 V=X C=RED;
SYMBOL3 V=X C=RED;
SYMBOL4 V=X C=RED;
SYMBOL5 V=X C=RED;
SYMBOL6 V=X C=RED;
RUN;
```