The graphical representation of structured multivariate data

Thesis

The Graphical Representation of Structured Multivariate Data

by

Michaela J. Cottee, B.Sc.(Hons), M.Sc.

A thesis submitted to The Open University for the degree of Doctor of Philosophy

Faculty of Mathematics & Computing, The Open University

December 1996

Abstract

During the past two decades or so, graphical representations have been used increasingly for the examination, summarisation and communication of statistical data. Many graphical techniques exist for exploratory data analysis (ie. for deciding which model it is appropriate to fit to the data) and a number of graphical diagnostic techniques exist for checking the appropriateness of a fitted model. However, very few techniques exist for the representation of the fitted model itself. This thesis is concerned with the development of some new and existing graphical representation techniques for the communication and interpretation of fitted statistical models.

The first part of this thesis takes the form of a general overview of the use in statistics of graphical representations for exploratory data analysis and diagnostic model checking. In relation to the concern of this thesis, particular consideration is given to the few graphical techniques which already exist for the representation of fitted models.

A number of novel two-dimensional approaches are then proposed which go part-way towards providing a graphical representation of the main effects and interaction terms for fitted models. This leads on to a description of conditional independence graphs, and consideration of the suitability of conditional independence graphs as a technique for the representation of fitted models. Conditional independence graphs are then developed further in accordance with the research aims.

Since it becomes apparent that it is not possible to use any of the approaches taken in order to develop a simple two-dimensional pen-and-paper technique for the unambiguous graphical representation of all fitted statistical models, an interactive computer package based on the conditional independence graph approach is developed for the construction, communication and interpretation of graphical representations for fitted statistical models. This package, called the "Conditional Independence Graph Enhancer" (CIGE), does provide unambiguous graphical representations for all fitted statistical models considered.

Acknowledgements

Above all, I would like to thank Professor David Hand for his support and guidance throughout the work described in this thesis, and for his assistance and forbearance during the past few years. It has been a privilege to work with him.

I would also like to thank Dr. Kevin McConway of the Open University for his contribution to the discussion of some of the work herein, and Dr. Joe Whittaker of Lancaster University for his suggestions. Particular thanks are due to Dr. Robert Hasson of the Open University for the practical computing advice he gave whilst CIGE was being implemented.

Special thanks are due to my parents, Don and Evelyn Cottee, who gave me the encouragement and support necessary to enable me to complete this thesis. Thanks are also due to Paul, June, Caroline and Dave for their friendship and support, and to all my friends and acquaintances who, each in their own way, have helped me make it to the end! In particular I would like to thank Martin for 'being there'. I would also like to acknowledge the contribution made by Arthur, who was just a kitten when I started all this.

Finally I would like to acknowledge the financial support of the Economic and Social Research Council, and to thank the Open University and the University of Hertfordshire for the use of their facilities.

Contents

List of Figures

List of Tables

1. Introduction to the Research Problem

1.1 Introduction

This thesis is concerned with the graphical representation of structured multivariate data or, in other words, with the graphical display of fitted statistical models.

To appreciate the need within statistics for graphical displays in general, and for graphical displays of fitted models in particular, it is first necessary to consider the role and importance of graphical representations within statistics to date. This is done in Section 1.2.

The nature of the research problem and how it is to be tackled in this thesis is then specified in Section 1.3.

Throughout this introductory chapter, I shall be considering statistical graphics in general. In subsequent chapters I shall discuss individual techniques and the particular uses to which they are most suited, such as exploratory data analysis, experimental design, diagnostic model checking, and the representation of fitted models. As shall be seen, in both this and subsequent chapters, very few techniques exist for the graphical representation of structured multivariate data in the form of fitted models, and this will be the concern of this thesis, as described in Section 1.3.

1.2 The Importance of Graphical Representations

In this section, I hope to demonstrate the importance of graphical representation techniques in statistics using evidence contained within the literature. I shall begin by briefly outlining the history of the use of graphical techniques in statistics from the 10th Century and through the work of Playfair. The past twenty-five years or so have seen a surge in the development and use of graphical techniques in statistics, and I shall pay particular attention to the developments which have taken place during this time. The various uses to which statistical techniques are put will be considered, and current areas of research shall be identified. I shall also make some suggestions as to the future role of graphical representations and consider directions in which developments may take place, such as virtual reality.

1.2.1 History

Although most graphs seem intuitively simple, their invention was neither simple nor obvious, according to Lewandowsky & Spence (1989b). The idea did not occur to the Greeks or Romans, nor even to great 17th century mathematicians like Newton and Leibnitz. The earliest attempt at the graphical depiction of empirical data has been traced by Beniger & Robyn (1978) back at least as far as the 10th or 11th Century, but it was only after the work of Playfair, who developed the bar-chart and the pie-chart for the display of economic data in the late 18th / early 19th Centuries that the use of graphs and charts for data display became commonplace (Costigan-Evans & MacDonald-Ross (1990)), although the histogram has been traced back to 1662 (Scott (1979)). Lewandowsky and Spence claim that since most graphs are Cartesian in nature, their development owes much to Descarte's *La Geometrie*, published in 1637.

Two excellent historical sources are the detailed appendix of historical developments in statistical graphics presented by Beniger & Robyn, and the book by Tufte (1983).

1.2.2 Development

Within the past 25 years a number of new and innovative graphical techniques have been developed. To quote Beniger and Robyn (1978): "Today, statistical graphics appear to be re-emerging as an important analytic tool, with recent innovations exploiting computer graphics and related techniques". The same statement is still true today, almost 20 years later.

Gnanadesikan (1981) also attributes the renaissance of graphical techniques in statistics to the development of easily and affordable and accessible graphics hardware, even though many widely used statistical software packages do not include the majority of available graphical tools. Another reason for the previous neglect of statistical graphics may have been the emphasis on more theoretical aspects of statistics.

Cleveland & McGill (1987) also suggest that statistical graphics is a newly activated area of statistics because of the 'computer graphics revolution' – high quality systems are now available at low cost, and software is being developed for graphing data.

Thus the surge of interest in the use of graphical representations in statistics has been coupled with an increase in the availability of computing power; both in terms of

reduced costs, and enhanced capabilities. Ball & Hall (1970), considered the implications of computer graphics and concluded that the use of interactive computer graphics systems should increase the ease of presenting and manipulating conceptual material since the interactive aspect allows the user to convey his needs to the computer rapidly, the graphic aspect allows the results of a computation to be conveyed rapidly in a quickly conceivable form, and the graphics computer allows the manipulation of graphics and the development of new graphical representations of complex data structures.

Today, statistics packages for personal computers (commonly for IBM-PC compatibles or Apple Macintosh computers) regularly incorporate graphics routines for representing data. Most usually, the graphics available are for the representation of raw data, in the form of pie-charts, histograms, and scatterplots — all well-known graphical techniques which were in existence before computing became so readily available. Other, more recent, techniques available for exploring the raw data usually include box-and-whisker plots, stem-and-leaf plots, and multi-dimensional scatterplots. Other graphical techniques are available, not for the exploration of the data, but for model diagnostics; for example, showing fitted regression curves, and plotted residuals. Examples of packages which are available, and which offer both computational routines and graphical displays, include Statgraphics, S-PLUS and SYSTAT. Personal computers provide graphics capabilities cheaply, and the displays obtained are adequate for most purposes. However, workstations, which are becoming more common, can provide very high resolution colour or monochrome graphics, and many of the packages for personal computers mentioned above have been implemented on such workstations. Conversely, many main-frame based statistics packages, such as SPSS, SAS, GLIM, Genstat and MINITAB, which tended to have quite crude graphics utilities, have become available in formats suitable for personal computers, and are increasingly incorporating more and more sophisticated graphics options. However, the quality of the graphics produced may still be restricted by limitations of the available hardware, especially printers.

In general, the increasing availability of fast high-resolution full-colour graphics displays and printing has encouraged the use of graphical representations in statistics. This is not to say, however, that pen-and-paper graphical representations no longer serve a purpose. Often a back-of-the-envelope type preliminary sketch can guide the statistician prior to a more formal analysis. However, it can be argued that a graphical representation can be obtained more accurately with the aid of a computer-based package, and where the

user is likely to analyse the data with the same package, little extra effort is required, whilst providing the facility for manipulation and good quality reproduction.

1.2.3 Uses

Mahon (1977) believes that failure on the part of the statistician to communicate his [*sic*] findings could mean that all of his work is wasted. According to Mahon, the three available media for communicating statistical information are words, tables and pictures, all of which may need to be used in combination for really effective communication. Mahon suggests that tables are best for communicating values, whilst graphs are best for communicating relationships. Although he warns that drawing graphs "is more of an art than a science", a good picture can provide insight and make a point very clearly. To quote: "A good picture can convey instantly, and memorably, a relationship that would otherwise require a laborious and easily forgotten explanation".

Fienberg (1979) goes so far as to claim that "graphical methods have played a central role in the development of statistical theory and practice". However, practices in statistical graphics are widely varied, and there is no theory of, or standards for, statistical graphics. Fienberg lists five benefits of using statistical graphics, originally presented by Schmid (1983), as follows:

1. Well-designed charts are more effective than other types of presentation in creating interest and in appealing to the attention of the reader.
2. Visual relationships, as portrayed by charts and graphs, are more clearly grasped and more easily remembered.
3. The use of charts and graphs saves time, since the essential meaning of large masses of statistical data can be visualised at a glance.
4. Charts and graphs can provide a comprehensive picture of a problem which makes possible a more complete and better-balanced understanding than could be derived from tables or text.
5. Charts and graphs can bring out hidden facts and relationships, and can stimulate and aid analytical thinking and investigation.

Thus uses for graphs and charts suggested by Schmid are illustration, analysis, and computation. Uses for graphs suggested by Tukey (1972) are: to show what has been

learnt using some other technique; to allow us to see what is happening above and beyond what is already known; to allow numbers to be conveyed; and for decoration.

Cox (1978) mentions that graphical techniques may have an important role in the teaching of statistics, to illustrate theoretical points, or in model formulation, including the testing of assumptions.

Gentleman (1977) also discusses the usefulness of computer graphics systems in teaching statistics. Graphics can be easily generated, and hard copies obtained, which will suitably demonstrate the concepts or techniques being taught. Such graphical displays can be included in text-books. Also, according to Gentleman, consultants can make use of interactive graphics in the analysis of data, and in presenting and explaining the results to the client.

Ball & Hall (1970) believe that three different populations of statisticians profit from developments in interactive computer graphics systems, these being:

1. ***Data analysts:*** Interactive computer graphics systems allow increased convenience in data analysis, especially in data manipulation. More complex and different techniques can be used, and more data exploration can be carried out cheaply. This, Ball & Hall suggest, could lead to a change in scientific research with much less emphasis on formal hypothesis testing.
2. ***Statisticians:*** Interactive graphic computer systems allow the development of new statistical tools. To quote: "Possibly...the interactive system will permit development of a new symbol set in which graphics is used much more intensively in the actual research toward developing and understanding statistical techniques. ...It seems possible that the development of new symbol sets that can be manipulated within an interactive graphic computer will cause statistical research to move in entirely new directions".
3. ***Statistics teachers:*** Interactive graphic computer systems provide a host of new teaching tools for communicating characteristics of statistics to students.

Cleveland (1984a) points out that graphs allow vast amounts of information to be summarised, and readily reveal quantitative patterns and relationships in data. As such they are analogous to written language, and vital for communication in science.

The surge in the availability and use of graphical methods in statistics, and the proliferation of techniques which have been developed within the past two decades or so, is indicative of the importance of graphical techniques in statistics. What statistician would plunge into an analysis without first checking the form of the data by some exploratory graphical means, and subsequently checking the results of the analysis by some diagnostic graphical means?

1.2.4 Current Research

The three currently active (overlapping) areas of research in graphical statistical methods are considered by Cleveland (1987) to be as follows:

1. **Methodology:** concerned with what information should be shown on a graph in order to explore the data or for model diagnostics. Dynamic methods allow direct manipulation of the graphical elements on the computer screen with near-instantaneous change.
2. **Computing:** concerned with building the interface between the statistician and the computational procedures, in order to produce the displays.
3. **Graphical perception:** aims to provide a scientific foundation for the evaluation of the methodology, and is concerned with the visual decoding of the information contained in the graph.

Cleveland also points out that, unlike numerical statistical procedures, there is no well-developed theory which provides criteria for the evaluation of graphical methods. To quote: "Inventing a graphical method is easy. Inventing one that works is difficult. Figuring out if a particular invention works is more difficult".

Even a quarter of a century ago, Ball & Hall (1970) believed that 'current' facilities for computing, display, and real time interaction had developed beyond our understanding of how to use them effectively in data analysis, and stated that a deeper insight is needed into the psychology of graphs, pictures and output media in general, both for use in interaction and communication.

1.2.5 The Future

Beniger and Robyn (1978) suggested that future innovations in statistical graphics were likely to follow developments in computer hardware and software, including colour computer graphics, computer animation, three-dimensional computer graphics, and even holography and mechanically controlled sound. It is certainly possible to report that colour graphics, animation and 3-dimensional graphics have played, and continue to play, a role in the development of new and existing graphical techniques. Even sound has been used (Mezrich et al (1984), Wilson (1991)). However, it now seems unlikely that holography could have a practical role to play in the development of statistical graphics, since comparable 3-dimensional representations can be achieved using modern interactive computer graphics.

Beniger and Robyn were probably not in a position, in 1978, to include "virtual reality" amongst their list of developing aspects of computer hardware and software. Virtual reality is a relatively new approach to interactive computer use which, with the aid of special goggles and sometimes also a 'data glove' or whole 'data suit', immerses the user in a 3-dimensional 'reality' created by the computer in which they can move and interact. Virtual reality has the potential to be used to move around inside and to explore representations of vast data sets. Worrall (1991) reports that just such a virtual reality system has already been created whereby the data is represented as a forest which the user can fly over. Use of a data glove with a virtual reality representation of the data may be a solution to the inadequacy, mentioned by Huber (1987), of current (2-dimensional) pointer devices used with conventional 3-dimensional representations.

To bring us back to reality, Fienberg (1979) concludes in his survey that there is a need for more and better use to be made of existing graphical techniques by statisticians and in journals, and for more experiments to be conducted on graphical methods (involving both statisticians and cognitive psychologists) for the development of a theory for statistical graphs.

Cox (1978) also suggests a need for the development of a theory of graphical methods which will bring the field together in a coherent manner, and provide a basis for dealing with new situations.

Cleveland (1984a) believes that statisticians could play a vital role in effecting an improvement in graphical communication in science, but five areas which would require more thought are as follows:

1. The assessment and development of graphical methods for data presentation.
2. The development of guide-lines to ensure clarity of graphs (both clarity of presentation and of explanation).
3. The study of human graphical perception; ie. of how people visually decode the quantitative information encoded on a graph.
4. Software design.
5. How people currently use graphs in science.

1.3 The Research Problem

Four very important components of any statistical analysis are: experimental design, data exploration, data modelling, and the communication and interpretation of the results. As was indicated in the preceding section and shall also be seen in the following chapters, graphical representations are of relevance for each of these components of statistical analysis. It is the concern of this thesis, however, that, whilst numerous graphical techniques exist for representation of the raw data, and are therefore of importance for determining which model it would be appropriate to fit to the data, and whilst several graphical techniques exist for diagnostic model checking and are therefore of importance for testing the appropriateness of the model which has been fitted to a given data set, very few techniques have been proposed and employed for the representation of the fitted model itself.

Thus the work in this thesis is concerned with the development of graphical techniques for the representation of fitted models. In particular, it is concerned with consideration of the few techniques which are already in existence and with the development of new and existing techniques for the representation of fitted models.

Any graphical technique which is developed for the display of a fitted model should also serve as a useful aid for the communication and interpretation of the model. It should therefore convey the terms which are contained in the fitted model, together with some indication of their significance. It is envisaged that techniques for the graphical

representation of fitted statistical models will be of particular benefit to users of statistics who may have only limited theoretical statistical knowledge, such as behavioural and social scientists, and for statistical consultants working in these areas. For example, the statistical consultant may use the graphical representation of the model to communicate features of the fitted model to the scientists, whilst the scientists themselves may use the graphical representation to assist in the interpretation of the fitted model and to communicate features of the fitted model to other scientists.

Any graphical technique which is developed should be made available in an easily implementable form, either as a static representation, which can be photocopied and incorporated into a journal, or report, etc., and/or as an interactive computer-implemented screen-based manipulable display (which may be printed out in its final form for inclusion in a journal or report).

1.3.1 Tackling the Research Problem

In the next chapter (Chapter 2), I shall outline some of the more common methods amongst the wide variety of graphical techniques which are available in statistics for the representation of raw data (generally termed Exploratory Data Analysis (EDA)). In Chapter 3 some of the techniques employed for diagnostic model checking are described. Together, these are the two areas in which graphical techniques are most frequently employed, and for which the most techniques exist. Mention will also be made of a few other, less common, uses of graphical representations in statistics, such as experimental design symbolisation, model selection procedures and multiple comparison procedures.

Chapter 4 is devoted to description and consideration of the few graphical techniques which exist in statistics for the representation of the fitted model itself. Such techniques include the well-known Analysis of Variance (ANOVA) interaction plots, and lesser known techniques such as Bond's technique for the representation of the terms in an ANOVA summary table, plus techniques for the representation of contingency tables and logit models. By far the most important technique which falls into this category, which will be considered in detail in later chapters as I develop new approaches to the representation of fitted models, is the conditional independence graph.

In Chapter 5, a number of issues in graphical perception and presentation are considered, which may have implications for the assessment of existing graphical techniques and for the development of new graphical representation techniques.

In Chapter 6, some attempt is made to develop some simple two-dimensional static representations of statistical models, but, as will be seen, the techniques developed are only of use in certain situations. In Chapter 7, the uses of conditional independence graphs in graphical modelling are described and their usefulness as a simple two-dimensional static technique for the representation of fitted statistical models is assessed. This leads to a more detailed consideration of their limitations, and the development of modifications to the conditional independence graph approach in Chapters 8, 9 and 10.

The remaining limitations of the conditional independence graph approach are overcome by the development of an interactive package, called the "Conditional Independence Graph Enhancer" (CIGE), which also incorporates the modifications developed in the preceding chapters. The features of the CIGE package are described in Chapter 11, and illustrated using a number of real and artificial example data sets in Chapter 12.

In Chapter 13, suggestions are made for further improvements and extensions to the conditional independence graph approach, as developed using CIGE, and in Chapter 14, an overall assessment is made of the success in fulfilling the research aims.

2. Graphical Techniques for Data Exploration

2.1 Introduction

There are very many techniques for the graphical representation of raw data. Indeed, this is probably the most prolific area of use of graphics within statistics. Such representations, of univariate or multivariate data, obtained either directly or indirectly through the use of dimension reduction techniques, are particularly useful for obtaining an appreciation of the structure in the data and as such can assist the interpretation of the data and suggest more formal approaches to the analysis of the data. Such representations can also be useful for the identification of outlying data values, for the examination of the distribution of data points (including the identification of clusters of points), and for the identification of data values which may need to be treated in a special way (eg. by transformation or discarding). As such, these techniques are useful for data exploration, and can therefore be said to fall into the area of statistics termed "exploratory data analysis" (EDA) by Tukey (1977).

Within this chapter, a number of graphical techniques of use for the exploration of univariate or multivariate raw data are described:

For univariate data, direct representation methods which can be used to provide a graphical display of the data include histograms and bar-charts, frequency polygons, pie charts, stem-and-leaf plots, and box-and-whisker plots (variations of which include the schematic plot, the variable-width box-plot and the notched box-plot). Modifications of all these techniques exist for the graphical representation of raw bivariate data, including the three-dimensional histogram, back-to-back stem-and-leaf plots and the rangefinder box-plot, although by far the most commonly used graphical representation technique for bivariate data is the scatterplot.

Direct representation techniques which may be used for the display of raw multivariate data, which I shall refer to as "pictorial techniques", include glyphs, stars and cartoon faces. Such pictorial techniques show considerable imagination on the part of their inventors, but may be of little practical use in the exploration of the raw multivariate data. Within this section, I shall also consider extensions to the bivariate scatterplot and the scatterplot matrix.

Dimension reduction techniques for multivariate data include graphical displays resulting from the use of "ordination techniques" such as principal components analysis, the biplot, correspondence analysis, multidimensional scaling and principal coordinates analysis. These techniques can be used to transform multi-dimensional data so that it may be displayed in two or a few dimensions. In reducing the dimensionality of the multivariate data, it is to be hoped that the structure within the data is preserved. Additional dimension reduction techniques include cluster analysis and Andrews' plots.

To conclude this chapter, some mention will be made of dynamic computer-implemented approaches to exploratory data analysis which incorporate some of the techniques described in the following sections.

No illustrations will be given of the techniques described in this chapter. It is assumed that, if the reader is not already familiar with a technique, then he or she will refer to the references given for further details and illustrations. Good general references for many of the graphical techniques for data exploration described in this chapter include Tukey (1977), Everitt (1978), Barnett (1981), Chambers *et al* (1983), Seber (1984), DuToit *et al* (1986), Digby & Kempton (1987), Krzanowski (1988), Everitt & Dunn (1991), Manly (1994), Daly *et al* (1995).

2.2 Univariate Data

2.2.1 Bar Charts and Histograms

Bar charts are constructed to show some value (eg. count, percentage, mean) for each category (ordered or unordered) of a discrete variable. A rectangular bar is drawn, of height proportional to the value to be displayed. The width of the bar does not usually convey any information, and should be the same for each category. The bars are usually drawn separated from each other.

Histograms are similar to bar charts, but are usually constructed for continuous variables. The continuous variable is divided into equally-sized intervals, such that every value falls into exactly one interval. As for the bar chart, a rectangular bar is drawn, of height proportional to the value (count, percentage, mean, etc.) to be displayed. Because of the continuous nature of the scale used to determine the classes, the bars are usually drawn adjacent to each other.

Histograms and bar charts are probably by far the most commonly used form of graphical representation. They are of use for displaying and summarising data, and in so doing can highlight the distribution of the data values and any outliers. However, the appearance of the display can be greatly affected by the choice of the size of the intervals (or bin-width). This in turn will affect the number of intervals and the number of observations falling into any interval. If there are too few intervals, this will not give a clear indication of the distribution of the data values, but if there are too many intervals, the display will not be smooth and can be misleading. Various methods have been proposed for determining the optimum number of intervals (Scott (1979)).

There are many possible variations on the histogram theme. For example, if the data can be divided naturally into categories, eg. males and females, the values for the different categories can be represented by stacked bars, eg. with the bottom bar corresponding to males and the top bar to females. Alternatively, the bars for the females can be placed alongside corresponding bars for the males (on a bar chart) or the bars can be made a fixed length and divided according to the percentage contribution by each sex, etc.

2.2.2 Frequency Polygons

The frequency polygon is essentially an alternative representation of the histogram or bar chart for ordered groups, where the represented values are counts. Instead of drawing bars for each group, a point is plotted corresponding to the top of the bar and these points are then joined together in order (Scott (1985)). To complete the polygon shape, the ends of the line are usually extrapolated to 0. The frequency polygon has the advantage of making the shape of the distribution of the data values clearer, without the distraction of the vertical bars.

A cumulative frequency polygon can be constructed by combining the frequency of occurrence for each interval with the frequencies of the preceding intervals, and plotting these cumulative frequencies in the form of a frequency polygon. The cumulative frequency polygon is of use for estimating the number (or percentage) of observations above or below a particular value.

2.2.3 Pie Charts

Pie charts are commonly used to represent counts or percentages. Thus the whole 'pie' (usually drawn as a circle) corresponds to the total or 100%, and wedges of the pie are drawn in proportion such that the area of each wedge corresponds to the count or percentage relating to a particular category. The pie chart is essentially an alternative display of the information contained in the bar chart, although Lewandowsky & Spence (1989b) argue that the bar chart is better than the pie chart if comparisons between categories are to be made.

2.2.4 Stem-and-Leaf Plots

The stem-and-leaf plot forms a display resembling a histogram of frequencies in which all of the original data values are retained. A good introduction to the stem-and-leaf plot is given by MacDonald (1982) and by Landwehr & Watkins (1985).

The stem-and-leaf plot is constructed by using the first part of each data value in each interval (eg. the tens: 0,1,...,9) as the stems to determine the positions of the data points along the (usually vertical) axis, and using the last part of each data value (eg. the units: 0,1,...,9) to give the leaves by writing each unit value contained in the data as part of a bar alongside the corresponding tens value. If the data takes some form other than tens and units, a different choice of stem and leaf may be more appropriate. If the range of data values is very small, it may be appropriate to repeat the stem values for different values of the leaves. For example: 10,11,...,14 and 15,16,...,19 would have separate stems and separate bars corresponding to the two sets of leaf values.

Thus the stem-and-leaf plot may be seen to resemble the histogram for counts, with the added feature that the bars preserve the actual data values. However, in the same way that the use of different bin-widths can affect the appearance and interpretation of a histogram (see Section 2.2.1), so the choice of stems can affect the appearance and interpretation of a stem-and-leaf plot. Jobson (1991) discusses the choice of the number of intervals for stem-and-leaf plots.

2.2.5 Box-and-Whisker Plots

In a box-and-whisker plot, or boxplot, the data themselves are not conveyed, but the locality, spread, skewness and outliers of the set of data are. The five summary statistics used to construct the plot are the median, the upper and lower quartiles, and the upper and lower extreme values. The construction of a box-and-whisker plot based on this five-figure summary is illustrated in Figure 2-1.

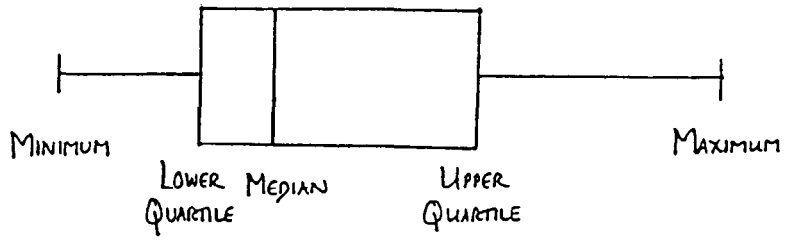

Figure 2-1: Construction of a box-and-whisker plot from five figure summary

A box-and-whisker plot allows a simple visual check to be made of the distribution of the data, and outliers can be identified, although Lewandowsky & Spence (1989b) claim that the boxplot is less effective than the stem-and-leaf plot for assessments of the distribution of data.

If two or more sets of data are to be compared, this can be done by constructing a box-and-whisker plot for each data set and plotting them in a manner in which they may be compared, for example according to the median value of each data set.

Many variations of box-and-whisker plots have been suggested, including the schematic plot. In a schematic plot, attention is drawn to outliers by representing all data values lying beyond chosen cut-off points as highlighted points. The whiskers are then drawn only as far as the most extreme values lying within the cut-off points. A common choice of cut-off point is median ± (1.5 × interquartile range).

McGill *et al* (1978) suggest three modifications to the standard box-and-whisker plot, based on their beliefs about the additional information which non-statisticians try to gain from boxplots. For example, it is their belief that non-statisticians, when shown a set

of boxplots, try to estimate the overall median by visualising an 'average' median. The three modifications are described below:

1. In a variable-width boxplot, the width of each boxplot is drawn proportional to the value of the square root of the corresponding sample size. This draws attention to size differences between samples and therefore allows better appraisal of the data.
2. In a notched boxplot, notches are drawn surrounding the median which indicate the size of the confidence interval around the median and so provide a rough indication of the significance of differences between samples. If the notches for the boxplots of two samples do not overlap, then the medians are significantly different at the level of significance used to construct the notches. If the end of the notch lies outside the box, then the notch is drawn protruding from the ends of the box.
3. In a variable-width notched boxplot, both of the above variations are combined, such that sample size and confidence intervals are seen simultaneously. However, since the confidence intervals are based on sample size, this introduces some redundancy.

2.3 Bivariate Data

Some of the univariate raw data representation techniques described in the previous section can be modified for the representation of bivariate raw data. However, the scatterplot is particularly suited to the representation of bivariate data. Extensions to some univariate representation techniques and the scatterplot are described below.

2.3.1 Histograms

A 3-dimensional histogram can be used for the display of bivariate data; for example, when frequencies have been cross-classified according to the categories or intervals of two variables. This is potentially of use for assessing the bivariate distribution of the data although in practice the 3-dimensional histogram can be difficult to read, especially if bars at the front are obscuring bars at the back. As for histograms constructed for univariate data, there are issues such as the choice of bin-width which may affect the appearance of the graphical representation.

2.3.2 Stem-and-Leaf Plots

It is possible to compare the distributions of the observations on two variables by placing two stem-and-leaf plots, one for each variable, back-to-back using the same stem.

2.3.3 Scatterplots

Probably the most commonly used and most useful technique for the representation of bivariate raw data is the scatterplot. Each of the two axes corresponds to one of the variables, and the observations are plotted in the coordinate space defined by the two axes, according to the values of each observation for the two variables. This technique is well-suited to continuous data, since exact data values can be conveyed (within the limits of the accuracy of the construction of the scatterplot). If one or both variables are discrete, there may be a greater problem with overlapping values than would be expected if both variables are continuous.

From a scatterplot, it is possible to identify visually outlying observations and clusters of observations. It is also possible to judge the apparent collinearity or correlation between the two variables, although dramatically different configurations of points may have the same correlation coefficient (as demonstrated by Anscombe (1973)). Thus a scatterplot can be looked at in terms of the location of individual points, or in terms of the overall shape. Smoothing techniques have been advocated to assist in detecting shape by reducing noise within a scatterplot (Cleveland & Kleiner (1975), Cleveland & McGill (1984b)).

Additional information may be conveyed on a scatterplot, such as the marginal distribution of each variable, by the superimposition of histograms or boxplots on the axes. Becketti & Gould (1987) describe the rangefinder box plot whereby six line segments are superimposed onto a scatterplot and display precisely the same information which would be contained in a box plot constructed for each of the variables individually. To construct a rangefinder boxplot, two central line segments are drawn on the scatterplot corresponding to the medians and intersecting at the cross-median values. The lengths of the horizontal and vertical central line segments delimit the interquartile range of the variables measured along the horizontal and vertical axes. Line segments are also drawn

on the left and right, and top and bottom, of the scatterplot where the whiskers of the box plot would terminate.

Scatterplots may also be used to display a fitted regression line (see Chapter 3); to assess bivariate normality, with the aid of superimposed elliptical contours corresponding to the bivariate normal distribution; to display data collected over time (not the concern of this thesis); and as a basis for enhancements of the data points to illustrate other aspects of the data (see Section 2.4).

2.4 Multivariate Data: Direct Representation Techniques

A number of representation techniques for raw multivariate data, which associate each observation with values on each of p measured variables, appear quite imaginative. I shall refer to these representations as "pictorial" techniques, for reasons which should become clear (particularly if one regards a picture as something which may be more decorative than useful!). Additional references for the techniques described in this section Everitt & Nicholls (1975) and Fienberg (1979).

2.4.1 Multidimensional Scatterplots

Trivariate raw data can be plotted on some sort of 3-dimensional structure, such as a 3-dimensional scatterplot contained within a cube. The 3-dimensional scatterplot is, however, more difficult to construct by hand and harder to interpret than a 2-dimensional scatterplot. Moreover, the technique cannot be directly extended to four or more variables.

Huber (1987) has considered the use of 3-dimensional scatterplots. Although he admits that 2-dimensional scatterplots may be of some use in identifying outliers and clusters of variables, and for identifying linearity or otherwise in the data, Huber suggests that 3-dimensional scatterplots are better for examining the structure of the data. For 2-dimensional scatterplots, Huber believes that the user can subconsciously relate the points to the axes, but not for 3-dimensional scatterplots. Thus 3-dimensional scatterplots may be more suited for the representation of data points normally located in 3-dimensional space.

To obtain a 2-dimensional scatterplot from a 3-dimensional scatterplot, one of the dimensions can be collapsed or, equivalently, as Huber suggests, the 3-dimensional

scatterplot can be rotated using real-time computer graphics (see Section 2.6). Alternatively, principal components analysis (see Section 2.5.1) can be applied to the 3-dimensional data to see which two 'principal component axes' will give the best 2-dimensional view.

2-dimensional scatterplots can be readily compared, if drawn on equivalent axes. However, to compare 3-dimensional scatterplots, it may be necessary to use interactive computer graphics to rotate the plots, or to enhance the plots using colours, lines, symbols, text, etc..

If one of the three variables is categorical, construction of a 3-dimensional scatterplot may result in some spurious clusters in the data. In this situation it would probably be better to construct a 2-dimensional scatterplot using the continuous variables and to superimpose information about the categorical variable onto each point, for example, by the use of shading or a symbol, or by drawing the points as circles with radius in direct proportion to the value of the third variable (sometimes called a 'bubble plot').

For three or more variables it is possible to construct a 'scatterplot matrix' or 'draughtsman's plot' formed by the 2-dimensional scatterplots drawn for all combinations of pairs of the variables. This approach is described by Carr, Littlefield, Nicholson & Littlefield (1987). For p variables, $p(p-1)$ scatterplots are required (of which half will be reflections in $x=y$). Representation of the data in this way may elucidate some 2-dimensional features of the data, such as outliers and clusters, but any higher-dimensional structure in the data may not be discernible. However, as Carr *et al* mention, the scatterplots could be enhanced to provide more information. If the scatterplots are drawn in an interactive graphical computing environment, it may be possible to highlight subsets of variables on one scatterplot which will automatically be highlighted in the other scatterplots. This technique is termed 'brushing scatterplots' and is described further in Section 5.6.

If there are only a few variables to be represented (say $3 \leq p \leq 6$) then, having constructed a 2-dimensional scatterplot using two of the variables, the remaining variables could be represented on this scatterplot, for example as straight lines emanating from the points representing the observed units, where the orientation of the line corresponds to the variable represented, the length of the line is proportional to the value of the observed unit

for that variable, and the direction of the line corresponds to the sign of the value. This variation is sometimes called a 'weather vane plot'.

Ball & Hall (1970) suggest another variation which involves tilting the line representing the third variable at an angle dependent on the value of the fourth variable. Other possible variations could make use of the glyphs and stars described below (Sections 2.4.2 and 2.4.3).

With all these variations on scatterplots, however, there is the potential difficulty of plotted information overlapping and becoming confusing for observed units which are close together.

2.4.2 Glyphs

In the construction of glyphs, proposed by Anderson (1960), a circle is drawn for each observed unit and a number of lines are drawn emanating from the top of the circle with as many lines as there are variables and each line corresponding to a particular variable. The length of each line is proportional to the value of the observed unit on that variable. The glyphs constructed for each observed unit can then be displayed together and examined for clusters, outliers, etc..

Within the literature, glyphs are typically constructed to represent five variables. With fewer variables, it may be more appropriate to use a scatterplot-based representation technique, and with more variables, the glyphs may appear quite messy. However, it is possible to reduce the number of dimensions to be represented by two by plotting the glyphs for each observed unit on a scatterplot, in positions determined by the values taken on the two variables excluded from the glyph. This may be particularly appropriate if one or two variables are related to location.

2.4.3 Stars

Stars are very similar in construction to glyphs, except that each observed unit is represented by a point, the lines emanating from the points are regularly distributed about the point (again there are as many lines as there are variables and the length of each line is drawn in proportion to the value of the observed unit on the variable represented), and the ends of the lines are joined together.

Within the literature, stars are typically constructed to represent six variables. With fewer variables the stars may resemble quadrilaterals or triangles. With many more variables, stars, like glyphs, may appear quite messy. As for glyphs, it may be appropriate to use stars with another representation technique, such as a scatterplot.

2.4.4 Cartoon Faces

Various versions of Chernoff's (1973) original cartoon faces have been proposed (eg. Flury & Riedwyl (1981)), but they all conform to the same basic principle. Thus a cartoon face is drawn for each observed unit which is comprised of a number of features whereby each feature (eg. shape of mouth, shape of eyes, shape of chin, location of nose, etc.) corresponds to a particular variable and takes on an appearance which is dependent on the value of that particular variable for the observed unit. The appearance of a particular feature may be derived by interpolating between two extremes of appearance, where these extremes correspond to the most extreme values taken by any of the observations on that variable. Most commonly, between 9 and 18 variables/features may be represented in this way although it is possible to represent more variables by increasing the number of features, or to represent fewer variables by holding a number of features constant for all faces.

Since people are experts in studying and reacting to faces, it is to be hoped that faces are an efficient technique for the representation of multivariate data and that by displaying the faces corresponding to each observation together, it will be a straightforward exercise for an observer to identify any clusters, outliers, etc., in the multidimensional data. However, very little is known about how people actually study and react to faces and it seems quite likely that different observers may react in different ways to different faces, perhaps concentrating more on some features than on others. Thus the effectiveness of Chernoff faces as a data representation technique may depend on which variables are assigned to which features (Chernoff & Rizvi (1975)).

Despite their failings, cartoon faces have enjoyed a certain popularity and are frequently described in books concerned with graphical representation techniques for multivariate data, yet cartoon faces are hardly ever used in practice. This popularity is probably due to the novelty of the technique, yet is seemingly out of all proportion to their usefulness as a graphical representation technique. Of course, the basic idea behind

Chernoff faces (the mapping of variable values to features of a pictorial graphical display) could be modified and applied to more abstract representations, thus avoiding some of the subjective problems attached to the examination of faces, but this does not seem to have been done in practice.

2.5 Multivariate Data: Dimension Reduction Techniques

Given a set of multivariate data in the form of a $(n \times p)$ data matrix $\mathbf{X}$, obtained by measuring n observed units on each of p variables, a p-dimensional graphical representation of $\mathbf{X}$ would be required to preserve the relationships between the variables. However, a lower-dimensional representation of the data may be desired in order to facilitate interpretation of the data and comparisons with other data sets. Since it is not possible to represent the data directly, except with the use of the techniques described in the previous section, a reduction in the dimensionality of the data to just two or three dimensions is usually sought, using projection techniques.

The techniques described in this section are computational techniques commonly used for the reduction of the dimensionality of multivariate data and which have the advantage of resulting in a graphical representation of the data in fewer dimensions. The techniques described include principal components analysis, the biplot, correspondence analysis, principal coordinates analysis, multidimensional scaling and cluster analysis. I shall also describe the lesser used technique of Andrews' plots.

The techniques described in this section differ from those described in Section 2.4 since, in reducing the dimensionality of the data, the raw data values are not necessarily preserved although it is to be hoped that relationships between the observations (or variables) are preserved. If the relationships in the data are preserved, then the graphical representations (typically constructed in two dimensions) resulting from these dimension reducing techniques may be of use in identifying outliers, clusters, and other features which exist in the high dimensional data.

The majority of the techniques described in this section are termed 'ordination techniques' since they are designed to preserve the geometric distance between the points in the original p dimensions in the lower-dimensional representation. It is the distance (usually Euclidean) between plotted observations which is usually of importance in the interpretation of the graphical representations.

None of the techniques described will be dealt with in a great amount of theoretical detail, since the interested reader who is not already familiar with these techniques can easily look the more mathematical details up elsewhere (see, for example, Manly (1994), Mardia *et al* (1979), Krzanowski (1988)). However, details will be given on the construction and interpretation of the graphical representation(s) which can be obtained using these techniques.

Less commonly encountered techniques for the graphical representation of raw multivariate data in a reduced number of dimensions, but which will not be described here, include factor analysis, multidimensional unfolding, orthogonal procrustes analysis, three-way scaling and proximity analysis. The interested reader is referred to Gower (1966, 1985) for a description of these techniques.

2.5.1 Principal Components Analysis

Perhaps the most commonly used dimension reduction technique is principal components analysis (PCA). The following description and discussion of PCA is based on the books by Krzanowski (1988), Manly (1994), Everitt & Dunn (1991), Everitt (1978), Digby & Kempton (1987) and Jolliffe (1986), the latter being the most comprehensive. I have not given a detailed mathematical derivation of PCA since this can be readily obtained from the references given or from other multivariate texts, such as Mardia *et al* (1979), Chatfield & Collins (1980) and Jobson (1992), any of which would be incomplete without a section on PCA.

For a multivariate data set consisting of p measurements on each of n observations, the only true representation can be in p-dimensional space. As has already been discussed, where $p > 2$, a series of 2-dimensional scatterplots constructed for each pair of variables cannot be relied upon to reveal the true structure within the data. PCA, in common with other dimension reduction techniques, can be used to find the 'best' (in some sense) projection of the data points onto a 2-dimensional plane. This might be expected to give a better representation of the structure in the data which can then be examined for clusters and outliers, etc..

In PCA, the 'best' projection of the data is one which minimises the sum of squares of the perpendicular distances between the observations and the plane, and in doing so preserves as much as possible of the variation in the original data. Of course, it is

not necessary to consider a plane — projection of the data onto any p^*-dimensional sub-space of the original p-dimensional space (where $p^* < p$) could be considered. However, it is usually preferable, in order to produce a readily constructed and interpretable graphical representation, to consider the projection with $p^* = 2$. Values of $p^* = 1$ or $p^* = 3$ may also be considered, the former being readily represented by a line, the latter requiring a 3-dimensional or enhanced scatterplot (see Section 2.4.1).

The projection of the data points onto the p^*-dimensional sub-space is obtained by considering the (n x p) data matrix **X** (n observations measured on each of p continuous variables), having sample covariance matrix **S**. The jth principal component is the linear combination of the original variables with coefficients given by the elements of the eigenvector corresponding to the jth largest eigenvalue of **S**. Note that the values of the coefficients are constrained, such that the jth principal component will have variance equal to the jth eigenvalue, and will be orthogonal to the ith principal component, for all $i < j$. There are p principal components in total. The total variation, explained by the set of p principal components, is therefore the sum of the eigenvalues, equivalent in value to the trace of **S**.

The first principal component obtained, corresponding to the largest eigenvalue of **S**, represents the line of least-squares best fit for the n p-dimensional observations, such that the perpendicular projection of the points onto this line will result in the best 1-dimensional representation of the data and the spread of the n points when projected onto this line will be a maximum. However, unless the observations are perfectly collinear, it is more usual to consider the first two or few principal components, which define a low-dimensional sub-space onto which the data points can be projected. For example, the second principal component obtained will be the line with the greatest spread of the points when projected onto it (though not as great as for the first principal component), subject to the constraint that it is orthogonal to the first. The third principal component similarly accounts for the next greatest amount of variance, and is orthogonal to the first two, and so on. By maximising the variance explained by the first few principal components, the configuration of the points in the sub-space defined by these principal components may be expected to give the best approximation to the original configuration of the data points in p-dimensional space, in terms of Euclidean distances between points, and thus to reveal the essential features of the multivariate data.

Principal component scores can be computed for each individual observation on each principal component. The coefficients for each variable being given by the

eigenvector corresponding to the eigenvalue corresponding to the principal component under consideration. It is these scores, for the first p^* principal components, which are plotted against p^* orthogonal axes to give the required representation of the observations in the p^*-dimensional sub-space defined by the first p^* principal components.

As has already been stated, plotting the data in the space defined by the first few principal components can be useful for detecting patterns such as clusters and outliers in the data. Outliers observed in the first few dimensions often correspond to individuals whose values on one or more variables are inflating the measured covariances or correlations. By plotting the data in the space of the last few principal components, outliers may be detected which are adding insignificant dimensions to the data. However, if an individual has a large score in one of the last few dimensions, this may indicate that they are poorly represented by the projection into p^*-dimensional space which may be a failing of the projection, not of the individual. As always, care is required in the interpretation of outliers.

If the variables are measured on different scales, or have widely differing variances, it is advisable to standardise the data in order to prevent variables from dominating the results of the PCA. This transformation of the original data is equivalent to using the sample correlation matrix **R** in place of the covariance matrix **S**. Choice of whether to use **S** or **R** is important, as different eigenvalues and eigenvectors, and therefore different sets of principal components and different representations will be obtained, with no obvious relationship between the two. If **R** is used, the sum of the eigenvalues will equal p since the variables will all have unit variance. Other transformations of the original data matrix may be used, such as giving variables different weights to reflect their relative importance, but iterative procedures are then required to identify the principal components.

If the points in p-dimensional space have some simple non-linear structure — for example, the points all lie on the surface of a sphere — then PCA is unlikely to detect this. Also, if there is little correlation among the original variables, PCA is unlikely to find a representation in a few dimensions.

To assess how well the p^*-dimensional projection of the data points approximates the original p-dimensional configuration of the n observations, the proportion of the total variance which is accounted for by the first p^* principal components can be calculated simply as the sum of the eigenvalues corresponding to the principal components

considered, divided by the total variance defined above. As a rule of thumb, if the p^*-dimensional representation accounts for 70–90% of the variance of the observations in the original p-dimensional space, then the p^*-dimensional representation may be regarded as providing a good representation of the data, such that the relative positions of the projected points very closely approximate the relative positions of the original points.

Another rule of thumb for determining the appropriate value of p^* is based on the graph known as a *scree diagram*. This involves plotting the ordered eigenvalues against the rank values $1,\ldots,p$ and joining the points. The resulting diagram will, typically, resemble a cross-sectional view of a cliff, with a steep descent followed by a shallow, near horizontal collection of values resembling scree. The value of p^* is chosen as the rank value where the steep decline ends. This may be regarded as the point beyond which there are no marked changes in the values of the eigenvalues, and these values are, hopefully, negligible.

Yet another rule of thumb involves choosing those principal components whose eigenvalue exceeds the average value of all eigenvalues (ie. which exceed 1 if the correlation matrix **R** is considered).

Different rules of thumb may lead to different decisions. More formal methods have been suggested by Jolliffe (1986).

The effectiveness of a 2-dimensional representation can also be assessed by superimposing a Minimum Spanning Tree (MST). The MST joins the points so that each point is connected and there are no closed loops in such a way that the sum of the 'lengths' of the lines, in the original p-dimensional space, is minimised. Observations which are close in p-dimensional space will be linked. Thus superimposing the MST on the 2-dimensional representation should, if the representation is effective, preserve links between points which are close. If points which are close are not linked, or if there are very long links, this indicates that there is some distortion in the 2-dimensional representation.

Occasionally, attempts are made to identify the principal components with underlying properties of the data represented. This is a subjective process — there is no reason why a purely mathematical procedure can be expected to produce meaningful dimensions, excepting that the resultant principal components will be linear combinations of the original variables. Moreover, different principal components may be obtained depending on whether **S** or **R** was used, and the coefficients obtained using **R** will be for standardised variables. However, if treated with due caution, interpretation, or *reification*,

of the principal components, by examination of a principal components plot or by interpretation of the values and signs of the coefficients, can further enhance the interpretation of the structure in the data. It is a common feature of PCA when applied to biological measurements, that the first principal component relates to 'size' whilst the second and subsequent components relate to aspects of 'shape'.

PCA is sometimes used as a starting point for Factor Analysis (Child (1970), Cooper (1983)) and a rotation of the axes employed to attempt to associate points more strongly with axes and so assist the process of reification. Factor analysis is most commonly applied to data arising in behavioural experiments, for example to identify dimensions of personality or intelligence. Of course, the rotated factors will no longer maximise the variance. Moreover, they may be allowed to become non-orthogonal during the rotation.

Thus PCA is a useful technique for the graphical representation of multivariate data in a few dimensions. It is a well-known and popular technique, due in part to the relative simplicity of the calculations involved and the ready availability of statistical software to carry out these calculations. PCA has a variety of applications (see Jolliffe (1986)), but its most common use is as a dimension-reducing technique. The representations obtained by plotting the principal component scores of the individuals against the orthogonal axes defined by the first few principal components are readily interpreted and, subject to the proviso that the number of dimensions chosen provide a good representation (it will be the best in that number of dimensions), should readily reveal any structure in the data such as outliers and clusters. If more than two principal components are required to obtain a representation which reveals an adequate amount of the total variation in the data, it will be necessary to use enhanced scatterplots, or the pictorial techniques described in the preceding section, as the results will still be intrinsically multivariate.

2.5.2 The Biplot

The following description and discussion of the biplot is based on Everitt (1978), Gower (1985), and Gower & Digby (1981). Other books and papers in which the biplot is described include Gabriel (1971), who introduced the biplot, Gabriel (1981), Seber (1984), Digby & Kempton (1987) and Gower & Hand (1995).

In a biplot, as in principal components analysis, the n observations in a $(n \times p)$ multivariate data matrix are represented by a set of points for which the inter-point distances are considered to indicate the relationships between the observations. However, unlike principal components analysis, the relationships between the p variables, obtained using the correlation or covariance matrix derived from the raw data, are represented in the same multidimensional space as the observations by a set of vectors. As in principal components analysis, in constructing a biplot a low-dimensional representation is then sought, preferably in just two dimensions, so that they may be easily represented. The term "biplot" refers not to the 2-dimensional representation which may be derived, but to the dual representation of the observed units and the measured variables.

A simple form of the biplot may be obtained by representing the n observations in the original data matrix $\mathbf{X}$ (where the elements of $\mathbf{X}$ have had the variable means subtracted) by projecting them onto the 2-dimensional plane defined by the first two principal components of $\mathbf{X}$. To represent the p observed variables on the same 2-dimensional plane, each of the p original axes in the original p-dimensional space (where each axis corresponds to one variable) could be projected onto the plane such that each axis/variable is represented by a unit vector passing through the origin. The appropriate coordinates for representing the p variables in 2-dimensional space are given by the first two columns of the matrix of principal component scores $\mathbf{H}$ — if these coordinates are plotted as points in the same space as the observations, and a vector drawn linking each point to the origin, then this will yield a 2-dimensional vector representation of the p variables superimposed upon a 2-dimensional point representation of the observations.

The above biplot is just one possible form. It can be interpreted with respect to how the plotted observed units relate to the vector representations of the variables — if an observation lies near to a vector but far from the origin, then that observation can be regarded as having a high (positive or negative) score on the variable represented by that vector.

In matrix terminology, the above biplot can be obtained by considering the singular value decomposition of $\mathbf{X}=\mathbf{LSH}$ (see Gower (1985) for details concerning the singular value decomposition). The covariance matrix $\mathbf{S}=\mathbf{X}'\mathbf{X}=\mathbf{HS}^2\mathbf{H}$, where the columns of $\mathbf{H}$ correspond to the component loadings used in principal components analysis. Thus the coordinates for the observations are the rows of $\mathbf{XH}=\mathbf{LS}$, and the coordinates for the variables are the rows of $\mathbf{H}$. If all the dimensions are considered, the rows of $\mathbf{LS}$ and of $\mathbf{H}$ therefore define the biplot, and the inner product of these two matrices reproduces $\mathbf{X}$. If

only the first r dimensions are to be plotted, it follows from the decomposition that these will give the best (in a least squares sense) rank r approximation to **X**. Thus the best rank 2 approximation to **X** using this simple form of biplot, will be of the form $\mathbf{X}_{[2]}=\mathbf{A}_{[2]}\mathbf{B}_{[2]}'$, where $\mathbf{A}_{[2]}$ is an $(n\times 2)$ matrix, the rows of which correspond to the coordinates for plotting the n observations in two dimensions, and $\mathbf{B}_{[2]}$ is a $(p\times 2)$ matrix, the rows of which correspond to the coordinates for plotting the p vectors representing the p variables.

An alternative form of biplot can be obtained for which (assuming that the biplot gives a good approximation to **X**) the lengths of the vectors, instead of being of unit length as in the biplot described above, are proportional in length to the relative variance of the variables (provided **X** has not been normalised to ensure that all variables have standard, unit, variance). To obtain this form of biplot, the coordinates of the observations are obtained by considering the rows of **L**, and the coordinates of the vectors representing the variables are obtained by considering the rows of **HS**, where the inner product of **L** and **HS** will reproduce **X**. This form of biplot has the following properties:

1. The standard deviation of a variable is equal to the length of its corresponding vector (thus the variance of a variable is equal to the square of the length of its corresponding vector), since the length of the vectors is related to $\mathbf{HS}^2\mathbf{H}'=\mathbf{X}$.
2. The covariance of two variables is given by the inner product of the pair of vectors representing the variables.
3. The correlation between two variables is given by the cosine of the angle between the two vectors representing the variables, such that highly correlated variables have nearly coincident vectors, and poorly correlated variables have nearly orthogonal vectors.
4. The distance between two observations is given by the distance between the points which represent these two observations as plotted on the biplot. However, the distances between the plotted points on this form of biplot are Mahalanobis, whereas they were Euclidean on the first form of biplot described.

Again, a 2-dimensional representation of the biplot can be obtained by considering the rank 2 approximation to **S**. To assess the usefulness of this approximation, the proportion of the total variance accounted for by the two dimensions can be calculated, based on the eigenvalues of **S**, as in PCA.

A third form of biplot, which is a compromise between the two forms already described, is proposed by Gower (1985), who believes that the first form of biplot deals well with the points representing the observations but poorly with the vectors representing the variables, whereas the second form of biplot deals well with the vectors representing the variables, but poorly with the points representing the observations. Gower therefore suggests using the rows of **LS** to give the coordinates for plotting the points representing the observations (as in the first form of biplot), and using the rows of **HS** to give the coordinates for plotting the vectors representing the variables (as in the second form of biplot). However, the inner product of **LS** and **HS** will not correspond to **X**.

Applications of the biplot technique for the representation of multivariate data have been considered by Gabriel (1971), who considers the application of biplots to principal components analysis in order to show inter-point distances and to indicate the clustering of variables, as well as to display the variances and correlations between the variables, and by Bradu & Gabriel (1978) who describe the use of the biplot as a diagnostic tool for tables of data.

Three-dimensional biplots are described in the paper by Gower (1990), although their display, in common with other 3-dimensional graphs such as the scatterplot (see Section 2.4.1), is not straightforward. Gabriel et al (1986) discuss the use of interactive computer graphics and colour to display 3-dimensional biplots.

Another extension to the biplot is the non-linear biplot (Gower & Harding (1988)). This is appropriate for use with non-metric data, of the sort used in PCO (see Section 2.5.4).

2.5.3 Correspondence Analysis

Correspondence analysis is an exploratory data analysis technique for the graphical display of multivariate categorical data, which includes two-way contingency tables. Gower & Harding (1988) describe correspondence analysis as equivalent to constructing biplots for two-way tables. The history of correspondence analysis can be traced back many decades, under a variety of different names, but its popularity is due essentially to the French school of statisticians headed by Benzecri, who term the technique *l'analyse des correspondances*.

The most comprehensive English-language exposition of correspondence analysis is contained in Greenacre (1984). A less theoretical, more practical, account is contained

in Greenacre (1993). Geometric aspects of correspondence analysis are described by Greenacre & Hastie (1987). The links between correspondence analysis and related techniques for quantifying and modelling categorical data, such as dual scaling, reciprocal averaging, optimal scaling and homogeneity analysis are discussed by Tenenhaus & Young (1985). Van der Heijden & de Leeuw (1985) and van der Heijden *et al* (1989) consider the complementary use of correspondence analysis and log-linear modelling for contingency table data. Illustrations of the use of correspondence analysis, are contained in the aforementioned books and papers, and also in Lauro & Decarli (1982), Greenacre & Vrba (1984), Hoffman & Franke (1986) and Higgs (1991).

Correspondence analysis scales the rows and columns of a ($n \times p$) rectangular categorical data matrix or contingency table $\mathbf{X}$ in *corresponding* units so that each may be displayed in the same low-dimensional space. This representation can be used to reveal structure and patterns inherent in the data. Each row of $\mathbf{X}$ can be represented exactly as a point in p-dimensional space, and each column of $\mathbf{X}$ may be represented exactly as a point in n-dimensional space.

To obtain a low-dimensional representation of $\mathbf{X}$ (say in two-dimensions), the first step is to rescale the original data matrix so that the sum of the elements is 1. This gives a new matrix, $\mathbf{P}$, called the correspondence matrix, with elements corresponding to the relative frequencies or probability densities. The row sums of $\mathbf{P}$ are used to construct a ($n \times n$) diagonal matrix $\mathbf{D}_r$, and the column sums of $\mathbf{P}$ are used to construct a ($c \times c$) diagonal matrix $\mathbf{D}_c$. These row and columns sums are called the *masses* and are equivalent to the marginal probability densities. The masses are used to weight each point in proportion to its frequency.

The row and column profiles of $\mathbf{P}$ are defined as the row and column elements of $\mathbf{P}$ divided by their respective masses. Thus the n row profiles in p-dimensional space are given by the rows of $\mathbf{R}=\mathbf{D}_r^{-1}\mathbf{P}$ and the p column profiles in n-dimensional space are given by the given by the rows of $\mathbf{C}=\mathbf{D}_c^{-1}\mathbf{P}$.

The problem now is to find a low-rank approximation to the original data matrix which optimally represents both the row and column profiles in k-dimensional sub-spaces (ideally k=2). This involves centering the correspondence matrix $\mathbf{P}$ and finding the generalised singular value decomposition.

The k-dimensional subspaces have a geometric correspondence that enables both row and column profiles to be represented on the same display. The geometric display of

each set of points highlights the nature of similarities and variation within the rows or columns, and the joint display highlights the correspondence between the row and column points. Distances between points within the rows or columns correspond to a chi-square metric, but distances between row and column points cannot be interpreted so readily.

Some attempt may be made to interpret the axes. A point will be related to an axis when it has a large mass and is close to the centroid, or is a very large distance from the centroid irrespective of its mass; ie. when the point has large *inertia*, which is based on the squared distance from the point to its centroid.

Again there is the problem of determining the most appropriate number of dimensions for the low-dimensional display. Two is most convenient for the construction of the display. The proportion of spatial variation in the data which is accounted for by each axis can be assessed in much the same way as in principal components analysis, based on the sum of the eigenvalues.

2.5.4 Principal Coordinates Analysis

Principal coordinates analysis (PCO) is a metric multidimensional scaling technique based on principal components analysis (PCA). PCA is applied to Euclidean distance measures obtained from the $(n \times p)$ data matrix **X**, but PCO is applied to a symmetric matrix **M** of order n, where the elements of **M**, m_{ij}, correspond to some measure of the association (ie. the similarity, dissimilarity, or distance (not necessarily Euclidean)) between the observations i and j. **M** may be observed directly, or derived from **X**. In all metric scaling techniques, which includes PCO, **M** is analysed to give an ordination defined by a set of n coordinates in p^* ($p^* < p$) dimensions, such that the Euclidean distance between the plotted points i and j approximates m_{ij}. The goodness of fit of this approximation can be assessed using criteria involving some simple function $f(m_{ij}, m_{ij}^*)$. Note that non-metric multidimensional scaling, described in the following section, uses more general goodness of fit criteria.

In principal coordinates analysis, it is assumed that coordinates can be found for the n points P_i (i=1,2,...,n) in no more than (n−1) dimensions, with Euclidean inter-point distances $\Delta(P_i,P_j)$ which exactly equal m_{ij}. Given these coordinates of the P_i, principal components analysis can then be used to obtain a n^*-dimensional ($n^* < (n-1)$) approximation by projecting the P_i onto the n^*-dimensional sub-space which is defined by

the first n^* principal components and which minimises the sum of squares of the distance of the P_i from this sub-space.

2.5.5 Multidimensional Scaling

Multidimensional scaling (MDS) is a technique originally due to Kruskal and to Shepard. The following description is taken from Everitt (1978). Methods for MDS are also described by Clarkson & Gentle (1986), as well as within other texts referred to in this chapter.

Principal coordinates analysis (PCO), described in the previous section, is alternatively known as metric MDS, or classical MDS. In classical MDS, the similarity measures used have cardinal properties. In contrast, in non-metric MDS, which will be considered in this section, the similarity measures used do not have any cardinal properties; rather they have only ordinal properties. Therefore non-metric MDS, henceforth referred to just as MDS, is used to obtain a coordinate representation of a similarity matrix using only the ordinal properties of the data contained in the matrix.

As with the other ordination techniques already described, in MDS we wish to represent the n observed units measured on p variables as n points plotted in p^*-dimensional space (where $p^* << p$) by finding a set of p^*-dimensional coordinates for each observed unit.

A graphical representation of a similarity matrix **A** should be such that a large (Euclidean) distance between observed units on the graph corresponds to a small similarity in **A** and, conversely, that a small distance corresponds to a large similarity. Thus if the similarities in the matrix **A** were to be ranked from smallest to largest, one would expect the corresponding distances on the graph to be ranked from largest to smallest. In other words, the Euclidean distances in the p^*-dimensional sub-space used in the graphical representation should be monotonically related to the similarities in **A**, where only the rank ordering of the similarities is of importance, not their values. The success of the representation in preserving the monotonicity of the data may be assessed by the calculation of what is termed *stress*. This is essentially a residual sum of squares, such that the lower the value taken by the stress, the better the representation.

In order to find a p^*-dimensional representation with minimum stress, one can begin with an arbitrary configuration of the points and then move them around so as to

reduce the stress value until no further improvement in stress can be made. This approach is due to Kruskal. However, there may be problems with local minima so it may be a good idea to repeat the procedure with a different starting configuration.

2.5.6 Cluster Analysis

Cluster analysis is the term given to a collection of techniques which attempt to form 'clusters' or groups of individuals or of variables from an ($n \times p$) data matrix **X**, such that individuals/variables in the same cluster are more similar, in some sense, than individuals/variables in different clusters. The techniques of cluster analysis are well described in the book by Everitt (1993).

Cluster analysis is not in itself a dimension reduction technique, unlike the ordination techniques described previously. However, superimposing the results of a cluster analysis can aid in the interpretation of the data displayed on any of the plots described previously. In addition to this, many of the tools of cluster analysis are intrinsically graphical, such as the dendrogram or the minimum spanning tree, both of which are described below.

Within this section, I shall be concerned with hierarchical clustering techniques. These may be agglomerative, whereby individuals or groups of individuals which are the most similar, in some sense, are fused together, thus proceeding from n individuals to a single cluster. Or these may be divisive, which takes the opposite approach, thereby progressing from a single cluster containing all n individuals, to the individuals themselves. Either approach will produce a hierarchical set of solutions in which clusters are formed by merging or partitioning other clusters. For any one solution, each individual will appear in one and only one cluster. Different results may be obtained depending on which measure of similarity/dissimilarity is used (eg. correlation coefficient, Euclidean distance, city-block metric, etc., according on the nature of the data); and depending on which clustering method is used (eg. nearest-neighbour single-linkage, furthest-neighbour complete-linkage, centroid clustering, median clustering, group average, or Ward's method). These methods differ in the way in which the distance between groups is defined. Different results may also be obtained depending on the number of clusters it is decided to accept as the final solution.

A scatterplot or principal components plot may be used initially to give insight into the data as to whether there may be natural clusters in the data, and how many, since

cluster analysis will always identify clusters!. The clustering results may be presented in the form of a *dendrogram*. This is a simple two-dimensional tree-like diagram showing the complete set of hierarchical cluster solutions and the distances at which clusters are merged or partitioned, and which may assist in determining the appropriate number of clusters to extract. The chosen solution may be superimposed on the original scatterplot or principal components plot, for example by enclosing clustered points, or by symbolic coding of points. However, if the original display is not very good, superimposing the solution may give the appearance of overlapping clusters. An alternative approach is to relate the solution to the minimum spanning tree (described in Section 2.5.1).

2.5.7 Andrews' Plots

This is a dimension reduction technique due to Andrews (1972), which is also described in Everitt (1978) and other multivariate texts.

Quite simply, each of the p-dimensional observations $\mathbf{x}$ is used to define a function of the form:

$$f_x(t) = \frac{x_1}{\sqrt{2}} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(2t) + x_5 \cos(2t) + \ldots$$

which is then plotted over the range $-\pi \leq t \leq \pi$. This will result in a set of lines drawn across the plot, one per observed unit.

Since this function representation preserves Euclidean distances, it is possible to use this technique to identify clusters, outliers, etc.. For example, points which lie close together in the original p-dimensional space will be represented on the plot by lines which remain close together for all values of t, whereas points which lie far apart in the original p-dimensional space will be represented by lines which stay apart at least for some values of t.

However, since the variables $x_1, x_2, \ldots$ are not equally weighted in the function, it may be useful to associate x_1 with the most important variable, and so on. An alternative approach is to apply principal components analysis (PCA) prior to constructing an Andrews' plot and to associate x_1 with the first principal component, etc..

Tests of significance have been derived for use with Andrews' plots, and confidence intervals can be constructed. See, for example, Goodchild & Vijayan (1974). Moreover, one-dimensional projections can be obtained for particular values of t.

One draw-back of Andrews' plots is that only a limited number of observations can be represented on the same diagram before the number of lines make it too confusing. One approach suggested by Everitt is to begin with a plot of all the observations in order to get an overall view of the data, and then to construct plots using subsets of the observations in order to get a clearer picture.

Another draw-back of Andrews' plots is that, owing to the composite nature of each observation's function, it is not possible to observe the effect of a single variable in isolation.

2.6 Dynamic Techniques for Exploratory Data Analysis

As Mezrich *et al* (1984) point out: "Our normal perceptual intake of information in our everyday environment is dynamic. The fact that data ... are portrayed statically is not because of the requirements of human information processing but rather due to the traditional limitations of technology". Indeed, the development and success of dynamic methodologies has been very much tied in with technological advances in both hardware and software.

The earliest attempts at dynamic graphics can be traced back to the 1970's, but I shall not consider those early crude systems (PRIM-9, CLOUDS, ORION (Friedman *et al* (1982)), PROMENADE (Ball & Hall (1970)) here. The development of dynamic graphics really took off in the 1980's, due to the work of Becker and his colleagues at AT&T Bell Laboratories (eg. Becker, Cleveland & Wilks (1987, 1988), Becker, Chambers & Wilks (1986), Cleveland & McGill (1988)). The development and use of dynamic graphics has also accompanied the decreasing price and increasing availability of powerful computing environments, such as workstations, which facilitate the use of dynamic graphics. Dynamic graphics are also available for Apple Macintosh – for example, MacSpin (Donoho et al (1986)).

The addition of dynamic capabilities to traditional static data display methods provides an enormous increase in the power of statistical methods to convey information about data. New, highly-interactive, dynamic methods have become possible, and capabilities which were difficult or impossible in a static environment are now simple and fast. This does not, however, mean that static methods will fall into disuse, rather that the collection of methods available will become much larger and richer. After all, the most

convenient way to present the results of a dynamic method is usually to select and print a suitable static view.

As Becker *et al* (1987, 1988) point out, dynamic graphical methods have two important properties: direct manipulation of graphical elements on a computer graphics screen and instantaneous change of these elements – the data analyst takes some action through manual manipulation of an input device, and something happens on the screen in real time. Dynamic statistical graphics tools are therefore of use for exploring and visualising structure in high-dimensional data (Young & Rheingans (1991)).

Becker *et al* (1987, 1988) review a collection of dynamic graphical methods for data analysis, including identification, deletion, brushing, scaling, rotation, and dynamic parameter control. These are considered in turn below:

Identification: It is not possible to routinely show the labels of all points because this would make an unintelligible mess on the screen. The user can either select a particular element on the screen and find out what its label is, or select a label and find the location of the corresponding element on the screen. It is also possible for the user to identify a subset of elements, and these will be highlighted. Alternatively, the display can cycle through different subsets. This avoids the necessity for different plotting symbols (colours, shapes, etc.) on a static display.

Deletion: Points can very easily be deleted from a graph with dynamic graphics, using the cursor.

Linking: If a point, or cloud of points, is selected on one display, the corresponding elements will be highlighted in related displays. This method is particularly useful with scatterplot matrices.

Brushing: Brushing is a dynamic method in which the data analyst moves a rectangle, known as the brush, around the screen using a mouse. Brushing can be used to support identification, deletion and linking in a transient, lasting or undone mode. More details of brushing, within the context of scatterplot matrices, are contained in Becker & Cleveland (1987) and Becker *et al* (1988). Brushing can also be applied to other graphical displays, such as 3-dimensional plots and histograms.

Scaling: An important aspect of a two-dimensional display is the aspect ratio (the physical length of the vertical axis divided by the length of the horizontal axis). The

aspect ratio can affect the appearance of any pattern in the data. Dynamic scaling permits the rapid consideration of different aspect ratios.

Rotation: For measurements on three variables, or on a subset of three variables, the data may be represented as a point cloud in 3-dimensions, but the screen, like a piece of paper, permits only 2-dimensional views. However, unlike a piece of paper and more like a movie, a computer display can convey a sense of the 3-dimensional structure by a rapidly changing sequence of 2-dimensional views. This gives an apparent rotation of the cloud about a specified axis. The use of stereo vision has been suggested to further enhance the 3-dimensional effect of the rotating cloud. Rotation is also described by Becker *et al* (1988).

Parameter Control: Dynamic methods can be used to control any parameter, discrete or continuous, that can affect a graphical display, eg. the subsets of variables, the aspect ratio, the axis of rotation, power transformations of the variables.

MacSpin (Donoho *et al* (1986)) has graphics capabilities implemented on an Apple Macintosh desktop computer which allows some of the features of Becker's dynamic graphics such as rotation of a point cloud to show a third dimension, identification of interesting points, highlighting of important subsets, animation to look at the effect of a fourth variable (such as time), transformation of the data, and marking subsets. Data can be readily imported and hard copy of screen displays readily obtained or incorporated into reports.

Haslett *et al* (1991) have also incorporated many of Becker's dynamic graphical methods into a system implemented on the Apple Macintosh computer which has been developed for the exploratory analysis of spatial data with an emphasis on identification and linking.

The use of animation, also advocated by Andrews (1981), may be considered to fall into the family of Grand Tour techniques (Asimov (1985), Buja & Asimov (1986), Buja *et al* (1986), Young & Rheingans (1991)). This is a set of projection pursuit-type methods, the basic purpose of which is to try and understand the structure of the data as an aid to interpretation.

Another, specific, example of the development and use of dynamic computer graphics is Stuetzle's (1987) Plot Windows for drawing and considering scatterplots and histograms.

Many computing issues must be considered when developing a dynamic method. Hardware must very often be pushed to its limits to achieve the requisite speed, and software environments can leave a large gulf between the conception of a new idea and its implementation. Two aspects of the hardware which have a large impact on the speed and ease of use of dynamic techniques are the bandwidth of the system (ie. the speed with which user input can be translated into image updates, which ideally should be perceived to be instantaneous) and the specific input/output devices used (eg. mouse, light-pen, colour printer). Software considerations include the (high level) environment, the (medium level) graphics library functions and the (low level) program components.

Interface design has also been shown (see Jones (1988)) to have a significant influence on aspects such as learning time, performance speed, error rates and user satisfaction. The need for specialised programming techniques to get adequate performance for dynamic displays means that implementations usually vary slightly from machine to machine and are therefore not readily portable. A survey of some requirements for and applications of interactive graphical systems for the analysis of data is contained in Andrews *et al* (1988).

As has already been seen, dynamic graphics are often used to enhance scatterplot matrices. Scatterplot matrices provide an effective tool for graphical exploratory multivariate data analysis. Carr *et al* (1986) and Carr *et al* (1987) consider interactive density representation and display techniques for enhancing scatterplot matrices for very large N, including brushing, subset selection and animation.

Huber (1987) considers his subjective experiences with 3-dimensional scatterplots or point clouds which may, as has already been indicated, be enhanced by the use of interactive dynamic graphics. One of the most important applications of interactive 3-dimensional graphics, according to Huber, is the finding of good 2-dimensional views. Also important is the consideration and comparison of 3-dimensional structures.

Dynamic graphics are considered, by Becker *et al* (1989) and Becker *et al* (1990), to be particularly useful for the visualisation of network data (eg. transportation data) consisting of a set of nodes, possibly with a geographical location on a map, and links between nodes. Statistical data may be associated with both nodes and links. With even a modest number of links and nodes, it may be impossible to encode the data values in a static display. Becker et al have therefore developed tools that make use of dynamic graphics to display and manipulate network data. Parameters which can be readily

manipulated, making use of a three-button mouse and on-screen boxes and sliders, are the statistics, levels, geography/topology, time, aggregation and size. It is also possible to see how the statistics change over time.

To illustrate the dynamic interaction essential for 3-dimensional graphics, Huber (1987) argues that videotapes are required. Static printing can only be done in the form of 2-dimensional views, although stereo pairs may be produced if the third dimension is essential. A perspective view of a 3-dimensional cube is usually useless. Colour slides may be obtained, but not readily reproduced. Despite the use of colour in their dynamic displays, Becker *et al* (1990) were unable to reproduce colour figures in their published report.

Higher dimensional data may be more readily examined for structure using dynamic graphic methods for 3-dimensional scatterplots and scatterplot matrices if dimension reduction techniques of the sort described in Section 2.5 are first applied to the data. This approach is considered by Weihs & Schmidli (1990) – their Online Multivariate Exploratory Graphical Analysis (OMEGA) strategy combines dimension reduction methods with dynamic graphics methods within the framework of an overall data analysis perspective, to provide a tool for exploring, understanding and forming hypotheses about the structure of multivariate data.

To be really useful, dynamic graphics routines should be embedded in a data analysis environment to permit data management, basic graphics, statistical modelling, data transformations, etc..

2.7 Summary

This chapter has been concerned with the description of some of the most commonly used graphical techniques for data exploration, or EDA. This is by far the most prolific area of research within statistical graphics.

It has been impossible to consider every one of the many hundreds of published chapters and articles relating to this area, and many less well known techniques have been omitted. In particular, no consideration has been given to statistical maps or to graphics of use for time series data.

The techniques which are considered in this chapter have been classified as techniques for univariate raw data, for bivariate raw data and for multivariate raw data,

the latter represented directly or transformed in some way to reduce the intrinsic dimensionality of the data. Although these techniques are not the main concern of this thesis, they serve to indicate the most appropriate model which may be fitted to the data. Moreover, many of the techniques described, including histograms and scatterplots, are of use for other statistical activities, as will be seen in Chapters 3 and 4. In particular, PCA and MDS will be employed elsewhere in this thesis, in Chapter 10.

Particular attention was paid, in Section 2.6, to dynamic techniques for the representation and exploration of multivariate raw data. With the increasing availability of computers with graphic facilities capable of supporting interactive real-time computer graphics, this is likely to be a popular and fruitful area for future research. Indeed, an interactive computer program is developed in Chapter 11 as part of the research described in this thesis.

BLANK PAGE IN ORIGINAL

3. Graphical Techniques for Model Fitting and Diagnostics

3.1 Introduction

In the preceding chapter, some of the very many graphical techniques which exist for the representation and exploration of the raw data were described. These techniques serve a useful purpose in revealing the structure within the data and suggesting the kind of statistical models which it would be appropriate to fit to the data. In this chapter, some of the graphical techniques which exist to assist in the fitting of models to data will be considered. As shall be seen, these are not so numerous as techniques for the representation of raw data.

In Section 3.2, some techniques for determining the distribution of the raw data will be described. This is important for assessing the appropriateness of the distribution of the data for the chosen model and can reveal the need for transformation of the data. Particular attention will be paid to probability plotting techniques which, as shall be seen, are of use not just for investigating the distribution of the data, but also for comparing distributions, estimating parameters and detecting effects.

Some graphical techniques of use with Analysis of Variance (ANOVA) will be considered. These include techniques for summarising the ANOVA design, prior to collecting and modelling the data according to the specified design. These are described in Section 3.3. In Section 3.4, some graphical techniques for the multiple comparison of means, used for determining which pairs, or subsets, of means differ are outlined.

Some graphical techniques of use with regression will also be considered. In Section 3.5 some graphical approaches to model selection within multiple regression, used as an aid to selecting a parsimonious model to fit to the data, are described, and in Section 3.6 the important role of graphical representations in regression diagnostics; ie. for checking the appropriateness of a fitted regression model, is discussed.

3.2 Assessing Distributional Assumptions

One of the simplest ways to examine the distribution of a set of raw data is by the construction of a histogram (see Section 2.2.1). This allows a crude assessment of the shape of the distribution of the data, which can be enhanced by superimposing a normal (or whatever) curve onto the graph. However, as was discussed in Section 2.2.1, the appearance of the histogram may be distorted, depending on the choice of interval, or bin, width.

An alternative technique is the boxplot (Section 2.2.5), though this is very crude, being based on just five summary statistics.

Another possible display is a symmetry plot (Gnanadesikan (1977), Jobson (1991)). This allows one to see whether a distribution is symmetric or not, although it says nothing about the precise form of the distribution. If a distribution is symmetric, then the ith lowest observation from the median, x_i, and the ith highest observation from the median, x_{n-i+1}, should each be the same distance from the median. Therefore a plot of $x_M - x_i$ against $x_{n-i+1} - x_M$ (where x_M is the median value) should result in a straight line with intercept 0 and slope 1. Departures from the expected line will indicate the nature of any asymmetry.

Better insight into the distribution of the data may be obtained through the use of more formal probability plotting techniques, which are described below.

3.2.1 Probability Plotting Techniques

Probability plotting is a visual method commonly employed to check the basic assumptions about the distribution of data and to detect departures from these assumptions. Good overviews of the techniques and uses of probability plotting are given by Wilk & Gnanadesikan (1968), Gerson (1975), Gnanadesikan (1977), Chambers *et al* (1983) and Jobson (1991), amongst others.

Probability plotting may be used to compare two distributions where both distributions are empirical, or one is theoretical.

The underlying assumption in the probability plotting of experimental data is that if a number of values are sampled independently from a population with some known frequency distribution and the ordered values are plotted against the corresponding

quantiles of this distribution, or against the quantiles of some other distribution, then the resulting points will lie approximately on a straight line passing through the origin with a gradient of 1. Of course, the determination of what can and cannot be considered a straight line is a subjective matter – measures of the goodness of fit have been advocated (Mage (1982), Gan *et al* (1991)).

A Percentile, or Percentile–Percentile, or P-P, plot can be constructed by finding the probability of different values or quantiles from the cumulative probability distribution for each distribution and plotting the probabilities for one distribution against the corresponding probabilities for the other. If the distributions are identical, this will result in a straight line through the origin with slope 1. Any variation from such a line would indicate that the two distributions were not identical. The P-P plot will be most sensitive to differences in the middle of the distributions, but not in the tails.

A Quantile, or Quantile-Quantile, or Q-Q, plot can be constructed in a similar way by considering the quantiles corresponding to different probabilities and plotting the quantiles for one distribution against the corresponding quantiles for the other. In this case a straight line plot will be obtained provided one random variable is a linear transformation of the other. Differences in mean and variance may be inferred from the intercept and slope of the graph respectively. If the two variables are from identical distributions, the slope will pass through the origin with gradient 1. The Q-Q plot is most sensitive to differences in the tails of the distributions rather than in the middle. A single straggler at the beginning or end of the plot may indicate an outlier. Curvature at the ends indicates kurtosis (ie. a greater concentration of points in the tails of one of the distributions). Convexity or concavity suggests a lack of symmetry. Kafadar & Spiegelman (1986) suggest that some curvature of the Q-Q plot is an inherent fault which makes linearity hard to quantify and propose the use of conditional Q-Q plots, though these must carry the risk that real effects may not be detected.

A set of values thought to be from a normal distribution can be plotted against the corresponding quantiles of the standard normal distribution. This is more commonly called a normal plot. This may be done directly, or using special normal probability plotting paper. A half-normal plot is one where, if $\mu=0$, the absolute values are ordered and plotted against order statistics of the standard half-normal distribution (see Daniel (1959), Sparks (1970), Zahn (1975a), Zahn (1975b)).

Choice of plotting positions has attracted some debate. When the n sample values are ordered from smallest to largest, it is necessary to decide on the percentiles to which they correspond in order to evaluate the quantiles of the relevant distribution. Note that the pth quantile will correspond to the $(100 \times p)$ th percentile of the data set. Probably the most widely used value for the percentiles is $p_i = (i-0.5)/n$. Other values have been suggested, although different plotting conventions can lead to different lines (Gerson (1975), Mage (1982)).

In general, claims Gerson, P-P and Q-Q plots between them provide reasonably sensitive indicators of distributional discrepancies, although if both distributions show the same departures, a straight line will still result. Q-Q plots tend to be the more favoured. P-P and Q-Q plots are only identical when the two variables are each from a Uniform (0,1) distribution.

When a half-normal plot is constructed, there may well be one or more values which are much bigger than expected and lie some distance from the line on which the other points lie. It is possible to test whether such points have been obtained by chance or whether they represent an outlier (Gerson (1975), Zahn (1975a, 1975b)).

Plotting techniques are also commonly employed for the Gamma distribution, which includes the chi-squared and exponential distributions as special cases (Wilk, Gnanadesikan & Huyett (1962)). A gamma plot, using quantiles of the chi-squared distribution, is useful for detecting the existence and form of heterogeneity when ordered variance estimates are plotted. However, the gamma distribution is complicated by the presence of a shape parameter which must be estimated before the probability plot may be constructed.

Hybrid plots can be constructed based on P-P or Q-Q plots, usually involving some function of the percentiles or quantiles (Wilk & Gnanadesikan (1968)).

In addition to testing distributional assumptions and revealing outliers, probability plots may assist the determination of appropriate models. For example, a plot can be done of the ordered mean squares obtained from an Analysis of Variance, in order to permit informal 'internal comparisons' (Wilk & Gnanadesikan (1964), Wilk & Gnanadesikan (1968), Gnanadesikan & Wilk (1970)). The precise approach depends on the degrees of freedom. If no real effect exists, the resulting plot will be linear, otherwise values will be found not to lie on the line. These values should be removed and the graph re-plotted to reveal any other effects. Also, derived data, such as the residuals from a multiple

regression, analysis of variance, or some other analysis, may be plotted to check distributional assumptions and for outliers (see Section 3.6).

For multivariate data, it may be tempting to construct a scatterplot for pairs of variable values to look for an elliptical pattern, but this will only be effective for large samples. For smaller samples probability plotting can be used to compare the variables in pairs, or one at a time against the normal distribution. Normal probability plotting techniques may, however, be extended to handle multivariate data (see Healy (1968), Everitt (1978)).

3.3 Experimental Design Symbolisation

Experimental design symbolisation techniques for use in Analysis of Variance (ANOVA) have been proposed by Winer *et al* (1991), Lee (1966), Wilkinson & Rogers (1973), and Taylor & Hilton (1981), amongst others. Of these techniques, those of Lee and of Taylor & Hilton seem to be particularly informative in that the experimental design implies which main effects and interaction effects can be tested for using a particular design.

The reader is assumed to be familiar with the terminology connected with ANOVA.

3.3.1 ANOVA Design Table

Perhaps the most commonly used experimental design symbolisation technique is the design table, mentioned in Lee (1966). An example of a design table for a three factor ANOVA design where factor A has 2 levels, factor B has 4 levels, factor C has 2 levels, and factor A is crossed with factors B and C but B is nested under C, is presented in Figure 3-1.

By looking at an ANOVA design table, it is possible to determine which factors are crossed and which are nested. If the subjects are themselves nested under one or more factors (in a between subjects design), then the subjects can be represented individually as different levels of a factor 'subjects'. Similarly in a repeated measures (within subjects) design, 'subjects' can be represented as a factor which is crossed with all combinations of the other factors. Mixed designs (involving both between and within subjects factors) require a combination of the two approaches in the construction of the design table.

		A1	A2
C1	B1		
	B2		
	B3		
C2	B4		
	B5		
	B6		

Figure 3-1: Design table for a three factor ANOVA design

However, if there are more than three factors in the design, it may not be easy to construct or interpret the design table for any experimental design.

Design tables have a practical use since they form a table of cells where each cell corresponds to a particular treatment combination. In this sense, design tables resemble contingency tables. In each cell, the subjects assigned to that particular treatment combination and/or the observations made on those subjects can be recorded.

3.3.2 Blocks Representation for Three Factor ANOVA

An experimental design symbolisation technique which I have developed from the ANOVA design table described in the previous subsection (3.2.1) specifically for use with three factor ANOVA is the blocks representation. Again, using this technique it is possible to determine visually which factors are crossed with, or nested under, which other factors. Again, consider an experimental design in which factor *A* has 2 levels, factor *B* has 6 levels, and factor *C* has 2 levels. Factor *B* is nested under factor *C*, and factor *A* is crossed with factors *B* and *C*. The blocks representation for this design will be as given in Figure 3.2.

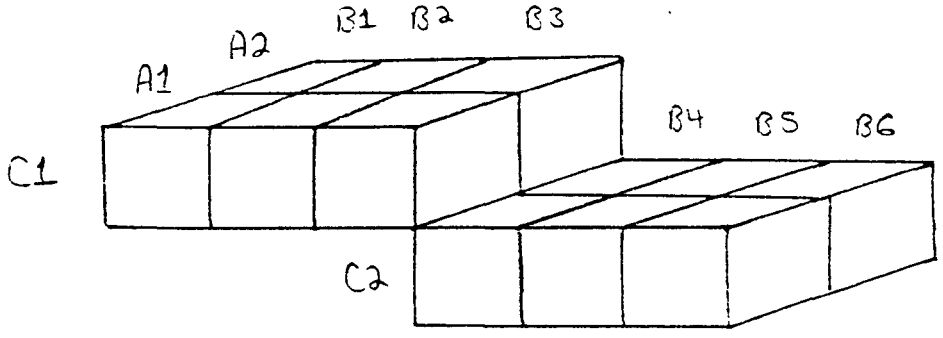

Figure 3-2: Blocks representation of example ANOVA design

An imaginary view from the top would indicate that factors *A* and *B* are fully crossed, as shown in Figure 3.3. Similarly, an imaginary view from the side would indicate that factors *A* and *C* are fully crossed, as shown in Figure 3.4. However, an imaginary view from the front would indicate that factor *B* is nested under factor *C* as shown in Figure 3.5.

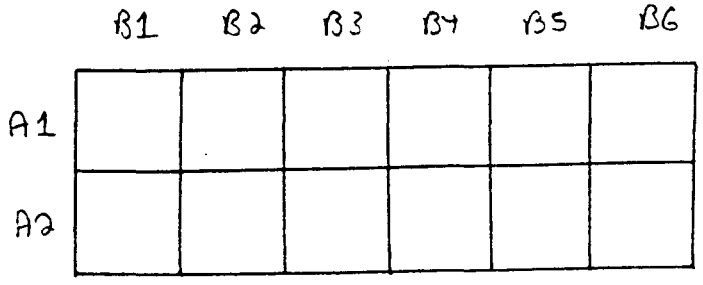

Figure 3-3: Top view of blocks representation

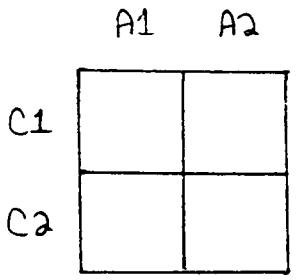

Figure 3-4: Side view of blocks representation

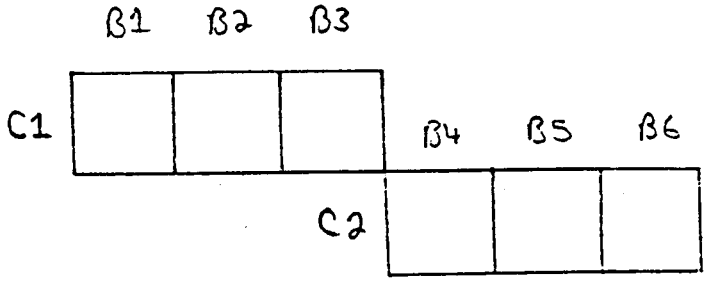

Figure 3-5: Front view of blocks representation

Obviously, the blocks representation is inherently 3-dimensional and could not easily be extended for the representation of designs involving more than three factors.

3.3.3 Symbolic ANOVA Design Representation

Winer *et al* (1991) suggest an experimental design symbolisation technique in which an alphabetic symbol represents a factor in the experiment, a line including a cross drawn between two factors indicates that the two factors are crossed, and a line including a circle drawn from one factor and pointing towards another factor indicates that the former is nested under the latter. An example of an experimental design involving four factors represented in this symbolic way is given in Figure 3-6. It can be seen that factor *A*

is crossed with each of factors *B, C* and *D*, that *C* is crossed with *D*, and that factor *D* is nested under factors *A* and *B*. Thus this technique can be used to convey the factors in the experiment and the relationships between these factors in a visual manner. It would be quite straightforward to add a subscript to each factor symbol in order to indicate the number of levels in each factor, and to add some kind of underscore to each factor symbol to indicate whether they are fixed or random factors.

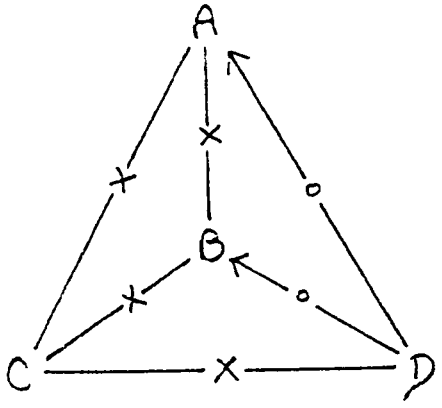

Figure 3-6: Symbolic representation of example ANOVA design

This symbolic ANOVA representation technique can be used to represent a greater number of factors than can be effectively achieved using the design table or blocks representation techniques described above. However, for a large number of factors, some crossing of the lines representing the relationships between factors would inevitably occur, leading to a messy representation.

3.3.4 Factor Relation Table for ANOVA

Lee (1966), in addition to the design table, also describes a "factor relation table" which conveys exactly the same information as Winer *et al*'s (1991) symbolic representation technique, but in tabular form. For example, the factor relation table presented in Figure 3-7 represents the same experimental design as represented in Figure 3-6. A 'x' in a cell of the table indicates that the two factors which form that cell are

crossed and parentheses are used to indicate the nesting of factors — for example, $D(A)$ indicates that factor D is nested under factor A.

	A	B	C	D
A	—	×	×	D(A)
B	×	—	×	D(B)
C	×	×	—	×
D	D(A)	D(B)	×	—

Figure 3-7: Factor relation table for example ANOVA design

One advantage of the factor relation table is that it is readily extendible to any number of factors. However, disadvantages of the factor relation table are that a lot of information is duplicated (since it is a symmetric table), and relationships between factors may not be as immediately apparent as in the more visual symbolic representation.

The factor relation table was developed by Lee for use with a programmable algorithm which can be applied to the table in order to derive the appropriate ANOVA model corresponding to the experimental design. In the process of deriving this model, an alternative design symbolisation is derived of the form, for example, $A \times B(C \times D)$, where the '×' indicates that the factors are crossed and the '(...)' indicates that the factor immediately outside the parentheses is nested under the levels of the factor or compound factor enclosed in the parentheses. This design symbolisation may be quite difficult to interpret for a complex design without the aid of a factor relation table, and it conveys no information about the number of levels of each factor, but it is in a form from which the appropriate ANOVA model can be derived using Lee's algorithm.

Wilkinson & Rogers (1973) present a design symbolisation similar to that of Lee, using '*' to represent crossed factors and '/' to represent nested factors. They too develop this syntax for use with programmable algorithms for model specification.

3.3.5 Structure Diagram Symbolisation for ANOVA

The structure diagram symbolisation of Taylor & Hilton (1981) is probably the most versatile experimental design symbolisation technique of those considered in this

section. It incorporates all of the best features of the other experimental design symbolisation techniques described in that all the factors are shown, the crossing and nesting relationships between the factors are shown, the number of levels of each factor are shown, and the appropriate ANOVA model corresponding to the experimental design, together with the necessary formulae for testing the terms in the model, can be derived using a series of rules presented by Taylor & Hilton and which are potentially programmable.

An example of a structure diagram is presented in Figure 3-8. In the structure diagram, the presence of (vertical) links between two factors indicates that the lower factor is nested under the upper factor, and the absence of any link between two factors indicates that the two factors are crossed. 'U' is used to represent the overall mean and all other factors are nested under this. 'E' is used to represent error and is nested under all other factors.

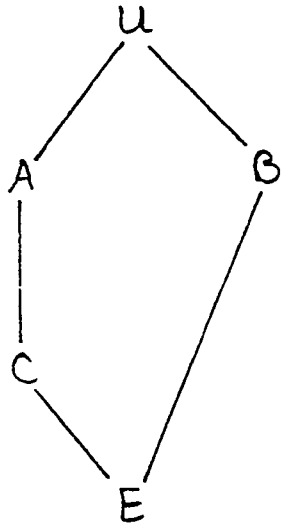

Figure 3-8: Example ANOVA structure diagram

The structure diagram representation of the experimental design not only conveys the relationships between the factors but, as has already been mentioned, can also be used to obtain the appropriate linear model corresponding to the experimental design. This is

achieved by writing down each factor as a main effect together with an interaction term for each combination of the crossed factors. Subscripts are written corresponding to each main effect symbol, with the subscripts of all nesting factors (ie. factors linked from above) written in brackets. The ANOVA model corresponding to the design represented by the structure diagram in Figure 3-8 is therefore

$$Y_{abce} = u + A_a + B_b + AB_{ab} + C_{(a)c} + BC_{(a)bc} + E_{(abc)e}$$

The appropriate ANOVA model is derived more easily and more quickly than if Lee's (1966) algorithm had been employed, and requires only the use of visual rules of thumb rather than the iterative application of Lee's algorithm. However, it is not apparent whether Taylor & Hilton's visual rules of thumb could easily be programmed and executed automatically. For a very large number of factors, with complex nesting relationships between them, it is likely that the structure diagram and the corresponding working diagram (described below) would become quite messy and difficult to interpret visually. This criticism was also made of Winer *et al*'s (1991) superficially similar symbolic representation technique (see Section 3.2.3).

The working diagram corresponding to the structure diagram in Figure 3.8 is presented in Figure 3.9. In the working diagram, the presence and absence of links between pairs of variables again indicates the presence of nesting or crossing relationships. The number in parentheses alongside each factor symbol indicates the number of levels of that factor, and the presence of a squiggly underline indicates that it is a fixed factor, else it is assumed to be a random factor. These additions are an improvement over Winer's symbolic representation technique, yet could readily be incorporated into it.

Using the working diagram, in which the mean effect is implicit, it is possible to derive the appropriate F-test in order to test the significance of each effect of interest contained within the ANOVA model derived from the structure diagram. In order to derive the appropriate F-test, a quite complex set of rules contained in Taylor & Hilton (1981) are used. These essentially involve partitioning the complete set of factors into various disjoint sets, according to the effect to be tested, and these are then used to develop the appropriate formulae for the degrees of freedom, the sum of squares, the expected mean squares and the F-tests, together with variance component estimates. It is claimed by Taylor & Hilton that, with practice, it is possible to derive the appropriate F-tests directly by visual inspection of the working diagram and without reference to the

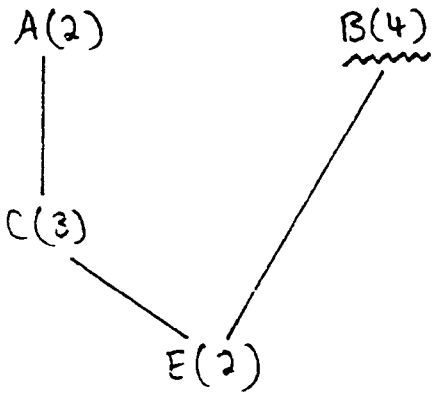

Figure 3-9: Example ANOVA working diagram

derived linear ANOVA model. Alternatively, it may be possible to program their rules as a computer implemented algorithm.

Thus Taylor & Hilton's structure diagram symbolisation (encompassing both structure diagram and working diagram representations) is the most sophisticated of all the ANOVA experimental design symbolisation techniques presented in this section. Like Lee (1966), the representation is related to a model building process, and in a computer implemented format it would allow easy examination of the effects upon the linear ANOVA model and upon the corresponding F-tests of adding or deleting factors, or of altering the relationships between factors. However, Taylor & Hilton's technique has the limitation that it can only be used with balanced ANOVA designs.

3.4 Multiple Comparison Procedures

One-way ANOVA is used to test hypotheses that all treatment means are equal. However, having rejected this hypothesis, the experimenter is often interested in determining which means differ and whether there are subsets of treatments having similar means. There are very many computational multiple comparison procedures (MCPs) available; for example, due to Dunnett, Tukey, Duncan, Scheffe, etc. (see, for example, Howell (1992)). Such tests usually test pairwise equality of the means.

Pairwise comparison of means in ANOVA can also be done graphically. For a set of k-treatment means, the means are plotted in increasing order of magnitude with uncertainty intervals, based on the standard error. Any pair is judged to be different if and only if their uncertainty intervals do not overlap. Such a procedure is a MCP only if the experimentwise error rate (ie. the overall probability of making at least one Type 1 error within the set of conclusions) is controlled. Such a plot is quick and easy to use, and simple to interpret. Gabriel (1978) presents a simple graphical procedure which can be used for the comparison of means of unequally sized samples. Confidence intervals for more complicated (ie. non-pairwise) contrasts can also be constructed using Gabriel's approach. Andrews *et al* (1980) and Hochberg *et al* (1982) each consider four graphical MCPs for pairwise comparisons, suitable for use with unbalanced ANOVA designs, which differ in their computational simplicity and efficiency.

Schweder & Spjotvoll (1982) present a graphical procedure called a P-value plot for pairwise comparisons, which is based closely on the half-normal plot described in Section 3.2.1. This permits estimation of the number of hypotheses which can be rejected, or the p-value which should be used. Suppose there are T hypotheses / pairwise comparisons, each with an associated p-value. Plot the number of p-values greater than p against 1-p. For large p, this should result in a straight line with slope equal to the unknown number of true null hypotheses. The null hypothesis should be rejected for the points deviating from the straight line, corresponding to small p on the right-hand side of the plot. Often the plot will show a gradual bend rather than a clear break, which will make interpretation less clear. It is suggested that there should be at least 15 p-values (ie. $T \geq 15$) in order for a straight line to be fitted.

An alternative approach is to partition the treatment means into groups using cluster analysis (see Section 2.5.6). The dendrograms generated by the hierarchical clustering algorithms give a graphical representation of the results and can be used to describe differences among the treatment means. Scott & Knott (1974) present a clustering method for grouping means in ANOVA. Calinski & Corsten (1985) propose two clustering methods to group treatments into homogeneous clusters. Tasaki et al (1987) consider six methods of clustering sample means. The main finding, perhaps not surprisingly, is that different clustering methods result in different dendrograms and therefore in different conclusions. It cannot be inferred that treatments in different groups are significantly different, nor does the fact that means are in the same group prove

equality. I would suggest that clustering means in ANOVA should therefore be regarded as an exploratory technique.

3.5 Model Selection Procedures

In multiple regression, with n observations on k independent variables $x_1,...,x_k$ and one dependent variable y, a number of different possible models may be fitted, using different subsets of the independent variables. Model selection procedures in multiple regression are described in general in Chapter 7. In this section I wish to consider one particular graphical approach, based on Mallow's C_p statistic (Mallows (1973)). For a regression based on p of the k independent variables, a plot of C_p against p for all possible regressions (of which there are 2^k) should indicate which regressions are adequate. For an adequate regression, C_p should be close to p (ie. lie on the line $C_p = p$), otherwise C_p will be much larger than p. Spjotvoll (1977) presents two alternative, but related, plots.

3.6 Regression Diagnostics

Diagnostic graphical methods are frequently used in regression analysis, where they commonly involve plotting some combination of the observed values of the dependent variable, the observed values of the independent variable(s), the fitted values of the dependent variable, the residuals (ie. observed minus fitted values of the dependent variable), and/or the regression line itself, in order to diagnose the appropriateness of the fitted model (as represented by the regression line) in relation to the data.

There is a vast literature in this area which I shall not attempt to review in detail since the use of graphical displays for regression diagnostics are not the main concern of this thesis. However, it seems to be appropriate to give an outline of some of the work which has been carried out in this area, since techniques used in regression diagnostics make considerable use of graphical representations.

Anscombe (1973) provides a good introduction to the use of scatterplots in regression analysis, giving details of the implications for the fitted model of various patterns which may emerge in a scatterplot. For example, in a scatterplot of the observed values of the dependent variable y against the measured values of the independent

variable x, where a simple regression model of the form $y_i=a+bx_i$ has been fitted, five (or more) patterns may be found as follows:

1. The plotted points may lie in a straight line.
2. The plotted points may lie on a smooth curve.
3. The y-values may be scattered about independently of the corresponding x-values.
4. Some combination of the above may be found.
5. Most of the plotted points lay on a straight line, but a few of the points are scattered a long way away.

The first of these patterns is the ideal scatterplot which shows that the model fits very well. The second pattern can usually be easily converted to the first by use of a suitable transformation of the data. The fifth pattern indicates the presence of outliers.

The residuals ($e_i = y_i - \hat{y}_i$) should also be plotted, either against the x_i values or against the fitted values of the dependent variable $\hat{y}_i$. The simplest approach to the analysis of residuals is to plot the frequency distribution of the set of residuals as a histogram or as a cumulative plot on normal probability paper (see Section 3.2.1). If the assumptions of the regression are close to being satisfied, then the residuals will appear to be distributed normally with zero mean and common variance for all x_i. In a plot of the residuals (against x_i or $\hat{y}_i$) the following four (or more) patterns may be found:

1. A few of the residuals are much larger in magnitude than the others, ie. there are outliers among the residuals.
2. The residuals may form a curve.
3. There may be a progressive change in the variability of the residuals accompanying an increase in the values of the fitted or independent values.
4. There may be a skewness or otherwise non-normal distribution of the residuals.

The latter three patterns can sometimes be removed simultaneously by a suitable transformation of the data. The second pattern can also sometimes be removed by the addition of a quadratic term into the regression equation.

Anscombe and Tukey (1963), Anscombe (1973), Denby & Pregibon (1987) and Chatterjee & Price (1991) provide general introductions to the examination and analysis of residuals.

To illustrate the usefulness of residual plots in regression diagnostics, Anscombe (1973) considered four fictitious data sets. Each of these data sets yields the same values in fitting a regression equation; also each data set has the same number of observations, the same mean of the x values, the same mean of the y values, the same regression line, the same regression sum of squares, the same residual sum of squares, etc.. However, scatterplots of the y-values against the x-values for the four data sets with the regression line superimposed highlight problems with all except one of the data sets which would not have been detected otherwise.

In a more complex regression analysis, for example with two independent variables x_1 and x_2, it is not practicable to construct a 3-dimensional scatterplot of y against x_1 and x_2 simultaneously. Instead, scatterplots can be constructed for each of the three possible combinations of pairs of variables, or a scatterplot can be constructed for two of the variables with the value of the third variable indicated by a suitable symbolic or numeric value (see Section 2.4.1). If there are many independent variables to be considered, however, it is not so straightforward to use scatterplots.

Atkinson (1985) devotes a book to graphical methods for diagnostic regression analysis, mostly based on the scatterplot. Atkinson indicates that regression diagnostics may be used for the following purposes:

1. To identify errors arising in the measurement or recording of independent or dependent variables.
2. To assess the adequacy of the linear model to describe the systematic structure of the data.
3. To assess the need to transform the dependent or independent variables.
4. To assess the error distribution within the data, and so assess the suitability of using least squares regression.

Some examples of residual plots are given, which can be used to detect certain departures from the fitted model or from the assumptions of the technique as follows:

1. Scatterplots of the dependent variable y against each of the independent variables $x_1, \ldots, x_n$.
2. Scatterplots of the residuals against each independent variable in the model — if a curvilinear relationship exists, this indicates that a quadratic term should be included in the model.
3. Scatterplots of the residuals against each independent variable not in the model — the presence of a relationship would suggest that the corresponding independent variable should be included in the model.
4. A scatterplot of the residuals against the predicted y values obtained from the fitted model — if the variance of the residuals increases with the fitted values, this suggests that the values of the dependent variable may need to be transformed.
5. A normal probability plot of the residuals — the residuals should look like a sample from a normal distribution.
6. If the data are collected in a particular time order then, even if time is not treated as an independent variable, the dependent variable and the residuals should be plotted against time — such plots may lead to the detection of previously unexpected patterns in the data due to time or due to other variables which are correlated with time.

Rather than construct residual plots, it is possible to detect the same effects by the calculation of suitable test statistics and the assessment of their significance. However, Anscombe claims that plots show the effects quickly and vividly (particularly if implemented on a computer), and so formal testing is made unnecessary.

Anscombe also discusses the use of residual plots in 2-way ANOVA. Having calculated the row and column means and the ANOVA equation, and the residuals obtained in fitting this equation, scatterplots can be constructed of the residuals against the fitted values, or a triple scatterplot constructed of the row effects against the column effects with the residuals encoded. Indeed, for any $R \times C$ table (including contingency tables) for which a set of main effects and interaction effects have been calculated, it should be possible, Anscombe suggests, to plot the residuals against the fitted values. Graphical methods also exist for assessing logistic regression models (Landwehr *et al* (1984)).

Probably the most common reason for the calculation of residuals is for the identification of outliers; ie. observations which have large residuals relative to the other observations. Outliers may occur due to an error in the measurement or recording of the

observations, or they may represent the correct observation of an unusual phenomenon. In the former case the value may be discarded, or assigned a modified value, or given a low weight, or be considered separately, or be transformed in some way together with the other observations.

Graphical procedures are likely to be more 'sensitive' to outliers than numerical procedures for the examination of residuals. For example, scatterplots are frequently used to examine residuals — outliers can be readily identified as isolated points, and peculiarities in the distribution of the residuals or associations between the residuals and the fitted values can also be detected visually.

3.7 Summary

In this chapter, a variety of graphical techniques for model fitting and model diagnostics have been described, including techniques for assessing distributional assumptions, experimental design symbolisation in ANOVA, multiple comparison procedures in one-way ANOVA, model selection, and regression diagnostics. Such techniques are of use for determining which model(s) it would be appropriate to fit to the data.

BLANK PAGE IN ORIGINAL

4. Graphical Techniques for the Representation of Fitted Models

4.1 Introduction

The main concern of this thesis is with the graphical representation of fitted models. As has already been seen, in Chapter 2, there are very many graphical techniques for the representation of the raw data prior to the fitting of a model and, in Chapter 3, a number of graphical techniques for checking the appropriateness of a fitted model were described. However, there are few graphical techniques already in existence for the representation of the fitted model itself. Those graphical model representation techniques which have been developed to date are described and considered below.

Representations for ANOVA models are considered in Section 4.2. These include the commonly employed interaction plot and extensions to this; a technique for the display of the ANOVA summary table; and a Venn diagram representation of Sums of Squares. Extensions are suggested for some of these techniques.

Representations for contingency table data are described in Section 4.3. These include graphical displays such as histograms, circle graphs, barycentric plots, cluster analysis and probability plots. Correspondence analysis is also revisited in the Section, having been described in some detail in Chapter 2.

Also in this chapter, two graphical representation techniques for logit models are briefly considered, in Section 4.4.

By far the most useful technique for the representation of a variety of fitted models, such as covariance selection models, log-linear interaction models and mixed interaction models (which include ANOVA, multiple regression, etc.), is the conditional independence graph. Because of the importance of this representation technique for the work described within this thesis, the conditional independence graph will be considered separately, in Chapter 7.

4.2 Representations for ANOVA Models

4.2.1 Interaction Plots

A commonly employed representation technique used with ANOVA is to construct an interaction plot, based on means. For example, given two factors A and B, a plot can be constructed showing the value of the mean for each level of factor A at each level of factor B (ie. the simple effects of B). The means plotted for the first level of factor A can then be linked, as can the means for the second level of factor A, and so on. Of course, the value of the mean for each level of factor B at each level of factor A (ie. the simple effects of A) could be plotted instead.

If the lines on the plot are parallel, this indicates that there is no interaction between factors A and B, but if the lines are not parallel, or if they intersect, this indicates that there is an interaction between the two factors. The extent of the departure of the lines from being parallel indicates the extent of the interaction. Thus interaction plots can be used as an exploratory technique, in the investigation of which model would best fit the data, but they can also be used to illustrate the model which has been fitted to the data.

Details of the construction and use of interaction plots can be found in most introductory statistics text-books which include ANOVA, particularly those intended for behavioural scientists, eg. Winer *et al* (1991), Howell (1992), Boniface (1995). An example of an interaction plot is given in Figure 4.1, based on data contained in Boniface (1995). It can be seen that the lines representing the simple effects of A are not parallel, in other words that the effect of B is not the same at every level of A. This indicates that there may be an interaction, the significance of which would need to be tested using numerical techniques.

The main effects of A and B can also be represented graphically, by constructing a bar chart showing the mean for each level of A (ie. the main effect of A) and a bar chart showing the mean for each level of B (ie. the main effect of B). This is illustrated in Figure 4.2.

In the vast majority of standard ANOVA texts, raw means are graphed in the manner described. However, Rosnow & Rosenthal, cited in Allison et al (1993), argue that the main effects should be removed and the residuals plotted, since an interaction

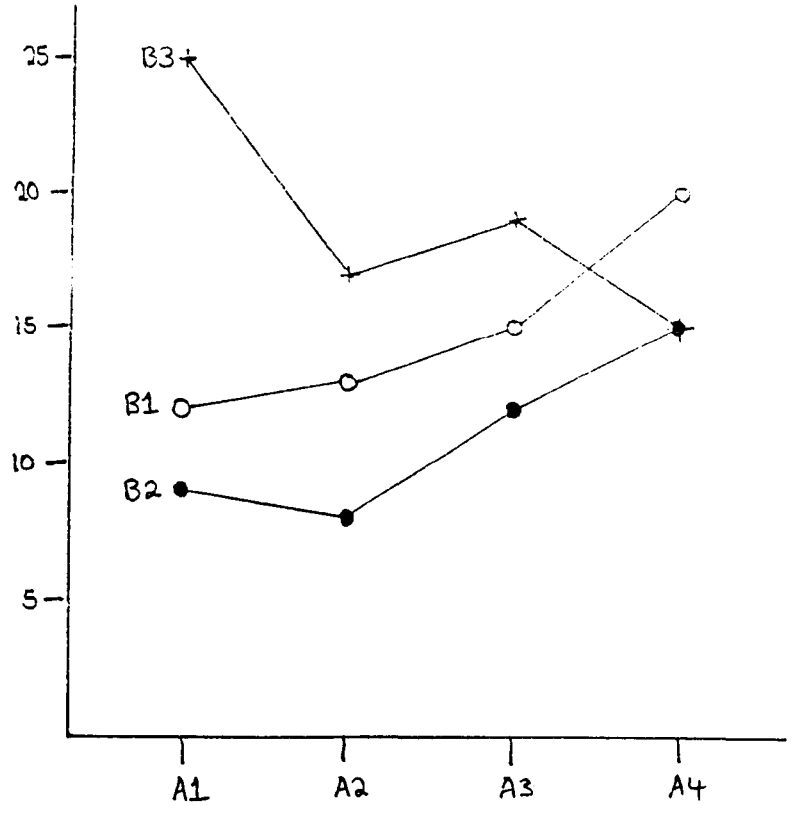

Figure 4-1: Interaction plot for example data set

effect is the multiplicative effect of two (or more) variables after controlling for the individual additive effects (ie. the main effects).

For an ANOVA experiment involving three factors, interaction plots are usually constructed for two of the factors at each level of the third. If a three-way interaction is present, the interaction between any two of the factors (ie. the simple interaction effects) will be different at different levels of the third factor. However, it is not always apparent whether the three-way interaction is present or absent and, depending on the choice of the two factors to be plotted at each level of the third, it can also be difficult to determine which two-way interactions are present or absent.

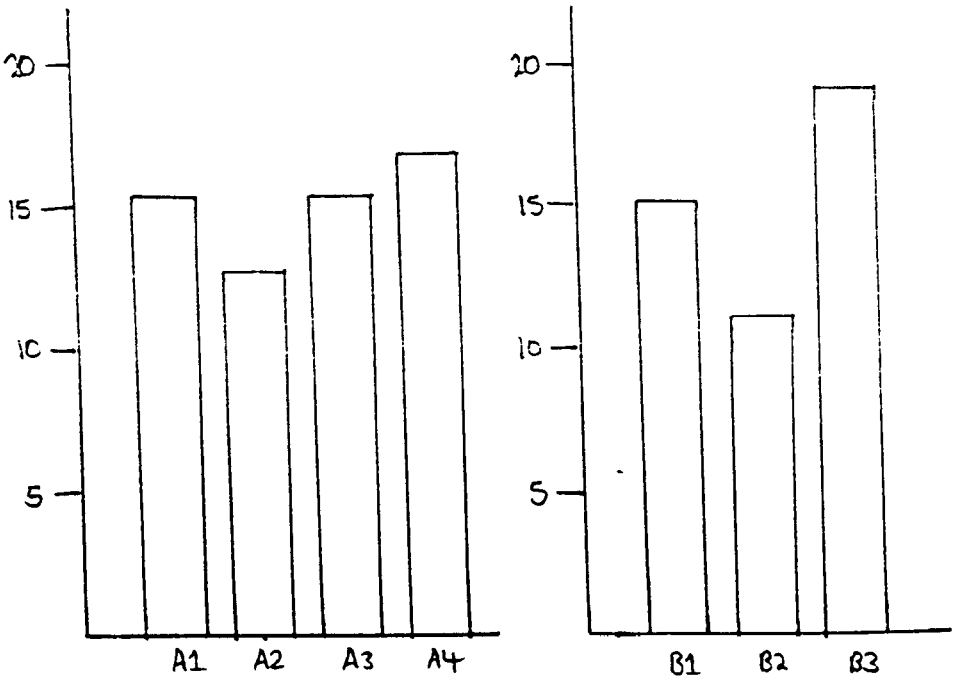

Figure 4-2: Bar charts of means

Monlezun (1979) has proposed a variation upon the standard interaction plot in order to represent interactions in three factor ANOVA. This approach involves plotting the differences between the values of the cell means, rather than the cell means themselves. The plots constructed depend on which effects are zero – Monlezun considers cases in which the three-way interaction is zero and none, one, or two two-way interactions are also zero.

For example, Figure 4.3 shows the usual interaction plot constructed for a three factor ANOVA (data taken from Monlezun). From this plot it is difficult to determine whether there is a three-way (*ABC*) interaction or not, but the viewer is likely to conclude that there is a three-way interaction. Compare this with Figure 4.4 which shows a plot suggested by Monlezun based on the differences between means for different levels of *A*. This plot indicates that there is no *BC* interaction based on the differences between means, from which it follows that there is no *ABC* interaction. (Note that there is a *BC* interaction based on the raw means).

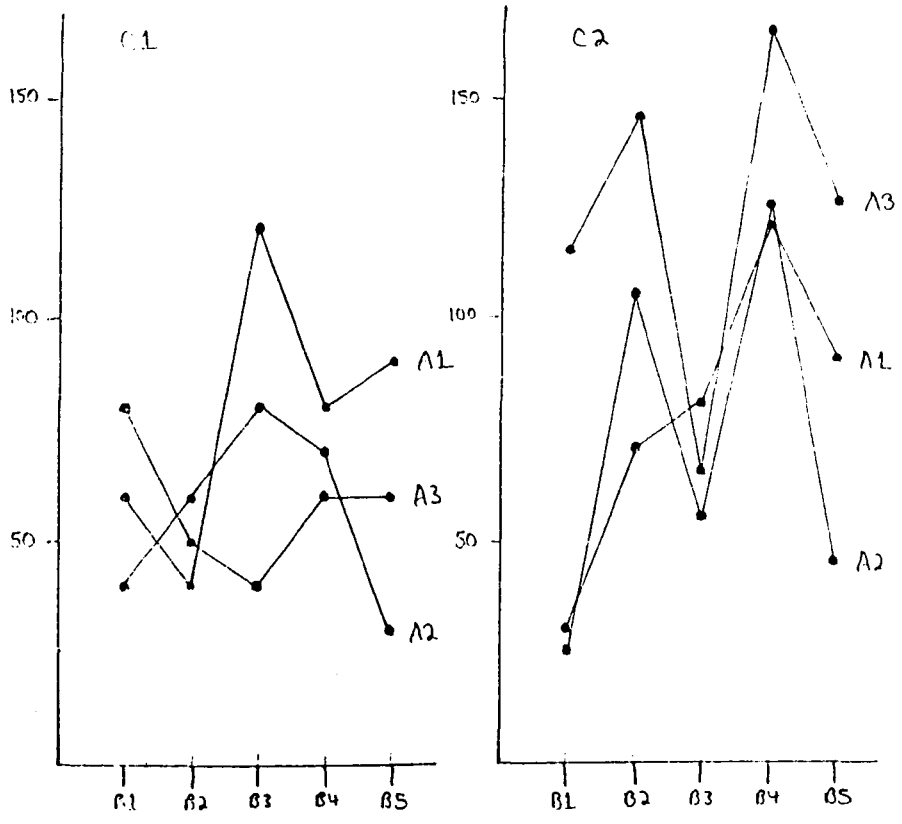

Figure 4-3: Interaction plot for 3 factor ANOVA example

Monlezun only considered data sets for which the three-way interaction was absent. I applied the technique to some contrived three-factor data sets with the three-way interaction effect present in some cases, and to others with various main effects and two-way interactions absent, in addition to the three-way interaction. I also extended the technique and applied it to some higher order data sets, plotting the differences between differences, in order to obtain a two-dimensional representation. In practice, the appropriate plots to construct are dictated by which interactions are present or absent in the data, since these in turn dictate which differences it is appropriate to calculate and plot. The plots were not straightforward to construct, and they were also very difficult to interpret. For the four-factor data, the difficulties of construction and interpretation were compounded.

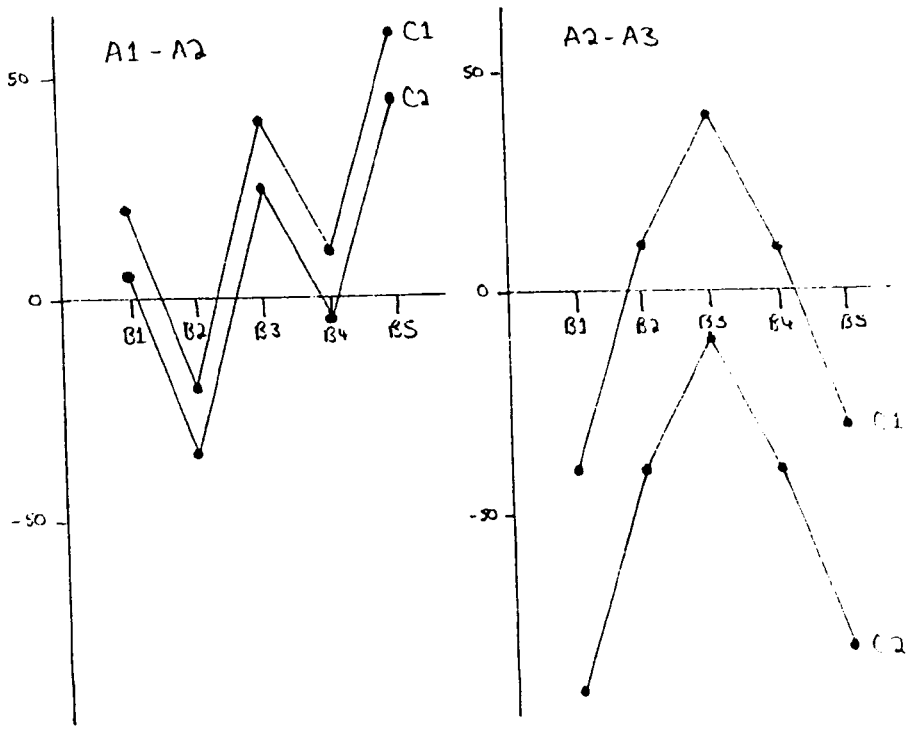

Figure 4-4: Monlezun plot for 3 factor ANOVA example

The two main criticisms I have of Monlezun plots are that the plots show what is not, rather than what is, in the model, and that a large number of plots are required for a given data set. The difficulty of interpretation of the plots, even for an experienced user with considerable knowledge of the construction and rationale of the plots, would seem to be such that it would be better simply to state the model. Thus I conclude that Monlezun's plots are not worthwhile pursuing.

Winer *et al* (1991) suggest the use of a three-dimensional representation for three-factor ANOVA. In Winer's representation, each factor is assigned to one of each of the three co-ordinate axes, and the levels of each factor are spaced along the corresponding axis. This is illustrated in Figure 4.5 with a diagram taken from Winer *et al* (1991). From such a representation it is possible to interpret the three-way interaction, the three two-way interactions, the three main effects, and the various simple main effects and simple interaction effects, depending on the viewing perspective, or cross-section, chosen (see

Winer *et al* for illustrations of these effects). However, in my experience, it can be very difficult to construct Winer's representation for three-factor ANOVA by hand, and such a technique is not readily extendible to higher dimensional models.

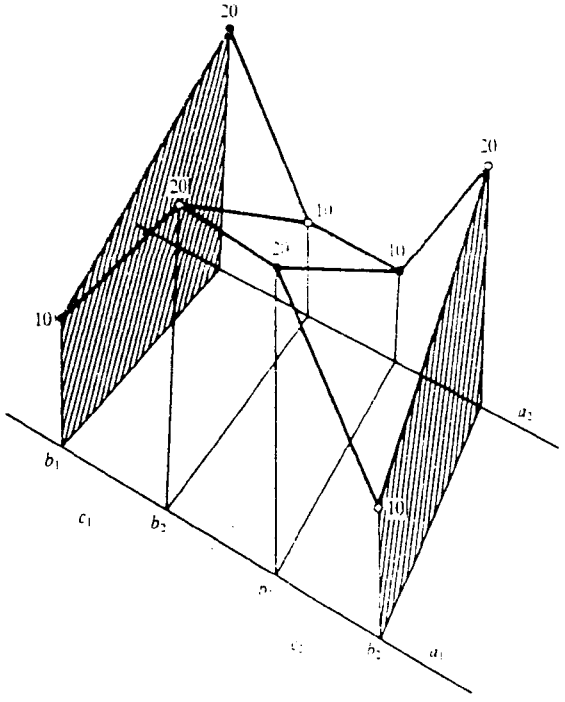

Figure 4-5: Three-dimensional representation of 3 factor ANOVA

All of the techniques considered above are for use only with balanced ANOVA designs. Paik (1985) has suggested the use of circles of diameter proportional to the cell sample size for use in the representation of a three-way (2×2×2) contingency table (see 4.3.3). Incorporation of this approach with ANOVA interaction plots may help to resolve paradoxes which can occur as a result of unequal sample sizes.

4.2.2 ANOVA Summary Table

Graphical techniques are available for representing elements of the ANOVA summary table. Consider the ANOVA summary table presented in Table 4.1, which

corresponds to the data contained in Boniface (1995) used to illustrate the interaction plot in the preceding section.

Source	SS	df	MS	F
A	126.67	3	42.22	15.64
B	640.00	2	320.00	118.52
AB	493.33	6	82.22	30.45
Error	129.38	48	2.70	
Total	1389.38	59		

Table 4-1: Example ANOVA summary table

The table itself presents useful information relating to the corresponding model. For example, the Sums of Squares (SS) values give a breakdown of the variation within the data and some indication of the size of the effects, the degrees of freedom (df) indicate the sample size and number of treatment levels, and the Mean Squares (MS) give the variance estimates used in the ANOVA. In particular, a numerical indication of the size of the main effects and the interaction effect is given by the F values, which can be tested for significance by comparison with appropriate critical values from tables.

Boniface (1995) presents a graphical technique for the representation of the SS values in an ANOVA summary table, based on the Venn diagram. A rectangle is constructed to represent the total SS and then circles are drawn (not to scale) to indicate the proportion of the total SS explained by the different effects. This is illustrated in Figure 4.6 for the summary table in Table 4.1.

By the use of separate circles, overlapping circles, and areas linking circles, Boniface extends the Venn diagram approach to illustrate concepts such as sequential SS, synergic SS, unique SS and adjusted SS. He also modifies this approach to separate out between subjects and within subjects sources of variation. However, Boniface's Venn diagram approach cannot be extended to three factor designs. I shall be considering an alternative Venn diagram approach for higher order designs in Chapter 6.

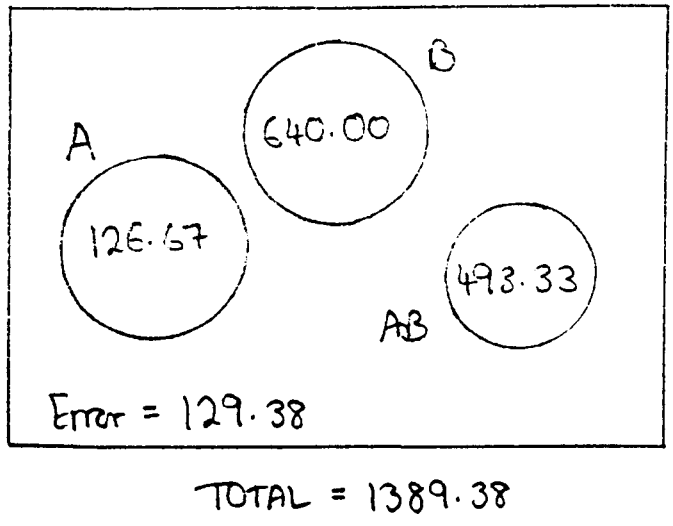

Figure 4-6: Venn diagram representation of Sums of Squares in ANOVA

Bond (1988) has proposed a straightforward but effective technique for the representation of an ANOVA summary table. A bar-chart is constructed showing each main effect and interaction effect, and the error 'effect' which is used to test the significance of the model effects. For each effect, a bar is drawn such that the height of the bar corresponds to MS_{effect} and the width of the bar corresponds to df_{effect}. Thus the area of the bar corresponds to SS_{effect}, since SS/df = MS. In this way, all of the information contained within an ANOVA summary table may be displayed concisely. Bond's technique is illustrated in Figure 4.7 for the data contained in Table 4.1 (ignore the vertical axis on the right-hand side of the graph for now).

I have extended Bond's technique so that each effect, as represented by a bar, can be assessed for significance. Since $F_{test}=MS_{effect} / Ms_{error}$ is significant if $F_{test} > F_{crit}$, this involves multiplying the value of the mean square error by the critical F value found from Kokoska & Nevison (1992) for the appropriate degrees of freedom. The values thus obtained can be indicated on or above the bar representing the effect, or on a right-hand axis of the bar-chart. This is illustrated in Figure 4.7 with $\alpha=0.05$. If the bar representing the effect extends beyond the appropriate critical value of $MS_{error} \times F_{crit}$, then the effect represented by the bar is significant at the level of significance used. In Figure 4.7, the

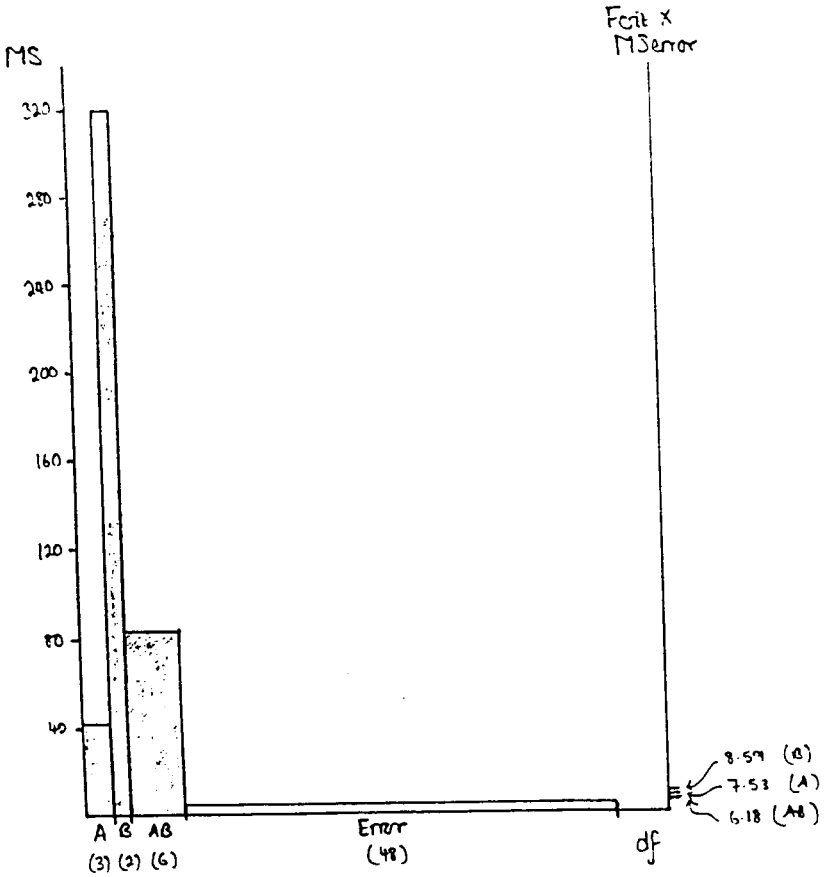

Figure 4-7: Bar chart representation of ANOVA summary table

main effects and the interaction effect are all significant. However, this extension does not indicate how confident one may be that a particular effect is or is not significant.

Other extensions of Bond's technique are also conceivable. For example, I have considered making the height of the bar equivalent to the calculated F-statistic, F_{test}, the width of the bar equivalent to the MS_{error}, and the area of the bar equivalent to MS_{effect}, since $F_{test}=MS_{effect}/MS_{error}$

4.3 Representations for Contingency Tables

4.3.1 Correspondence Analysis

The technique of correspondence analysis has already been described in Section 2.5.3 of Chapter 2. In that section it was seen how correspondence analysis can be used to obtain a graphical representation of the structure within a contingency table. Such a display can be used to communicate and interpret a model which has been fitted to the contingency table data.

4.3.2 Histograms

Boardman (1977) suggests constructing bar charts to show row percentages by column categories for a two-way contingency table, to show how the distribution of the row categories vary according to the column category. This is illustrated in Fig 4.8 for the contingency table contained in Table 4.2, taken from Boardman (1977). Cohen suggests a modification in which the width of each bar is proportional to the row total, giving additional information about marginal distributions.

	Hair Colour				
Eye Colour	Black	Brunette	Red	Blonde	Total
Brown	68	119	26	7	220
Hazel	15	54	14	10	93
Green	5	29	14	16	64
Blue	20	84	17	94	215
Total	108	286	71	127	592

Table 4-2: Example 2-way contingency table

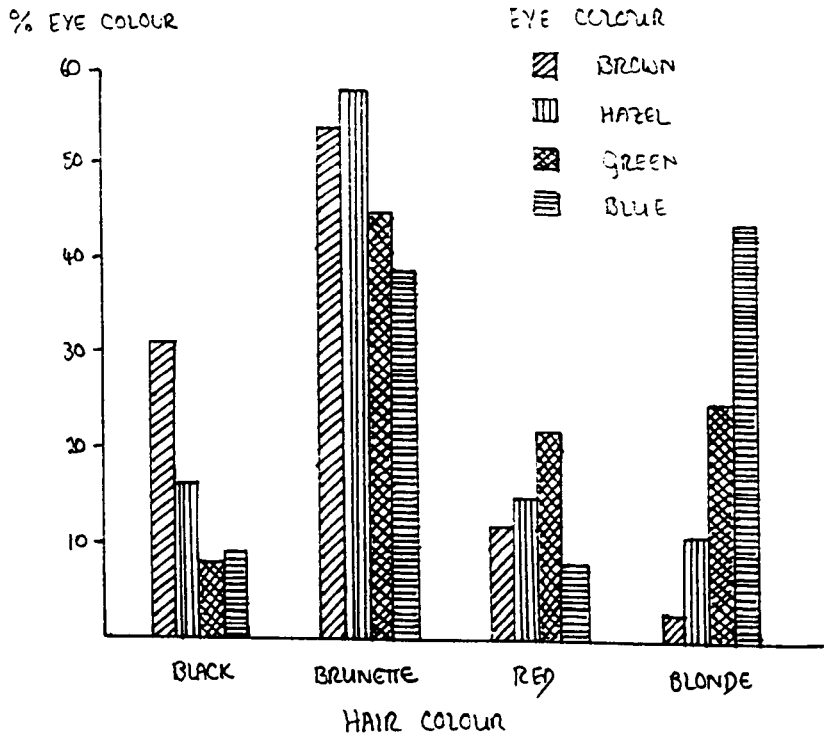

Figure 4-8: Bar chart representation of 2-way contingency table data

Dallal and Finseth (1977) suggest the use of dual histograms to display graphically the cell entries in an R×2 contingency table. The dual histogram is equivalent to the back-to-back stem-and-leaf plot described in Section 2.3.2 of Chapter 2, with the vertical stem formed by the R categories of the contingency table and the two sets of leaves drawn as horizontal bars to represent the two column categories, with length proportional to the frequencies in the corresponding row categories. The dual histogram can be used in an exploratory manner to detect symmetry in the column frequencies, and departures from symmetry. However, if there were a larger number of counts in one column category than in the other, this would need to be taken into account.

An R×2×2 contingency table can be represented by using double dual histograms, whereby a dual histogram is constructed for each category of one of the two dichotomies, and then superimposed. Thus one dichotomy is represented by the left and right sides of the stem, whilst the other dichotomy is represented by each dual histogram.

Boardman (1977) and Cohen (1980) suggest graphical displays for contingency tables which resemble Bond's (1988) technique for the representation of the ANOVA summary table (see Section 4.2.2).

For each cell of the contingency table, a bar is drawn on a bar-chart where the height of the bar is proportional to $U_{ij}=(O_{ij}-E_{ij})/\sqrt{E_{ij}}$ (where O_{ij} is the observed frequency in the *ij*-th cell, and E_{ij} is the expected frequency under the null hypothesis). Cohen suggests that the width of the bar should be proportional to $\sqrt{E_{ij}}$, and the area of the bar proportional to $(O_{ij}-E_{ij})$, to incorporate valuable information about sample sizes. The U_{ij} is therefore a measure of the contribution of the cell to the overall chi-squared statistic. This is illustrated in Figure 4.9 (from Cohen (1980)) for the data contained in Table 4.2.

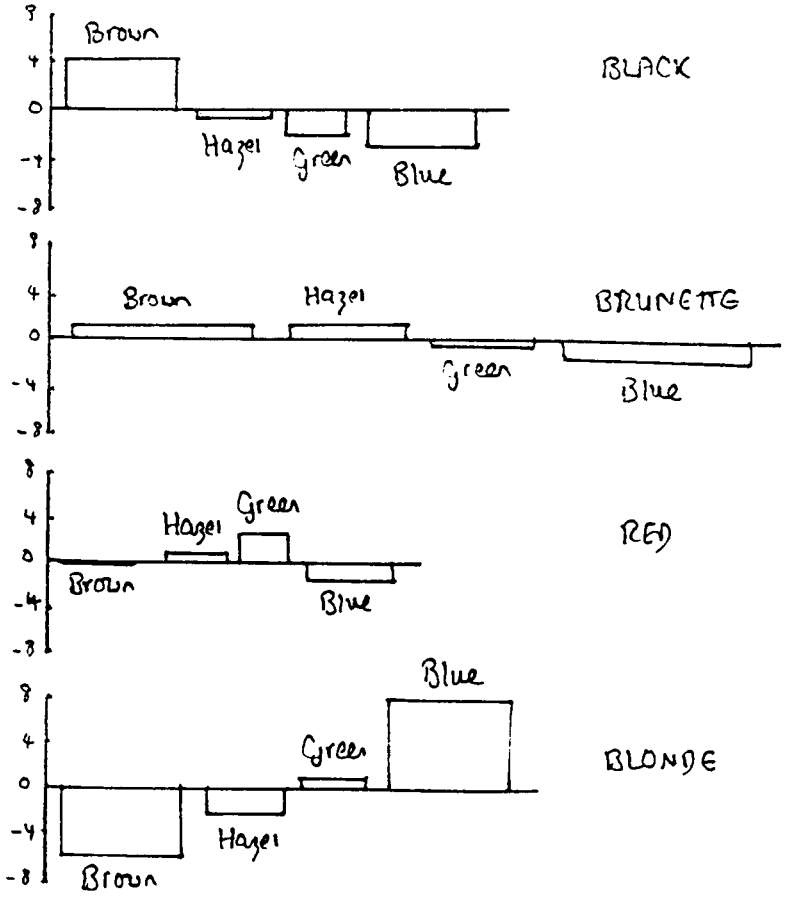

Figure 4-9: Bar chart representation of 2-way contingency table

4.3.3 Circle Graphs

Paik (1985) presents a simple graphical representation for a 2×2×2 contingency table which could also be used for the representation of logit analysis and unbalanced two-way ANOVA. It involves drawing a graph not dissimilar to the interaction plots used in ANOVA (see Section 4.2.1), with circles superimposed on the plotted points. Each circle is drawn with its area proportional to the size of the sample sub-group for that particular plotted point. Different circle graphs can be drawn by varying the plotted positions of each variable and the corresponding circle sizes.

Using the circle graph it is possible to discern relationships existing within the contingency table, and both within group and overall correlations can be seen allowing for the sample sizes as represented by the circles. In this way, use of the circle graph can highlight the occurrence of Simpson's Paradox, which only becomes apparent when sample sizes are taken into account. The circle graph drawn in Figure 4.10 for the data for three classifying variables *A*, *B* and *C* contained in Table 4.3, taken from Paik (1985), provides an illustration of Simpson's paradox.

	A1		A2	
	B1	B2	B1	B2
C1	550	1250	2950	800
C2	1450	2750	1050	200
	2000	4000	4000	1000

Table 4-3: Table of data

4.3.4 Barycentric Plots

For R×2 and R×3 contingency tables, Snee (1974) suggests plotting the observed proportions P_{ij} of frequencies for each row (where $P_{ij}=n_{ij}/N_i$ for the *i*-th row and *j*-th column) as points in a (C−1)-dimensional simplex with barycentric coordinates (a one-dimensional simplex corresponds to a bar-chart; a two-dimensional simplex corresponds to an equilateral triangle). Points which cluster together correspond to homogeneous rows, although it may be helpful to calculate and draw a confidence interval for each point. In

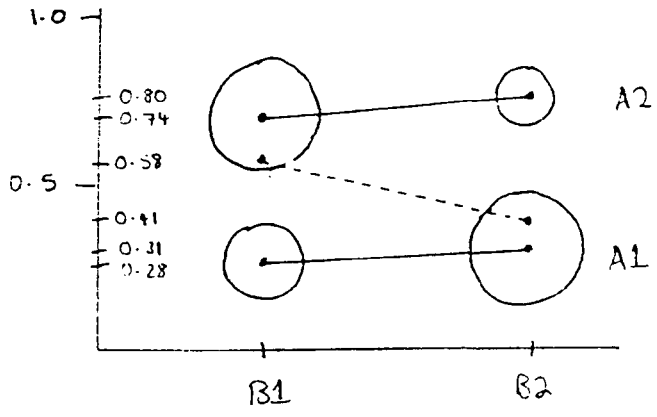

Figure 4-10: Example Circle Graph

this way the user may determine which rows of the contingency table are contributing to a significant chi-squared statistic.

For an Rx4 contingency table, the rows could be plotted as points in a three-dimensional simplex (ie. a tetrahedron), but the confidence regions are not easily displayed, and the points become difficult to interpret. This leads Snee to suggest that for Rx4 and higher-order tables, the points could be plotted in the two-dimensional sub-space defined by the first two eigenvectors obtained in a principal components analysis applied to the original coordinates. Confidence regions can also be projected onto this two-dimensional sub-space. However, Cohen (1980) claims that this might result in loss of information about the structure of the data, and could be confusing to a non-statistician. Figure 4.11 gives the two-dimensional barycentric plot, with confidence intervals, for the (4x4) contingency table data contained in Table 4.3 (figure from Snee (1974)).

A by-product of this graphical display is that it illustrates the inverse relationship between sample size and the size of the confidence regions, which may help the user to decide visually whether a larger sample size is required.

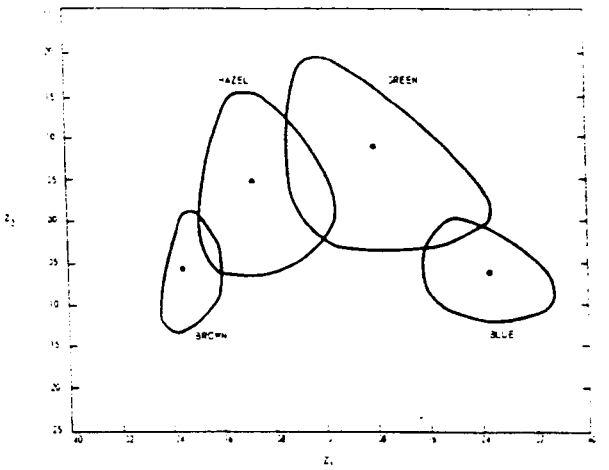

Figure 4-11: Example barycentric plot

4.3.5 Cluster Analysis

With Snee's barycentric plots, described in the preceding section, if *min*(R,C)>3, principal components analysis needs to be carried out in order to reduce the dimensionality of the data. Since this may lead to loss of information about the structure in the data, and may be confusing to a non-statistician, Cohen suggests using cluster analysis, which does not reduce the dimensionality of the data.

The distance measure is taken to be the chi-squared distance between the row (or column) categories, although other distance measures could be used to study the similarities or differences between the categories, depending on the purpose of the analysis. A clustering method (see Section 2.5.6) can then be applied to the symmetric matrix of distance measures — which is constructed either for the rows (to study the relationships between the rows) or for the columns (to study the relationships between the columns), and a dendrogram obtained. Where two very similar categories are merged in the dendrogram, it may be possible to collapse the contingency table across these two categories and to calculate a new distance matrix.

The clustering procedure is particularly helpful in cases where *min*(R,C) is large.

4.3.6 Probability Plots

Cox and Lauh (1967) and Fienberg (1969) have developed graphical methods for the display of functions of a contingency table, based on the half-normal probability plot, for the visual determination of the cells exhibiting interaction. Fienberg's technique is for use with two-way, ie. ($r \times c$), contingency tables, whereas Cox & Lauh's is for use with multidimensional contingency tables where there is a binary response variable. Cohen (1980) presents a gamma probability plot which is claimed to be a simpler alternative to Fienberg's plots. Any deviant points on this plot correspond to cells which contribute to the significant chi-squared statistic.

4.4 Representations for Logit Models

In this section, two graphical representation techniques for logit models are briefly outlined — one by Karger (1980), and the other by Long (1987). Because logit models are not of concern anywhere else in this thesis, details and illustrations have been omitted. The interested reader is referred to the original papers.

Karger (1980) describes a graphical representation technique for 2-way tables of ratios. This is intended to show how the two categorical variables defining the table determine the values of a dependent dichotomous variable. The graphical display is claimed to have the following features:

- It provides a graphical display of the ratio values.
- It permits easy visualisation of the way in which the row and column factors affect the ratios.
- It provides a graphical representation of the results of suitable significance tests, which will provide statistical confirmation of the nature of some effects which are visually apparent.

A rectangle is constructed to represent the *ij*-th cell, having the following features:

- The area of the rectangle is proportional to the corresponding cell value.

- The base of the rectangle x_{ij} equals the column effect under the hypothesis of independence.
- If rectangles are drawn of different heights for the same row, this provides evidence of an interaction effect.
- If there is independence the height of the rectangle y_{ij} is proportional to the row effect.

To avoid the risk of spurious visual effects and to improve the efficiency of the graphical representation as a communication device, statistical information about the nature of the effects can be incorporated. This can be achieved by carrying out a suitable significance test on the data and then shading in those rectangles representing cells for which the corresponding interaction terms have been found to be significant. In order to further increase the visual impact, the rectangles may be shaded according to the size of the corresponding p values. A similar method of shading could be used to communicate the significance of the main effects in the case where there is complete independence of the independent variables.

The R×C rectangles can be arranged according to the rows and columns of the original tables, thus maintaining the basic structure of the original lay-out, or rearranged according to size to indicate the presence of an interaction effect.

Karger suggests that the same approach could also be used for the graphical representation of 2-way contingency tables. The width of the rectangle $x_{ij}=P_{ij}$ and the height of the rectangle $y_{ij}=P_{ij} / P_{.j}$, where P_{ij} is the ij-th cell proportion and $P_{.j}$ is the j-th marginal proportion.

Long (1987) proposes the use of a graphical method, called an "Effects Plot", to represent the magnitudes and statistical significance of all effects in a multinomial logit analysis and which allows easy visual interpretation of the relative magnitudes and directions of effects both within and between the independent variables. Currently, most (non-graphical) applications of logit analysis are limited to making statements about the direction and significance of effects without reference to their magnitudes. This is because of the problems introduced by non-linearity of effects and the large number of coefficients to be considered.

Using an effects plot, it is possible to summarise the magnitudes, direction and statistical significance of the effects of all of the independent variables on all combinations of the dependent categories. Comparisons can also be made of the relative

magnitudes of the effects between different independent variables, as well as within a single independent variable, indicating the differentiation of the different dependent outcomes by that particular variable.

The effects plot can be further supplemented by inclusion of numerical details of the significance of the chi-squared test of the hypothesis that all of the logit coefficients for a given independent variable are zero. In other words, that the given independent variable has no effect on the odds of a particular outcome.

Also, the effects plot allows the user to assess the rankings of the dependent categories, with respect to the ability of an independent variable to differentiate between their odds of occurring.

Thus, Long concludes that adoption of the effects plot in multinomial logit analysis can significantly increase the amount of information obtained in the analysis, and allows a simple presentation of complex results in a form easily conveyed to people unfamiliar with the technical details of logit analysis. Long suggests that the technique could be adapted for use with other statistical techniques which involve nominal or ordinal dependent variables.

4.5 Summary

The main concern of this thesis is with the graphical representation of fitted models. In Chapter 2, some of the many graphical techniques of use for the examination of the raw data in order to determine which model it may be appropriate to fit to the data were described, and in Chapter 3 some graphical techniques for assessing the appropriateness of the fitted model were considered. This chapter has been concerned with the few techniques which exist in statistics for the graphical representation of the fitted model itself; in particular for use with ANOVA models and contingency tables.

None of the techniques considered in this chapter fulfil the aims of this thesis, which is to obtain a graphical technique of use for the representation of all fitted models. However, in Chapter 7 another representation technique which has been developed for use with fitted models, namely the conditional independence graph, will be considered in relation to some simple pen-and-paper techniques which I develop in Chapter 6. As shall be seen, the conditional independence graph goes part-way towards meeting the research aims and will be developed further in the remainder of the thesis.

BLANK PAGE IN ORIGINAL

5. Issues in Graphical Perception and Graphical Presentation

5.1 Introduction

Statistical graphs have been in use for two centuries, since the work of Playfair. Despite their long history and the ubiquity of graphs, little is known about how people perceive and process statistical graphs, and numerous appeals for more empirical investigations have been made (Lewandowsky & Spence (1989a)). Such concerns are not restricted to pen-and-paper techniques — Broersma & Molenaar (1985) are particularly concerned that, although there has been a steady development of software for graphical data analysis, many of the methods for displaying data offered by statistical computer packages are deficient and not based on a clear scientific foundation or on systematic experimentation.

It can be argued (eg. Cleveland (1984b), Kosslyn (1985), Lewandowsky & Spence (1989b)) that graphical displays are so effective because they exploit the natural perceptual, cognitive and memory capabilities of human beings. Human beings are well equipped to recognise and process visual patterns, vision being the dominant human sense. However, very little is known about how graphical displays are processed, there being few experimental studies of how people process graphs. Very often intuition is relied upon to decide whether a graphical display is good or bad. Similarly, advice on use and guidelines for construction are very often without empirical foundation. Graphs are a vital part of communication but the design of graphs for data analysis and presentation is largely unscientific. As Fienberg (1979) states: "Although advice on how and when to draw graphs is available, we have no theory of statistical graphics, nor ... do we have a systematic body of experimental results to use as a guide".

This chapter will consider the main theoretical and experimental results contained within the graphical perception literature (Section 5.2), together with published recommendations for the construction of graphical representations (Section 5.3). These will then be considered (Section 5.4) in connection with my research aims.

5.2 Issues in Graphical Perception

Cleveland (1984) states that the main criterion for judging a graph should be how well people extract the quantitative information portrayed on it.

The earliest research on graphical perception, carried out in the 1920's, tended simply to test whether one type of graph (eg. a pie chart) is better than another (eg. a bar chart) when carrying out judgement or inference tasks (see, for example, Fienberg (1979)). More recently researchers, such as Cleveland & McGill (1984, 1985, 1987), have begun to consider the cognitive processes which operate when people decode the information presented in a graph. Two key questions to be addressed by researchers are, according to Simkin & Hastie (1987):

1. How is the information from a graph represented mentally?
2. What mental processes intervene between early vision and the establishment of the mental representation, operate on the representation to infer non-obvious properties, and operate on the representation and the inferences to generate a task-appropriate response?

Of course, other researchers have continued to focus on aspects of or comparisons between particular techniques by means of controlled experiments; for example: Chernoff & Rizvi (1975) considered the permutation of features in Chernoff faces, Wainer & Francolini (1980) considered the use of colour in two-variable colour maps, Cleveland, Diaconis & McGill (1982) considered the effect of scale changes on judgements of correlation in scatterplots, Cleveland & McGill (1983) considered colour-caused optical illusions on statistical maps, Cleveland, Harris & McGill (1983) revisited Cleveland's earlier work and also considered the judgement of circle sizes, Broersma & Molenaar (1985) compared the effectiveness of stem-and-leaf plots and boxplots for the graphical perception of distributional aspects of data, Lewandowsky & Spence (1989a) considered the use of different symbol types in scatterplots, and Kelly (1993) used a 3x3x3 Latin Square design to compare tables, graphs and text, concluding that the latter is inferior with no significant difference between graphs and tables. However, as Broersma & Molenaar (1985) state, from experiments of this kind it is always hard to make statements that will hold in a more general context.

One way to study graphical perception, according to Cleveland (1987), is to invoke theoretical and experimental results from the more general field of visual perception. Another way is to run specially designed controlled experiments. Cleveland & McGill (1984a) combine the results of their own reasoning and experimentation with the results of psychophysical experiments and the theory of psychophysics.

Cleveland & McGill's (1984a, 1985, 1987) approach, or "paradigm" (also described by Kolata (1984)), is based on the consideration of "human graphical perception". They define graphical perception as the visual decoding of the information encoded on graphs. The aim of their work is to identify the elementary perceptual tasks which are carried out when people extract quantitative information from graphs, and to order these tasks according to how accurately people perform them. This is intended to provide a vocabulary with which to account for or predict performance in simple graph-perception tasks, and guide-lines for graph construction.

Ten elementary codes identified by Cleveland & McGill, which correspond to both textural and geometric aspects of graphs, are: position along a common scale, position along non-aligned scales, length, angle, slope, area, volume, density (ie. amount of black), colour saturation and colour hue. Cleveland & McGill describe how these elementary codes are judged in order to extract quantitative information for a variety of graph forms. They claimed that, for most graphs, a viewer must perform one or more of these mental-visual tasks in order to extract the values of the variables represented. The power of a graph therefore lies in its ability to enable the viewer to take in the quantitative information displayed, to organise it, and to see patterns and structure not readily revealed by other methods of displaying the data.

On the basis of theory and experimentation, both informal and formal, the ten elementary codes have been ordered from most to least accurately judged as follows:

1. Position along a common scale.
2. Position along non-aligned scales.
3. Length.
4. Angle.

4.-10. Slope.

6. Area.
7. Volume.

8. Density.
9. Colour saturation.
10. Colour hue.

This is the version of the list as contained in Cleveland & McGill (1987). This ordering of the codes varies from those presented in the earlier papers by Cleveland & McGill, as more experimentation has been carried out. Still more experimentation is required to confirm the ordering above, particularly for items 8–10.

Simkin & Hastie (1987) are concerned that there is an interaction between graph type and judgement type, believing that people have expectations about the type of information that will be contained in different graph types. In other words, that the ordering of the elementary codes by Cleveland & McGill may depend on the analytic task. Simkin & Hastie assess performance by consideration of both the accuracy and speed of the extraction of quantitative information.

The ultimate aim of Simkin & Hastie is to develop a vocabulary of elementary mental processes that can be combined to build information-processing models of performance in graph-perception tasks. Processes considered by Simkin & Hastie are simple anchoring, scanning, projection, superimposition and scanning operators. These are considered to be sufficient to write instructions to enable another person to perform the experimental tasks, or even to write computer-program models to perform the tasks. They consider that there may be distinct consistent schemata associated with different graph types (a schema being a generic cognitive structure, learned from past experience and stored in long-term memory, that guides a perceiver in organising incoming information into a complex knowledge representation) which make use of the elementary mental processes they have identified.

Lewandowsky & Spence (1989a) conducted two behavioural experiments to study the effect of symbol type and the expertise of the observer in discriminating strata in scatterplots. They also considered both accuracy and speed, showing that "measuring response latency in addition to accuracy is essential in research on graphical perception". Not surprisingly, some symbol types were processed more quickly than others, with different colours being processed fastest, and confusable letters being processed slowest. Identical patterns of accuracy were found, but not until response time was restricted. A surprising difference was found between experts and novices, such that the experts were

slightly more accurate but slower. When processing time was restricted, however, the superior accuracy of experts was confirmed.

As well as response time and error rate, other methods, described by Kruskal (1982), may be used to compare two graphical methods. These include recall and more aesthetic considerations such as user preference. Kruskal is critical of the fact that very few papers have been published in which subjects have been asked to do graphical statistical tasks more complex than simply judging size, answering factual questions, etc.. However, such advantages as insight, understanding and discovery are difficult to assess.

Broersma & Molenaar (1985) are concerned that the results of graphical perception experiments will depend to a large extent on the design of the experiment (eg. the instruction given and the face validity of the exercise), features of the subjects (eg. their experience and motivation), the stimuli used, the specific display form, and on many other circumstances that may make it really difficult to carry out a properly controlled experiment with results that are readily generalisable.

Although there is some evidence that a science of graphic presentation has started to emerge (Allison et al (1993)), stimulated largely by the work of Cleveland and his colleagues, much still remains to be learned, particularly as regards cognitive aspects of the processing of graphs including the role of short term and long term memory (Lewandowsky & Spence (1989b)). Pittenger (1995) observes that most of the research on perception and the use of graphs has been published by statisticians in statistical journals – psychologists have written very little in this area, despite their extensive use of statistical graphics!

5.3 Issues in Graphical Presentation

A rational set of standards for graphical presentation should be based on theory and experimentation, yet graph design for data analysis and presentation is largely unscientific. As Wainer & Francolini (1980) state: "The search for rules for effective graphical display, whether for the purpose of communication, exploration, or reconstitution, has been hampered by the lack of a cohesive body of experimental evidence regarding the parameters of efficacious graphical display".

At present we have only a few results from the graphical perception literature – for example, Cleveland & McGill (1984a) claim that graphs should employ elementary codes

as high in the ranked list presented above as possible. On the basis of the derived ordering of the elementary perceptual tasks, Cleveland & McGill have analysed some of the more common forms of graphs and have suggested replacements were appropriate. For example, they suggest replacing divided bar-charts by dot-charts and replacing pie-charts by bar-charts.

Several sets of simple, intuitive, suggestions have been made that should improve the clarity of most graphs. For example, Cox (1978) suggests the following:

1. The axes should be clearly labelled with the names of the variables and the units of measurement.
2. Scale breaks should be used for false origins.
3. Comparison of related diagrams should be made easy, for example, by using identical scales of measurement and placing diagrams side by side.
4. Scales should be arranged so that systematic and approximately linear relations are plotted at roughly 45° to the x axis.
5. Legends should make diagrams as nearly self-explanatory, that is, independent of the text, as is feasible.
6. Interpretation should not be prejudiced by the technique of presentation, for example, by superimposing thick smooth curves in scatter diagrams of points faintly reproduced.

Wainer (1984) gives twelve rules for bad data display. Rather than quote the rules for the bad display of data, I have modified them to give twelve principles for good data display as follows:

1. The less information carried in the display, the worse it is. The amount of information in a display can be measured using Tufte's (1983) "data density index" which is equivalent to the number of numbers plotted per square inch. A good graphical technique can convey a large amount of information in a small space.
2. Maximise the "data-ink ratio". Also devised by Tufte, this measures the amount of ink used in graphing the data to the total amount of ink in the graph. A low data-ink ratio means a lot of 'chart-junk' which can obscure the data or features of the data.
3. A set of numbers having both magnitude and order should be represented by an appropriate visual metaphor, such that magnitude and order of the metaphorical representation matches that of the numbers.

4. Length should not be used as the visual metaphor when the observer is more likely to perceive area. It is possible to calculate Tufte's measure of "perceptual distortion", which is the perceived change in magnitude divided by the actual change.
5. The perception of the graph can be modified misleadingly by the choice of interval or scale (particularly for time series data), and by showing parts of the data outside of the context of the rest of the data.
6. Changing the scale in mid-axis can make large difference look small, and make linear changes look linear, and vice versa. This can therefore be misleading.
7. A graph should not emphasise the trivial and ignore the important.
8. To allow comparisons to be made between graphs, these graphs should start from a common base.
9. Ordering graphs and tables alphabetically can obscure structure in the data which would have been made obvious if the display had been ordered by some aspect of the data.
10. Labels should not be illegible, incomplete, incorrect or ambiguous.
11. Excessive decimal places should not be shown in tables since they may imply greater accuracy than is justified, and may not be very clear. Similarly, excessive dimensions should not be used in graphical representations — for example, using both length and area may be ambiguous, and colour should be used with caution, owing to inconsistencies in human perception.
12. Consider techniques which have been used in the past, and which may be better suited to a current purpose than any alternative techniques which you may invent.

Kosslyn (1994) has published a "how-to book" on graph design which offers step-by-step recommendations for the construction of simple graphs, and which also addresses the perceptual and cognitive principles underlying his recommendations. In a chapter concerned with how people lie with graphs, he presents a set of recommendations, rather in the way that I have converted Wainer's rules for bad data display into recommendations above. However, Kosslyn's list is too long to reproduce here.

Other authors have concerned themselves with the bad display of data; for example, Huff's (1973) classic book contains many examples of misleading graphics, as does Reichmann (1961).

Tufte (1983, 1990) has studied graph usage, particularly within the media, and found widespread misrepresentation of data and gross errors, together with some shining examples of the use of graphics over the past 200 years. Based on his experience in the field, rather than experimental assessments, Tufte presents a number of recommendations and criteria for making "good" graphs. For example, Tufte criticises the use of "chart junk" – the embellishment or decoration of graphs often employed by the media – and believes that nothing should be included in the graph that is not absolutely necessary for the display of the data, ie. that the "data-ink ratio" (the ratio of ink used for data to ink used for other parts of the graph) should be maximised. Other criteria have already been presented above, such as "perceptual distortion" and "data density index".

Other books which contain a survey of, and recommendations for, graph construction are Bertin (1983) and Schmid (1983). Bertin even goes so far as to develop a taxonomy of graphical components and to introduce a grammar for the description of graphs so that, in theory at least, graphs may be unambiguously reduced to a brief grammatical description and subsequently reconstructed.

Cleveland (1984a) presents a short list of guidelines, both for journals and for authors, to prevent the sorts of errors he detected in a detailed survey of the use of graphs in the journal *Science*. In common with the guidelines contained in Tufte, Schmid, etc., however, there is no compunction for such guidelines to be adhered to.

Fienberg (1979) reports that, with the rapid growth of graphical presentations, has come a concern for the need of recognised standards. Such concern is not altogether new, and in fact may be traced back to 1853, but has met with only limited success to date. A Joint Committee on Standards for Graphic Presentation (1915) published 17 basic rules for graphical presentations, although the types of displays considered reflect the kinds of graphs that would have dominated publications at that time. Other attempts to formulate graphical standards were made in 1936 and 1941 (Bachi (1975)).

Some aspects of graphical presentation must, however, surely be consistent. For example, Wainer (1990) finds three important areas of agreement between Playfair and Tukey, despite the 200 years which divide these pioneers of graphical design:

1. Impact is important.
2. Understanding graphs is not always automatic.
3. A graph can show us things easily that might not have been seen otherwise.

5.4 Summary

In Section 5.2, what little work has been carried out on graphical perception has been considered. Some of this work (eg. Cleveland & McGill (1984a), Simkin & Hastie(1987)) has been concerned with the development of a theory or model of graphical-information processing. Ideally, such work would lead to a clear set of standards or guidelines for graphical presentation. Other work has been directly concerned with the effectiveness of graphical presentation techniques (eg. Lewandowsky & Spence (1989a)). However, there is as yet insufficient evidence available to guide the construction of new techniques for the graphical display and communication of statistical data, which is the concern of this thesis. As has been seen, in Section 5.3, what advice has been offered has been for the most part subjective and intuitive.

Having considered the available evidence and advice, it seems necessary for me to attempt to develop graphical representation techniques for the display of structured multivariate data in the time-honoured way of intuition and subjective insight. However, it is my intention to assess the effectiveness of any technique which seems to be particularly successful by means of a graphical perception experiment which, in the light of the work by Simkin & Hastie (1987) and Lewandowsky & Spence (1989a), should assess both latency and accuracy.

BLANK PAGE IN ORIGINAL

6. Some Two-Dimensional Approaches

6.1 Introduction

In this chapter I describe a number of novel pen-and-paper approaches to the problem of obtaining a graphical representation of a given fitted model, each of which attempts to indicate the interactions which are present in the model. The techniques developed have been divided into three sections. In the first section (Section 6.2), various ways of combining points in two-dimensions will be considered. In the following section (Section 6.3), an idea based upon Venn diagrams will be explored. In the final section (Section 6.4), a new type of graphical display, the Topological–Magnitude graph, is described, which takes two forms; the first of which (the Topological Graph) indicates the interactions which are present in the model, and the second of which (the Magnitude Graph) illustrates the nature of an interaction and can be used to communicate additional information about the size of these interactions. However, few of the techniques which I develop can be used for the representation of any given fitted model, and those which can do not always result in an aesthetically pleasing representation, so some attempt will be made to assess the shortcomings of each technique.

In this and subsequent chapters I shall be concerned only with hierarchical fitted models and so shall begin by giving a definition of a hierarchical model together with an explanation of the notation adopted. I shall then go on to define and illustrate the concept of the generating class of a model, which provides a unique and concise way of summarising the interactions in a given fitted (hierarchical) model. I will then describe the use of the 'Implication Diagram', which is based on a lattice structure, to determine and represent visually all the terms in a fitted model having a given generating class, before going on to describe the two-dimensional approaches in the following sections.

6.1.1 Hierarchical Models

A hierarchical model is one for which if an effect of a particular order is included in the model, then all lower order effects involving the variables contained in this higher order effect are also included in the model. For example, if the second order (three-way) interaction term involving the variables A, B, C, which I shall denote ABC, is included in

the model then the first order (two-way) interaction terms AB, AC, BC, and the zero order interaction terms (ie. the main effects) A, B, C are also included in the model, together with any constant term. Examples of models which are typically (but not necessarily) hierarchical, and which can therefore be considered in this way, include log-linear interaction models and ANOVA models.

6.1.2 Generating Class

The generating class of a (hierarchical) model is the set of terms contained in the given model whose presence in the model is not implied by any other terms contained in the model, but which between them imply the presence of all the other terms in the model.

For example, consider a model containing all main effects and some interaction effects involving the four variables A, B, C and D, which shall be expressed as

$$A+B+C+D+AB+AC+BC+CD+BCD.$$

The generating class of this model would be

$$\{[BCD]\ [AB]\ [AC]\},$$

since the BC and CD interactions are implied by the presence of the BCD interaction, as are the B, C and D main effects (which are also implied by the BC and CD interactions which were themselves implied by the BCD interaction), and the presence of the A main effect is implied by both the AB and AC interactions, which also imply the presence of the B and C main effects. However, no terms in the model imply either the AB, AC or BCD interactions, hence the generating class given is derived.

6.1.3 Implication Diagram

Lattice diagram type representations have been used in a model selection context (for example, by Whittaker & Aitkin (1978), Cottee (1987) and Whittaker (1988,1990)) to show hierarchical relationships between different models. I have extended this idea in order to show the hierarchical relationships between the different terms in a single model. Use of this lattice diagram representation enables easy determination of the terms in the model whose presence is implied by the elements of the generating class of the model; hence the lattice has been termed an "Implication Diagram".

The Implication Diagram is constructed with all the interactions contained within the model which are of the same order located on the same row, and a link is drawn

between two interactions of different orders but in adjacent rows if the presence of one interaction in the model implies the presence of the other. The links are drawn as arrows from the higher order interaction term to the lower order interactions which are implied by this term by the exclusion of each of the variables in turn. Thus an interaction term involving n variables will have links drawn as arrows pointing to n lower order terms each involving $n-1$ variables and formed by dropping a single variable from the original interaction term.

Given the generating class of the model, the elements of the generating class are written on the appropriate rows according to the order of the interactions represented by the elements. Then, taking each element in turn, the interactions (or main effects) implied by an element are written down on the row below, and the links drawn in. Then the interactions implied by these interactions are written down on the lattice, and so on until all possible interactions and all the main effects implied by the elements of the generating class have been included.

To illustrate the Implication Diagram approach, consider the generating class

$$\{[ABC]\ [ACD]\ [BD]\}$$

The corresponding implication diagram will be as given in Figure 6-1, and by considering each of the interactions and main effects in the diagram, which will correspond to each of the terms in the model, the model having the given generating class can be determined to be as follows:

$$A+B+C+D+AB+AC+AD+BC+BD+CD+ABC+ACD$$

6.2 Two-Dimensional Combinations of Points

6.2.1 Introduction

All of the techniques described in this section have in common the following features:

- Variables are represented by points located freely in two dimensions (ie. on the plane), with one point (or vertex) per variable.
- These points are combined in some way in order to represent the generating class of the fitted hierarchical model which it is intended to represent.

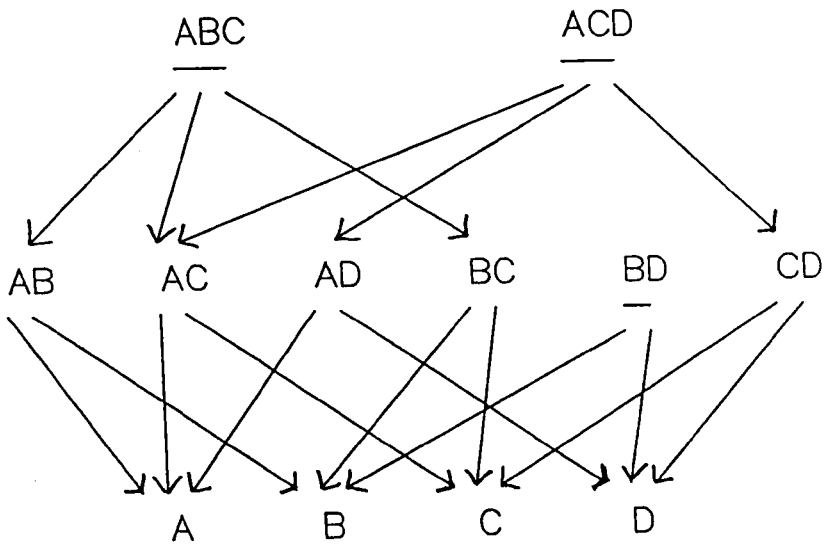

Figure 6-1 Implication Diagram for model with generating class {[ABC] [ACD] [BD]}

- If the generating class of the model may be represented by the technique without ambiguity, it becomes possible to determine all of the interactions contained in the model, and these may be read from the graph.

The most intuitive way of combining two points is to draw a line (or edge) between them on the graph. The presence of an edge in a graph can then be interpreted as indicating that the two variables corresponding to the two vertices joined by the edge are associated and that the nature of this association would be a two-way interaction, whereas the absence of an edge would indicate that the two variables are not associated. Since my concern is with hierarchical models, if the two-way interaction is zero then all higher order interactions involving this pair of variables will also be zero.

To represent a three-way interaction, all possible edges could be drawn between the three vertices corresponding to the three variables involved in the three-way interaction, resulting in the complete graph or sub-graph involving these vertices. This is illustrated in Figure 6-2. The three two-way interactions and the three main effects can also be determined from this graph.

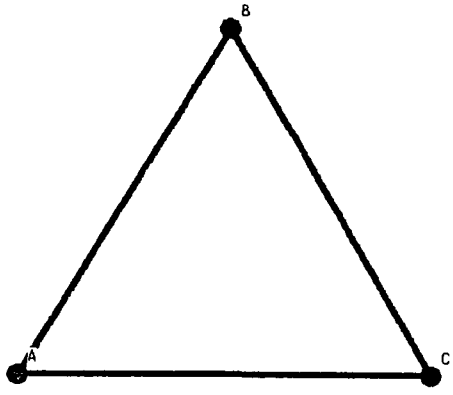

Figure 6-2 Complete graph on three vertices drawn to represent a three-way interaction

However, suppose that this approach is to be used to represent a model in which the three two-way interactions *AB*, *AC* and *BC* are non-zero, but there is no three-way interaction *ABC*. How could this be represented on a graph so as to avoid confusion with the complete (sub-) graph drawn on three vertices which represents the *ABC* interaction?

What follows in the rest of this section is a description of a number of approaches which I have developed, all of which involve the combination of points on the plane, either by lines, shapes, or shadings, in an attempt to represent, in an unambiguous manner, the two-way, three-way, and higher-order interactions which may be contained in fitted hierarchical models. In practice, however, for most of the techniques described there may be some models which cannot be represented without ambiguity. However, every technique described is of use for the representation of some models (which will differ between techniques), and an attempt will be made to define the limitations of each technique. The selection of possible approaches presented is by no means exhaustive.

In later chapters, a number of the ideas explored in the development of these techniques will be considered further.

6.2.2 Links Between Vertices

This is the approach outlined in Section 6.2.1, in which a line is drawn between two vertices if the corresponding pair of variables are associated. As has already been

mentioned, it would not be possible, using this technique, to distinguish between the model with generating class {[*AB*][*AC*][*BC*]} and the model with generating class {[*ABC*]}, owing to the pair-wise definition of the edges. It could be made a convention that a complete sub-graph on three vertices always corresponds to a three-way interaction, in which case it would not be possible to represent the model with generating class {[*AB*] [*AC*] [*BC*]} using this technique.

Suppose that we wished to represent a model having three three-way interactions in its generating class; eg. {[*ABC*] [*ABD*] [*BCD*]}. Using links between vertices the graph shown in Figure 6-3 would be obtained. However, this graph appears to represent the model with generating class {[*ABCD*]}. Note that to represent a four-way interaction, the complete graph would be drawn on four vertices so as to imply the presence of the three-way interactions as well — if the four vertices were simply joined in a quadrilateral, this would represent a model having just four two-way interactions in its generating class. But even if the viewer was told that the four-way interaction is zero in the model represented in Figure 6-3, the graph appears to represent the model with generating class {[*ABC*] [*ABD*] [*ACD*] [*BCD*]} and is indistinguishable from the graphs which would be drawn to represent other models involving four variables and having three three-way interactions in their generating class; as well as being indistinguishable from certain other models, such as the model with generating class {[*ABC*] [*BCD*] [*AD*]} and the model with generating class {[*ABD*] [*ACD*] [*BC*]}. Thus it is impossible, using links between vertices in this way, to represent the model {[*ABC*] [*ABD*] [*BCD*]} without ambiguity.

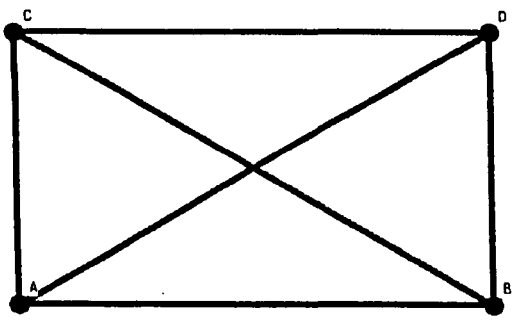

Figure 6-3 Graph drawn using links between vertices to represent the model with generating class {[ABC] [ABD] [BCD]}

As shall be seen in Chapter 7, the representation of models by graphs drawn using links between vertices is not dissimilar to the conditional independence graph approach used in graphical modelling. In the interpretation of conditional independence graphs, the cliques (maximally complete sub-graphs) of a given conditional independence graph are taken to correspond to the elements of the generating class of the model represented. Thus it is acknowledged that certain models, such as {[*AB*] [*AC*] [*BC*]} and {[*ABC*] [*ABD*] [*BCD*]} are 'non-graphical' and cannot be represented using this approach.

6.2.3 Representation of Interaction Type by Line Styles

In the previous section, in which the use of links between vertices was proposed in order to obtain a two-dimensional representation of the generating class of a given model, the same style of link was drawn between vertices irrespective of the order of the interaction to be represented. By the use of different line styles corresponding to different interaction types, it becomes possible to represent a larger number of models without ambiguity.

For example, if a dashed line is used to join pairs of variables involved in three-way interactions and the usual continuous line is used to represent two-way interactions, it becomes possible to distinguish between the model with generating class {[*AB*] [*AC*] [*BC*]} and the model with generating class {[*ABC*]}, as shown in Figure 6-4. However, the graph shown in Figure 6-5, which has been drawn to represent the model with generating class {[*ABC*] [*ABD*] [*BCD*]}, is still indistinguishable from the graphs which would be drawn to represent other models involving three or four three-way interactions on four variables. The model with generating class {[*ABC*] [*BCD*] [*AD*]} would, however, now be distinguishable, as shown in Figure 6-6. Also, the model with generating class {[*ABCD*]} could be represented without ambiguity, by the use of a different line style to correspond to a four-way interaction. Alternatively, different thicknesses of lines could be used in place of different styles, or different colours could be used, although this latter approach would not result in a representation which could be readily reproduced.

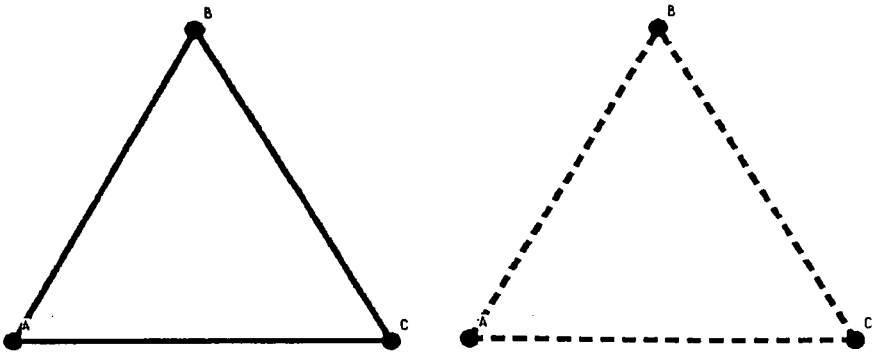

Figure 6-4 Graphs drawn using line styles corresponding to interaction type to distinguish between the model with generating class {[AB] [AC] [BC]} (left) and the model with generating class {[ABC]} (right)

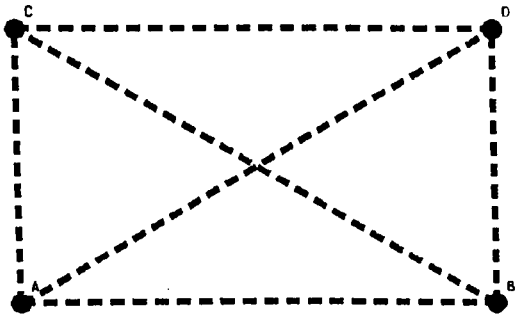

Figure 6-5 Graph drawn using line styles corresponding to interaction type to represent the model with generating class {[ABC] [ABD] [BCD]}

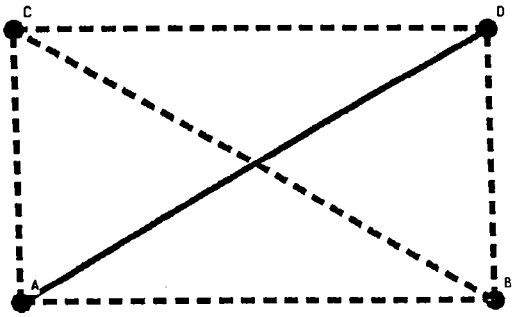

Figure 6-6 Graph drawn using line styles corresponding to interaction type to represent the model with generating class {[ABC] [BCD] [AD]}

Thus using this approach, in which different interaction types are represented by different line styles, it is possible to represent a larger number of models uniquely than could be achieved using just one line style, but there may still be problems of ambiguity for some models. I shall, however, be considering the use of different line styles for the representation of interactions further in Chapter 9, within the context of conditional independence graphs.

6.2.4 Use of Line Styles to Distinguish Between Interactions

If, instead of using different line styles corresponding to different orders of interaction as described in Section 6.2.3, different line styles were to be used corresponding to the different interactions contained in the generating class of the model, it should be possible to identify the elements of the generating class, and therefore the model itself, without ambiguity. As before, the line styles may be formed by different styles, by different widths, or by different colours, etc..

For example, Figure 6-7 shows the graphs corresponding to the model with generating class {[*AB*] [*AC*] [*BC*]} and the model with generating class {[*ABC*]}. Figure 6-8 shows the graph corresponding to the model with generating class {[*ABC*] [*ABD*] [*BCD*]}, and Figure 6-9 shows the graph corresponding to the model with generating class {[*ABC*] [*BCD*] [*AD*]}. All of these models can be readily determined from the graphs displayed.

Thus using this approach, it would seem to be possible to represent any given model without ambiguity. However, the use of many different line styles could be confusing. For example, the different line styles used may not, if there are many of them, be readily distinguished and, unlike the use of different line styles corresponding to different orders of interaction, the line style used does not convey any information about the order of the interaction represented. Moreover, there can be confusion where more than one line is drawn between a pair of vertices when two or more interactions have an edge in common. This situation is illustrated in Figures 6-8 and 6-9.

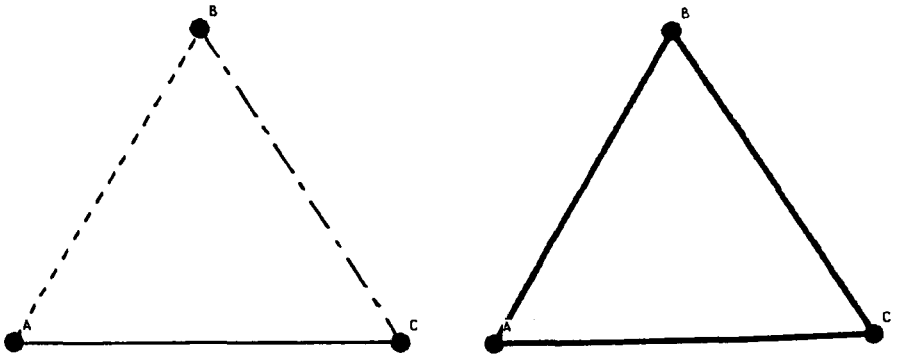

Figure 6-7 Graphs drawn using line styles corresponding to different interactions to distinguish between the model with generating class {[AB] [AC] [BC]} (left) and the model with generating class {[ABC]} (right)

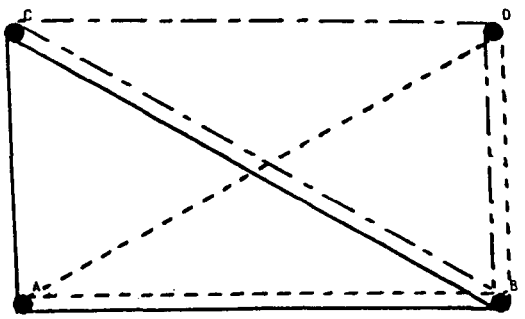

Figure 6-8 Graph drawn using line styles corresponding to different interactions to represent the model with generating class {[ABC] [ABD] [BCD]}

6.2.5 Shading of Areas

Shading of Interacting Areas

Rather than focusing on the joining of vertices by lines, with this approach it is proposed that the fundamental shape formed by the variables/vertices involved in each interaction (eg. no shape for two-way interactions (two vertices linked by a single edge), a triangle for three-way interactions (three vertices linked by three edges), a square or a

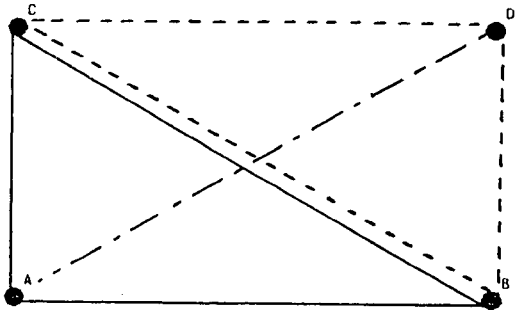

Figure 6-9: Graph drawn using line styles corresponding to different interactions represent the model with generating class {[ABC] [BCD] [AD]}

triangle for four-way interactions (four vertices linked by six edges), etc.) be filled in using a unique shading style.

For example, in Figure 6-10 the graphs corresponding to the model with generating class {[*AB*] [*AC*] [*BC*]} and the model with generating class {[*ABC*]} are shown. The shading of the triangle in the right-hand graph indicates the presence of the three-way interaction. In Figure 6-11, the graph for the model with generating class {[*ABC*] [*ABD*] [*BCD*]} is shown, and in Figure 6-12, the graph for the model with generating class {[*ABC*] [*BCD*] [*AD*]} is shown, with the areas corresponding to the interactions in the generating classes shaded. It is possible to determine the model which is represented in each case by inspection of the different shapes, each shaded in a unique manner, and the vertices forming the boundary of these shapes. If a line does not form the boundary of a shaded region (as for the line AD in Figure 6-12), then this line corresponds to a two-way interaction which forms part of the generating class.

It can be seen that the models represented using the shading of interacting areas can be determined quite readily, whereas they would have been indistinguishable from other models if links between vertices had been used (see Section 6.2.2 above). Thus with this approach it would seem that it is possible to represent a larger number of models uniquely.

However, the main disadvantage of this approach is that there can only be so many unique and readily distinguishable monotone shades which could be layered one on top of the other and still be discernible. Use of a particular shading style could only be

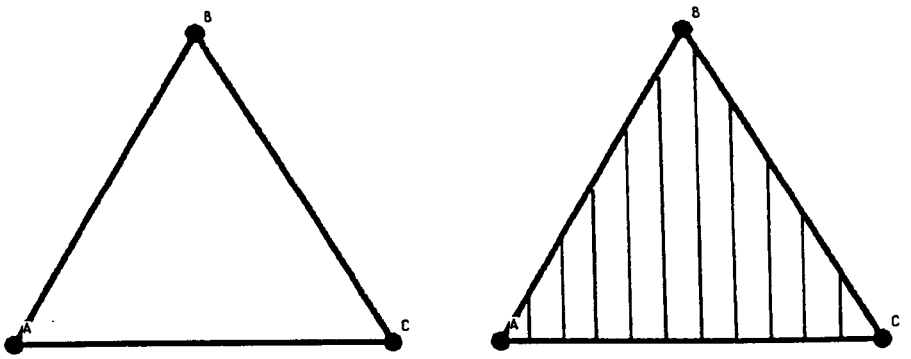

Figure 6-10 Graphs drawn using shading of interacting areas to distinguish between the model with generating class {[AB] [AC] [BC]} (left) and the model with generating class {[ABC]} (right)

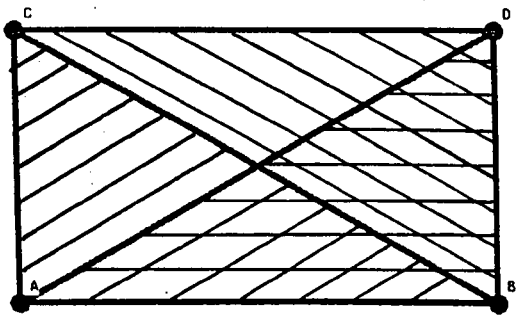

Figure 6-11 Graph drawn using shading of interacting areas to represent the model with generating class {[ABC] [ABD] [BCD]}

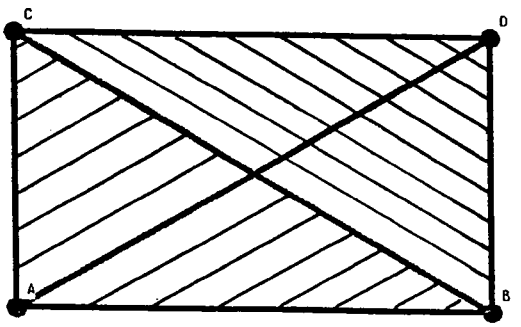

Figure 6-12 Graph drawn using shading of interacting areas to represent the model with generating class {[ABC] [BCD] [AD]}

duplicated for non-adjacent non-overlapping areas. It would not be possible to use solid shading or coloured shading because of the problems that would result in the overlapping areas.

Shading of Non-Interacting Areas

In this approach, any enclosed area of the graph which is *not* involved in a three-way or higher-order interaction is shaded, whereas the other areas (which are therefore involved in three-way or higher-order interactions) are left unshaded. This is therefore the opposite approach to that described above.

Figure 6-13 shows the two graphs corresponding to the model with generating class {[*AB*] [*AC*] [*BC*]} and the model with generating class {[*ABC*]}. Knowledge that the shaded areas do not form part of any interactions makes it possible to correctly determine the generating class of each model represented. However, this approach would only work for certain graphs. In some graphs there would be no regions which are not involved in at least one interaction and could therefore be shaded. See, for example, Figures 6-14 and 6-15, which show the models with generating class {[*ABC*] [*ABD*] [*BCD*]} and {[*ABC*] [*BCD*] [*AD*]} respectively.

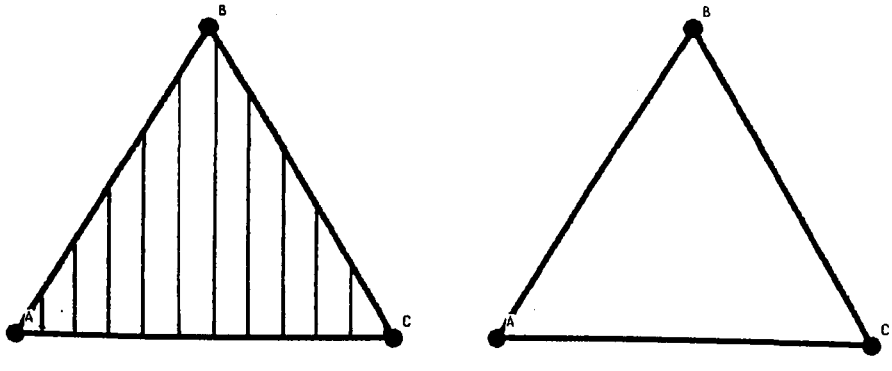

Figure 6-13 Graphs drawn using shading of non-interacting areas to distinguish between the model with generating class {[AB] [AC] [BC]} (left) and the model with generating class {[ABC]} (right)

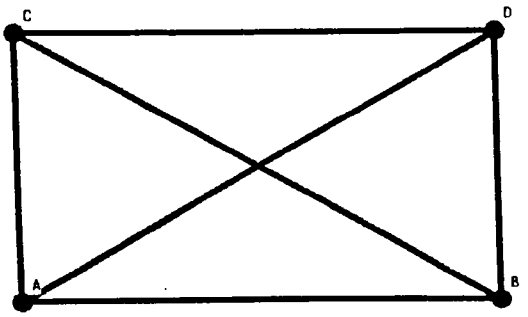

Figure 6-14 Graph drawn using shading of non-interacting areas to represent the model with generating class {[ABC] [ABD] [BCD]}

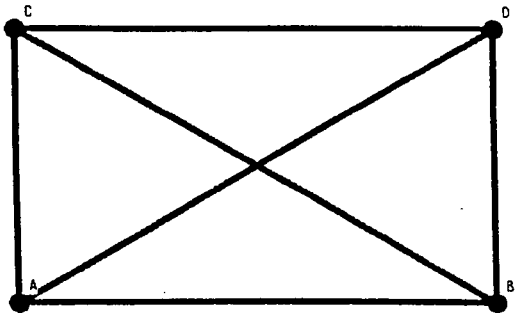

Figure 6-15 Graph drawn using shading of non-intersecting areas to represent the model with generating class {[ABC] [BCD] [AD]}

Figure 6-16 shows an alternative representation of the model with generating class {[*ABC*] [*BCD*] [*AD*]} which is represented in Figure 6-15. Because of the different lay-out of the vertices, it is now possible to determine correctly the model represented. However, it would not be possible to rearrange the vertices to represent the model with generating class {[*ABC*] [*ABD*] [BCD]} without ambiguity. Thus this approach is of use for the representation of some models, but the success of the approach depends on the nature of the interactions involved in the model, and may also depend on the lay-out of the vertices adopted.

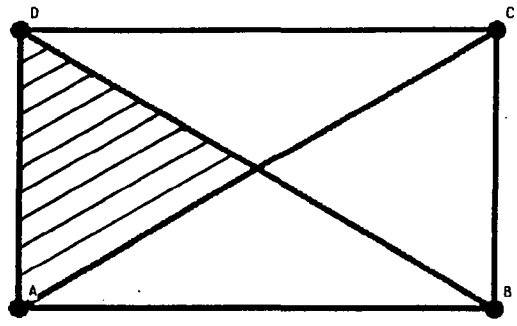

Figure 6-16 Graph drawn using shading of non-interacting areas to represent the model with generating class {[ABC] [BCD] [AD]}, using alternative lay-out of vertices

6.2.6 Enclosure of Vertices by Boundaries

This approach involves surrounding the vertices representing the variables involved in the interactions in the model by some boundary or closed shape, rather than linking the vertices in some way and/or shading areas. Thus the model with generating class {[*AB*] [*AC*] [*BC*]} can be readily distinguished from the model with generating class {[*ABC*]}, since in the former case the three pairs formed by the vertices will be enclosed within three boundaries, whereas in the latter case, the triplet formed by the vertices will be enclosed within a single boundary, as shown in Figure 6-17. Also, the model with generating class {[*ABC*] [*ABD*] [*BCD*]} and the model with generating class {[*ABC*] [*BCD*] [*AD*]} can be represented without ambiguity as shown in Figures 6-18 and 6-19 respectively.

Although this technique could be extended to any number of vertices and any size of interaction, for a large number of interactions the boundaries of the shapes will inevitably overlap, making the graph quite messy and difficult to read. Different types of shapes could be used corresponding to the order of interaction enclosed, so that it would be possible to know at a glance how many vertices should be enclosed within a particular shape. For example, one might use oblongs for two-way interactions, triangles for three-way interactions, squares for four-way interactions, etc., but for a large number of vertices it may be impossible to draw regular shapes and avoid enclosing vertices which should not be enclosed.

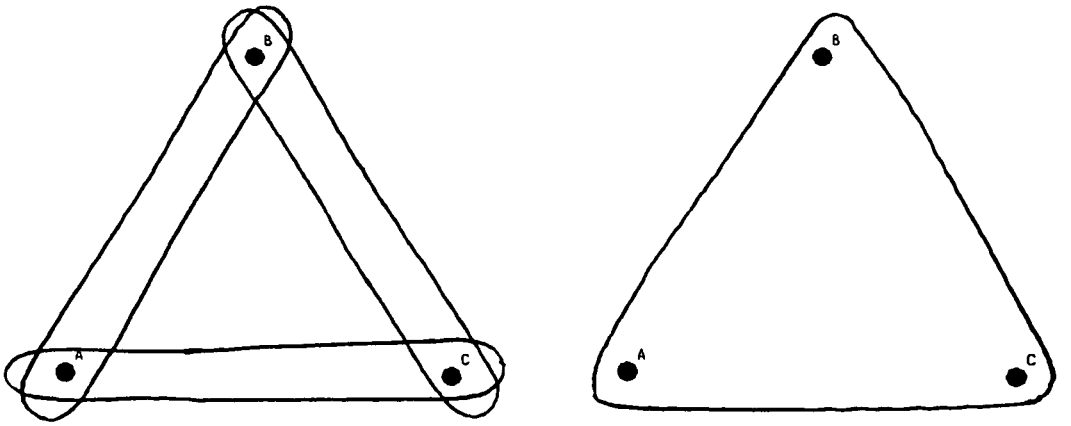

Figure 6-17: Graphs drawn using enclosure of vertices by boundaries to distinguish between the model with generating class {[AB] [AC] [BC]} (left) and the model with generating class {[ABC]} (right)

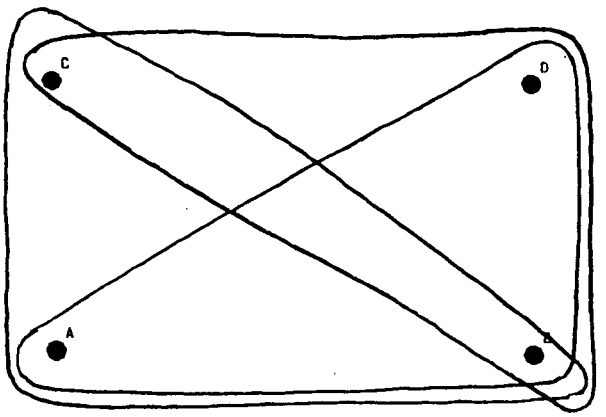

Figure 6-18: Graph drawn using enclosure of vertices by boundaries to represent the model with generating class {[ABC] [ABD] [BCD]}

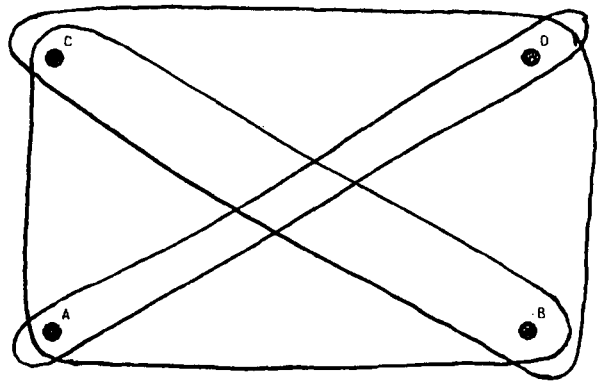

Figure 6-19: Graph drawn using enclosure of vertices by boundaries to represent the model with generating class {[ABC] [BCD] [AD]}

Thus it is unlikely that this approach could give a pleasing and legible representation for anything other than simple models involving a few interactions on a few variables.

6.2.7 Disjoint Interacting Areas

This approach is similar to the enclosure of vertices by boundaries described above (Section 6.2.6). However, the shapes used are always the characteristic line, triangle, or square, etc., corresponding to the order of the interaction to be represented, which are obtained by using links between vertices (see Section 6.2.2). These shapes or areas are drawn in close proximity to the vertices representing the variables involved but in such a way that no two areas have a coincident edge, even if they have a pair of variables/vertices in common. In this way, it is possible to identify the interactions in the generating class of the model represented according to the number and nature of the disjoint areas.

For example, in Figure 6-20 the model with generating class {[*AB*] [*AC*] [*BC*]} is represented by three separate lines and so it can be readily distinguished from the model with generating class {[*ABC*]} which is represented by a single triangle. Figure 6-21 shows the model with generating class {[*ABC*] [*ABD*] [*BCD*]}, and Figure 6-22 shows the model with generating class {[*ABC*] [*BCD*] [*AD*]}.

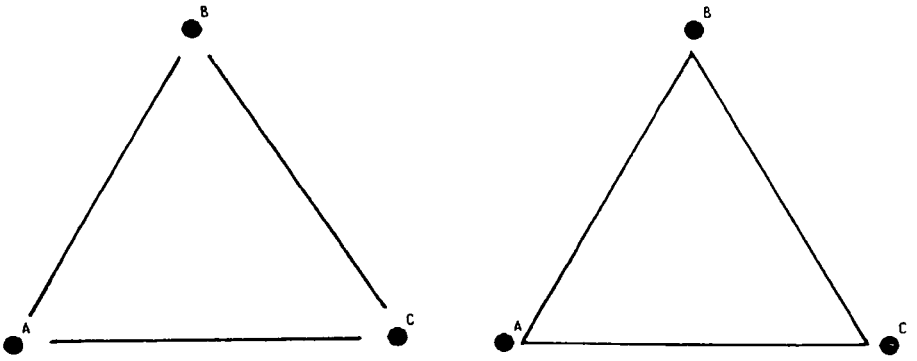

Figure 6-20: Graphs drawn using disjoint interacting areas to distinguish between the model with generating class {[AB] [AC] [BC]} (left) and the model with generating class {[ABC]} (right)

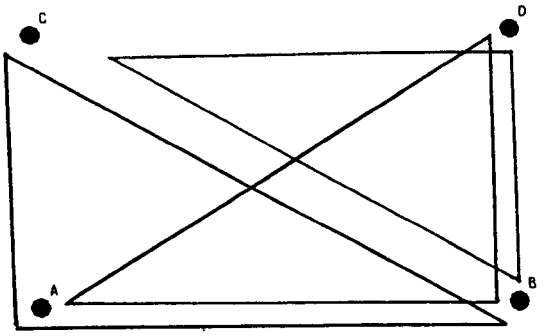

Figure 6-21: Graph drawn using disjoint interacting areas to represent the model with generating class {[ABC] [ABD] [BCD]}

Thus the use of disjoint interacting areas provides a unique representation of the generating class of a model since it avoids some of the ambiguity which can arise as a result of common edges and the indeterminability of non-interacting areas. However, if there are many vertices and many interactions in the generating class of the model, the representation can look quite messy. Also, because the end-points of the areas are disjoint, it may not be immediately apparent that several interactions have a common vertex/variable. Moreover, for a large number of variables it may not be possible to locate

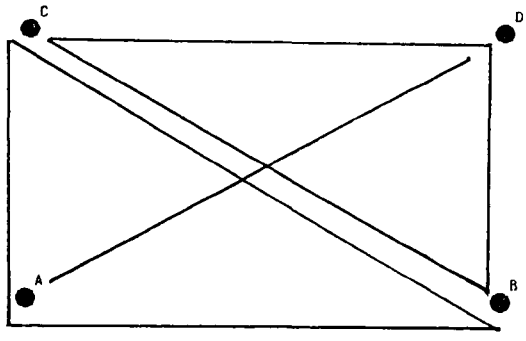

Figure 6-22: Graph drawn using disjoint interacting areas to represent the model with generating class {[ABC] [BCD] [AD]}

the vertices of the graph in such a way that they may be enclosed using the characteristic shape corresponding to the interaction to be represented.

6.2.8 Separate Display of Generating Class Elements

The model with generating class {[*AB*] [*AC*] [*BC*]} can always be distinguished from the model with generating class {[*ABC*]} if the elements of the generating class of the models are displayed separately, as shown in Figure 6-23. Similarly, the model with generating class {[*ABC*] [*ABD*] [*BCD*]} can be determined uniquely, as can the model with generating class {[*ABC*] [*BCD*] [*AD*]}, as shown in Figures 6-24 and 6-25 respectively.

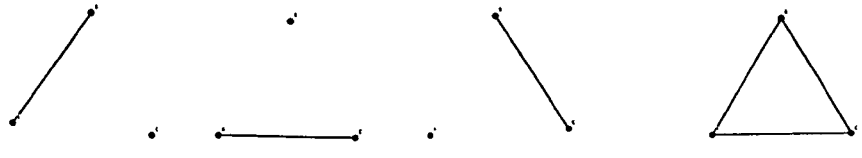

Figure 6-23: Graphs drawn using separate display of generating class elements to distinguish between the model with generating class {[AB] [AC] [BC]} (left) and the model with generating class {[ABC]} (right)

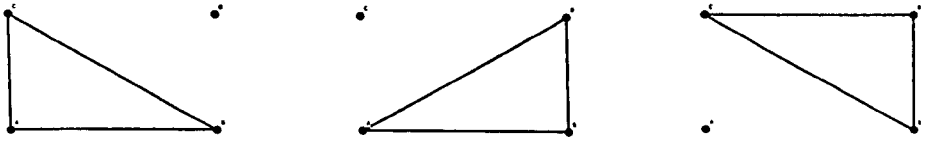

Figure 6-24: Graph drawn using separate display of generating class elements to represent the model with generating class {[ABC] [ABD] [BCD]}

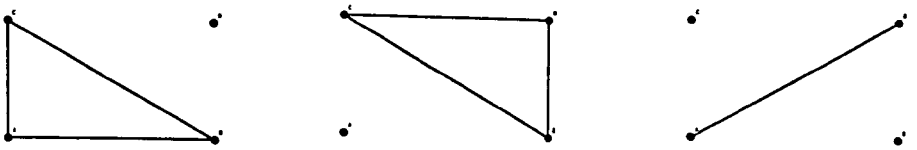

Figure 6-25: Graph drawn using separate display of generating class elements to represent the model with generating class {[ABC] [BCD] [AD]}

Because each element of the generating class of a model corresponds to a single interaction, the elements can be represented as simple links between vertices (as described in Section 6.2.2) without ambiguity. By representing the elements of the generating class separately, there can be no ambiguity and it is therefore possible to represent any model uniquely. This approach is considered again in Chapter 11.

6.2.9 Summary

Of all the techniques suggested above for the two-dimensional combination of points in order to represent the generating class of a fitted model, the least successful, due to the small number of models which can be represented without ambiguity, are the use of links between vertices (Section 6.2.2), the representation of interaction type by line style (Section 6.2.3), and the shading of areas corresponding to interacting or non-interacting areas (Section 6.2.5). Moderately successful techniques, which could potentially be used to represent any model without ambiguity, although they can result in quite messy representations for a large number of interactions, are the use of different line styles to

distinguish between different interactions (Section 6.2.4), the enclosure of vertices by boundaries (Section 6.2.6), and the use of disjoint interacting areas (Section 6.2.7).

The only wholly successful approach of those considered for the representation of the generating class of a fitted model would appear to be the separate display of the generating class elements (Section 6.2.8). However, it could be argued that this approach is really a three-dimensional combination of points since in viewing the separate two-dimensional displays of each of the elements, time is used as a third dimension. Moreover, it could be argued that it would be simpler to write down the generating class of the model than to display it graphically in this way. Nonetheless, I shall be returning to this approach in later chapters, and also to the use of links between vertices and the representation of interaction type by line style since these are intuitively attractive, and at least partially successful, approaches.

6.3 Venn Diagram Approach

Venn diagrams are commonly used to represent relationships between sets of objects graphically. Closed curves, such as circles, are used to represent the sets, and if the curves intersect each other, then the intersecting areas represent sub-sets of objects which have the characteristics of two or more of the original sets in common. Each characteristic for each set may be considered as being binary; ie. the objects either possess the characteristic (in which case they are included in the set) or they do not (in which case they are not included in the set), for each of the characteristics considered.

If a Venn diagram is constructed for which each circle corresponds to a variable, then the intersecting areas may be regarded as corresponding to interactions between these variables, as shown in Figure 6-26 for three variables A, B and C. [Note that this differs from the approach taken by Boniface (1995), described in Chapter 4, who uses the intersecting areas to illustrate the effect of adjustment for a covariate]. Any given model involving these three variables could be represented using this Venn diagram by shading the areas in the diagram which correspond to the main effects and interaction effects contained in the model. The area outside of the circles but enclosed by the boundary (which represents the Universe) can be regarded as representing the constant term, but can be left unshaded. Thus for the model with generating class $\{[ABC]\}$ all seven enclosed areas would be shaded, whereas for the model with generating class $\{[AB]\ [AC]\ [BC]\}$ the

central area corresponding to the three-way interaction *ABC* would be left unshaded. However, traditional Venn diagrams cannot be used to display satisfactorily more than four sets or variables.

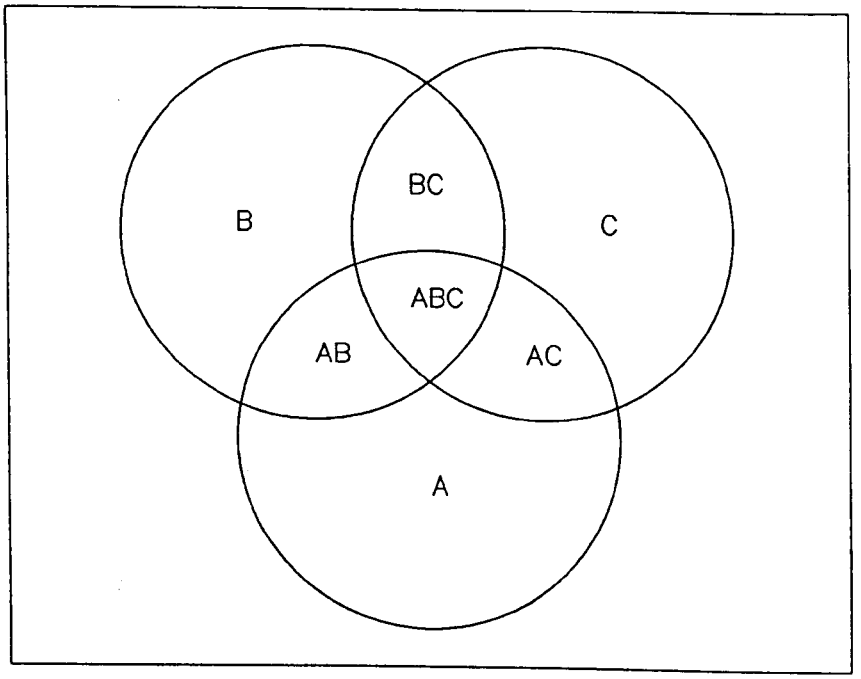

Figure 6-26: Traditional Venn diagram representation for three variables A, B, C

Edwards (1989) has designed a new kind of Venn diagram which can be applied to any (moderate) number of variables. The Universe is taken to be the surface of a sphere, and the boundary of the first set is taken to be the equator of the sphere (eg. above the equator if the object has the characteristic, or below the equator if not). The second and third sets are obtained by intersecting two meridians at right angles to the equator (eg. the 0–180 meridian and the 90–270 meridian). The fourth set will cut the equator four times, swinging 'north' and 'south' to give four semi-circles centred on the four intersections of the meridians with the equator. The boundary of the fifth set will then consist of eight such alternating semi-circles, centred on the eight equatorial crossings. The sixth set has sixteen such alternating semi-circles centred on the sixteen equatorial crossings, and so on *ad infinitum*.

This representation of the sets on the surface of the sphere can be projected onto a flat surface. Edwards chooses to project the sphere from one of the poles and to reinstate an artificial (rectangular) boundary to the Universe. An alternative projection suggested by Edwards, and which I have chosen to consider, involves drawing the equator as a straight line around which successive curves weave. As for the traditional Venn diagram, each combination of sets differs from its neighbour by the inclusion or exclusion of a single thing.

Suppose that a given model contains the N variables A, B, ..., N. The first of these (A) can be represented by the equator (ie. a straight line if drawn on a flat surface). Above the equator, say, will correspond to inclusion of A, and below will correspond to exclusion (equivalent to the constant parameter, ϕ, of the model). The first curve to be woven around the equator (resulting in one semicircle above, and another below, the equator) can be taken to represent the next variable B. Again, above this curve may correspond to inclusion of B, and below to exclusion. Thus, from 'top' to 'bottom' there are four areas corresponding to AB, A, B, and ϕ. The next variable, C, will be represented by the second curve to be woven round the equator, which will form four semi-circles, and result in areas corresponding to ABC, AB, AC, BC, A, B, C, and ϕ. Further curves can then be woven around the equator, until all N variables have been represented. Figure 6-27 shows the corresponding (Edwards-style) Venn diagram for three variables A, B, C.

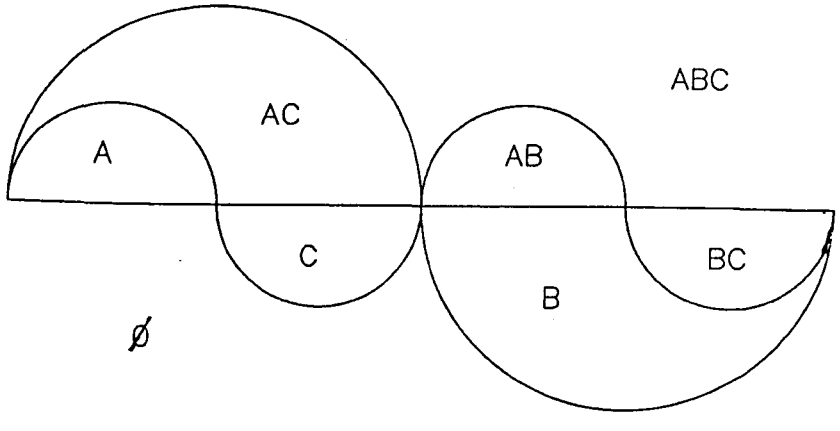

Figure 6-27: Edwards-style Venn diagram representation for three variables A, B, C

In moving from one area of the Venn diagram to another, a variable is either included or excluded. Which particular variable is dropped or added will correspond to the variable represented by the boundary which has been crossed. Thus certain parallels can be seen between the Venn diagram in Figure 6-27 and the corresponding Implication Diagram, shown in Figure 6-28. However, whereas the Implication Diagram conveys a sense of dropping variables by moving downwards, in the Venn diagram a variable may be dropped in a downwards *or* a horizontal direction (as can be seen in Figure 6-27), since it is sometimes necessary to move across three boundaries to drop just one particular variable (adding and then dropping, or dropping and then adding another variable at two of the boundaries).

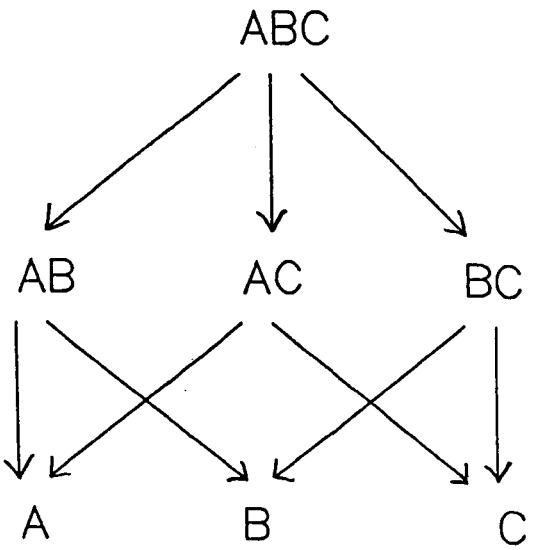

Figure 6-28: Implication Diagram corresponding to the Edwards-style Venn diagram representation for three variables A, B, C

As mentioned already, the Venn diagram approach can be used to represent the terms in a fitted model by shading in the areas corresponding to the terms contained in the model. However, there are a number of disadvantages in using Edwards-style Venn diagrams for many variables, as follows:

- A Venn diagram is not ideally suited to the representation of the generating class of a model by the shading of the areas corresponding to the elements of the generating

class, since it is not straight-forward to determine all the terms in the model implied by the elements of the generating class from the diagram (since, as stated above, it is sometimes necessary to move across more than one boundary to drop a particular variable).

- Because of the diminishing size of the semi-circles as more variables/curves are added, different areas of the Venn diagram which correspond to interactions of the same order, may be different sizes. This may have the effect of seemingly giving more weight to some interactions than to others of the same order.
- Because of the different sizes of the different areas representing the same order of interaction, quite similar models may not appear similar from their Venn diagram representation. For example, in Figure 6-29, three Venn diagrams are shown corresponding to the three similar models with generating classes {[*AB*] [*C*]}, {[*AC*] [*B*]}, and {[*BC*] [*A*]}, yet these models look quite different if the Venn diagram representations are compared.
- Similarly, models with generating classes consisting of all interactions of a particular order, such as {[*AB*] [*AC*] [*BC*]} or {[*A*] [*B*] [*C*]} would not appear as symmetric as they would in, for example, an Implication Diagram (see Figure 6-30).

These criticisms, although illustrated only on models with three variables, would also apply to models with four or more variables.

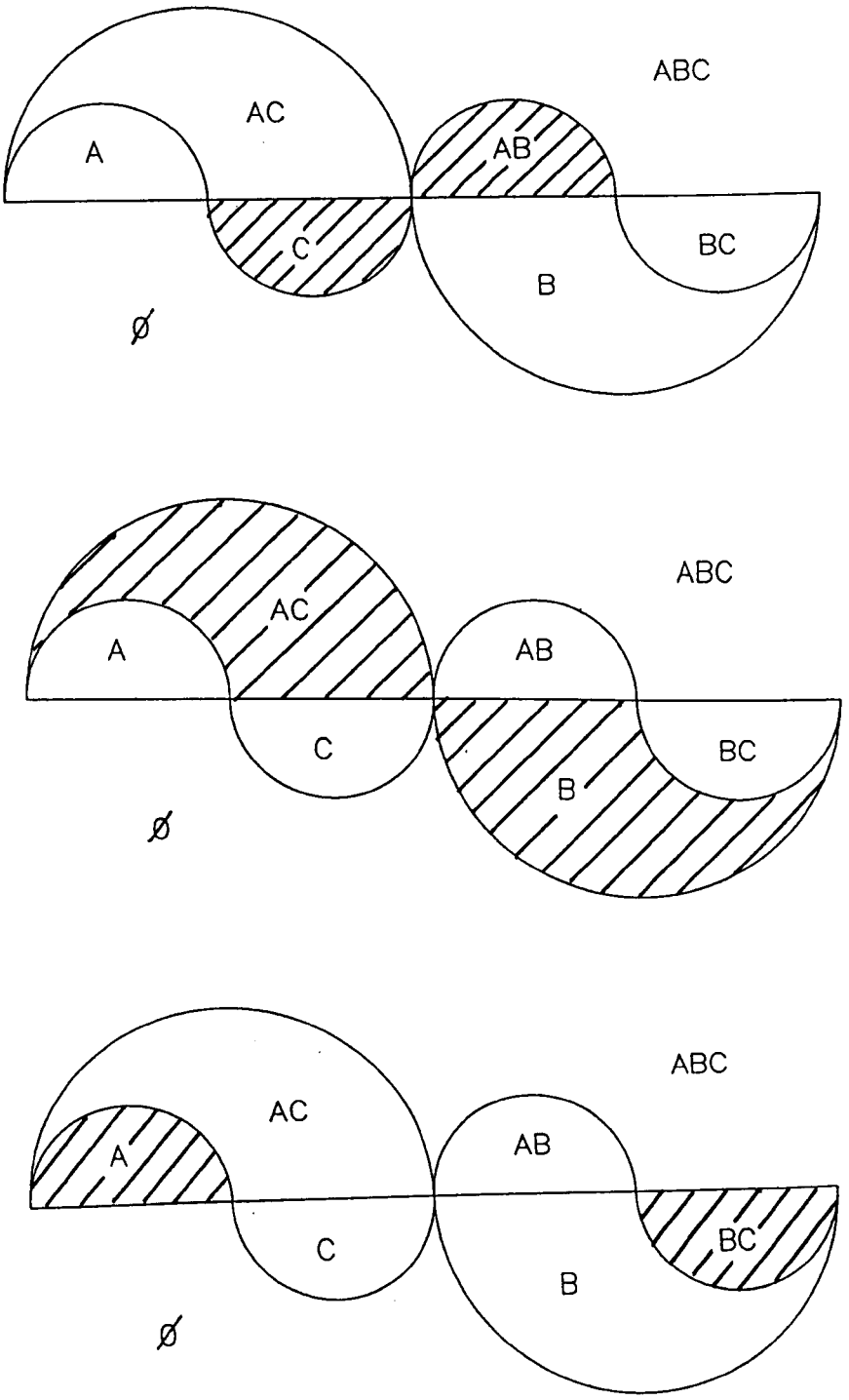

Figure 6-29: Venn diagram representations drawn for three models with generating classes {[AB] [C]} (top), {[AC] [B]} (middle), and {[BC] [A]} (bottom)

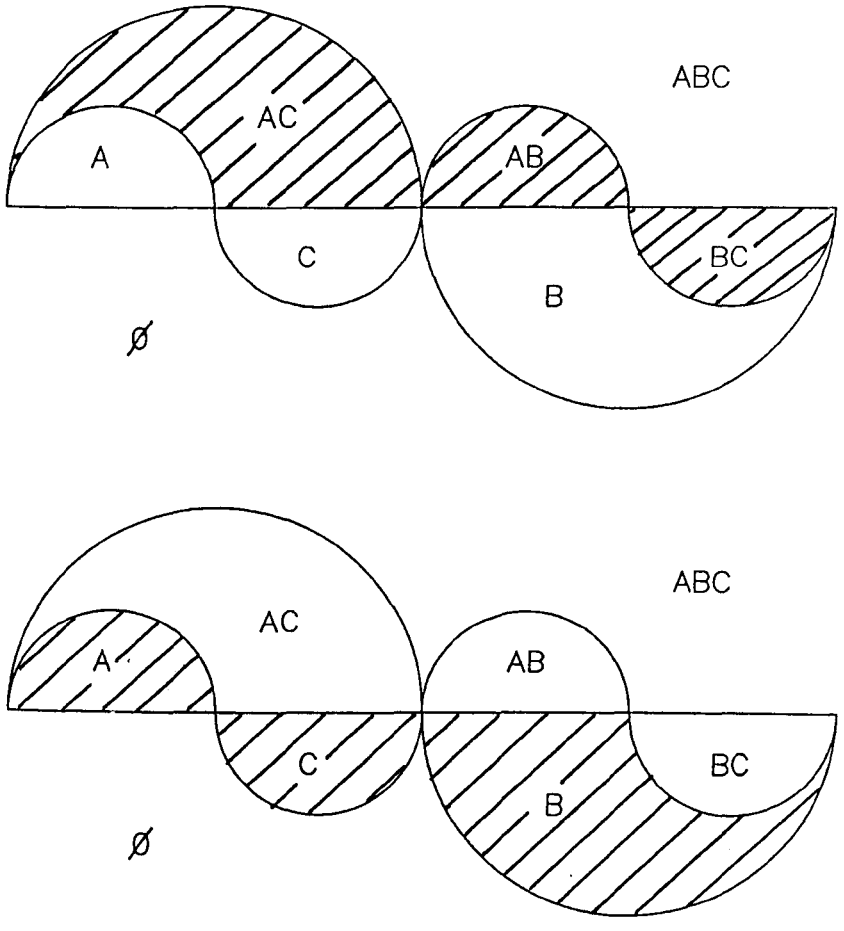

Figure 6-30: Venn diagram representations drawn for the models with generating classes {[AB] [AC] [BC]} (top) and {[A] [B] [C]} (bottom)

6.4 Topological–Magnitude Graphs

6.4.1 Introduction

Suppose that the hierarchical model $A+B+C+AB$ has been fitted to a particular data set in which the variables A, B, and C are discrete variables or factors, where A has two levels, B has three levels and C has two levels, and that the different levels for each of the

four effects in the model (ie. the three main effects and the interaction effect) take the values given in Figure 6-31.

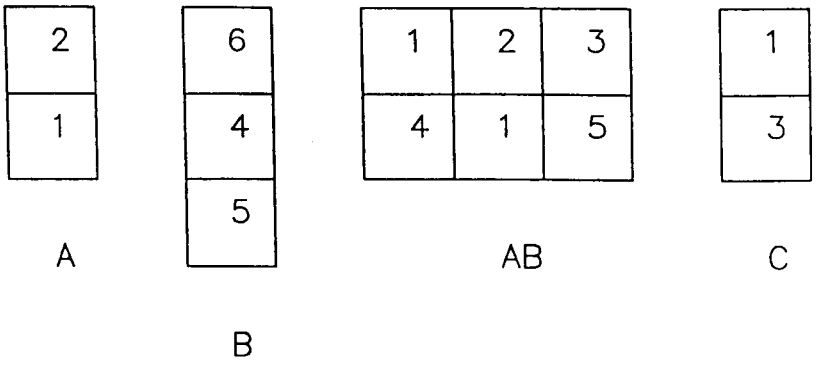

Figure 6-31: Tables showing the values taken by different levels of variables A, B, and C for the main effects and interaction effect in the model A+B+C+AB

The overall value for a particular combination of the levels of the variables (eg. $A=i$, $B=j$, $C=k$) can be determined by summing the appropriate cells in each of the four tables in Figure 6-31. For example:

$$(A=1, B=1, C=1) \rightarrow 2+6+1+1 = 10$$

$$(A=2, B=2, C=1) \rightarrow 1+4+1+1 = 7$$

etc..

In the following sections, two graphical techniques are developed for the representation of hierarchical models like the one just described. The particular features of the model which I hope to communicate using these two techniques are as follows:

- The presence or absence of the effects in the model, to be communicated by means of a *Topological Graph.*
- The magnitude of the effects present in the model, to be communicated by means of a *Magnitude Graph.*

In addition, the graphical techniques developed should serve to illustrate the additivity of the effects present in the model and should illustrate what is meant by 'interaction'; eg. that given the presence of an AB interaction, this implies that the value

for a particular level of variable *B* depends on the level of variable *A* which is considered, and vice versa.

In the following sections, I shall derive the construction of the Magnitude Graph (Section 6.4.2), from which the construction of the Topological Graph follows (Section 6.4.3). Some general criteria are presented (Section 6.4.4) for the construction and interpretation of Topological–Magnitude Graphs (a generic term used for the graph type) which involves the development of a novel algebraic system for the expression of the generating class, which is related to the two types of links used in the construction of the graphs. A number of examples are presented (Section 6.4.5) to illustrate the construction and interpretation of Topological–Magnitude Graphs using this algebraic expression of the generating class. However, as will be seen, two additional types of links are required in order to be able to represent the generating class of any given model successfully. The criteria for the construction of Topological–Magnitude Graphs are therefore revised before the usefulness of the Topological–Magnitude Graph approach is assessed (Section 6.4.6).

Following on from the development of the Topological–Magnitude Graphs, two extensions are suggested. The first of these (Section 6.4.7), called the *Proportional Graph*, is an alternative representation technique for the communication of the magnitudes of the effects present in a given model. The second extension (Section 6.4.8) is based on the Proportional Graph and can be used to communicate confidence intervals for the effects in the model.

6.4.2 Construction of the Magnitude Graph

Consider, for example, the model *A*+*B*, which has generating class {[*A*] [*B*]}. Since there is no interaction effect, the effect of variable *B* will be the same irrespective of the level of *A* which is considered, and *vice versa*. Suppose that each variable has two levels. The values of the effects of the levels could be presented, for each variable, in a two-box table of the type used in Figure 6-31. Given two two-box tables (one to represent each of the variables *A* and *B*), some technique is required to link the two tables which indicates that the value for the chosen level of variable *B* is independent of the level of variable *A* which is chosen. The suggested representation is a horizontal link as shown in Figure 6-32.

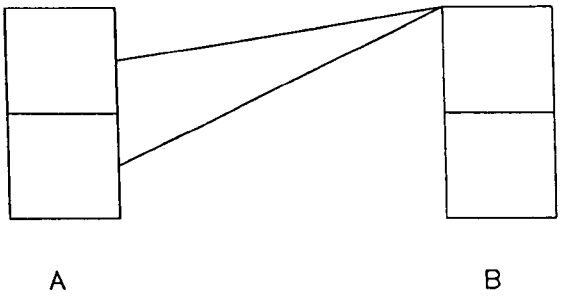

Figure 6-32: Use of horizontal link in Magnitude Graph drawn to represent the model with generating class {[A] [B]}

Consider now the model $A+B+AB$, which has generating class $\{[AB]\}$. Here the effect of the chosen level of variable B will depend on the chosen level of variable A, since there is an AB interaction. If each variable has two levels, two two-box tables could be drawn of the type used in Figure 6-31 to represent the main effects of A and B, and a four-box table could be drawn to represent the interaction effect AB, but having one table per effect contained in the model could lead to a very large number of tables for a model involving a large number of variables and/or a large number of interactions. Moreover, for three-way and higher-order interaction effects involving more than two variables, the construction of two-dimensional tables is not straight-forward. A simplified representation of the generating class of the fitted model is therefore desired, which reduces the number of tables required without leading to loss of information. It is suggested that only two sets of tables are needed in practice, one corresponding to A and the other to B, provided that some technique is used to link the two sets of tables which indicates that the value for the chosen level of variable B is dependent upon the level of A which is considered. The suggested representation is a diverging link as shown in Figure 6-33.

In the Magnitude Graph for the model with generating class $\{[A]\ [B]\}$, presented in Figure 6-32, the top-most box in the table representing variable A will contain the value of the main effect corresponding to the first level of A and the bottom-most box will contain the value of the main effect corresponding to the second level of A. Similarly, the top-most box in the table representing variable B will contain the value of the main effect

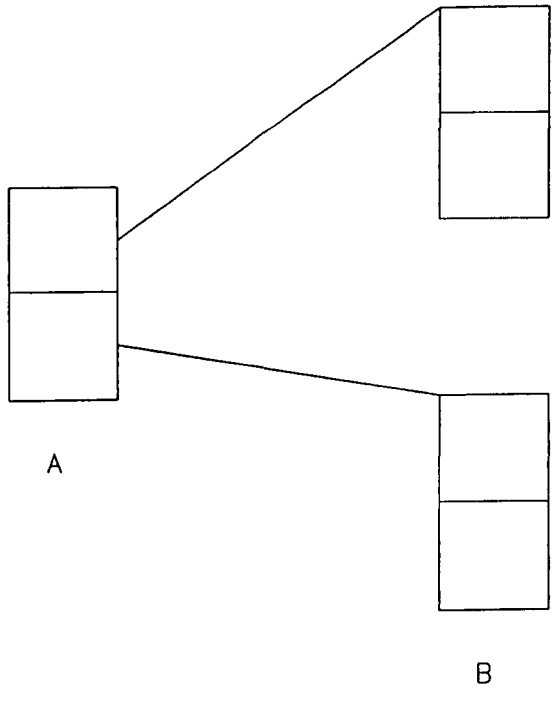

Figure 6-33: Use of diverging link in Magnitude Graph drawn to represent the model with generating class {[AB]}

corresponding to the first level of *B* and the bottom-most box will contain the value corresponding to the second level of *B*.

In the Magnitude Graph for the model with generating class {[*AB*]}, presented in Figure 6-33, the top-most box in the table representing variable *A* will again contain the value of the main effect corresponding to the first level of *A*, and the bottom-most box will contain the value corresponding to the second level of *A*. However, the top-most box in the top-most table representing variable *B* will contain the value of the main effect corresponding to the first level of *B plus* the value of the interaction effect corresponding to the interaction of the first level of *A* with the first level of *B*. The top-most box in the bottom-most table representing variable *B* will also contain the value of the main effect corresponding to the first level of *B*, *plus* the value of the interaction effect corresponding to the interaction of the second level of *A* with the first level of *B*. Similarly, the bottom-most boxes will contain the value of the main effect corresponding to the second level of

B plus the value of the interaction effect corresponding to the interaction of the second level of *B* with the first or second level of *A*, as appropriate.

The Magnitude Graph representation of the model with generating class {[*AB*]} could be used to represent the model with generating class {[*A*] [*B*]}, but because there are no interaction effects in the model corresponding to {[*A*] [*B*]}, the top-most and bottom-most tables for *B* would contain the same entries for each level of *B*. By using the two different forms of links, the suggested Magnitude Graph representation of {[*A*] [*B*]} makes it clear that the effect of *B* is the same irrespective of the level of *A* which is chosen, and the suggested Magnitude Graph representation of {[*AB*]} makes it clear that the effect of *B* will differ according to the level of *A* which is chosen.

The two graphs drawn and described above (in Figures 6-32 and 6-33) could have been constructed by consideration of variable *B* first, and variable *A* second. In this way, the graphs would have been constructed as shown in Figures 6-34 and 6-35. Comparing the Magnitude Graph shown in Figure 6-32 with that shown in Figure 6-34, and comparing the Magnitude Graph shown in Figure 6-33 with that shown in Figure 6-35, it would appear that a failing of the Magnitude Graph representation technique is that, since the Graphs are constructed in a left-to-right manner, there is a loss of symmetry between the variables. For example, it appears that *A*+*B* is not the same as *B*+*A*, even though the tables for *A* in Figures 6-32 and 6-34 will contain the same values, as will the tables for *B*. Similarly, it appears that *A*+*B*+*AB* is not the same as *B*+*A*+*BA* — although it is clear that the effects of the levels of one variable depend on which level of the other variable is chosen, the two sets of tables will have different values in each graph. In Figure 6-33, the left-hand table will contain values corresponding to the *A* main effect, and the right-hand table will contain values corresponding to the *B* main effect plus the *AB* interaction effect. In Figure 6-35, the left-hand table will contain values corresponding to the *B* main effect, and the right-hand table will contain values corresponding to the *A* main effect plus the *AB* interaction effect. If these graphs were to be used to communicate these models to somebody who is statistically naive, then the symmetry of the relationships would have to be made explicit.

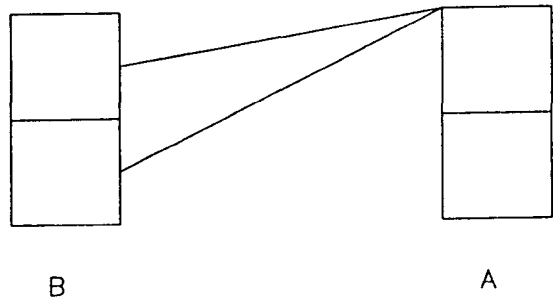

Figure 6-34: Alternative Magnitude Graph drawn to represent the model with generating class {[A] [B]}

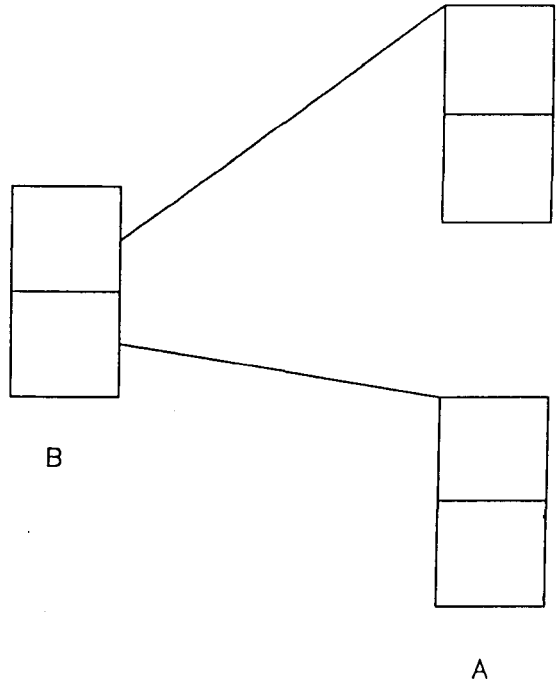

Figure 6-35: Alternative Magnitude Graph drawn to represent the model with generating class {[AB]}

6.4.3 Construction of the Topological Graph

The two basic forms of the Magnitude Graph described in the previous section are useful for the communication of the numerical values of the effects present in a given model, since the appropriate numbers can simply be written in the appropriate boxes. However, the user may wish simply to communicate the effects implied by the generating class which are present in the model without attaching numerical values to these effects. In other words, there is a need for a graphical representation which communicates the topology of the effects, but not their magnitude.

The Topological Graph representation can be derived from the Magnitude Graph representation simply by removing the tables of boxes from the Magnitude Graph (which would otherwise be unnecessary clutter), but retaining the links between the vertices which, in replacing the tables, correspond to the variables in the model. For example, the Topological Graph representation for the model with generating class {[*A*] [*B*]} will use the horizontal link as shown in Figure 6-36 (compare this with the Magnitude Graph in Figure 6-32 in the previous section), and the Topological Graph representation for the model with generating class {[*AB*]} will use the diverging link as shown in Figure 6-37 (compare this with the Magnitude Graph in Figure 6-33 in the previous section).

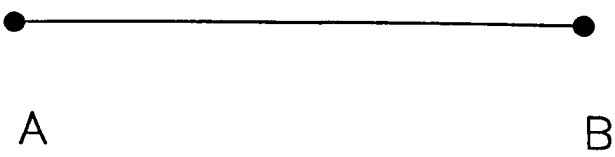

Figure 6-36: Use of horizontal link in Topological Graph drawn to represent the model with generating class {[A] [B]}

Although the choices of links used in both the Magnitude Graph and Topological Graph representations are intended to correspond in a particular way to the effects in the model which they represent, in practice the choice of links is arbitrary and other types of links could be adopted provided that the conventions adopted for the meaning of the links were applied consistently wherever the links were used.

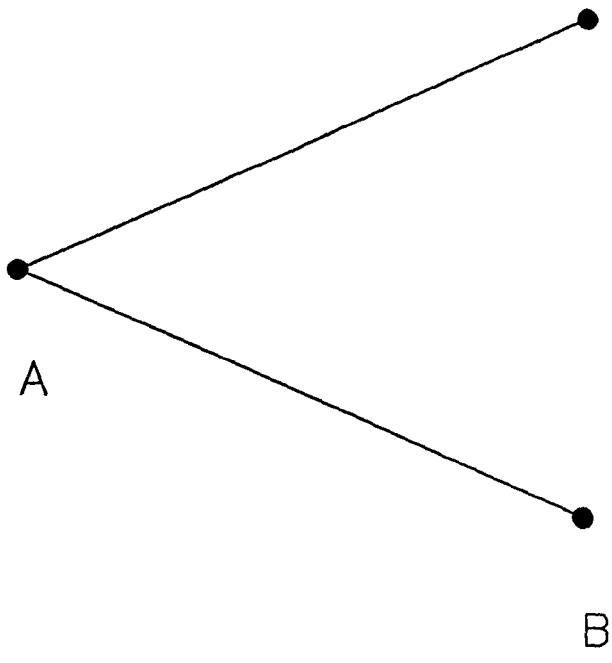

Figure 6-37: Use of diverging link in Topological Graph drawn to represent the model with generating class {[AB]}

6.4.4 Criteria for the Construction of Topological–Magnitude Graphs

Given the generating class of a model, the Topological and Magnitude Graphs corresponding to the model may be derived using the two types of links presented in the preceding sections (Sections 6.4.2 and 6.4.3) by following the criteria for construction presented in this section. Conversely, for a given Magnitude Graph or Topological Graph, the generating class of the model represented can be determined by use of the same criteria. As has already been seen, the Topological Graph can be determined from a given Magnitude Graph, but it would not be possible to derive the Magnitude Graph from the Topological Graph without additional information about the number of levels of the variables and the magnitudes of the effects.

Some conventions are necessary in order to derive the criteria for the construction of the Topological–Magnitude Graph corresponding to a given generating class. As has

already been indicated, these conventions are also required for the interpretation of the Graphs. The conventions adopted involve some simple 'algebra', and are best described by means of some examples, as presented below.

Consider the model with generating class {[*AB*] [*ACD*]}. The generating class is to be re-expressed as *AB*+*ACD* (where the '+' has a subtly different usage than in the full expression of the model: *A*+*B*+*C*+*D*+*AB*+*AC*+*AD*+*CD*+*ACD*). This expression of the generating class is now factorised, as in conventional algebra, to give *A*×(*B*+*CD*), where the '×' represents multiplication with the elements in the brackets (this is not usually made explicit in conventional algebra). If the factorisation of the variables is carried to the extreme, the expression *A*×(*B*+*C*×(*D*)) is obtained.

The two types of links used in the Topological–Magnitude Graphs described in the preceding sections corresponded to the model with generating class {[*A*] [*B*]} (Figures 6-32, 6-34 and 6-36) and to the model with generating class {[*AB*]} (Figures 6-33, 6-35 and 6-37). In algebraic terms, these models can be re-expressed as *A*+*B* and *A*×(*B*) respectively.

To construct the Topological–Magnitude Graph corresponding to a given generating class, the algebraic expression is considered and, reading from left to right, a vertex (or vertices) is drawn in the graph whenever a variable is encountered, and a horizontal link corresponding to '+' or a diverging link corresponding to '×' is drawn in as appropriate. Following these rules for the model with generating class {[*AB*] [*ACD*]}, which, as has been shown, can be expressed algebraically as *A*×(*B*+*C*×(*D*)), the Topological Graph displayed in Figure 6-38 is obtained.

The derivation of the generating class of the model represented by a given Topological (or Magnitude) Graph is carried out in the same manner – reading from left to right, the variables are written down, together with a '+' or a '×', according to the vertices and types of link (horizontal or diverging) encountered. Provided it is remembered that a '×' always precedes a set of brackets, the correct algebraic expression of the generating class is obtained from which the generating class itself is readily determined.

Given the equivalence between the two types of links used in Topological and Magnitude Graphs, the Magnitude Graph for the model with generating class {[*AB*] [*ACD*]} would be constructed in the same way as the Topological Graph presented in Figure 6-38. Suppose, for instance, that each of the four variables in the above example

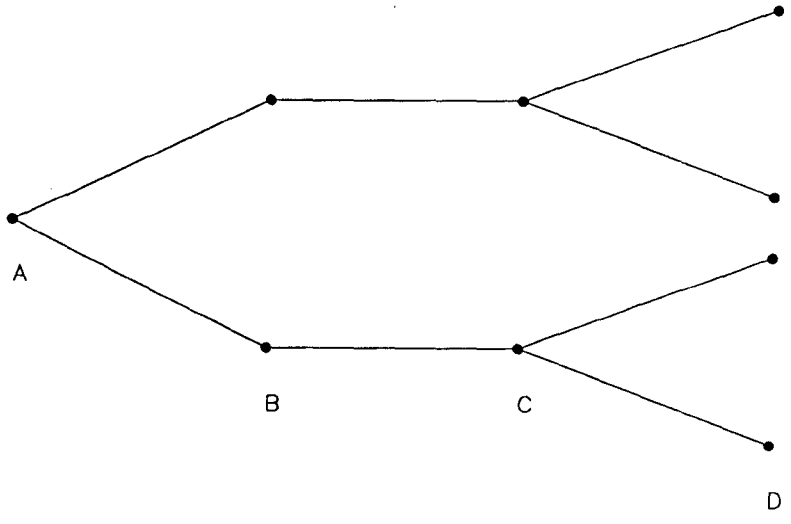

Figure 6-38: Topological Graph drawn to represent the model with generating class {[AB] [ACD]}

has two levels; the Magnitude Graph for this model would then be as given in Figure 6-39. From this graph, it can be seen, given the conventions for the 'meaning' of the links used and the boxes, that the values of A will depend on the levels of B (ie. there is an AB interaction), that the values of C will depend on the levels of A but not on the levels of B (ie. there is an AC interaction, but not a BC interaction), and that the values of D will depend on the levels of C, which itself depends on the levels of A, but the values of D do not depend on the levels of B (ie. there is an ACD interaction). Hence it can be confirmed that the Magnitude Graph does indeed represent the model with generating class $\{[AB][ACD]\}$.

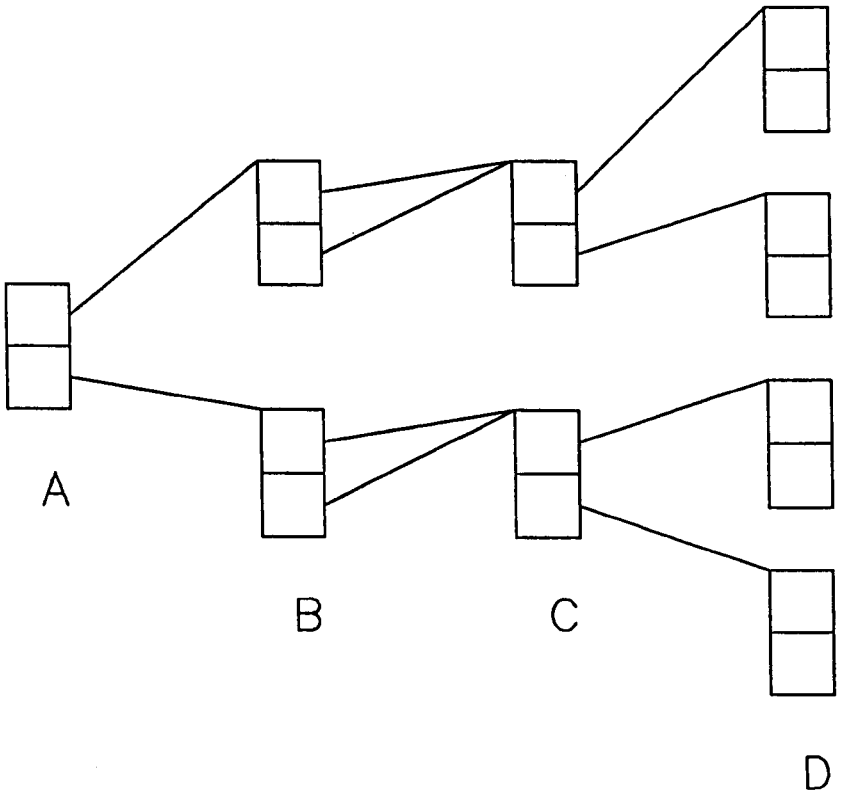

Figure 6-39: Magnitude Graph drawn to represent the model with generating class {[AB] [ACD]}

6.4.5 Some Example Topological–Magnitude Graphs

For all of the models considered in this section it is assumed, in order to simplify the construction of the Magnitude Graph, that the variables involved each have just two levels. In the more general case where a variable A has i levels, a variable B j levels, a variable C k levels, etc., then boxes drawn for the main effects of A, B, and C would have i, j, and k levels respectively. For an AB interaction re-expressed as $A\times(B)$, i boxes would be drawn corresponding to B, each with j levels. Similarly, for a BC interaction re-expressed as $B\times(C)$, j boxes would be drawn corresponding to C, each with k levels; and so on.

Consider the model with generating class {[*A*] [*BC*]}. The algebraic re-expression of this model would be $A+B\times(C)$. The corresponding Topological and Magnitude Graphs would be as presented in Figure 6-40.

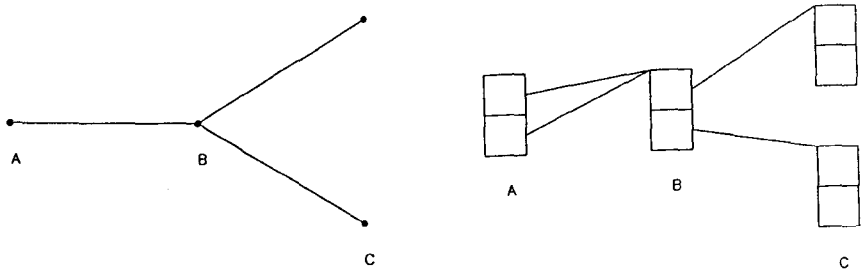

Figure 6-40: Topological and Magnitude Graphs drawn to represent the model with generating class {[A] [BC]}

Consider the model with generating class {[*AB*] [*AC*]}. The algebraic re-expression of this model would be $A\times(B+C)$. The corresponding Topological and Magnitude Graphs would be as presented in Figure 6-41.

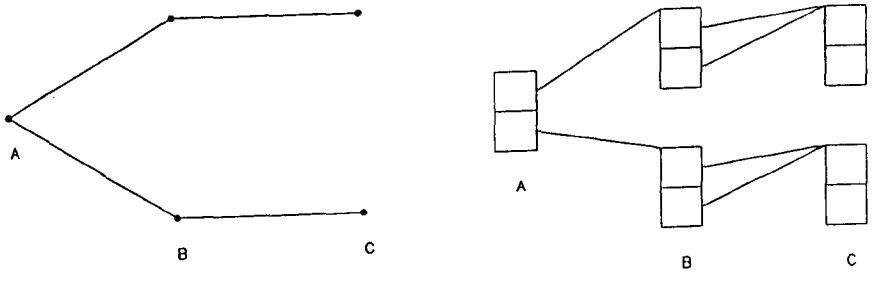

Figure 6-41: Topological and Magnitude Graphs drawn to represent the model with generating class {[AB] [AC]}

Consider the model with generating class {[*ABC*]}. The algebraic re-expression of this model would be $A\times(B\times(C))$. The corresponding Topological and Magnitude Graphs would be as presented in Figure 6-42.

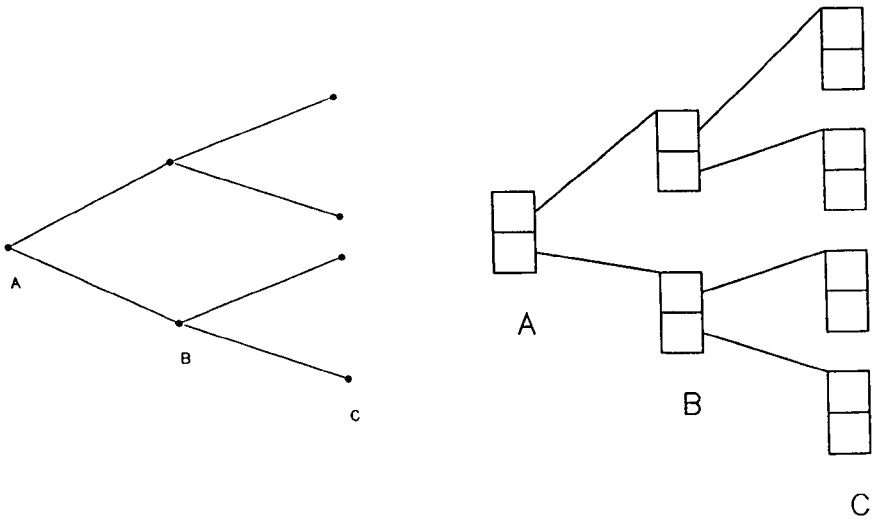

Figure 6-42: Topological and Magnitude Graphs drawn to represent the model with generating class {[ABC]}

Consider again the model with generating class {[*A*] [*BC*]} shown in Figure 6-40. An alternative and equivalent algebraic expression of this model would be $B\times(C)+A$. If the Topological and Magnitude Graphs implied by this expression were to be constructed following the criteria presented in the previous section (Section 6.4.4) relating the algebraic expression to the types of links used, the graphs obtained would be as presented in Figure 6-43. However, if these graphs were to be interpreted using these same criteria but relating the types of links to the algebraic components, the expression $B\times(C+A)$ would be derived, which corresponds to the model with generating class {[*AB*] [*BC*]}, which is not the model I wished to represent.

A problem arises because, when interpreting the graphs algebraically, it is not apparent when any brackets which have been opened by the diverging link corresponding to '×' should be closed. Until now, an assumption has been made when deriving the algebraic expression corresponding to a given graph that all brackets remain open until the far right-hand side of the graph is reached, when they are closed. Provided the algebraic expression of the generating class of the model can be written in such a way that all brackets close at the far right-hand side, the corresponding graph can be constructed

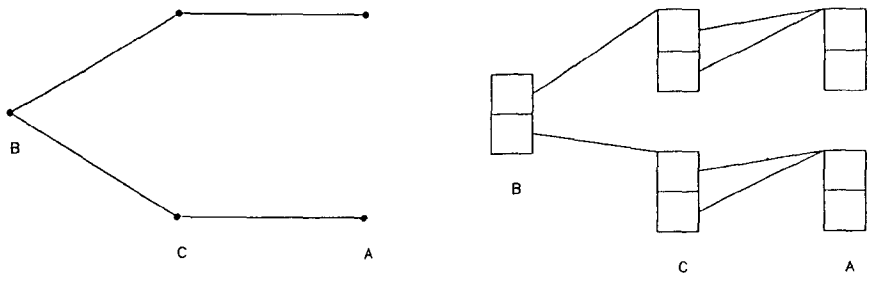

Figure 6-43: Alternative Topological and Magnitude Graphs drawn to represent the model with generating class {[A] [BC]}

and interpreted in the manner described. However, there are some generating classes which it is impossible to express algebraically in this way.

Consider, for example, the model with generating class {[*AB*] [*CD*]}. The algebraic expression of this model is of the form $A\times(B)+C\times(D)$, and it is impossible to rearrange this expression so that both sets of brackets close on the far right-hand side. It would, therefore, be impossible to construct the Topological or Magnitude Graph corresponding to this model in the usual manner such that it would be interpreted correctly. The adoption of another type of link corresponding to the closure of the preceding brackets followed by a '+' (ie. corresponding to ')+' in the algebraic expression) is therefore required. A converging link is proposed, as illustrated in Figure 6-44.

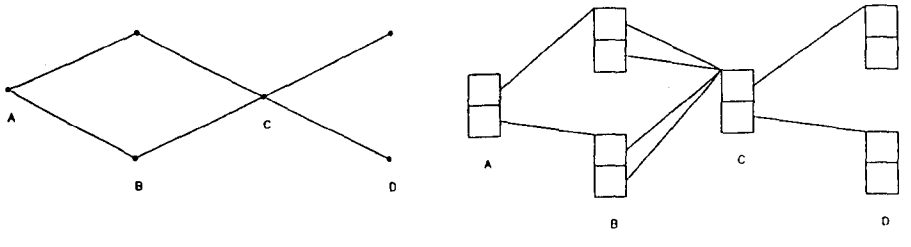

Figure 6-44: Use of converging link in Topological and Magnitude Graphs drawn to represent the model with generating class {[AB] [CD]}

In the Magnitude Graph presented in Figure 6-44, it is apparent that the value of B depends on the level of A (ie. that there is an AB interaction), and that the value of D depends on the level of C (ie. that there is a CD interaction), but the value of C does not depend on the level of A or B. Thus the graph does represent the model with generating class $\{[AB]\ [CD]\}$. Alternatively, the links could be interpreted according to the criteria relating the links to the algebraic expression in order to derive the expression $A\times(B)+C\times(D)$, from which the correct generating class is readily derived.

Up until now, there has only been one vertex or column of vertices in the Topological Graphs or one box or column of boxes in the Magnitude Graphs corresponding to each variable in the model represented, since each variable has appeared only once in the algebraic re-expression of the model.

Consider the model with generating class $\{[AB]\ [BC]\ [CD]\}$. In algebraic terms this model can be re-expressed as $B\times(A+C)+C\times(D)$ or $C\times(B+D)+A\times(B)$. In either case, it is impossible to avoid duplication of one of the variables. In constructing the Topological and Magnitude Graphs corresponding to this model, two separate components could be drawn, one corresponding to $B\times(A+C)$ and the other corresponding to $C\times(D)$, for example, or the converging link described above could be used. However, both of these approaches would involve duplication of the variable in the Topological–Magnitude Graph. I therefore propose the adoption of a fourth type of link to correspond to the duplication of a variable. Use of this link requires the algebraic expression of the model to be written such that the repeated variable appears on both sides of a ')+' relationship. A crossing link is proposed, as illustrated in Figure 6-45 which shows the Topological and Magnitude Graphs corresponding to the expression $B\times(A+C)+C\times(D)$.

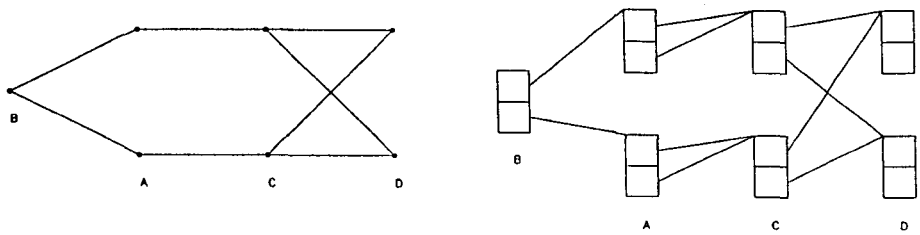

Figure 6-45: Use of crossing in Topological and Magnitude Graphs drawn to represent the model with generating class {[AB] [BC] [CD]}

From the Magnitude Graph presented in Figure 6-45 it can be seen that the value of A depends on the level of B (ie. that there is an AB interaction) and that the level of C depends on the level of B but not on the level of A (ie. that there is a BC interaction). Also, the value of D depends on the level of C, but not on the level of A or B (ie. there is a CD interaction). Hence the generating class of the model represented can be determined correctly from the Magnitude Graph or by application of the criteria for interpreting the links to either the Topological or the Magnitude Graph.

There are some models, however, for which two or more variables will be duplicated in the algebraic expression of the generating class. Consider, for example, the model with generating class {[AB] [AC] [BC]}. This can be represented algebraically as $A\times(B+C)+B\times(C)$, $B\times(A+C)+A\times(C)$, or $C\times(A+B)+A\times(B)$. In the construction of the Topological and Magnitude Graphs, the crossing link can again be used to correspond to the ')+' relationship, but it is impossible to avoid repeating one of the two duplicated variables. For example, the Topological and Magnitude Graph representation corresponding to $A\times(C+B)+B\times(C)$ will be as shown in Figure 6-46. The question of what values will appear in the boxes of a Magnitude Graph in which one of the variables is repeated is dealt with in Section 6.4.6.

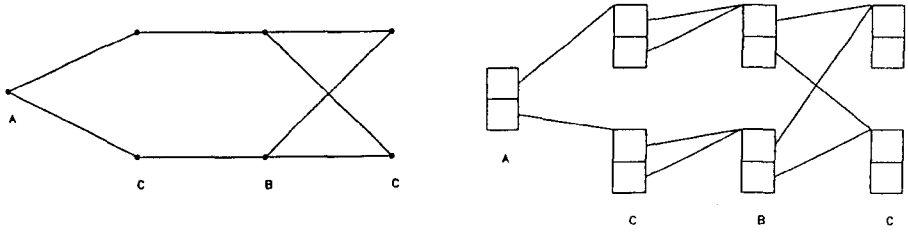

Figure 6-46: Use of crossing link and duplication of variable in Topological and Magnitude Graphs drawn to represent the model with generating class {[AB] [AC] [BC]}

In interpreting the Magnitude Graph shown in Figure 6-46 from left to right, it can be seen that the value of B depends on the level of A (ie. that there is an AB interaction) and that the value of C depends on A but not on B (ie. that there is an AC interaction). However, the crossing link and the duplication of B indicates that the value of B does depend on the level of C (ie. that there is a BC interaction). Hence the generating class of the model represented can be determined correctly from the Magnitude Graph or by

application of the criteria for interpreting the links to either the Topological or Magnitude Graph.

6.4.6 Revised Criteria for the Construction of Topological–Magnitude Graphs

In the preceding section (Section 6.4.5), two new links were introduced to give a total of four link types of use in the construction of Topological–Magnitude Graphs. By the application of criteria relating these links to the symbols used in the algebraic re-expression of the generating class of the model, it is now possible to construct and correctly interpret the Topological–Magnitude Graph for any given model. These links are listed in Table 6-1 together with their representation in Topological and Magnitude Graphs and their algebraic equivalent. Any brackets which remain open are automatically closed at the far right-hand side of the algebraic expression.

The generating class of a model can be represented using only the horizontal and diverging types of links if the most parsimonious algebraic expression of the generating class does not involve duplication of any the variables and all brackets close on the far right-hand side. This is the case if there is only one element in the generating class — for example, as in $\{[ABC]\}$; if the elements of the generating class have no variables in common and there is at most one interaction effect — for example, as in $\{[A]\ [BC]\}$; or if two or more of the elements of the generating class have just one variable in common — for example, as in $\{[AB]\ [BC]\}$ or $\{[AB]\ [BC]\ [BD]\}$.

The generating class of a model can be represented using only the horizontal, diverging and converging types of links if the most parsimonious algebraic expression of the generating class does not involve duplication of any of the variables but not all the brackets close on the far right-hand side. This is the case if the elements of the generating class have no variables in common and there are at least two interaction effects — for example, as in $\{[AB]\ [CD]\}$.

The generating class of a model may be represented using all four types of lines (horizontal, diverging, converging, and crossing) if the most parsimonious algebraic expression of the generating class involves duplication of at least two different variables — for example, as in $\{[AB]\ [AC]\ [BC]\}$ or $\{[ABC]\ [CD]\ [AE]\}$. In the former case, every pair of elements has a least one variable in common and it will be impossible to avoid constructing the Topological–Magnitude Graph without duplication of at least one of the variables. In the latter case, there is at least one pair of elements without a variable in

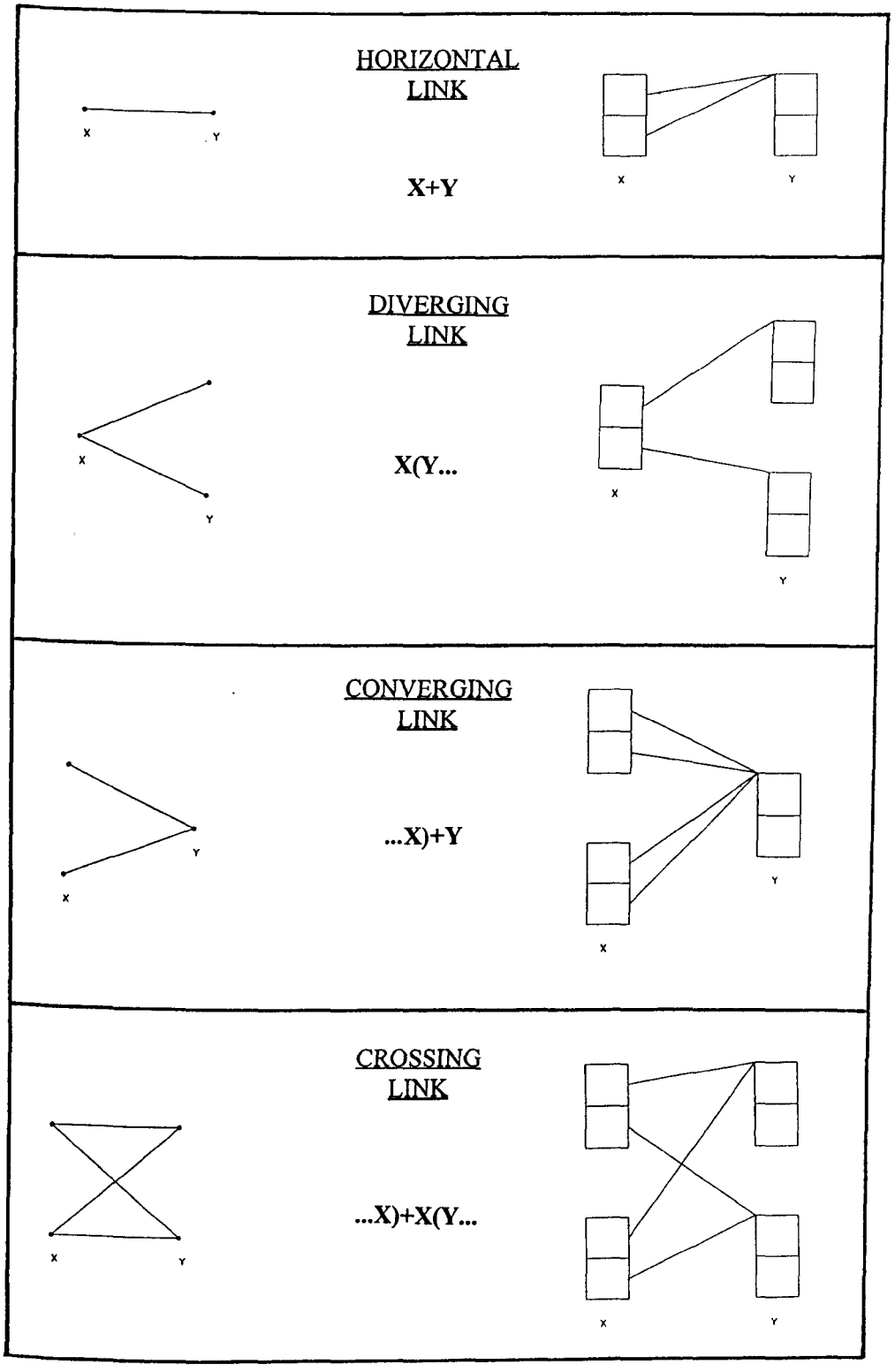

Table 6-1: List of links used in Topological-Magnitude Graphs, together with their algebraic interpretation

common and it is again impossible to construct the Topological–Magnitude Graph without duplication of any of the variables.

Consider again the hierarchical model $A+B+C+AB$ having generating class $\{[AB][C]\}$ which was first considered in Section 6.4.1 and for which the different effects in the model take the values presented in Figure 6-31 according to the levels of the variables considered. The Magnitude Graph corresponding to this model, with the appropriate values placed in the appropriate boxes, is as presented in Figure 6-47.

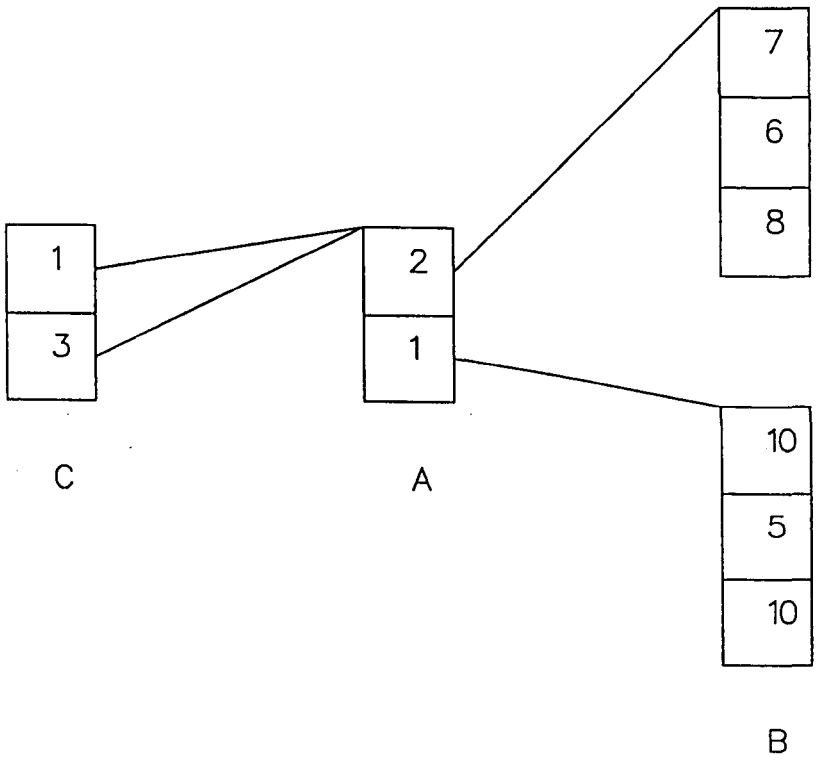

Figure 6-47: Magnitude Graph drawn to represent the model with generating class {[AB][C]} with example values

The boxes representing variables C and A contain the values corresponding to the main effects of C and A, whereas the boxes representing variable B contains the values corresponding to the main effect of B *plus* the values corresponding to the interaction effect of AB. For a more complex Magnitude Graph, in which one (or more) of the variables is duplicated, only one of the occurrences would incorporate the values corresponding to the main effect of that variable with an interaction effect; the other

occurrence(s) of the variable would contain values corresponding to the appropriate interaction effect only.

The compounding of the values in the Magnitude Graph may be regarded as a disadvantage to its use. One solution would be to separate the values within the boxes, according to the effects which they correspond to, but this could result in a lot of clutter which could make the graph difficult to read.

An alternative representation of the magnitudes of the effects in the model involves replacing the terms in the Implication Diagram (see Section 6.1.3) by n-box or $n \times m$-box tables containing the values corresponding to the main effects and interaction effects. This is illustrated in Figure 6-48 for the example model with generating class $\{[AB]\ [C]\}$ and values as given in Figure 6-31. Note that the arrowheads have been omitted from this representation owing to the change in usage of the Implication Diagram. Normally an Implication Diagram would be constructed as an aid to determining which terms are present in a model having a given generating class and the arrowheads are used to highlight the terms in the model which are 'implied' by other terms in the model. In the usage of the Implication Diagram illustrated in Figure 6-48, the main concern is with the display of the numerical values corresponding to each term in the model — it is useful to retain a link between effects having variables in common in order to highlight the existence of higher-order interaction effects, but the direction of the link is not meaningful in this context. However, as has already been mentioned, for interactions involving three or more variables, it would be difficult to construct multi-dimensional tables.

A more attractive solution would involve separating out the main effects and interaction effects. To achieve this, I propose a new type of graph, based on the Topological–Magnitude Graph, called the 'Proportional Graph', which is described below.

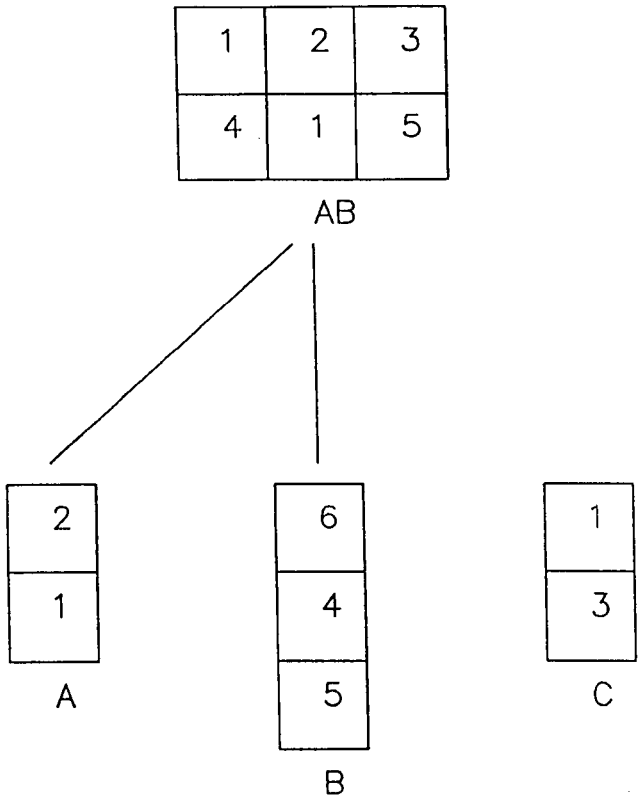

Figure 6-48: Implication Diagram drawn to represent the model with generating class {[AB] [C]} with example values

6.4.7 The Proportional Graph: An Alternative Representation for Magnitudes

The Proportional Graph takes the same basic form as the Topological Graph except that only the horizontal and diverging links are needed, irrespective of the generating class of the model, and these links are drawn corresponding to the levels of each variable and to the levels of any interaction effects. Diverging (ie. sloping) lines correspond to the main effects and horizontal lines correspond to the interaction effects, and the length of each line is proportional to the value of the effect it represents. Thus the relative sizes of the effects in the model can be determined; lines of the same length will correspond to a constant (non-zero) effect, and lines omitted from the graph (ie. having zero length) will correspond to zero effects.

For example, the Proportional Graph corresponding to the model with generating class {[*AB*] [*C*]}, having the Magnitude Graph presented in Figure 6-47, is presented in Figure 6-49. The lines have been drawn to scale such that 0.75cm corresponds to a value of 1.

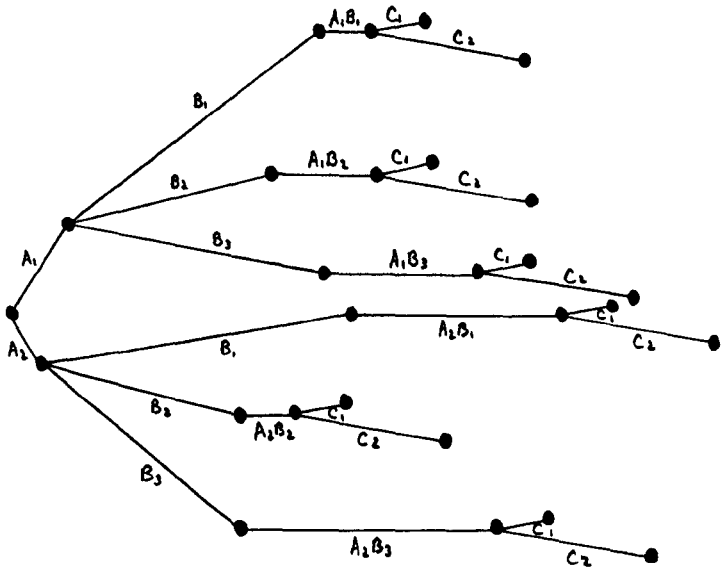

Figure 6-49: Proportional Graph drawn to represent the model with generating class {[AB] [C]}

To interpret the Proportional Graph, a series of consecutive lines corresponding to the levels of interest of the variables are followed from left to right and the total length of the path followed will be proportional to the overall value for the combination of levels of the variables considered. For example, since the Proportional Graph in Figure 6-49 has been drawn to scale, it can be determined that

$$A_1+B_1+A_1B_1+C_1 = 10$$

$$A_2+B_2+A_2B_2+C_1 = 7$$

as found in Section 6.4.1.

In the Topological–Magnitude Graphs, a set of criteria is required for the construction and interpretation of the graphs, which is based on an algebraic expression of the generating class, in order to obtain a condensed graphical representation for any given model. However, the Proportional Graph does not provide such a condensed representation since lines are drawn corresponding to every main effect and interaction

effect in the model. The order of the lines in the Proportional Graph is immaterial since they are always labelled, and for this reason one could use just a single link style, although the use of diverging and horizontal links makes the distinction between main effects and interaction effects clearer. For example, the basic structure of the Proportional Graph (with every line of constant length, and assuming that each factor has two levels) for the model with generating class {[*AB*] [*AC*] [*BC*]} is presented in Figure 6-50. Comparing this graph with the Topological and Magnitude Graphs presented previously in Figure 6-46, it can be seen that the Proportional Graph provides a less condense representation of the model. Thus the Proportional Graph is only recommended for use if it is necessary to communicate the magnitude of each effect separately.

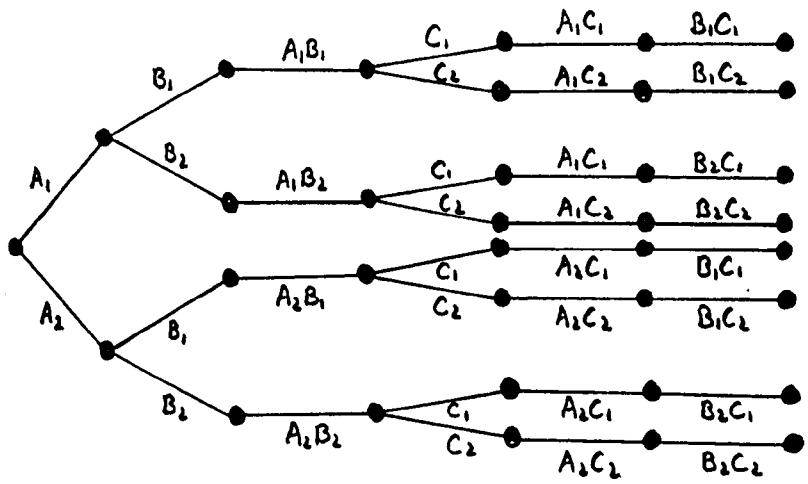

Figure 6-50: Structure of Proportional Graph drawn to represent the model with generating class {[AB] [AC] [BC]}

6.4.8 A Representation Technique for Confidence Intervals Based on the Proportional Graph

Because the Proportional Graph represents every main effect and interaction effect in the model, it can be used to display additional information such as confidence intervals for the effects. This can be done by drawing a vertical bar corresponding to each (horizontal) line in the graph of length proportional to the value of the confidence interval.

For example, the Proportional Graph presented in Figure 6-49 is presented again in Figure 6-51 with hypothetical confidence limits superimposed at the ends of the lines corresponding to each effect.

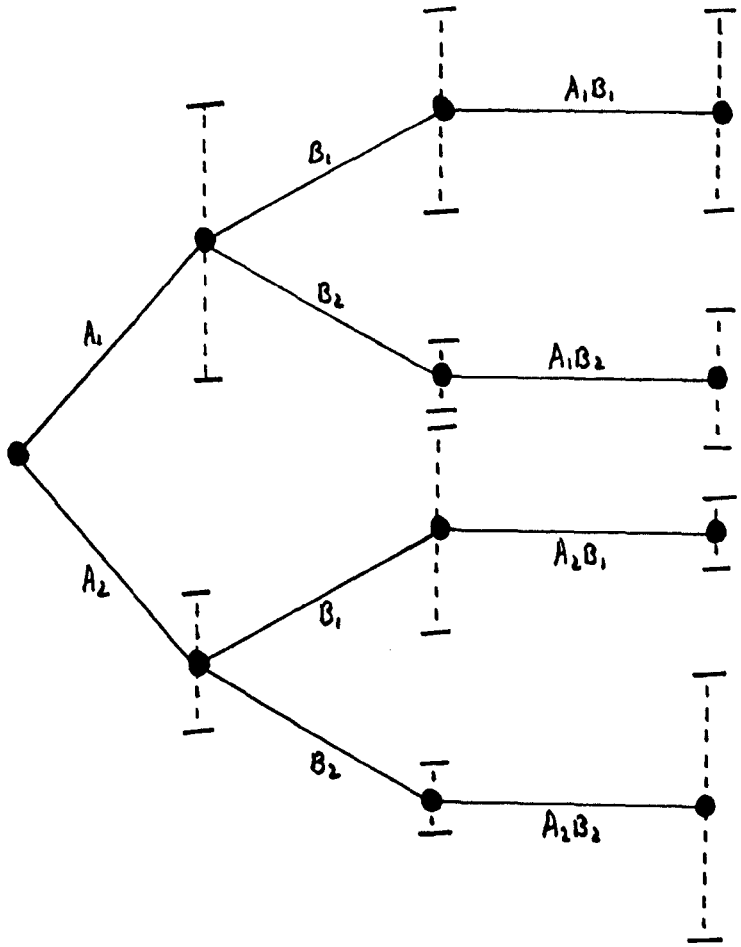

Figure 6-51: Proportional Graph drawn to represent the model with generating class {[AB] [C]} with hypothetical confidence limits super-imposed

Information about the confidence limits of the effects could also be included in the boxes of the Magnitude Graph but this may lead to unnecessary clutter.

6.4.9 Conclusions

In this section, a novel technique for the representation of the generating class of a model, called the Topological–Magnitude Graph, was presented. This technique can be used for the graphical communication and interpretation of a model, in terms of the topology of the effects (ie. the effects which are present) and the magnitude of the effects (ie. the size of the effects which are present). The relationships between the two graph types and the generating class of the model, in terms of the derivation of one from another, is illustrated in Figure 6-52.

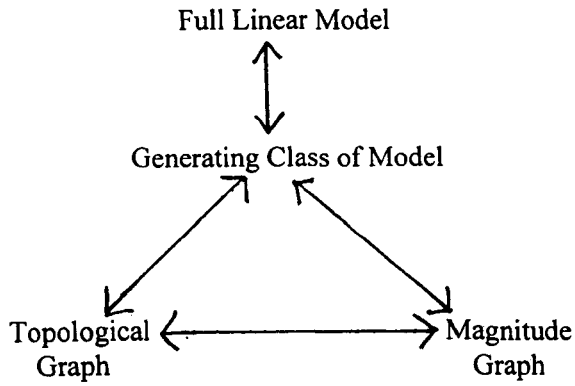

Figure 6-52: Figure showing the relationships between the Topological and Magnitude Graphs and the generating class of the model

This technique requires the use of an algebraic re-expression of the generating class which relates the generating class to four types of links used according to a set of criteria which has been described in some detail. The algebraic expression which is derived has implications for the number of and nature of the link styles required, and whether any variables must be duplicated in the representation. The actual style of the links used in the graph, and the equivalent symbols used in the algebraic expression of the model are not important, although those suggested have an intuitive attraction. Any link or relationship used must be used in a consistent manner. A certain amount of knowledge about Topological–Magnitude Graphs is therefore required to be able to interpret the graphs effectively, but this form of criticism is applicable to most graphical techniques.

The Topological–Magnitude Graph approach is applicable to both (hierarchical) ANOVA and log-linear interaction models, such that it is claimed (without proof) that the generating class of any model, and therefore any model (since the generating class of a model is unique) can be represented without ambiguity. Very complex graphs would be obtained for models containing very many variables and interactions between these variables, but this is likely to be true of all representation techniques.

Although the Topological–Magnitude Graphs illustrate the additivity and interaction of effects, a potential criticism of the technique is that the left to right nature of the construction of the graphs leads to some asymmetry in the representation of models, since the order in which the variables are considered may affect the appearance of the graph. Also, in an interaction, the value taken by one variable appears to be dependent on the level of another, but the converse relationship (ie. that the value of the second variable is similarly dependent on the level of the first) does not appear to be true (although it is). Another criticism of the Topological–Magnitude Graph is that the compounding of effects within the boxes of the Magnitude Graph may be a draw-back, depending on the use to which the Magnitude Graph is to be put, although this can be circumvented by use of the Proportional Graph, which is described and suggested as a basis for the graphical communication of other information such as confidence intervals.

6.5 Summary

In this chapter, three different approaches to the graphical representation of a fitted statistical model have been developed. All three approaches have in common the fact that they are novel two-dimensional techniques, designed to illustrate which terms are present in the model represented.

The first approach, based on combinations of points, resulted in a number of different methods for linking points in order to indicate interactions in the model. Of all the methods developed, however, only one (involving the separate display of the generating class elements) was wholly successful in permitting an unambiguous representation of the model terms for all fitted models. This method in particular, together with certain of the other methods considered, will be developed further in subsequent chapters.

The second approach, based on the Edwards-style Venn diagram, was found to have a number of shortcomings. This approach will not be considered any further in this thesis.

The final approach, based on an original concept termed the Topological–Magnitude Graph, was found to be useful not only for indicating which effects are present in the model, but also for indicating the size of these effects. A set of criteria has been presented for the construction of the Topological and Magnitude Graphs from the generating class of any given model. Although this approach will not be considered further within the context of this thesis, it is believed to be a useful technique for the graphical representation of fitted models.

In the following chapter, the conditional independence graph approach to the representation of a particular class of models (the so-called graphical models) will be considered. As will be seen, there are certain similarities between the conditional independence graphs described and the first of the methods considered in this chapter (namely, the use of links between vertices for the two-dimensional combination of points).

7. The Conditional Independence Graph Approach

7.1 Introduction

In the preceding chapter, an attempt was made to develop some new graphical techniques for the representation of structured multivariate data. The first set of techniques described, involving the two-dimensional combination of points (Section 6.2), is probably the simplest of the approaches considered, both to construct and to comprehend. The basic notion of drawing two-dimensional graphs with vertices representing variables and links representing associations between variables is intuitively simple but quite profound. Indeed, this particular approach is not wholly novel, for graphs of this sort form the foundations of a relatively new area of statistics: graphical modelling.

In this chapter I wish to describe the graphical modelling approach to statistical data analysis, in the hope that the concomitant graphical representations of use in the fitting of graphical models may provide a starting point for the development of further techniques for the representation of structured multivariate data.

The graphical representation technique used in graphical modelling, which closely resembles the two-dimensional combination of points involving vertices and edges between associated variables/vertices presented in Section 6.2.2 of the preceding chapter, is known as the conditional independence graph, or independence graph. As the name suggests, the construction and interpretation of the conditional independence graph is based on the important notion of conditional independence, which is considered briefly in Section 7.2. As will be seen, many statistical models may be interpreted in terms of conditional independence and therefore represented by conditional independence graphs.

Having considered conditional independence in general terms, I shall present some necessary graph theoretic concepts used in graphical modelling before going on to describe the construction and interpretation of conditional independence graphs in Section 7.3.

In Section 7.4 of this chapter, and in the remainder of this thesis, I shall be considering two types of statistical model having a conditional independence interpretation and commonly considered in graphical modelling; namely the covariance selection model for continuous data, and the log-linear interaction model for discrete data, typically used in contingency table analysis. However, there are other measures of

association which can be used in contingency table analysis, such as the odds ratio, and I shall also consider this. For the sake of completeness I shall also make some mention of models for mixed data and of causal modelling. Most other statistical models fall into these categories; eg. models for regression analysis, analysis of variance, logit analysis, logistic regression, etc.. However, because graphical modelling theory is not so tractable for mixed data models, they are not considered again until Chapter 13.

Although this thesis is concerned with the representation of fitted statistical models, and not with the process of fitting statistical models, mention has already been made in Chapters 2 and 3 of a number of graphical representations which may be of use for this purpose; therefore some mention will also be made, in Section 7.5, of the techniques of graphical modelling. This provides an opportunity to describe the statistical packages GLIM and MIM, which are used and referred to in subsequent chapters.

Finally, but most importantly, I shall describe, in Section 7.6, the limitations of the conditional independence graph as a technique for the representation of fitted statistical models in relation to my research aims. This has lead to the development of modifications to the conditional independence graph which are described in Chapters 9 and 10. However, even using the suggested modifications there are still limitations to the conditional independence graph approach. I have solved these by the development of an interactive computer package called the "Conditional Independence Graph Enhancer", or CIGE. The features of CIGE are described in Chapter 11 (and in more detail in Appendix A), and illustrated using real and example data sets in Chapter 12.

Graphical modelling was really born with the seminal paper by Darroch, Lauritzen & Speed (1980), anticipated by Speed (1978), which unified conditional independence theory and Markov graphs. In the years since, the field has expanded quite rapidly, primarily due to the output of a small number of enthusiastic researchers. The technique is becoming more widely known, but to date there has only been one text which succeeds in bringing together all the various aspects of graphical modelling, which have previously only been described in journal articles or technical reports; this being the book by Whittaker (1990), although a less theoretically detailed book has recently been written by Edwards (1995). The books by Whittaker and Edwards form the basis of the description of the conditional independence graph approach presented below, but where other sources from the ever-increasing graphical models literature have been used, these have been acknowledged.

7.2 The Importance of Conditional Independence

The basic concept that underlies the theory of graphical modelling is that of conditional independence. Although not always obvious, conventional statistical techniques often have a conditional independence interpretation.

For example, Dawid (1979), in a paper pre-dating the conditional independence graphs literature, explains how conditional independence forms a conceptual frame-work for much of the theory of statistical inference, as well as probability theory. It is Dawid's assertion that many areas of statistics, which otherwise appear to be unrelated, may be unified in terms of conditional independence. Indeed, pre-empting the development of graphical modelling, Dawid claims that "rather than just being another useful tool in the statistician's kit-bag, conditional independence offers a new language for the expression of statistical concepts and a framework for their study".

Whittaker (1990) shares Dawid's contention that conditional independence can provide a unifying theoretical frame-work for multivariate statistics, but states that graphical modelling and independence graphs are just one manifestation. As will be seen in this chapter, graphical modelling uses notions of conditional independence to unify covariance selection models, log-linear models, path analysis, and, more recently, models for a mixture of continuous and discrete data.

Knuiman (1978) argues that identification of the conditional independences between variables or subsets of variables offers ease of interpretation for high-dimensional data involving a large number of interactions.

To denote that 2 random variables X and Y are independent, Dawid uses the notation $X \perp\!\!\!\perp Y$. This situation arises if any information we know about Y does not affect the distribution of X. If $X \perp\!\!\!\perp Y$, then $Y \perp\!\!\!\perp X$.

If X and Y are independent given that a third variable $Z=z$, this is expressed $X \perp\!\!\!\perp Y \mid Z$. This states that the conditional distribution of X, given Y and Z is in fact completely determined by Z alone, Y being superfluous once Z is known. If $X \perp\!\!\!\perp Y \mid Z$, then $Y \perp\!\!\!\perp X \mid Z$.

It is possible that $X \perp\!\!\!\perp Y \mid Z$ without $X \perp\!\!\!\perp Y$, and *vice versa*. This is termed Simpson's Paradox (Simpson (1951) and is discussed further in Section 7.4.2 below.

Slightly different notations to that suggested by Dawid exist, and are considered in Section 7.3.2 below.

7.3 Conditional Independence Graphs

It is necessary, for a full understanding of the role of conditional independence graphs in graphical modelling, to have some knowledge of the graph theoretic terms which are commonly used, so I shall begin by defining and illustrating the necessary terms. I shall then consider briefly the concept of conditional independence and go on to give a general description of the construction of conditional independence graphs, given the set of conditional independence relationships identified in a set of data. However, the identification of the conditional independence relationships within a given continuous or discrete data set will not be dealt with until Section 7.4. I shall conclude this section by defining the Markov properties of conditional independence graphs, which are of importance in the interpretation of the graphs.

7.3.1 Graph Theory

In graph theoretic terms (see, for example, Harary (1969) or Wilson (1985)), a *graph* is a mathematical object, $G(V,E)$, or G, which is formed by a set of *vertices* $\{V\}$, and a set of *edges* $\{E\}$ formed by unordered pairs of vertices (although in a directed graph, the vertices may be ordered). A *drawing* of a graph is a *mapping* of the graph onto a *surface*. In graph theoretic terms, the vertices are mapped onto *nodes* and the edges are mapped onto *arcs*. I shall be less rigorous in the use of graph theoretic terms and refer to the drawing of the graph (which is always mapped onto the plane in order to obtain a convenient representation of the conditional independence graph) as a graph, and to the nodes (drawn as filled circles) as vertices and arcs (drawn as straight lines) as edges. In this section, I shall only be concerned with simple undirected graphs, although directed graphs, in which the edges are formed by arrows, do have a role in graphical modelling and will be briefly described in Section 7.4.4. A *simple* graph is one which has no multiple edges or loops.

In all drawings of graphs, the following assumptions are made:

- Adjacent edges (ie. edges with the same vertex in common) never cross.
- Two non-adjacent edges cross at most once.
- No edge crosses itself.

- No more than two edges cross at a single point in the plane.
- There is no more than one edge between a pair of vertices.
- There are no loops (ie. edges having the same vertex at both ends).
- All edges are undirected.
- There are a finite number of vertices and edges.

These assumptions are also applicable to conditional independence graphs, for which most of the assumptions are satisfied by drawing the edges of the graph as straight lines.

An example of a conditional independence graph is presented in Figure 7-1. This will be used to illustrate the terms which are introduced in the following paragraphs.

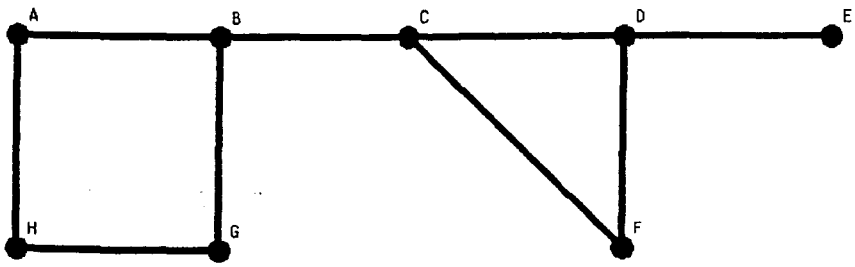

Figure 7-1: An example conditional independence graph

Two vertices are *adjacent* if there is an edge between them. For example, vertices *A* and *B* in the diagram are adjacent. Two edges are adjacent if there is a vertex between them. For example, edges A–B and B–C are adjacent.

A *path* is a sequence of vertices with an edge between each successive pair. For example, in the diagram there are two paths between vertices *B* and *F* (*B–C–F* or *B–C–D–F*). The path is a *cycle* if the start and end-points are the same. For example, in the diagram there is a cycle *C–D–F–C* and a cycle *A–B–G–H–A*. It is a *chordless cycle* if no pair of variables other than successive ones in the path are adjacent. In the diagram, *A–B–G–H* form a cycle which is chordless, but if there was an edge between *A* and *G* and/or an edge between *B* and *H*, then the cycle would no longer be chordless.

Two vertices i and j are *connected* if there is a path from i to j. For example, in the diagram vertices G and F are connected, since there is a path between them (eg. G–B–C–F). If all pairs of vertices are connected, as in the example graph, then it is a *connected graph.*

A *separating subset* of vertices separates two vertices i and j if every path joining the two vertices contains at least one vertex from the separating subset. For example, the variables $\{C,D,F\}$ form a separating subset between the variables B and E in the diagram. A separating subset separates two subsets of vertices a and b if it separates every pair of vertices $i \in a$ and $j \in b$. For example, the variables $\{C,D,F\}$ form a separating subset between the subsets $\{A,B,G,H\}$ and $\{E\}$.

If a is a subset of vertices, the *neighbours* of a are all those vertices adjacent to a vertex in a which are not themselves in a. For example, the neighbours of the subset $\{C,D,F\}$ are $\{B,E\}$.

The *sub-graph* of a is obtained by deleting all the vertices not in a from the graph, together with all edges which do not join two elements of a. A graph or sub-graph is *complete* if all of the vertices are joined to every other vertex. A *clique* is a subset of vertices which induces a complete sub-graph, but the addition of a further vertex would render the graph incomplete; ie. a clique is a *maximally complete sub-graph.* In the example diagram, the variables $\{A,B,G,H\}$ form an incomplete sub-graph, whereas the variables $\{C,D,F\}$ form a complete sub-graph, which is also a clique. The variables $\{C,D\}$ by themselves form a complete sub-graph, but this is not a maximally complete sub-graph and therefore not a clique, since the sub-graph remains complete when the vertex F is included.

The most important feature of the conditional independence graph is that it highlights non-adjacent variables and cliques. As shall be seen, this is of importance in the construction and, in particular, in the interpretation of the graph.

7.3.2 Construction of Conditional Independence Graphs

For the three-variable case where A and B are conditionally independent given C, I shall adopt the notation $A \perp B \mid C$, used in the papers by Whittaker (1988), Lauritzen (1979) and Edwards & Kreiner (1983), amongst others. Two alternative notations are $A \perp\!\!\!\perp B \mid C$, employed by Dawid (1979), Whittaker (1990) and Edwards (1995), and in more recent papers by Edwards, Lauritzen, and Wermuth (eg. Edwards (1990), Lauritzen & Wermuth

(1989), Wermuth (1985, 1988) and Wermuth & Lauritzen (1983,1990)), and $A \otimes B | C$, employed by Darroch, Lauritzen & Speed (1980). The extension of the notation adopted to multivariate cases should become apparent.

In the construction of a conditional independence graph, one vertex is drawn per variable, located freely on the plane. No edge is drawn between a pair of vertices if, and only if, the corresponding pair of variables are conditionally independent given the rest. Thus, because the pair-wise definition of conditional independence used corresponds naturally to the pair-wise definition of an edge, the edges in the graph illustrate the patterns of dependence or association within the data.

Consider the 4 different conditional independence graphs which it is possible to construct for the three variables A, B and C, shown in Figure 7-2. In general, for k variables it is possible to construct $2^{\binom{k}{2}}$ different conditional independence graphs.

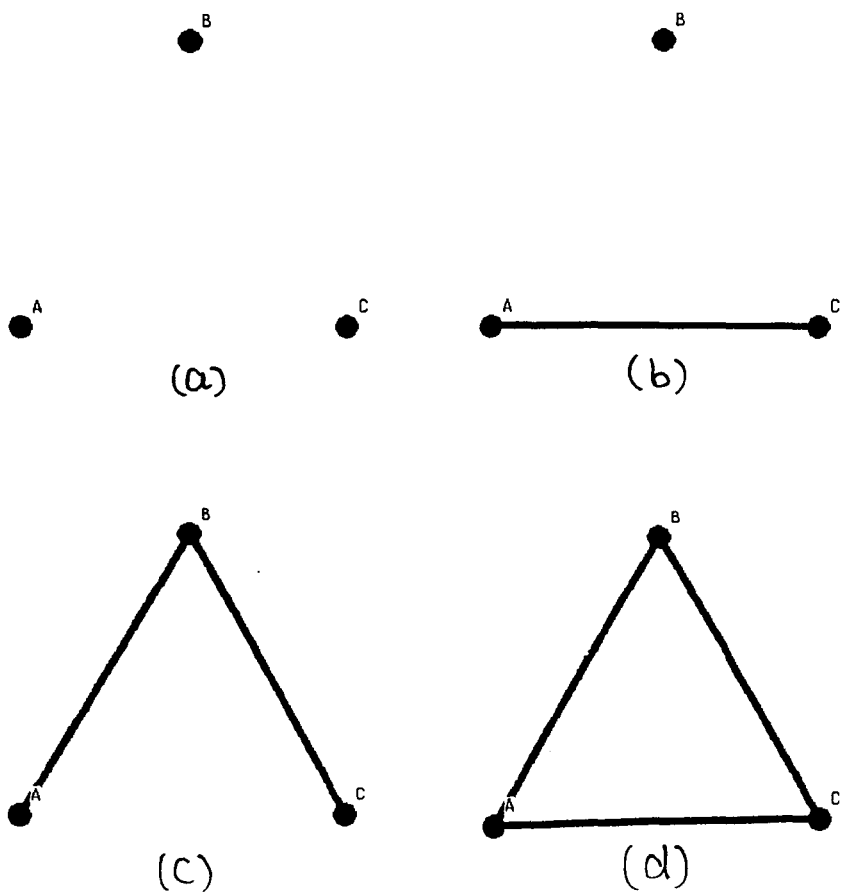

Figure 7-2: Four conditional independence graphs constructed for three variables A, B, C

The conditional independence interpretation of the four graphs is as follows. The first graph corresponds to mutual independence, and the conditional independence statements which can be read from the graph are: $A \perp B \mid C$, $A \perp C \mid B$, and $B \perp C \mid A$. The second graph corresponds to partial independence, and the conditional independence statements which can be read from the graph are: $A \perp B \mid C$ and $B \perp C \mid A$. The third graph corresponds to conditional independence, and the only conditional independence statement which can be read from the graph is $A \perp C \mid B$. The fourth graph corresponds to no independence, and no conditional independence statements can be read from the graph.

7.3.3 Markov Properties of Conditional Independence Graphs

The usual interpretation of a conditional independence graph holds that two variables i and j which are not joined by an edge (ie. which are not adjacent) are conditionally independent given the 'rest' of the variables. ie. $i \perp j \mid \{\text{rest}\}$. However, conditional independence graphs have been shown (eg. Darroch, Lauritzen & Speed (1980), and Speed (1978)) to have Markov properties, which have implications for the interpretation of the graphs. The three Markov properties of conditional independence graphs and their implications are described below.

Pair-wise Markov Property: Non-adjacent pairs of variables are independent conditional on the remaining variables. This is the property which essentially defines the construction and interpretation of conditional independence graphs.

Local Markov Property: Any variable is independent of all the remaining variables conditional only on the adjacent variables (ie. the boundary).

Global Markov Property: Any two subsets of variables separated by a third are independent conditional only on the variables contained in the third subset.

These three properties are in fact equivalent. Proofs are contained in Whittaker (1990), and make use of the separation theorem which is presented below.

Separation theorem: Non-adjacent variables are independent given their separating subset alone. In other words, if every vertex in subset b is separated from every vertex in subset c by the vertices in subset a, then $\{b\} \perp \{c\} \mid \{a\}$. If a is the minimal

separating subset, this implies that some of the conditioning variables may be redundant in an independence relationship.

7.4 Models of Interest

Two types of model of primary interest to graphical modellers are covariance selection models for continuous data and log-linear interaction models for discrete data. Wermuth (1976a) showed that these two model types are in fact analogous. Recently, advances have been made in the consideration of models for mixed data and although, because of their theoretical complexity, these will not be my main concern, they are described for the sake of completeness. Covariance selection models and log-linear interaction models may be regarded as special cases of models for mixed data having just one type of data: continuous or discrete. Within graphical modelling, models for continuous multivariate normal data (ie. covariance selection models) are also known as *graphical Gaussian models*, models for discrete multinomial data (ie. log-linear models) are also known as *graphical log-linear models*, and models for mixed data are also known as *graphical conditional Gaussian models* or *mixed interaction models*. Within the section on mixed models, I shall also be considering other types of graphical models, such as *graphical chain models* for directed data.

7.4.1 Covariance Selection Models for Continuous Data

Standard analyses of continuous data, when summarised by a correlation or covariance matrix, commonly employ techniques which examine linear combinations of the variables; see, for example, the multivariate ordination techniques described in Section 2.5 of Chapter 2. However, the technique of covariance selection described in Dempster (1972) and Knuiman (1978) takes an alternative approach which, as the name suggests, involves selecting elements of the (inverse) covariance matrix, and setting them to be zero. Because of the importance of covariance selection models within graphical modelling in general, and their relevance for the remainder of this thesis in particular, I shall provide a brief overview of covariance selection models in this section.

In multivariate normally distributed, or Gaussian, data consisting of p continuous variables, which can be represented as a correlation or covariance matrix, there are a

possible $p(p-1)/2$ pairwise associations. As in all of statistical model fitting, a way is required which permits the modeller to analyse these associations in order to explain and easily interpret the data.

Dempster (1972) introduced covariance selection as a technique for reducing the number of parameters (ie. the number of elements of the inverse covariance matrix) to be estimated. This is achieved by setting elements of the inverse covariance matrix, known as *concentrations*, to zero. This is equivalent to setting partial correlations to zero. Thus if the element σ_{ij} of the matrix Σ is the covariance of variables x_i and x_j, then the element σ^{ij} of the inverse matrix Σ^{-1} is the concentration of variables x_i and x_j, and $\rho_{ij.K}$ is the partial correlation.

Iterative fitting procedures involving the use of likelihood ratio tests and based on the calculation of deviance were derived by Dempster for determining which concentrations could be set to zero, but there is an alternative "naive" covariance selection modelling approach which I describe below and shall be applying in later Chapters.

Wermuth (1976a) noted the connections between log-linear models and covariance selection models, both of which belong to the exponential family of statistical models, and both of which can be interpreted in terms of conditional independence. Essentially, if a concentration in the inverse covariance matrix is zero, then the corresponding variables are conditionally independent given the rest. If a concentration is non-zero, then the pair of vertices corresponding to these variables will be joined by an edge in the corresponding conditional independence graph. The pattern of non-zero and zero concentrations is sometimes referred to as an adjacency matrix.

It should be noted that a covariance selection model may be specified completely by the set of pairwise non-zero associations, or two-way interactions. Therefore, unlike log-linear models (see below), we do not need to restrict our attention to hierarchical models, nor is it possible that a covariance model may not be represented uniquely by its corresponding independence graph – all covariance selection models are graphical models.

Naive Covariance Selection Modelling

This is a simplistic but effective approach to covariance selection modelling, described in Whittaker (1988, 1990).

1. Estimate the population covariance matrix **V** by its maximum likelihood estimate: the sample covariance matrix **S**. Alternatively, estimate the population correlation matrix by the sample correlation matrix **R**.
2. Calculate the inverse of the sample covariance matrix, $\mathbf{S}^{-1}$. Alternatively, calculate the inverse of the sample correlation matrix, $\mathbf{R}^{-1}$. The diagonal elements of either matrix are interpretable in terms of partial variances, and indicate how well the corresponding variable can be predicted from the other variables. For example, the diagonal elements of the inverse correlation matrix correspond to $1/(1-\mathbf{R}^2)$, where **R** is the multiple correlation coefficient between the corresponding variable and the rest.
3. Re-scale either inverse to have unit values on the diagonal. The off-diagonal elements will now correspond to the negatives of the partial correlation between the two corresponding elements i and j, partialled on the rest, K ie. $-\rho_{ij.K}$.
4. Set 'sufficiently small' off-diagonal elements of the scaled inverse matrix to zero. This may be assessed by calculating the edge exclusion deviance, e.e.d = $-N \ln\{1-(\rho_{ij.K})^2\}$ and comparing to χ^2 value with 1 degree of freedom.
5. Draw the resulting independence graph on the basis that no edge is included in the graph if the corresponding partial correlation coefficient has been set to zero.

This approach may be criticised in that a whole set of edges is removed at once, instead of removing the least significant edge, recalculating the edge exclusion deviances, removing the least significant edge and so on in a backwards model selection procedure (see below) until no more edges can be removed. However, Whittaker *et al* (1988) argue that either approach usually results in the same set of edges being included in the graph.

7.4.2 Log-Linear Interaction Models for Discrete Data

The seminal paper by Darroch *et al* (1980) considered the use of graphical models within the context of log-linear models. Log-linear models, which are linear models of the logarithms of the expected cell counts in a multidimensional contingency table, are well described in many texts. Probably the definitive text is Bishop, Fienberg & Holland (1975), although a more accessible account is contained in Everitt (1992). Because of the importance of log-linear models within graphical modelling in general, and their

relevance for the remainder of this thesis in particular, I shall provide a brief overview of log-linear interaction models in this section. This overview will be based on the approach taken by Everitt (1992) and Edwards (1995). Alternative sources are Kastenbaum (1974), Upton (1978), Upton (1986)

Suppose, for example, we have N observations on three discrete variables A, B and C, having $\#A$, $\#B$ and $\#C$ levels respectively. We can form a three-way contingency table of counts, having individual cell counts n_{ijk}, where i=1,...,$\#A$, j=1,...,$\#B$, k=1,...,$\#C$. The cell probability (ie. the probability that an observation falls in a given cell) can be written as p_{ijk} where $p_{ijk} = p_{i..}p_{.j.}p_{..k}$ (ie. the product of the corresponding row, column and layer totals). Hence the expected cell count $m_{ijk} = Np_{ijk}$. If the cell counts are assumed to be independent, then the joint probability distribution of the table of counts is given by the multinomial distribution. It is assumed that the categories of the variables are unordered, and that the marginal counts of the table are not fixed by the design.

The simplest linear model for a three-way table expresses the (natural) logarithm of the cell probabilities as:

$$\ln(p_{ijk}) = u + u_iA + u_jB + u_kC \quad \text{(i)}$$

where the u's are unknown parameters. Analogous to ANOVA, u relates to an overall mean effect and u_i, u_j, u_k to main effects. This model states that A, B and C are completely independent, ie. $A \perp B \mid C$, $A \perp C \mid B$ and $B \perp C \mid A$. This corresponds to Figure 7.2a.

A more complex log-linear model is:

$$\ln(p_{ijk}) = u + u_iA + u_jB + u_kC + u_{jk}AC \quad \text{(ii)}$$

Extending the analogy with ANOVA, this model states that there is a two-way interaction between A and C, although $u_{ij}AB$=0 and $u_{ik}BC$=0 implying that A and B, and B and C, are independent, ie. $A \perp B \mid C$ and $B \perp C \mid A$. This corresponds to Figure 7.2b.

Another possible model for the three variables is:

$$\ln(p_{ijk}) = u + u_iA + u_jB + u_kC + u_{ij}AB + u_{jk}BC \quad \text{(iii)}$$

This model states that there is a two-way interaction between A and B and between B and C, although $u_{ik}AC$=0 implying that A and C are independent, ie. $A \perp C \mid B$. This corresponds to Figure 7.2c.

Another model is:

$$\ln(p_{ijk}) = u + u_iA + u_jB + u_kC + u_{ij}AB + u_{ik}AC + u_{jk}BC \quad \text{(iv)}$$

This model states that there are two-way interactions between every pair of variables, thus there are no conditional independences. This corresponds to Figure 7.2d.

The full model is:

$$\ln(p_{ijk}) = u + u_iA + u_jB + u_kC + u_{ij}AB + u_{ik}AC + u_{jk}BC + u_{ijk}ABC \quad \text{(v)}$$

This model states that there is a three-way interaction, in addition to two-way interactions between every pair of variables, thus there are no conditional independences. This also corresponds to Figure 7.2d.

Attention is restricted to hierarchical log-linear models. The term hierarchical means that if a term in the model is set to zero, then all higher-order terms in the model involving the same subset of variables is also zero. For example, if $u_{ij}AB = 0$, as in (ii) above, then $u_{ijk}ABC=0$. The main reason for focusing attention on hierarchical models is their ease of interpretability compared with non-hierarchical models.

Because we are restricting attention to hierarchical models, it is possible to infer the presence of lower order interaction terms in the model from the higher order interaction terms. This gives a short-hand way of expressing a model, simply by writing down the list of terms corresponding to the maximal interaction terms in the model, from which all other terms may be inferred. This is termed the *generating class* of the model. For example, the five models considered above may be written as (i){[A] [B] [C]}, (ii) {[B] [AC]}, (iii) {[AB] [BC]}, (iv) {[AB] [AC] [BC] and (v) {[ABC]} respectively.

The independence graph is constructed by connecting any pair of vertices with an edge if there is a term in the generating class of the log-linear model which implies a two-way interaction between the variables represented by the vertices. As can be seen, two models may have the same independence graphs (eg. models (iv) and (v) above) where they contain the same two-way interactions. Thus log-linear models do not always have a unique representation. Where the cliques of the graph correspond the generating class of the model (as in model (v) above), the model is said to be *graphical*. Where the cliques of the graph do not correspond to the generating class of the model (as in model (iv) above), the model is said to be *non-graphical*. Pairwise conditional independences which may be inferred from the graph wherever there is no edge are equivalent to zero two-way

interactions. The same conditional independence relationship hold therefore for both models (iv) and (v), and model (iv) may be regarded as a sub-model of model (v). I shall return to the problem of the non-uniqueness of the independence graph in Chapter 9.

Edwards & Kreiner (1983), Whittaker (1990) and others refer to a graph constructed in this way as an "interaction graph", but in actuality it is identical to the independence graph. I wish to use the term "interaction graph" in a specific sense in Chapter 9 and subsequent chapters, and so will continue to refer to graphs drawn in this context as independence graphs.

The analysis of contingency tables by graphical models is considered in detail in Darroch *et al* (1980), Edwards & Kreiner (1983).

Simpson's Paradox

The fact that only the so-called graphical models can be represented by the independence graph illustrates one inadequacy of concentrating on the pairwise associations between variables. Simpson's Paradox (Simpson (1951), Hand (1979), Paik (1985)) illustrates another inadequacy of only studying pairwise associations.

Consider model (iii) above, for which $A \perp C \mid B$, indicating that A and C are conditionally independent given B. If the three-way contingency table were to be collapsed over the levels of B, we would find that A is no longer independent of C. This is a consequence of Simpson's Paradox, which occurs whenever there is a change in the nature of the association between the marginal and the conditional distributions This change may also be a reversal in the direction of the association.

Asmussen & Edwards (1983) discuss collapsibility within a multidimensional contingency table.

Fitting Log-Linear Interaction Models

Log-linear interaction models can be fitted using statistical packages such as GLIM and MIM, which are described in detail in Section 7.5. This section outlines some of the basic principles involved in the fitting of log-linear models.

The goodness of fit of a log-linear model may be measured by its deviance. This is based on the difference in the maximised log-likelihoods for the full, or saturated model and the model of interest M_0. This can written as:

$$G^2 = 2 \Sigma n_{ijk} \ln(n_{ijk}/m'_{ijk})$$

where $m'_{ijk} = N p'_{ijk}$, where the p'_{ijk} are the maximum likelihood estimates of p_{ijk} for the model M_0, obtained from the cell counts. Assuming that the hypothesised model of interest M_0 is true, then the deviance is asymptotically $\chi^2_{[k]}$ distributed, where the k degrees of freedom is the difference in the number of free parameters between the two models.

Alternatively, the deviance difference between two nested models M_0 and M_1 (where M_0 is nested under M_1) may be computed as:

$$d = 2 \Sigma n_{ijk} \ln(m_1'_{ijk} / m_0'_{ijk})$$

where $m_0'_{ijk}$ and $m_1'_{ijk}$ are the maximum likelihood estimates for M_0 and M_1 respectively. Assuming M_0 is true, then the deviance difference is asymptotically $\chi^2_{[k]}$ distributed, where k is the difference in the number of free parameters between M_0 and M_1.

In either case, the sub-model will be accepted in preference to the full or fuller model only if the difference in deviance is non-significant.

Note that there is a special class of log-linear models, called decomposable models, that have direct estimates that can be used in place of the maximum likelihood estimates in the above formulae. Decomposable models are characterised by being graphical models with triangulated graphs (ie. no cycles of length ≥ 4 without a chord). Thus the graph can be decomposed into a set of complete subgraphs having direct estimates. Decomposable models are considered in some detail in Darroch *et al* (1980) and Lauritzen *et al* (1984).

The following table (Table 7-1), taken from Darroch *et al* (1980), gives some idea of the number of models which may be fitted for 2, 3, 4 or 5 variables. Darroch *et al* actually present the graphs corresponding to all decomposable and hierarchical non-decomposable graphical models of dimension ≤ 5. As can be seen, for even a small number of variables, there are very many possible models. The topic of model selection will be returned to in Section 7.5.

	Dimension			
Type	2	3	4	5
Interaction	8	128	32768	2147483648
Hierarchical	5	19	167	7580
Graphical	5	18	113	1450
Decomposable	5	18	110	1233

Table 7-1: Number of models of given type for different numbers of variables

Other Measures of Association for Discrete Data

For simple two-way contingency tables with r rows and c columns, the most familiar measure of association is probably the Pearson X^2 statistic based on the differences between observed (O_{ij}) and expected (E_{ij}) frequencies such that

$$X^2 = \Sigma \Sigma (O_{ij}-E_{ij})^2 / E_{ij}$$

This statistic approximately follows a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom against which it may be compared to assess the independence of the classifying variables.

For 2×2 tables, a variety of statistics may be computed such as the relative risk, the odds, and the odds ratio. Consider, for illustration, the table presented in Figure 7-3.

		Variable *I*		
		Level 1	Level 2	
Variable *J*	Level 1	*a*	*b*	*a+b*
	Level 2	*c*	*d*	*c+d*
		a+c	*b+d*	*N*

Figure 7-3: Illustrative 2x2 contingency table

The odds of an individual being in Level 1 of *I* given that they are in Level 1 of *J* is *a/b*. The odds of an individual being in Level 1 of *I* given that they are in Level 2 of *J* is *c/d*. The odds ratio, alternatively known as the cross-product ratio, is given by $(a/b)/(c/d) = ad/bc$.

Alternatively, the odds of an individual being in Level 2 of *I* given that they are in Level 1 of *J* is *b/a*. The odds of an individual being in Level 2 of *I* given that they are in Level 2 of *J* is *d/c*. The odds ratio is now given by $(b/a)/(d/c) = bc/ad$.

If there is no association between *I* and *J*, then the proportions will be the same irrespective of the levels, ie. *a/b* will equal *c/d*. In this situation, the odds ratio will take the value 1, no matter how it has been calculated. Alternatively, one may consider the natural log of the odds ratio, which will take the value 0. Departures from these values indicate association between the classifying variables. An odds ratio less than one will result in a negative log odds ratio, whereas an odds ratio greater than one will result in a positive log odds ratio. Swapping the levels of one of the variables will give 1/(odds

ratio), as shown above. Ideally one would obtain a confidence interval for a calculated odds ratio in order to assess whether it is significantly different from 1.

For a 2×2×2 table, the odds ratio can be computed for the association between two of the variables at each level of the third. If the ratio of these two odds ratios is then computed, this will indicate whether there is a three-way association or not.

There are other measures of association not considered here. The interested reader is referred to Bishop *et al* (1975), Freeman (1983), Everitt (1992).

7.4.3 Models for Mixed Data

Many classical statistical methods involve both discrete and continuous variables — for example, ANOVA, MANOVA, ANCOVA, multiple regression, logistic regression and structural equation models — and may potentially be modelled using graphical modelling methods extended to mixed data (Lauritzen (1989), Edwards (1990), Wermuth & Lauritzen (1990)), although this has rarely been done in practice.

Graphical models for mixed data combine the models for continuous data and the models for discrete data previously described. Given a set Δ of p discrete variables, and a set Γ of q continuous variables, the probability that the discrete variable $I=i$ is p_i, and the distribution of the continuous variable Y given $I=i$ is multivariate normal. Hence the whole data set is described by the conditional Gaussian distribution. If $\Delta=\varnothing$, this reduces the problem to the class of covariance selection models, whereas if $\Gamma=\varnothing$, this reduces the problem to the class of graphical log-linear models.

Hierarchical models for mixed data, also called mixed interaction models or graphical association models, are defined by a sum of interaction terms and, rather like log-linear interaction modelling, mixed interaction models can be defined by setting higher-order interaction terms to zero.

Model formulae for mixed models can be expressed in a short-hand way in three parts, separated by slashes. For example, if $p=q=1$, $\Delta=\{A\}$, $\Gamma=\{Y\}$, one possible model is $A\ /\ AY\ /\ AY$. The first part of the model corresponds to the discrete generators, the second part to the linear generators, and the third part to the quadratic generators. Rules restrict the set of permissible formulae. For example, for each linear generator there must be a corresponding discrete generator, and for each quadratic generator and for each continuous variable there must be a corresponding linear generator. The letters $A,B,C,...$

are conventionally used to represent discrete variables; ...,*X,Y,Z* to represent continuous variables. If there are no continuous variables, then the discrete generators give the log-linear interaction model. If there are no discrete variables, then the two-factor quadratic generators give the covariance selection model.

If no discrete generator contains *AB*, then $A \perp B \mid \{\text{rest}\}$; If no linear (or quadratic) generator contains *AX* then $A \perp X \mid \{\text{rest}\}$; if no quadratic generator contains *XY* then $X \perp Y \mid \{\text{rest}\}$. Therefore to construct the independence graph for a given mixed model, simply connect with an edge those vertices that occur together in the same generator. When constructing the independence graph corresponding to a mixed model, dots (ie. filled in nodes) are conventionally used to represent the discrete variables, and circles (ie. empty nodes) to represent the continuous variables. For example, Figure 7.4 shows the independence graph corresponding to the mixed (homogeneous) model *AB / AX, BX, AY, BZ / XY, XZ.*

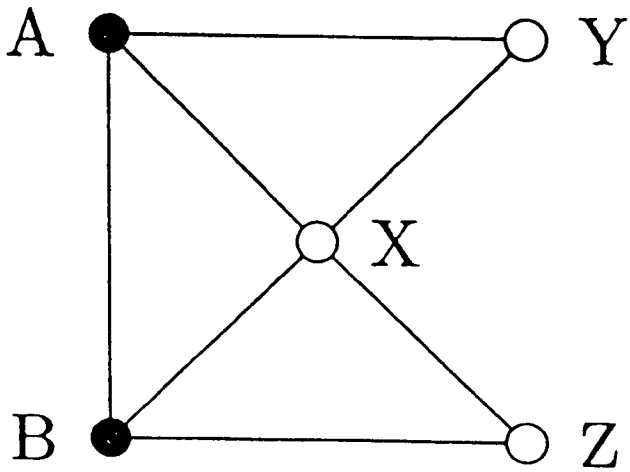

Figure 7-4: Conditional independence graph for mixed model AB / AX, BX, AY, BZ / XY, XZ

To find the formula of the graphical mixed model corresponding to a given graph, consider the cliques of the graph. The discrete generators are given by the cliques involving discrete variables only. The linear generators are given by the cliques involving a single continuous variable. The quadratic generators are given by the cliques involving continuous variables only (for a homogeneous model) or the cliques involving all variables (for a heterogeneous model). The same ambiguities arise as for log-linear

models in that two hierarchical models for mixed data may have the same independence graph, so only the graphical model may be determined from the graph.

Maximum likelihood estimation of the model parameters of a mixed interaction model is possible, and expressions can be derived for the deviance. Thus model fitting may be carried out a similar way to that already described for log-linear interaction models.

A detailed account of hierarchical interaction models is contained in Edwards (1990).

7.4.4 Directed Graphs

Directed graphs can be used to study causal mechanisms or the relationships between explanatory and response variables. The origins of the study of causal mechanisms using directed graphs lie with path analysis in the 1920's, but they now have applications in probabilistic expert systems with an emphasis on Bayesian inference. Edges are drawn on the graph as arrows corresponding to the direction of influence or causality. The graphs should be acyclic ie. with no loops. An early consideration of directed graphs within the context of graphical models is contained in Wermuth & Lauritzen (1983).

Chain Graphs, leading to the formulation of graphical chain models, are an attempt to combine directed and undirected graphs into a single framework (Wermuth (1985), Wermuth (1988), Lauritzen & Wermuth (1989), Wermuth & Lauritzen (1989), Lauritzen (1989)) assuming the conditional Gaussian joint distribution for mixed (continuous and discrete) variables. Within this framework, the set of vertices is divided into subsets called chain components. All edges between the vertices within a chain component are undirected; all edges between chain components are directed. As usual, a missing edge between any two vertices, or subsets of vertices V and W means $V \perp W | \{\text{rest}\}$.

7.5 Model Selection Procedures

Given the large number of different log-linear models it is possible to try fitting to a table of data, even for a modest number of classifying variables (see Table 7-1 above), it is useful to use some sort of strategy or model selection procedure to reduce the number of models which you need to try fitting, but which will result in a good-fitting (if not the best-fitting) parsimonious model for the data.

The commonest strategies involve the application of some sort of stepwise model selection procedure. These involve starting with a model and adding and/or deleting terms (or edges, if a graphical model is desired) until a model is obtained which satisfies some criterion. Similar strategies are employed for the stepwise selection of models in multiple regression.

A backwards elimination procedure typically begins with the saturated model, and terms are tested and removed one at a time subject to the criterion that their removal does not cause a significant change to the fit of the model. A forwards selection procedure typically begins with a basic model (no terms, or no interaction terms) and terms are added and tested one at a time subject to the criterion that their addition leads to a significant improvement in the fit of the model. Other procedures allow for a combination of these two approaches, such that terms which are added may later be removed if the addition of further terms renders them obsolete, and terms which are removed may later be added. A procedure will terminate when no further terms can be added or removed.

The above model selection procedures are not specific to graphical modelling. Further details of general model selection procedures for the fitting of log linear interaction models and/or covariance selection models, some of which incorporate simultaneous test procedures, are contained in Wermuth *et al* (1976), Whittaker & Aitkin (1978), Benedetti & Brown (1978), Aitkin (1980), Edwards & Havranek (1985), Upton (1991). However, it is possible to restrict attention to graphical models and to consider the addition or deletion of edges, and the corresponding terms, at each step. Graphical modelling is described in Wermuth (1976b), Edwards & Kreiner (1983), Havranek (1984). Having obtained a good fitting graphical model, one can then go on to consider any non-graphical models having the same conditional independence structure, ie. having the same graph (Lauritzen (1979)). Whittaker (1984) considers initially focusing attention

on decomposable models in an exploratory analysis of a multi-way contingency table in order to further simplify the calculation required..

The technique of graphical modelling, meaning the fitting of graphical models to data, serves three purposes as listed by Whittaker (1990):

1. **Interpretation.** Graphical modelling can be used to describe or explain the inter-relationships between several variables.
2. **Simplification.** Graphical models are easy to fit and interpret. They can be used to condense the data set without eliminating or obscuring any interesting relationships.
3. **Unification.** Graphical modelling provides a unifying framework for continuous data, discrete data, and mixed data.

The above model selection procedures can also be applied to fit models to mixed data.

7.5.1 Modelling with MIM

MIM is a command driven PC-program designed to carry out graphical modelling for continuous, discrete or mixed data. It is, in fact, the only comprehensive package written to date for carrying out graphical modelling. Its main features are described below. Full details, examples of applications and a copy of the software are contained in Edwards (1995).

Data can be read in to MIM, directly or from a file, as raw data, counts, means or covariances and manipulated. Model formulae can be specified in terms of the members of the generating class (or generators). MIM will report whether the specified model is graphical, whether it is decomposable and whether it is collapsible. Models may be changed by the deletion of two-way interactions from or the addition of two-way interactions to the current model. This corresponds directly to the deletion of edges from or the addition of edges to the conditional independence graph. Parameter estimates for the specified model can be obtained using the fit command. This will give the deviance and degrees of freedom for the specified model. For decomposable models, MIM can obtain the maximum likelihood estimates directly. For non-decomposable models an

iterative algorithm is employed. An asymptotic likelihood ratio test may be used to compare the current model with a specified base model (where the current model is a sub-model of the base model) – the deviance difference between the two models is treated as being asymptotically chi-squared distributed, with degrees of freedom as the difference in the number of free parameters between the two models.

It is also possible to test, in the same way, whether an edge may be deleted from the graph of the current model – this is equivalent to testing whether the corresponding pair of variables are conditionally independent. It is also possible to apply stepwise model selection procedures (see above). The default is backwards selection based on chi-square tests of the deviance difference between successive models. Forward selection may alternatively be used.

Standard input, output and on-screen display facilities (of graphs, data, models and summary statistics relating to the modelling procedure) are provided.

7.5.2 Modelling with GLIM

GLIM (Generalised Linear Interactive Modelling) is a package developed under the auspices of the Royal Statistical Society for the fitting of generalised linear models. Details of the use of GLIM are contained in Aitkin *et al* (1989). In common with SAS and SPSS, GLIM can be used to fit log-linear interaction models to contingency table data, since these fall into the category of generalised linear models, but it cannot be used to fit covariance selection models to continuous data. None of these packages offers procedures directly related to graphical modelling. Whittaker (1982) describes the fitting of graphical log-linear models using the GLIM language, and incorporates the notation of Wilkinson & Rogers (1973) previously described in Chapter 4. MIM resembles GLIM. GLIM has been used to fit the log-linear model considered in Chapter 12.

7.6 Limitations of the Conditional Independence Graph Approach

Whittaker (1990) himself states that "The independence graph of a set of variables conveys a vivid but terse description of their pattern of interaction" and goes on to suggest that the graphs could be augmented by attaching a number to each edge which corresponds to the strength of the bond represented by that edge.

Whittaker (1988) and Whittaker *et al* (1988) highlight the problem of fitting models and constructing graphs for large numbers of variables. For small numbers of variables, it is quite straightforward to construct the corresponding conditional independence graph by hand, but for larger numbers of variables, the placement of the numerous vertices and edges may not be so straightforward, if one wished to avoid having numerous crossovers and curves. Whittaker (1988) suggests two approaches to this.

The first approach suggested is to apply principal components analysis (PCA) to the matrix of negative partial correlations obtained through application of the naive covariance selection procedure. By plotting the vertices/variables in the space defined by the first two principal components, and drawing in edges between pairs of variables having a significant edge exclusion deviance, a graph is obtained which has the added feature that the lengths of the edges between vertices correspond to the strength of the association between pairs of variables. Of course, this does not necessarily result in a graph without numerous crossovers, but this approach provides a starting point for encoding the strength of association between pairs of variables, and will be considered in more detail in Chapter 10.

The second approach suggested by Whittaker for the construction of independence graphs for large numbers of variables is based on the idea of "blocking". This involves grouping together strongly associated subsets of variables, having a complete sub-graph between them, to form a single block or vertex. An edge between two blocks implies the presence of inter-block edges between all variables in the two blocks, whereas the absence of a link between two blocks implies that there are no inter-block edges between any of the variables in the two blocks. Blocks of variables can be determined by visual inspection of the graph, or of the negative partial correlation matrix or matrix of edge exclusion deviances, or by application of a hierarchical clustering technique such as average linkage cluster analysis to either matrix. Although it seems sensible, as an aid to interpretation, to cluster together strongly associated subsets of variables to simplify the edge structure between and within subsets of vertices, it has the disadvantage of loss of information in relation to the aim of wishing to communicate the associations between the variables contained in the model.

Both of these approaches are applicable only to covariance selection models for continuous data, and cannot readily be extended to multi-way contingency tables or mixed data. Furthermore, both approaches are intended as exploratory aids, and could be

combined by imposing the results of a cluster analysis on a graph obtained through the application of PCA. However, none of these approaches have been fully developed by Whittaker. In Chapter 10, the PCA approach will be considered in more detail and developed as a technique for the encoding of strength of association in conditional independence graphs. However, the use of blocking of variables in conditional independence graphs, although suggesting an approach to the summarisation of the structure in the data, seems to result in a gross simplification of the association structure between variables, and will not be pursued any further. The problems of constructing graphs for large numbers of variables with the minimum number of cross-overs will, however, be considered in the following chapter.

7.7 Summary

In this chapter, the use of the conditional independence graph as a simple two-dimensional technique for the representation of fitted statistical models has been described and assessed. This technique has been considered here rather than in Chapter 4 owing to its similarities with some of the two-dimensional approaches developed in Chapter 6.

Although the conditional independence graph has certain limitations, discussed in Section 7.6, which preclude its use as a representation technique for all fitted models, it does provide a starting point for the development of a graphical representation technique which does fulfil the aim of this thesis, and will be the concern of the following chapters.

8. The Crossing Number Problem

8.1 Introduction

A conditional independence graph (as described in Chapter 7) containing many vertices and many edges may, if there are a large number of cross-overs between the edges, appear quite messy (making the graph unacceptable on aesthetic grounds) and be difficult to read (making the graph unacceptable on the grounds of illegibility). Numerous cross-overs may make it difficult to follow the links between vertices which are far apart, and the user may erroneously mistake points at which two or more edges cross as being additional vertices, at least in the case of graphs consisting of huge numbers of edges and vertices. This section is therefore concerned with the minimisation of the number of cross-overs as an aid to legibility.

In three dimensions it would always be possible to construct an independence graph with zero cross-overs, but my research aims are concerned with the derivation of a two-dimensional solution. A three-dimensional solution could only be examined on a computer screen or by construction of a three-dimensional model, but a two-dimensional representation could be constructed by hand or with the aid of a computer and then subsequently be 'dumped' from the screen and printed out for inclusion in reports, etc.

Examination of the graph theory literature and personal correspondence (Guy (1988)) revealed that the problem of representing a graph with the minimum number of cross-overs had only been solved in a small number of specific cases. These results are described in the following sections.

8.2 Graph Theoretic Results

It is assumed that the reader is now familiar with the graph theory concepts presented in Section 7.3.1 of the previous chapter. Some additional concepts which are required for the understanding of the graph theoretic results presented in this chapter are presented below.

A *planar* graph is one which can be *embedded* in the plane; a *non-planar* graph is one which cannot. Thus whenever a non-planar graph is drawn in the plane, some of its edges will cross. The *crossing number* $\nu(G)$ of a graph G is the minimum number of

crossings with which G can be drawn in the plane. A drawing of G having the minimum number of crossings is said to be *optimal*. By definition, for a planar graph $\nu(G)=0$, whereas for a non-planar graph $\nu(G)\neq 0$. If a graph G is a sub-graph of a graph H, then $\nu(G)\leq\nu(H)$.

Erdos & Guy (1973) state: "Almost all questions that one can ask about crossing numbers remain unsolved". Similarly, Kainen (1974) states: "Very few results are known for crossing numbers".

For a complete graph K_p with all $\binom{p}{2}$ possible edges drawn between p vertices, it has been shown (see, for example, Guy (1971)) that:

$$\nu(K_p) \leq \frac{1}{4}\left[\frac{p}{2}\right]\left[\frac{p-1}{2}\right]\left[\frac{p-2}{2}\right]\left[\frac{p-3}{2}\right]$$

where [ ] means 'the largest integer not greater than'. Lower bounds have been suggested by Guy (1972).

It is conjectured (by Saaty (1964), amongst others) that equality holds in the above equation for all values of p, but exact results have only been obtained for $1\leq p\leq 10$, by Guy (1971, 1972). These known values of $\nu(K_p)$ are presented in Table 8-1.

p	1	2	3	4	5	6	7	8	9	10
$\nu(K_p)$	0	0	0	0	1	3	9	18	36	60

Table 8-1: Table giving the known crossing number results for complete graphs

For p=5 and p=6, there are unique optimal drawings of K_p. For p=7 there are five non-isomorphic (ie. distinct, in a graph theoretic manner) optimal drawings. For p=8 there are three, and for p=9 there are thought to be about 200 optimal drawings. The optimal drawings of K_p for $5\leq p\leq 8$ are presented in Erdos & Guy (1973) and Guy (1971, 1972). Harary & Hill (1962/63) present an optimal drawing of K_9.

A planar complete graph (ie. a graph for which $\nu(K_p)=0$ in Table 8-1) can be drawn with the edges as straight line segments and with no cross-overs (proved by Fary (1948)). It is not necessarily the case, however, that non-planar complete graphs can be drawn with the edges as straight line segments with the number of cross-overs given in the table. By convention, conditional independence graphs are usually drawn with straight edges, thus the *rectilinear crossing number*, $\overline{\nu}(G)$, of a graph G should be considered.

The rectilinear crossing number is the minimum number of crossovers with which a graph G may be drawn in the plane using a straight line segment for each edge. It has been shown by Jensen (1971) that:

$$\bar{\nu}(K_n) \leq \left\lceil \frac{7n^4 - 56n^3 + 128n^2 + 48n\left\lceil \frac{n-7}{3} \right\rceil + 108}{432} \right\rceil$$

and equality has been conjectured by Guy (1971).

It should be apparent that $\nu(G) \leq \bar{\nu}(G)$. The rectilinear crossing number has only been determined, by Harary & Hill (1962/63) (verified by Guy (1972)), for complete graphs K_p for which $1 \leq p \leq 9$. These results are presented in Table 8-2. It can be seen from this table that $\bar{\nu}(K_p) = \nu(K_p)$ for $1 \leq p \leq 7$ and $p=9$. For $p \geq 10$ it can be shown that $\bar{\nu}(K_p) > \nu(K_p)$. For example, it has been conjectured, by Guy (1971, 1972), that $\bar{\nu}(K_{10}) = 63$.

p	1	2	3	4	5	6	7	8	9
$\bar{\nu}(K_p)$	0	0	0	0	1	3	9	19	36

Table 8-2: Table giving the known rectilinear crossing number results for complete graphs

For complete bipartite graphs $K_{m,n}$ (ie. graphs having $m+n$ vertices and nm edges joining each of the m vertices to each of the n vertices), it has been shown that:

$$\nu(K_{m,n}) \leq \left\lfloor \frac{m}{2} \right\rfloor \left\lfloor \frac{m-1}{2} \right\rfloor \left\lfloor \frac{n}{2} \right\rfloor \left\lfloor \frac{n-1}{2} \right\rfloor$$

Equality is conjectured by Guy (1971), amongst others, but has only been shown to hold for $1 \leq \min(m,n) \leq 6$. Some examples of known results are shown in Table 8-3. For the smallest complete bipartite graph for which the crossing number is not known, ie. $K_{7,7}$, it is thought that $\nu(K_{7,7}) = 77$, 79, or 81.

m	1	2	3	4	5	6
$\bar{\nu}(K_{m,m})$	0	0	1	4	16	36

Table 8-3: Table giving known crossing number results for bipartite graphs for which $m=n$

It has been conjectured that all bipartite graphs can be drawn with the minimum number of crossovers as given in Table 8-3 with the edges drawn as straight line segments. ie. that $\nu(K_{m,n}) = \overline{\nu}(K_{m,n})$.

A given conditional independence graph will rarely be a complete graph (corresponding to no independence relationships), nor a bipartite graph. Typically, a conditional independence graph will be more akin to a 'general' graph $G(n,k)$ with n vertices and k edges. For such graphs it is conjectured (Erdos & Guy (1973)) that:

$$\frac{c_1 k^3}{n^2} < g(n,k) < \frac{c_2 k^3}{n^2}$$

where $g(n,k)$ is the minimum value of $\nu(G)$ taken over all possible graphs of $G(n,k)$ and c_1 and c_2 are constants.

From Euler's theorem, which relates the number of vertices, edges and *faces* of a plane graph (see, for example, Wilson (1985) or Harary (1969)), it is known that $g(n,3n-6) = 0$ and that $g(n,3n-5) = 1$. In general it is thought, by Erdos & Guy (1973) and Guy (1972), that:

$$g(n,k) = k - 3n + 6 \quad \text{for} \quad 3n-6 \le k \le \min(4n-8, \binom{n}{2})$$

except that $g(7,20)=6$ and $g(9,28)=8$.

The above results for a graph $G(n,k)$ are not known for the rectilinear case.

A few highly specific results and bounds have been found for other graphs, such as n-partite graphs (ie. having all possible edges drawn between n sets of vertices, where $n \ge 3$), but these will not be considered here. Similarly, crossing number results for other surfaces, such as the sphere, the Klein bottle, and the torus, are not considered here since they are not felt to be of relevance to the consideration of conditional independence graphs.

It would therefore seem to be the case that, with a few exceptions, no theoretical results exist for the determination of the minimum number of cross-overs for any given graph. Those results which have been obtained are for particular types of graphs (some complete graphs, some bipartite graphs, and a few other special types of graphs) which may be rarely encountered in practice. The majority of conditional independence graphs may be expected to be incomplete graphs, which do not fall into any of these categories.

8.3 Algorithmic Approaches

An alternative approach is to attempt to find an algorithmic approach to the problem of constructing a graph having the minimum number of cross-overs. The graph theory literature contains a number of algorithms for testing the planarity of a graph and for the construction of a planar graph. Some of these techniques are outlined in Section 8.3.1. However, although these algorithms can be used to *identify* non-planar graphs, they have not been developed for the *construction* of non-planar graphs. It may, however, be possible to modify the algorithms for this purpose, and I have made some suggestions as to how this might be achieved so as to result in the minimum number of cross-overs.

A review of the electronics literature, which might be considered to be a major application of graph theory in the design of Printed Circuit Boards, proved to be surprisingly fruitless. The problem in circuit board design, mentioned by Read (1979) and Gibbons (1985), is to construct an electrical circuit (or circuits) on the plane, ideally using the minimum amount of wire, but with the minimum number of cross-overs, since these can lead to interference problems. However, in the 1960's, when circuit designers were just beginning to come to grips with these problems, the nature of circuit designs changed dramatically with the development of multi-layered circuit boards. With such boards, problems with cross-overs never arise, since there is a third dimension into which the circuit can be directed.

The only algorithm to emerge from the electronics literature of relevance to the problem of constructing a two-dimensional graph with the minimum number of cross-overs is that of Nicholson (1968), which is described in Section 8.3.2.

8.3.1 Algorithms for Planar Graphs

One of the most fundamental characterisations of a planar graph is that it does not contain a sub-graph which is *homeomorphic to* (ie. essentially identical to) K_5 or $K_{3,3}$, these being the smallest complete and bipartite graphs, respectively, which are non-planar. This is a theorem due to Kuratowski, contained in Wilson (1985), amongst others, and can be used as a necessary and sufficient criterion for testing planarity (see, for example, Tutte (1963) and Gibbons (1985)). However, as Read (1970) states, this theorem does not

provide a practical algorithm for testing the planarity of a given graph, nor does it provide a practical algorithm for the construction of a planar drawing of a planar graph.

Read (1970, 1979) and Gibbons (1985), among others, present a number of algorithms which have been developed to test for the planarity of a given graph and for the construction of planar graphs. I describe three of these approaches below, together with suggestions as to how they might be modified for the construction of non-planar graphs.

One of the algorithms described by Read is due to Weinberg (1972). This involves construction of a spanning tree of the graph, which is drawn in the plane in any convenient manner. Then an attempt is made to add the remaining edges so that no crossings occur. This may not be possible, even if the graph is planar, depending on the position in which the tree was originally drawn. However, by rotating parts of the graph, it may be possible to accommodate the edges. This method is quite readily extendible to the construction of non-planar graphs with the minimum number of cross-overs. However, the algorithm is not suited for computer implementation although it can be implemented by hand quite readily.

A different approach is taken by Fisher & Wing, again described by Read. This involves starting with a circuit of the graph. If the graph is planar, this circuit can be deformed into a circle. If the vertices of the circle are removed, and any edges adjacent to these vertices also removed from the graph, this will leave a number of connected components. First of all, the planarity of each of these components must be determined and planar representations obtained. Then it must be determined whether each component needs to be placed inside or outside the circle to ensure that none of the link edges between the vertices in the circuit and the components of the graph cross-over when they are redrawn. If the graph is non-planar then this cannot be achieved. This algorithm reduces the problem to a number of similar problems involving smaller graphs and can be readily programmed, although fairly slow. If one wished to modify this algorithm for the construction of a non-planar graph, then one would try to place the components of the graph so as to minimise the number of cross-overs.

Another algorithm cited by Read (1979) is due to Lempel, Even & Cederbaum. This involves starting with a single vertex and building up a drawing of the graph in the plane by the addition of vertices and edges. This results in a fast computer-implementable algorithm.

If there are less than nine edges or less than five vertices, or if $E \leq 3V-6$, where E is the number of edges and V is the number of vertices, then the graph must be planar (due to Euler's theorem). For incomplete graphs involving just 5, 6, 7 or 8 vertices and any number of edges, I have considered how one might construct graphs having the minimum number of cross-overs by considering the optimal drawings of complete graphs having these numbers of vertices.

The complete graph on 5 vertices, K_5, has just one cross-over. If the graph to be constructed have five vertices and at least one edge missing, then by mapping one of the missing edges in the graph to be constructed onto one of the edges involved in the cross-over in the complete graph, a representation of the graph is obtained which has no cross-overs. This follows from Euler's theorem, since $E=3V-6$ in this case. The complete graph on 6 vertices, K_6, has three cross-overs. With more than 6 edges missing, it is possible to map the missing edges to edges in the complete graph in order to obtain a planar graph. However, if there are between 3 and 6 edges missing, it may or may not be possible to map the missing edges to edges in the complete graph to give a planar graph, depending on the vertices involved. If there are less than 3 edges missing, it may be possible to map the missing edges to the edges in the complete graph to minimise the number of cross-overs, but it will not be possible to construct a planar graph. For the complete graph on 7 vertices, K_7, at least 13 edges must be missing for a planar graph, and if fewer than 6 edges are missing this will always result in a non-planar graph. For the complete graph on 8 vertices, K_8, the corresponding figures are at least 20 edges missing for planarity and fewer than 10 edges missing for non-planarity. However, by careful mapping of the missing edges onto the edges involved in the cross-overs in the complete graphs, it should be possible to minimise the number of cross-overs. In theory this approach is readily programmable, but only appropriate for use with the handful of complete graphs for which the optimal drawings are known. For more than 8 vertices, the number of optimal drawings is not known.

Although I have presented a number of algorithms, due to Read, for the construction of planar graphs and have suggested how they may be modified for the construction of non-planar graphs, and I have also suggested an algorithm based on optimal drawings of complete graphs, these algorithms do not guarantee the 'best' solution in terms of the minimum number of cross-overs. The problem of knowing what

the best solution is would still remain since, as stated in the previous section (Section 8.2), there are no useful theoretical results for incomplete graphs.

8.3.2 Nicholson's Algorithm

Nicholson (1968) presents an algorithm, based upon permutation of the edges and vertices of a graph, which purports to result in the minimum, or otherwise a near-minimum, number of cross-overs. Hashimoto & Noshita (1971) explain why it is not always possible to obtain a representation having the minimum number of cross-overs.

Nicholson's algorithm involves taking a given graph, defined by a set of vertices and edges, and deforming it so that the vertices lie along a horizontal line with the edges between the vertices forming semi-circles above or below this line. This same linear representation is used by Saaty (1964) in considering the minimum number of intersections in complete graphs.

Given this linear representation of the graph, the edges and the vertices are then permuted until the number of crossings is a minimum. However, for a graph with n vertices and m edges, there are $2^{m-1}(n-1)!$ possible permutations of the vertices and edges. Nicholson's algorithm reduces the number of permutations to be tested by optimising the initial placement of the vertices and edges along the line, and by controlling the permuting of the vertices and edges in relation to a given function.

Nicholson's algorithm does not always result in the minimum number of cross-overs, but it does perform quickly and well. For large regular graphs, Nicholson obtained results which were equal to or within 2 cross-overs of the known results or lower bounds, and for random graphs the algorithm performed far better than Monte Carlo methods.

I have implemented the algorithm in a simple FORTRAN program, so that the algorithm can be readily applied to a graph by inputting the number of vertices and the connections which exist between pairs of vertices in the graph. The output consists of an ordering of these vertices, and a listing of those edges which must be drawn "above" an imaginary horizontal line drawn through the vertices in their final order, and those which must be drawn "below" this line. This solution is not necessarily unique.

The solution can be translated into two dimensions in order to obtain a more conventional representation of the graph, having the minimum, or a near-minimum, number of cross-overs. The vertices can be located freely on the plane, subject to the constraint that it must be possible to draw a non-intersecting imaginary line between the

vertices in the order in which they occur in the final permutation. The edges are then drawn either side of this imaginary line, subject to the constraint that all edges "above" the horizontal line which would be drawn through the vertices in the one-dimensional solution are drawn to one side of this imaginary line, and that all edges "below" the horizontal line which would be drawn in the one-dimensional solution are drawn to the other side of this imaginary line. This is equivalent to deforming the linear graph representing the solution.

There is obviously considerable freedom in the choice of positions of the vertices. One constraint which may be imposed for the construction of conditional independence graphs is that the edges between the vertices should all be straight lines. This may reduce the amount of freedom available in the location of the vertices, and make the graph harder (or even impossible) to construct by hand. It would almost certainly be impossible to modify the programmed algorithm to always give a graph with the minimum, or a near-minimum, number of cross-overs which could be constructed in two dimensions with straight-lined edges.

8.4 Summary

This chapter has been concerned with the construction of conditional independence graphs with the minimum number of cross-overs as an aid to legibility.

Some graph theoretic results for crossing numbers were found which may be of use for the construction of a small number of special, typically complete, graphs, but for the construction of more general incomplete graphs, no useful theoretical results exist.

It was instead decided to try an algorithmic approach to the construction of graphs having the minimum number of cross-overs. Most algorithms in the graph theory literature have been developed for the identification and construction of planar graphs (ie. graphs having no cross-overs), but suggestions were made as to how these algorithms could be modified for the construction of a non-planar graph with, hopefully, a small number of cross-overs.

A review of the electronics literature uncovered a readily programmable algorithm due to Nicholson for the construction of a graph having the minimum, or a near-minimum, number of cross-overs. However, the one-dimensional solution with curved edges which is obtained would not necessarily correspond to a two-dimensional

conditional independence graph with straight edges having the same number of cross-overs.

Thus it is concluded that no straightforward theoretical or algorithmic solution exists for the problem of constructing any conditional independence graph with the minimum number of cross-overs. I shall, however, suggest a way around this problem in Chapter 11.

9. Edge Coding of Interactions in Conditional Independence Graphs

9.1 Introduction

As has already been stated (in Chapter 7), in a conditional independence graph constructed to represent a hierarchical log-linear model, the absence of an edge indicates that the pair of variables corresponding to the two vertices which are not linked are conditionally independent, given the levels of the other variables. If there is an edge, this indicates that the variables are associated, but provides no further information about the nature of this association. In particular, no information is given about whether this association is the same or different between levels of the other variables; ie. about the order of any interactions (other than two-way) involving these two variables. It is, of course, not appropriate to talk about different 'levels' of variables, nor of higher order associations, in the case of covariance selection models for continuous data; hence this chapter is restricted to the consideration of (hierarchical) log-linear models for discrete contingency table data.

In the remainder of this chapter, an edge coding scheme for conditional independence graphs for log-linear models will be described, which provides additional information about the nature of the associations between pairs of variables. The use of edge codes is based on the representation of interaction type by line styles described in Section 6.2.3 of Chapter 6. Through the use of the edge coding scheme it is possible to represent a larger number of models without ambiguity than can be achieved with the existing form of the conditional independence graph. However, those situations in which models can still not be represented without ambiguity are discussed. The work contained in this chapter was originally presented in Cottee & Hand (1989).

9.2 Interpretation of Interactions in Conditional Independence Graphs

The independence graph drawn in Figure 9-1 corresponds to the complete graph drawn on three vertices. Since there are no edges missing from this graph, no conditional independence statements may be made. The edges in the graph imply that A and C are not

independent given the levels of B, nor are A and B independent given the levels of C, nor B and C given the levels of A. What cannot be determined from the graph in Figure 9-1, however, is whether the association between A and C, for example, has the same (non-zero) value for each level of B, or whether the AC association differs between levels of B. If the AC association has the same value for each level of B, then the graph corresponds to a model with no ABC interaction, and thus corresponds to the model with generating class $\{[AB]\ [AC]\ [BC]\}$. If the AC association does differ between levels of B, then the graph corresponds to the model containing the ABC interaction, which has generating class $\{[ABC]\}$.

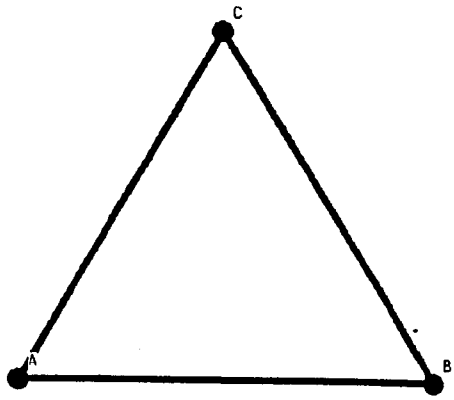

Figure 9-1: Independence graph for {[AB] [AC] [BC]} and {[ABC]}

For a graphical model, the cliques (ie. the maximally complete sub-graphs) of the independence graph correspond to the elements of the generating class of the model. Thus, in the context of graphical models, the graph in Figure 9-1 represents the graphical model with generating class $\{[ABC]\}$, and there is no independence graph which corresponds uniquely to the model with generating class $\{[AB]\ [AC]\ [BC]\}$, since this is not a graphical model.

A non-graphical model may be regarded as a sub-model of the graphical model having the same independence graph, since the conditional independence relationships which hold for the graphical model, and which can be read from the independence graph, will also hold for any sub-models. However, for a given independence graph, it would not

be possible to determine whether the edges in the graph are intended to correspond to associations in the graphical model, or to associations in a non-graphical sub-model.

9.3 Coding of Edges to Distinguish Between Models Involving Three Variables

As has been seen, for an independence graph drawn on three vertices the absence of an edge implies that there is no association between two variables at any of the levels of the other variable. The presence of an edge may imply the presence of a two-way association, if that association is the same for each level of the third variable. Alternatively, an edge may imply the presence of the three-way association, if the two-way association represented by that edge differs between levels of the third variable. The presence of the three-way association would, in turn, necessitate the presence of all two-way interactions implied by the three-way interaction, since I am concerned with hierarchical models.

I therefore propose using a new edge 'code', drawn as a dashed line, to represent the situation when the corresponding two-way association is different at each level of the third variable, and will retain the existing edge 'code', a continuous line, to represent the situation when the corresponding 2-way association is the same at each level of the third variable. Using these edge codes, Figure 9-2 shows the conventional independence graph which now represents the model with generating class {[*AB*] [*AC*] [*BC*]}, and the graph in Figure 9-3 represents the model with generating class {[*ABC*]}.

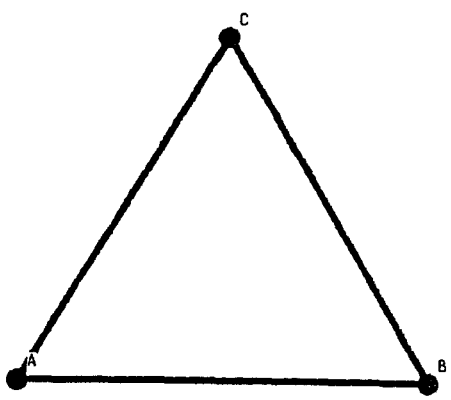

Figure 9-2: Interaction graph for {[AB] [AC] [BC]}

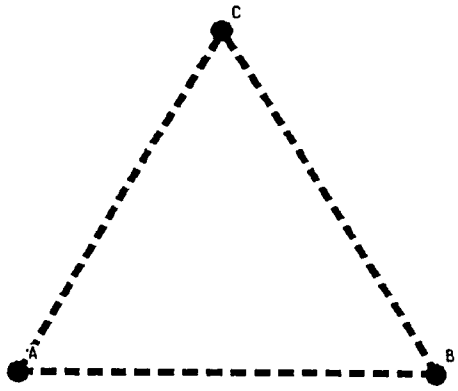

Figure 9-3: Interaction graph for {[ABC]}

Note that there are logical constraints on the use of the edge codes. For three variables it would not be possible to use both dashed edges (indicating the presence of the 3-way interaction) and continuous edges (indicating the absence of the 3-way interaction) in the same graph on three vertices/variables.

In using different edge codes to distinguish between different interaction types, my primary concern is with the display of the conditional interactions in the model (ie. the values of the pair-wise interactions, conditional on the levels of the other variables), rather than with the display of the pair-wise conditional independencies. The graph drawn using the edge codes will therefore be referred to as a *conditional interaction graph*, or *interaction graph* for short.

9.4 Coding of Edges to Distinguish Between Models Involving Four or More Variables

9.4.1 Four Variables

The notion of an interaction graph, using different edge codes to distinguish between different types of interaction in order to distinguish between different models, can be extended to four or more variables.

For example, Figure 9-4 shows an independence graph (ie. not using the edge codes) on four vertices. The corresponding graphical model is {[*ABD*] [*ACD*]} and the same diagram would be used to represent all of the following: {[*AB*] [*BD*] [*CD*] [*AC*] [*AD*]}, {[*ACD*] [*AB*] [*BD*]}, {[*ABD*] [*AC*] [*CD*]}, as well as {[*ABD*] [*ACD*]}. That is, from an independence graph it is not possible to distinguish between these four models. However, as Figures 9-5, 9-6, 9-7 and 9-8 show, using interaction graphs (with edge codes) all four of these models are quite distinguishable.

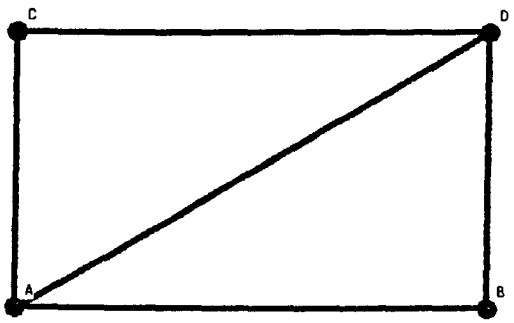

Figure 9-4: Independence graph for {[ABD] [ACD]}

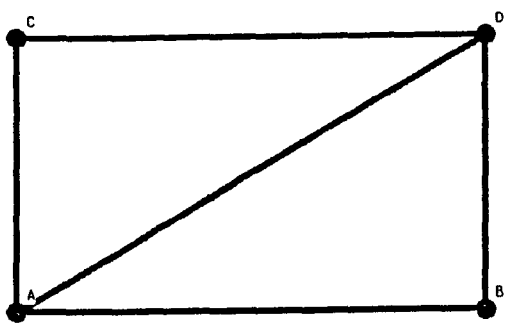

Figure 9-5: Interaction graph for {[AB] [AC] [AD] [BD] [CD]}

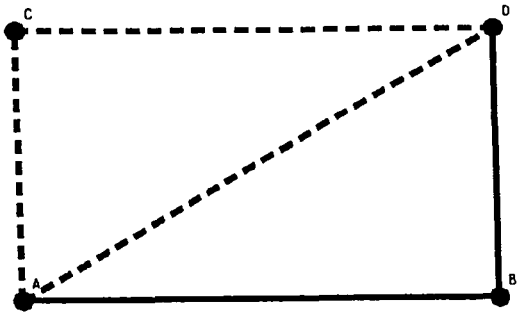

Figure 9-6: Interaction graph for {[ACD] [AB] [AD]}

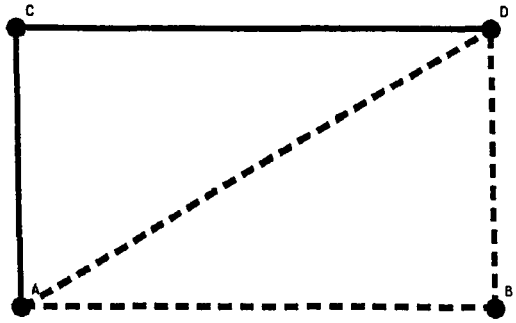

Figure 9-7: Interaction graph for {[ABD] [AC] [CD]}

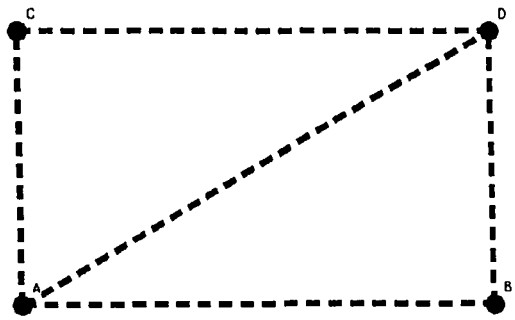

Figure 9-8: Interaction graph for {[ABD] [ACD]}

To distinguish the model with generating class {[*ABCD*]} from models involving only 3-way interactions and which have all pair-wise associations present, it is necessary to introduce another edge code. An alternately dashed-dotted line is used to indicate such a 4-way interaction. Figure 9-9 shows the model {[*ABCD*]}, while Figure 9-10 shows the graph representing the models with at least three 3-way interactions present, ie. {[*ABC*] [*ABD*] [*ACD*] [*BCD*]}, {[*ABC*] [*ABD*] [*ACD*]}, {[*ABC*] [*ABD*] [*BCD*]}, {[*ABC*] [*ACD*] [*BCD*]}, and {[*ABD*] [*ACD*] [*BCD*]}.

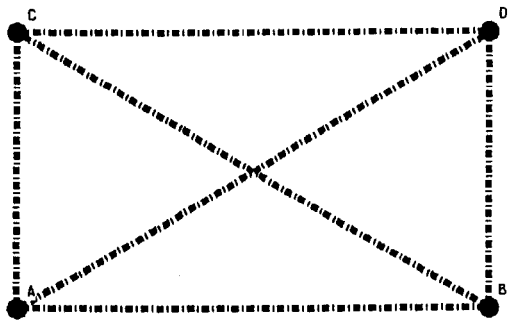

Figure 9-9: Interaction graph for {[ABCD]}

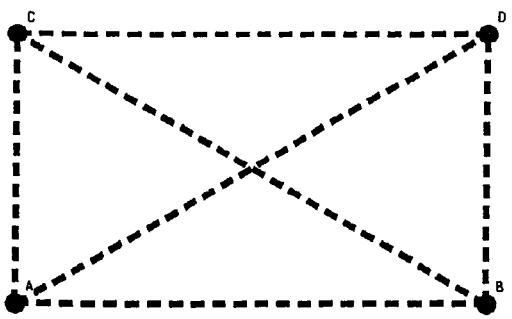

Figure 9-10: Interaction graph for models with at least three 3-way interactions present

As noted above, in a graphical model the cliques of the independence graph correspond to the elements of the generating class of the model represented. Non-graphical sub-models of the graphical model may have the same independence graph as

this model, thus the independence graph will not allow us to distinguish between these models. As can be seen in Figure 9-10, different models may also have the same interaction graph. A similar convention is therefore proposed for interaction graphs; ie. that the set of complete sub-graphs implied by the edge codes used in the interaction graph correspond to the elements of the generating class of the model represented. In this way, the graph of Figure 9-10 is taken to represent the model {[*ABC*] [*ABD*] [*ACD*] [*BCD*]}, even though models with only three of the 3-way interaction terms in the generating class have the same interaction graph. It is suggested that the model represented by the interaction graph be referred to as a *graphical interaction model*.

9.4.2 More Than Four Variables

In the previous sub-section, a graphical interaction model was defined as one for which each complete sub-graph of the interaction graph, implied by the edge codes, corresponds to an element of the generating class of the model. A non-graphical interaction model is therefore a sub-model of the graphical interaction model having the same interaction graph. Consider, for example, the interaction graph drawn on five vertices in Figure 9-11. This graph has been constructed to represent the model with generating class {[*ABCD*] [*ABCE*] [*ABDE*] [*CDE*]}, using the usual edge codes. However, the graphical interaction model corresponding to this graph, with the elements of the generating class of the model corresponding to the complete sub-graphs implied by the edge codes in the graph, is {[*ABCD*] [*ABCE*] [*ACDE*] [*ABDE*] [*BCDE*]}. The model represented is therefore a (non-graphical) sub-model of the graphical interaction model corresponding to the graph.

For three and four vertices, the complete sub-graphs of the interaction graph were each formed by a single type of edge code. For example, a continuous edge between two vertices corresponded to a 2-way interaction in the generating class, three dashed edges between three vertices corresponded to a 3-way interaction, and six dashed-dotted edges between four vertices corresponded to a 4-way interaction.

For five or more vertices, the complete sub-graphs corresponding to 3,...,(n–2)-way interactions (where n is the number of vertices) may not be formed by a single type of edge code. For example, in Figure 9-12 the *ADE* sub-graph is formed by two different edge codes — a dashed line for the *AE* and *DE* edges, and a dashed-dotted line for the *AD* edge. This situation arises as a result of the pair-wise nature of the edges: if an edge

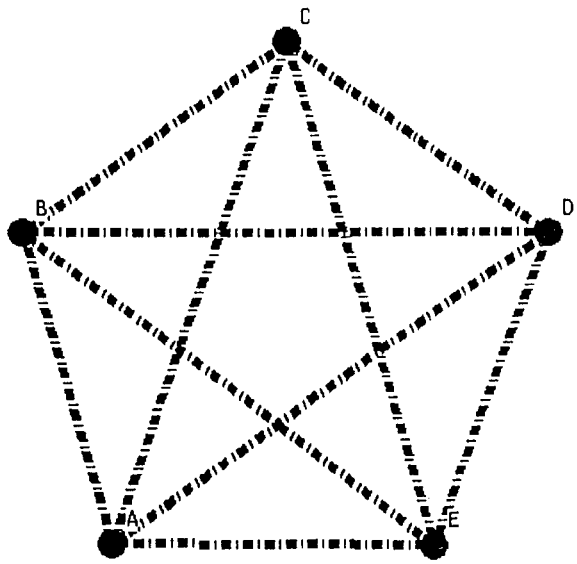

Figure 9-11: Interaction graph for {[ABCD] [ABCE] [ABDE] [CDE]}

(corresponding to a 2-way association) is contained within at least two elements of the generating class which are of different orders, only one code can be used to represent the edge. The code chosen is the one corresponding to the highest interaction containing that edge. Therefore, in Figure 9-12, the *AD* edge is contained within both the *ADE* and *ABCD* sub-graphs, and the dashed-dotted line code is used, corresponding to the 4-way interaction.

In general, if there are at least two terms of different orders, each greater than 2-way, in the generating class of the model, with at least one edge in common, then the complete sub-graphs of terms other than the highest are not formed using a single type of edge code.

The above is only applicable to interaction graphs drawn on five or more vertices. If there are three vertices, then there is either one 3-way interaction, or up to three 2-way interactions. If there are four vertices, then there is either one 4-way association, or some combination of 3-way and 2-way interactions. However, the generating class of a (hierarchical) model with four variables will never include a 2-way interaction which is contained within a 3-way interaction, and so only one edge code is ever needed for each complete sub-graph.

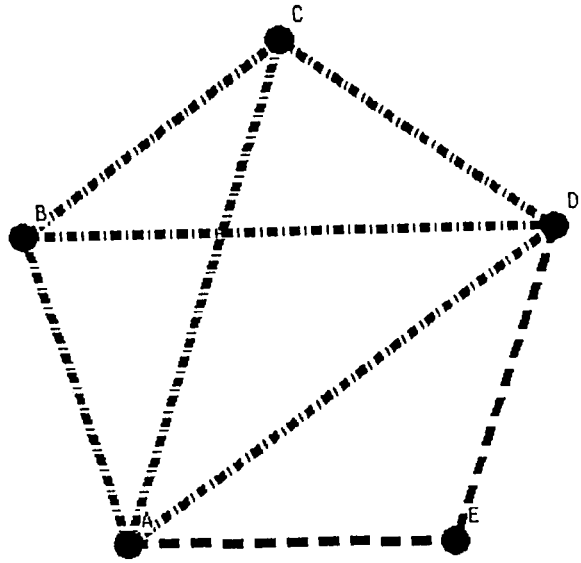

Figure 9-12: Interaction graph for {[ABCD] [ADE]}

For an interaction graph drawn on five or more vertices, a complete sub-graph on s vertices must have at least one edge corresponding to the s-way interaction, else the sub-graph will correspond to a higher-order interaction. The other $\frac{1}{2}(s(s-1))-1$ edges may correspond to s-way or higher order interactions.

Use of a particular code for an edge indicates that the association between the two variables joined by that edge is different between the levels of p other variables. For a continuous line, corresponding to a 2-way interaction, p=0; for a dashed line, corresponding to a 3-way interaction, p=1; and for a dashed-dotted line, corresponding to a 4-way interaction, p=2. Therefore, a given edge will form part of a complete sub-graph on p+2 vertices, corresponding to a (p+2)-way interaction in the generating class of the model represented.

In Figure 9-12, the dashed edge between A and E, for example, implies that there is at least one other variable X which will form a complete sub-graph on three vertices AEX. The other edges AX and EX must correspond to 3-way or higher associations. In this case X=D, and the model represented corresponds to the graphical interaction model.

In Figure 9-13, the dashed edge between C and E implies that there is at least one other variable X which will form a complete sub-graph CEX. The other edges CX and EX

must correspond to 3-way or higher order interactions. In this case X=A, B, or D, and the other two edges, in each case, correspond to 4-way interactions. The graphical interaction model corresponding to this graph is therefore {[$ABCD$] [$ABDE$] [CDE] [BCE] [ACE]}, and the model represented by the interaction graph is a non-graphical interaction sub-model of this model.

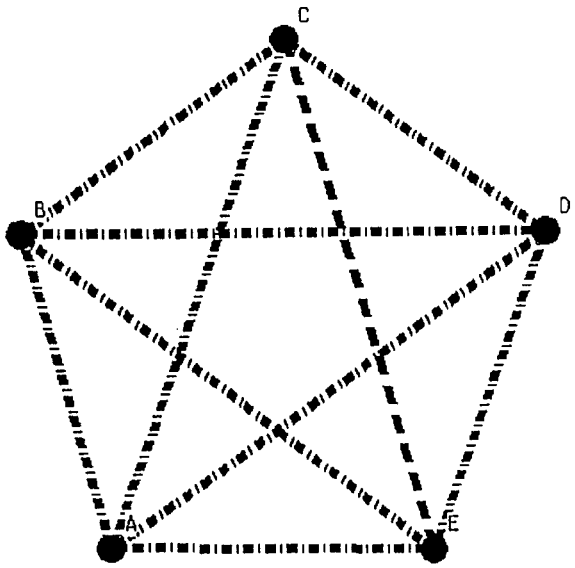

Figure 9-13: Interaction graph for {[ABCD] [ABDE] [ACE]}

In general, if there are n vertices the largest possible order interaction will be n-way, so that a maximum of (n–1) edge codes will be required. However, if the highest order interaction in the generating class, corresponding to the largest clique of the graph, involves s variables, then a maximum of (s–1) edge codes will be required, corresponding to the (2,...,s)-way associations (where $s \leq n$). A complete sub-graph on s vertices will have ½(s(s–1)) edges, so if s is very large it may become difficult to determine the complete sub-graphs corresponding to a particular edge. However, sparseness of such graph structures has been stressed as likely (for example, by Lauritzen and Spiegelhalter (1988)), so in practice not many edge codes may be needed, and the complete sub-graphs may not be too large.

9.5 Summary

In this chapter, an edge coding scheme is proposed for use with conditional independence graphs for discrete data. By adopting this edge coding scheme, a conditional interaction graph is obtained which is more informative than the corresponding conditional independence graph, and which can be used to provide an unambiguous visual summary for a larger number of models than can be achieved using the independence graph. In other words, the interaction graph is useful for the representation of some non-graphical as well as graphical models. However, there are still some non-graphical models which cannot be represented through the use of interaction graphs without ambiguity.

The interaction graph approach is considered again in Chapter 11 where a solution to the problem of ambiguity is proposed which can be applied to both independence and interaction graphs.

10. Encoding Strength of Association in Conditional Independence Graphs

10.1 Introduction

In the preceding chapter (Chapter 9), some consideration was given to the problem of how the nature of an interaction could be incorporated into conditional independence graphs by the use of an edge coding scheme. This chapter is also concerned with the incorporation of additional information in conditional independence graphs. The concern of this chapter is not, however, with the encoding of the *nature* of the interaction, but with the encoding of the *strength* of the association.

In encoding the nature of the association, it was only appropriate to consider models for discrete data (ie. hierarchical log-linear models) since models for continuous data (ie. covariance selection models) involve only pair-wise associations. However, in encoding the strength of the association, it is more straightforward to consider models for continuous data since the pair-wise associations correspond uniquely to edges in the conditional independence graph, whereas for models for discrete data the edges in the graph may correspond to more than one interaction. Thus this chapter is mainly concerned with the encoding of the strength of associations in conditional independence graphs for continuous data, although some consideration will be given to the discrete data case.

In the first two sections (Sections 10.2 and 10.3), measures of strength of association will be considered for continuous data and discrete data respectively, although these have also been considered in Chapter 7. I shall then go on to consider, in Section 10.4, the encoding of strength of association by distance, and in Section 10.5 will consider the encoding of strength of association by edge style. This leads on to the execution of a graphical perception experiment, which is described in detail in Section 10.5.1. Finally, encoding the sign of the association is considered in Section 10.6.

10.2 Measures of Strength of Association for Continuous Data

For continuous multivariate Normal (MvN) data, which is the concern of the bulk of this chapter, associations are pair-wise and correspond directly to the edges in the

conditional independence graph, such that zero pair-wise associations correspond to the missing edges in the graph.

As was indicated in Chapter 7, the strength of each association may be inferred from the value of the partial correlation $\rho_{ij.K}$ between a pair of variables i and j partialling on the rest of the variables K, or from the value of the edge exclusion deviance (e.e.d.). These two values are monotonically related as:

$$\text{e.e.d.} = -N \ln(1-(\rho^2_{ij.K}))$$

where N is the number of observed units.

In general, if the e.e.d. is sufficiently small (ie. $< \chi^2_{[1]} = 3.841$, $\alpha = 0.05$), thus corresponding to a small partial correlation coefficient, then no edge is drawn between the corresponding pair of variables in the graph. However, the greater the absolute value of the partial correlation, the greater the value of the e.e.d. and the stronger the association between the corresponding pair of variables.

10.3 Measures of Strength of Association for Discrete Data

In a conditional independence graph constructed for discrete data, each edge may form part of a two-way association, or of one or more three-way associations, and/or of one or more four-way associations, etc. Thus measures of strength of association which are related to the parameter values of the log-linear model (for example, their value or significance) could not readily be encoded in the graph since a single edge may take on multiple values, and three or more edges would be involved in three-way or higher order interactions.

There is a concept comparable to the edge exclusion deviance used in covariance selection models for continuous data which can be used with discrete data — namely the exclusion deviance based on the maximum likelihood test statistic. This is used within the context of graphical modelling to investigate the effect of removing each edge in turn from the graph corresponding to a fitted graphical model. Because of the computational complexity in the step-wise fitting of numerous graphical models, the exclusion deviance is usually found by using MIM or GLIM to fit the $\binom{k}{2}$ graphical models which can be derived by dropping one edge from a graph having k edges. A comparable concept to the matrix of edge exclusion deviances for continuous data would be a matrix of exclusion

deviances for the edges formed by each pair of variables in the final fitted graphical model. A missing edge would have zero deviance; otherwise the greater the deviance, the stronger the association.

Note that for continuous data, edge exclusion deviances may be non-significant but are unlikely to be exactly zero. For discrete data, the values of the edge exclusion deviances depend on the base model considered — this may be the saturated model, in which case some deviances may be non-significant but are unlikely to be exactly zero; or this may be a model fitted using a step-wise procedure, perhaps using an iterative procedure starting with the saturated model and dropping one edge at a time according to the value and significance of its exclusion deviance. The exclusion deviances of the missing edges in a model fitted in this way will be zero, whilst the deviances of the remaining edges may no longer resemble the magnitude of these edges in the saturated model from which the fitted model was derived. Moreover, when deleting each edge in turn, it is always the graphical model corresponding to the resultant graph which is then fitted — non-graphical models do not have a place in the model-fitting procedure.

For small contingency tables, typically 2×2, one well-known measure of association is the cross-product ratio, or *cpr* (see Bishop *et al* (1975), Whittaker (1990)). Although there are other measures of association available, based on the χ^2 statistic (see, for example, Everitt (1992)), these are more an index measure for assessing the significance of the association exhibited in a given table or for comparing the degree of association in different tables and, unlike the cross-product ratio, do not constitute a direct measure of the association between the variables in the table.

For a 2×2 table of proportions as shown in Figure 10-1:

$$cpr = \frac{p_{11}p_{22}}{p_{12}p_{21}}$$

If the row and column variables (*A* and *B* in Figure 10-1) are interchanged, the value of the cross-product ratio remains unchanged. If the rows of *A*, or the columns of *B*, alone are interchanged, the new value for the cross-product ratio will be the reciprocal of the previous value.

For perfect independence, *cpr*=1; otherwise the cross-product ratio may take positive values up to ∞. It is more usual, however, to consider the natural logarithm of the cross-product ratio. For perfect independence *ln cpr*=0; otherwise the natural log of the

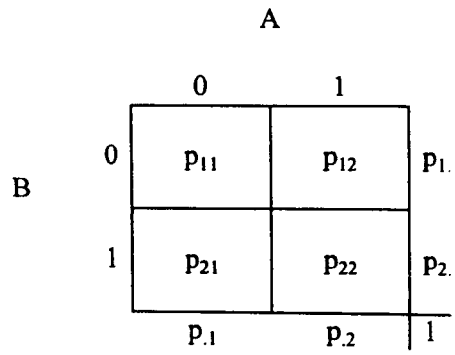

Figure 10-1: Figure showing the proportions in each cell used in the calculation of the cross-product ratio for a 2x2 contingency table

cross-product ratio may take positive or negative values to ∞. The natural log of the reciprocal of the cross-product ratio, corresponding to an interchange of the rows or columns of one of the variables, will take the same value as the natural log of the cross-product ratio of the untransformed contingency table, but will have the opposite sign.

The cross-product ratio is a strictly pair-wise concept for use with variables having two levels each. However, for a 2×2×2 table, for example, it is possible to calculate the cross-product ratio for two of the variables conditioning on the levels of the third variable. For larger tables one might consider collapsing the table across the levels of some of the variables to give a single 2×2 table or a series of 2×2 tables for which the cross-product ratio can then be calculated. However, in collapsing contingency tables and considering only the margins, erroneous conclusions may be reached due to misleading associations resulting from the act of collapsing the tables. This phenomenon is commonly known as Simpson's paradox (Simpson (1951)).

The cross-product ratio can also be directly related to the interaction terms of the log-linear model for two or three variables (see Whittaker (1990)).

Although the following sections concerning the encoding of strength of association in conditional independence graphs will be illustrated using continuous data examples, for which the strength of the associations between variables can be readily calculated as either the negative partial correlation coefficient or the edge exclusion

deviance, it is possible to adapt the approaches described for use with pair-wise measures of strength of association for discrete data (ie. the exclusion deviance or the cross-product ratio). The problem of representing the parameter values of the interactions in a hierarchical log-linear interaction model for discrete data is deferred until Chapter 11.

10.4 Encoding Strength of Association by Distance

Whittaker (1988), and Whittaker, Iliakopoulos & Smith (1988), consider the application of principal components analysis (PCA) to graphical modelling for use with large numbers of variables, as has already been mentioned in Section 7.6 of Chapter 7. Use of PCA in this context, in providing a rationale for the placement of the vertices, makes it possible to interpret the graph in terms of the strengths of the associations between pairs of variables and may be applied to any number of variables.

In order to encode strength of association in a conditional independence graph, PCA is applied to the matrix of negative partial correlations in the following way:

1. Calculate the negative partial correlation matrix (by application of the naive covariance selection procedure described in Chapter 7).
2. Calculate the edge exclusion deviances as $\text{e.e.d.} = -N \ln(1-(\rho^2_{ij.K}))$.
3. Apply PCA to the negative partial correlation matrix.
4. Plot the vertex corresponding to each variable in the space defined by the first two principal components.
5. Draw edges in the graph corresponding to those pairs of variables for which the edge exclusion deviance exceeds the critical value.

In this way, the strength of each association is encoded by the distance between pairs of vertices/variables.

Whittaker (1988) illustrated this approach by applying PCA to the matrix of negative partial correlations obtained using Jeffers' (1967) pit-prop data. This data set consists of a matrix of correlations between 13 variables and will be considered again in more detail in Chapter 12. By calculating the matrix of negative partial correlations and applying PCA in the manner described above, the conditional independence graph presented in Figure 10-2 is obtained. Of the 78 possible edges in the complete graph

drawn on 13 vertices, 25 edges correspond to a significant edge exclusion deviance (ie. $\geq \chi^2_{[1]} = 3.841$, $\alpha=0.05$) and are therefore included in the graph. It would be possible to use a higher level of significance, for example $\alpha=0.01$, for which $\chi^2_{[1]} = 6.635$, in which case there would be 14 edges in the resultant independence graph.

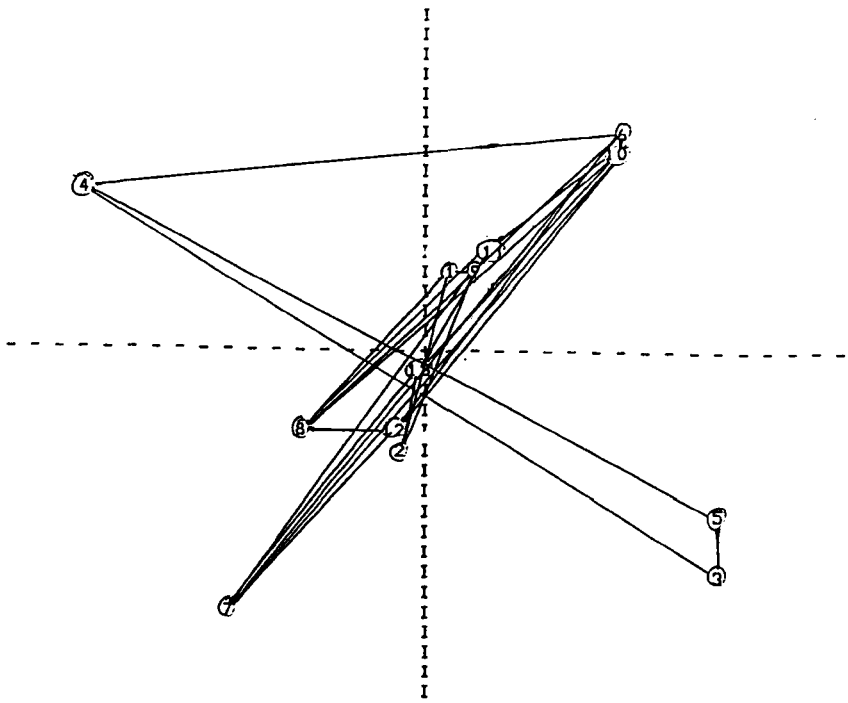

Figure 10-2: Pit-prop data: PCA plot constructed for the negative partial correlation matrix

One would anticipate that the closer a pair of vertices/variables are plotted to each other, the stronger the association between them. Thus only short edges would be expected in the graph, since longer edges would correspond to pairs of variables with insignificant edge exclusion deviances. However, in the graph in Figure 10-2, there are some long edges which correspond to pairs of variables having large negatively signed negative partial correlations. Since a large negative partial correlation will have a large (always positive) value for the corresponding edge exclusion deviance irrespective of the sign of the negative partial correlation, these long edges are significant and are therefore included in the graph.

I have considered further the application of PCA for the encoding of strength of association in conditional independence graphs. PCA has been applied to the correlation matrix and the partial correlation matrix, and to the absolute values of the correlation matrix and the partial correlation matrix for Jeffers' pit-prop data, and the results plotted in the space determined by the first two principal components using SPSS. The result of applying PCA to the negative partial correlation matrix has already been presented in Figure 10-2. The result of applying PCA to the absolute partial correlation matrix is presented in Figure 10-3.

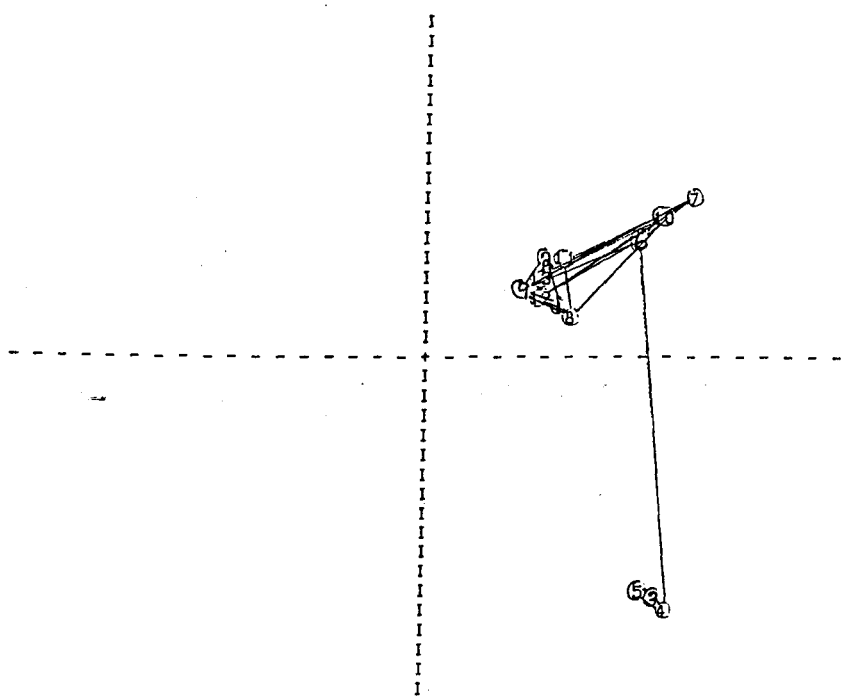

Figure 10-3: Pit-prop data: PCA plot constructed for the absolute partial correlation matrix

Since the correlation matrix is inverted in order to obtain the matrix of negative partial correlations, the absolute values of which are monotonically related to the values of the edge exclusion deviances, it seems more appropriate to consider the partial correlation matrices than the correlation matrices. Furthermore, it seems more appropriate to consider the absolute partial correlation matrix because of its relationship to the matrix of edge exclusion deviances, which means that very short edges should correspond to

highly significant/strongly associated edges and that the problem of very long edges corresponding to very strong (negative) associations does not arise.

The graphs obtained in applying PCA to the correlation matrices are not presented here, but similarities were found between the graphs for the correlation matrix and the partial correlation matrix, and between the graphs for the absolute correlation matrix and the absolute partial correlation matrix. For the correlation matrix and the negative partial correlation matrix (Figure 10-2), a mixture of long and short edges was found in the resultant independence graphs, whereas for the absolute correlation matrix and the absolute partial correlation matrix (Figure 10-3), the edges in the resultant independence graphs were mainly very short. In the latter two plots, there is a tendency for the shorter edges to correspond to stronger associations (which have larger values for the edge exclusion deviance), but one needs to beware the low proportion of the total variation in the data explained by the first two dimensions (38% for the graph corresponding to the negative partial correlation matrix, and 40% for the graph corresponding to the absolute partial correlation matrix). All four plots showed some grouping of the variables.

PCA has also been applied to Poole's kangaroo skeleton data set, presented in Andrews & Herzberg (1985). This data set consists of 18 measurements made on each of 148 kangaroo skeletons. Again, this data set is considered in more detail in Chapter 12. For the purposes of this analysis, no distinction was made between the different sexes or of the different species of kangaroo, and missing values were replaced by the intra-species intra-variable mean prior to the calculation of the correlation matrix. The independence graphs obtained by plotting the vertices in the space defined by the first two principal components for PCA applied to the negative partial correlation matrix and the absolute partial correlation matrix are presented in Figures 10-4 and 10-5 respectively.

Again, it can be seen from Figure 10-4 that application of PCA to the negative partial correlation matrix results in some very long edges corresponding to edges with a large negative partial correlation and high edge exclusion deviance, and it is quite difficult to discern any structure in the data. However, the first two principal components account for just 27% of the variation in the data (eight principal components would be required to account for 70%).

The independence graph obtained by application of PCA to the absolute partial correlation matrix (see Figure 10-5) also accounts for a very small proportion of the total variation in the data (28%, with nine principal components required to account for 70%).

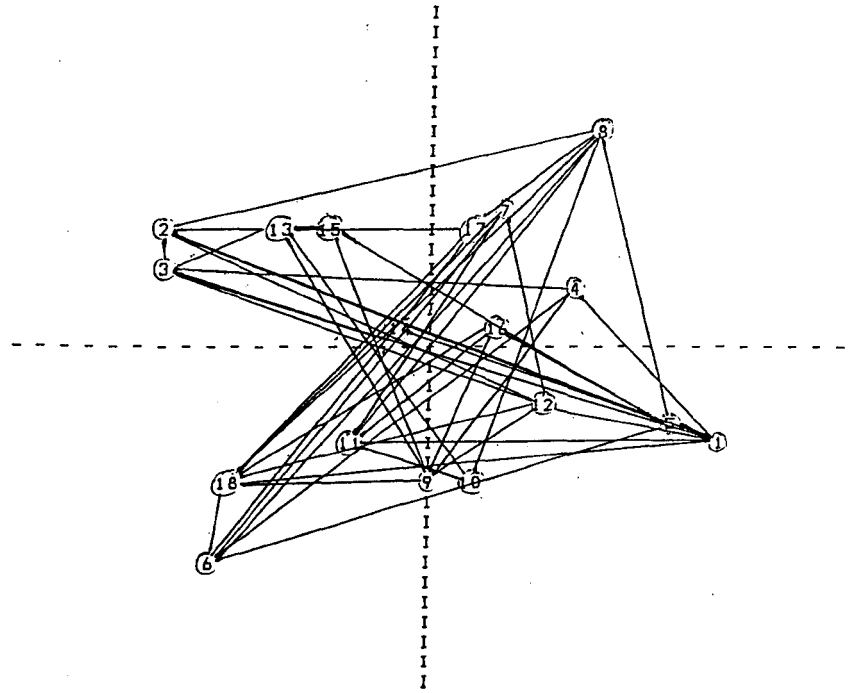

Figure 10-4: Kangaroo skeleton data: PCA plot constructed for the negative partial correlation matrix

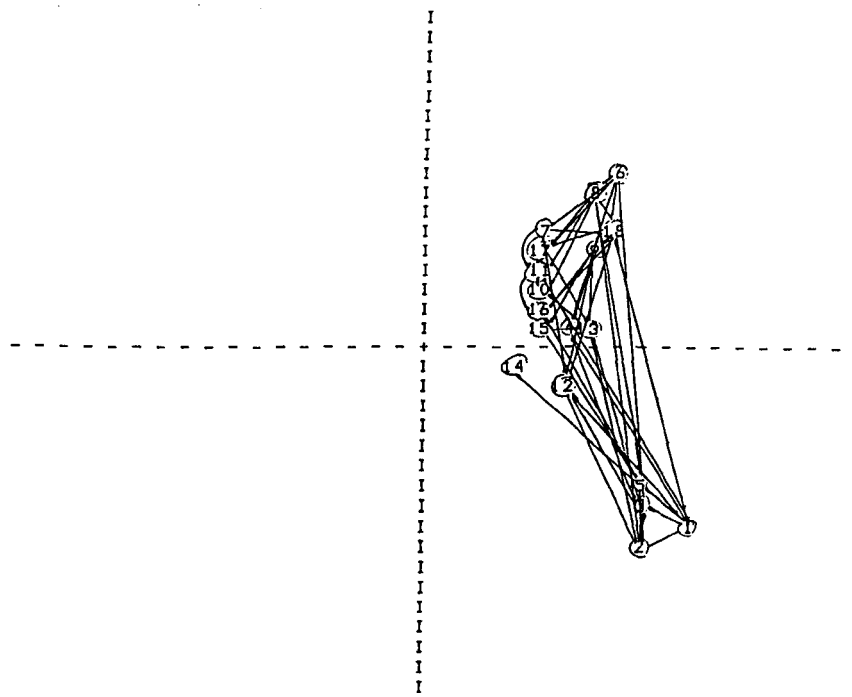

Figure 10-5: Kangaroo skeleton data: PCA plot constructed for the absolute partial correlation matrix

However, some structure is now apparent in the data — there are shorter edges and some grouping of the variables.

I have also applied classical multidimensional scaling (MDS) to Jeffers' pit-prop data set and Poole's kangaroo data set, using the same four matrices (correlation matrix, absolute correlation matrix, negative partial correlation matrix, and absolute partial correlation matrix), although only the results obtained using the two forms of the partial correlation matrix will be considered here. The results were plotted in the first two dimensions determined by the principal coordinates using SPSS.

For Jeffers' pit-prop data, the independence graph obtained by applying MDS to the negative partial correlation matrix, presented in Figure 10-6, is quite messy, with a mixture of long lines and irregularly dispersed points. The plotted graph obtained by applying MDS to the absolute values of the matrix of partial correlations, presented in Figure 10-7, is far more attractive, and suggests some structure within the data. Similarly, for Poole's kangaroo data, a messy representation is obtained using the negative partial correlation matrix, presented in Figure 10-8, but a quite pleasing representation is obtained using the absolute partial correlation matrix, presented in Figure 10-9. However, in both plots obtained by applying MDS to the absolute partial correlation matrix, the "horse-shoe effect", which is a common artefact in MDS, has occurred.

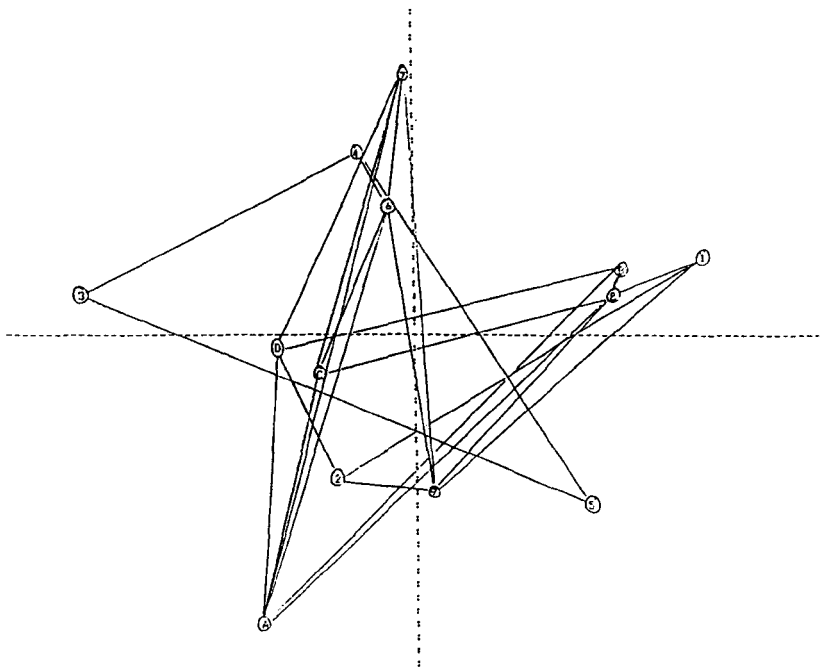

Figure 10-6: Pit-prop data: MDS plot constructed for the negative partial correlation matrix

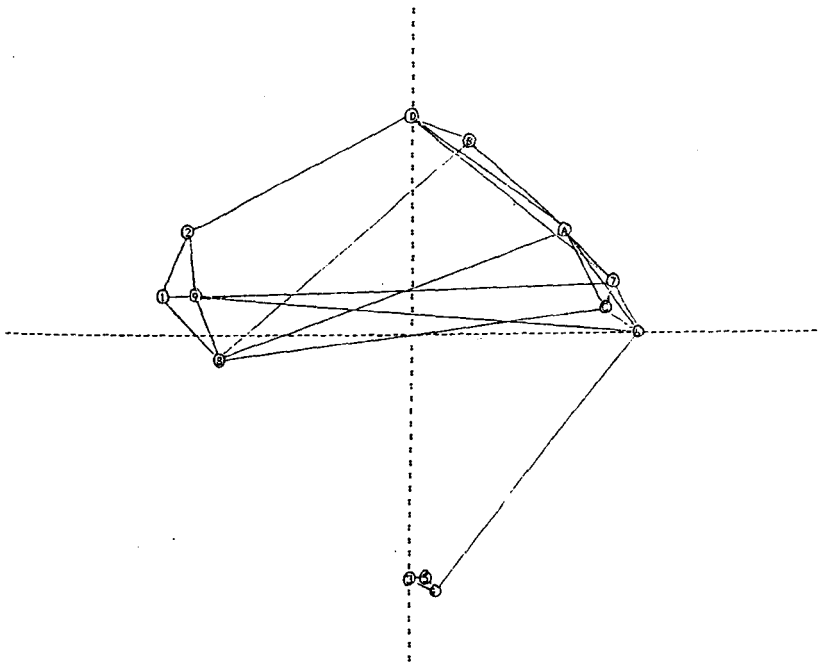

Figure 10-7: Pit-prop data: MDS plot constructed for the absolute partial correlation matrix

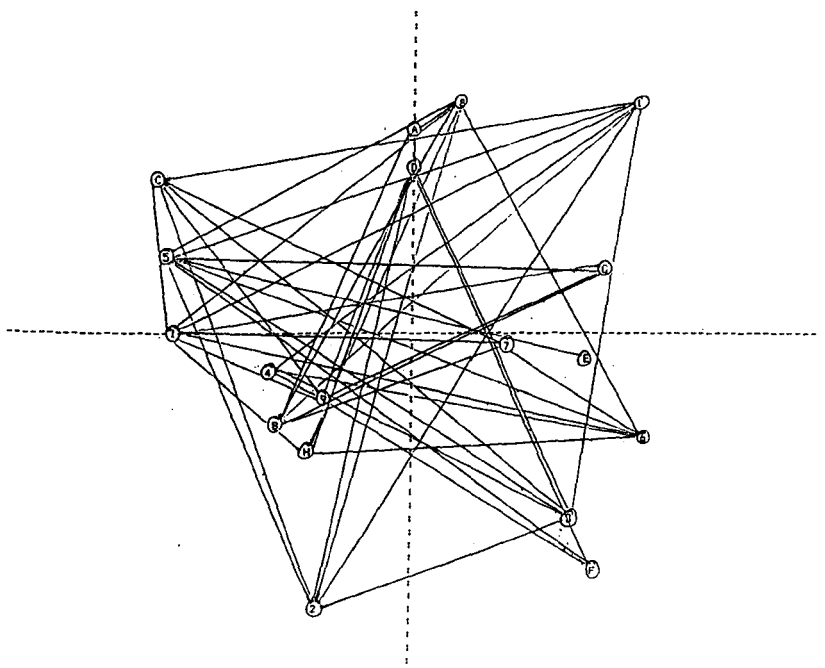

Figure 10-8: Kangaroo skeleton data: MDS plot constructed for the absolute partial correlation matrix

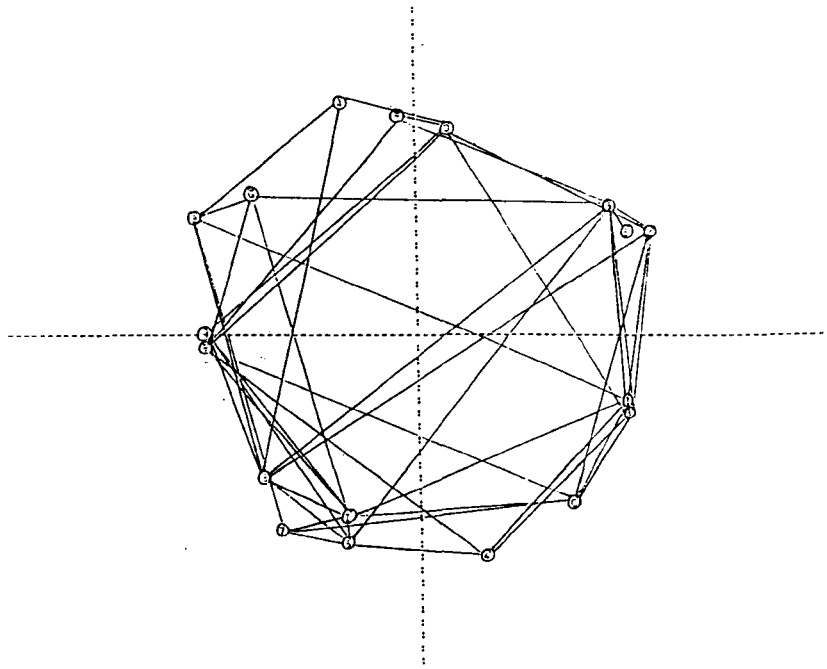

Figure 10-9: Kangaroo skeleton data: MDS plot constructed for the absolute partial correlation matrix

Thus it would appear to be the case that it is possible to construct an independence graph in which the strength of association between pairs of variables is encoded by the length of the edges by applying PCA or MDS to the absolute values of the partial correlation matrix, and plotting the results in the first two dimensions. The distances between the vertices are in inverse proportion to the strength of association, as measured by the absolute value of the partial correlation or the edge exclusion deviance (the two are monotonically related), such that, in general, the closer together two variables joined by an edge, the stronger the association between them.

10.4.1 Problems with Encoding Strength of Association by Distance

Although it would appear to be the case that it is possible to construct an independence graph in which the strength of association between pairs of variables is encoded by the length of the edges by applying PCA or MDS to the absolute values of the partial correlation matrix, and plotting the results in the first two dimensions, in practice a

PCA or MDS solution usually exists in more than two dimensions. For example, for the pit-prop data example, five principal components are required in order for the PCA plot to represent 70% of the total variance, and for the kangaroo data example, eight or nine principal components are required. By using just the first two principal components, less than 40% of the total variance is explained for the pit-prop data, and less than 30% for the kangaroo data. Similar problems apply to the MDS solutions. In showing a PCA or MDS solution in just two dimensions, the relationship between strength of association and length of edge may not be preserved unless the first two dimensions account for most of the variation. Thus it may be the case for the graphs obtained by applying PCA and MDS to the pit-prop data and the kangaroo data sets that variables which are very strongly associated may, on occasion, be plotted further apart than variables which are less strongly associated.

Although the application of PCA (and also of MDS) also has the advantage, cited by Whittaker (1988), of providing an approach to the construction of independence graphs for large number of variables; in practice, as can be seen from the figures presented in this section, the graphs obtained are not necessarily aesthetically pleasing, in that some vertices may be clustered together so closely that it is difficult to determine the edges which are present between them. Also, the graph may contain multiple crossovers, which may lead to a messy graph and make it difficult to determine which pairs of vertices are actually linked. One may wish to relocate some of the vertices in order to make the edges of the graph more readily discernible, but this would destroy the relationship between the length of the edge and the strength of the association.

It therefore seems to be desirable to seek an alternative way of encoding strength of association in the edges of the graph which is not dependent on the length of the edges, so that the user is free to relocate the vertices of the graph if they wish.

10.5 Encoding Strength of Association by Edge Style

As was seen in the previous section, it is possible, and intuitively attractive in the continuous case, to encode the pairwise strengths of association by distances between pairs of variables. However, in practice it was found that the resulting representation may not be visually pleasing. What is desired, therefore, is some technique whereby the strengths of associations are encoded in the graph without using distance, so that vertices

may be relocated to provide a visually pleasing representation without destroying the encoding of the strengths. Two possible approaches are suggested:

- **Width of edges:** such that the wider an edge, the stronger the absolute value of the association between the pair of variables linked by that edge. No edge corresponds to zero association, a thin edge corresponds to a weak (non-zero) association, and a thick edge corresponds to a strong association.
- **Shading of edges:** such that the darker an edge, the stronger the absolute value of the association between the pair of variables linked by that edge. No edge corresponds to zero association, a faint edge corresponds to a weak (non-zero) association, and a dark edge corresponds to a strong association.

Width was used in preference to area, since area is also dependent on length. Grey-tone shading of edges was used in preference to colour, in keeping with the aim of developing a monotone graphical representation which may be readily reproduced, and because different colours have no intuitive ordering. Even though different colours can be ordered according to physical criteria such as wavelength, other criteria such as intensity or brightness or even the viewer's subjective preference may influence the viewer's perception of the different colours.

It is the case that PCA or MDS could be used in conjunction with either of these techniques for the encoding of strength of association by edge style, perhaps as an initial exploratory approach to the construction of the conditional independence graph and to the examination of the structure within the data. If edges in the PCA or MDS solution were to be relocated, the encoding of the strength of association in the edge styles used would still be preserved, even though the distances between the vertices would no longer be meaningful.

10.5.1 A Graphical Perception Experiment

An experiment was conducted to determine which of the two suggested approaches for encoding strength of association by edge style — width of edges or grey-tone shading of edges — is the most effective.

This was intended to be a pragmatic experiment, to provide information about which of the two methods facilitates the fastest and/or the most accurate interpretation of

the information about the strengths of the associations encoded in the graphs. The experiment was not, however, intended to identify or explain any differences in the cognitive processing of widths and grey-tone shading. Although it may be expected that one edge style is found to be faster and/or more accurate than the other and that such a result may be due to differences in cognitive processing of the two styles, it would not be possible to eliminate the possibility that there may have been differences in the manner in which the two edge styles were presented. Thus the experiment described in the remainder of this section can not be regarded as a graphical perception experiment of the kind described in Chapter 5.

In the remainder of this section, the experimental design used is described and the results obtained are presented and discussed.

Experimental Method

It was decided to present the experimental graphs to the subjects on a computer-based display using a monochrome SUN workstation with a mouse pointer, since it was intended that the method found to be 'best' (in terms of speed and/or accuracy) for encoding strength of association would be incorporated in the "Conditional Independence Graph Enhancer" computer package, which has been written for SUN workstations and which will be described in more detail in Chapter 11.

In practice it was found that only about five different widths or grey shades could be used (in addition to 'no edge') in order for the edges to be readily distinguishable whilst avoiding the use of excessively thick lines. This limitation is due to the finite size of pixels on the computer screen. Table 10-1 shows the relationship between the level of the strength of association to be represented (0–5) and the degree of screen-based resolution available for representing widths or grey-tone shading; the widths being expressed by the number of pixels and the grey-tone shading being expressed as some percentage grey (the ratio of black pixels to white pixels). 'Root grey' is the SUN default grey-tone, falling between 25% and 50% grey.

Twelve graphs were used in the experiment, having 5, 6, 7 or 8 vertices and 0%, 25% or 50% of edges missing, as summarised in Table 10-2. This was felt to give a reasonable and representative number of edges and vertices. For each graph, the edges were randomly allocated a value between 0 and 5 inclusive corresponding to the different

Level	Width (pixels)	Grey-tone shading
0	(no edge)	(no edge)
1	1	25% grey
2	3	root grey
3	5	50% grey
4	7	75% grey
5	9	black

Table 10-1: Table showing the relationship between level of strength of association and degree of screen-based resolution

possible levels of strength of association as given in Table 10-1, but the allocation of the values was constrained by the number of missing edges to be allocated a value of 0, as summarised in Table 10-2. The lay-out of the vertices for each of the 12 graphs was determined by displaying each graph on the computer screen with the edges un-encoded. The graphs were then manipulated interactively by the experimenter by moving the vertices around the screen until the graph appeared complicated without being illegible (ie. with a large number of long edges but without overlapping vertices). For the graphs with few vertices and few edges this tended to involve creating as many crossovers as possible, whereas for the graphs with very many vertices and many edges this tended to involve removing several crossovers. Checks were subsequently made, once the edge styles had been incorporated, to ensure that there was a unique answer to each question in the experimental task, and the graphs were modified if necessary.

No. of vertices	Total no. of edges	No. of edges missing		
n	$n(n-1)/2$	0%	25%	50%
5	10	0	3	5
6	15	0	4	8
7	21	0	5	11
8	28	0	7	14

Table 10-2: Table showing number of vertices present and number of edges absent for each of the 12 basic graphs

Each of the twelve graphs was to be presented twice to each subject, once with the edges drawn using the different widths corresponding to the numerical value of the level assigned to each edge, and once with the edges drawn using the different grey-tone shades corresponding to the same numerical values. This was to control for the possibility that, if different graphs had been used for the two different edge styles, the graphs used for one

edge style may have been easier to interpret than those used for the other edge style. Each subject was therefore shown 24 graphs in total, in addition to two practice graphs (one for each edge style).

A modified 2×2 'treatment'×'period' cross-over design was employed (Jones & Kenward (1989), Clayton & Hills (1987)) with 'treatment' corresponding to the edge style and 'period' corresponding to whether a particular edge style (width or grey-tone shading) was presented first or second for any given graph. In a standard 2×2 cross-over design, half the subjects would have been presented with the graphs with the strength of association encoded by widths before being presented with the equivalent graphs with the grey-tone edge styles, for all twelve graphs, and the other half of the subjects would have received the grey-tone graphs before the equivalent widths graphs. Such a design enables the testing of carry-over and period effects, in addition to the treatment effect. In other words, having already been presented with a graph employing one edge style, does this make the experimental task easier when presented with the same graph with the other edge style? If yes, is the effect the same irrespective of the order of presentation of the two graph types? If yes to the latter then, although there may be a period effect, any carry-over effect is the same, and the grey-tone graphs can be compared directly with the equivalent widths graphs irrespective of the order of presentation of the graphs and irrespective of any period effect.

It was decided that the subjects should not receive all twelve graphs with one edge style first, but should receive one edge style first for six of the graphs, and the other edge style first for the other six graphs. The twelve basic graphs were therefore divided into two groups of six. To do this, the graphs were ordered according to the number of edges, or to the number of vertices in the case of a tie (see Table 10-2), and then assigned one by one to alternate groups. These two groups will be referred to as Group A and Group B. Half the subjects were to receive the Group A graphs with the grey-tone edge style first, and the Group B graphs with the widths edge style first. The other half of the subjects were to receive the Group A graphs with the widths edge style first, and the Group B graphs with the grey-tone edge style first. These two groups of subjects will be referred to as SG1 and SG2 respectively. The graphs were presented in a different random order for each subject, subject to the constraint regarding which edge style is presented first for each graph for each group of subjects, and also to the constraint that the same graph could not be presented twice in succession.

The experimental set-up was tried and tested using two pilot subjects who did not participate in the main experiment. Twelve subjects were used in the main experiment. All of the subjects were volunteers who were either post-graduate students or members of the academic staff of the Open University, and most were members of the Mathematics or Science Faculties. Some of the subjects were members of the Statistics Department, whereas the other subjects had differing amounts of statistical knowledge. However, given the nature of the task, and the detailed instruction and practice provided (for what is, in any case, a novel statistical technique), it was felt that the statisticians would have no advantage over the non-statisticians. There was no evidence to contradict this assumption. Six of the subjects were allocated to SG1, and six to SG2. Given the twelve subjects S1–S12 in order of recruitment, this allocation was made alternately, such that odd numbered subjects were allocated to group SG1 and even numbered subjects were allocated to group SG2.

For each graph, three questions were asked. These questions were each intended to be of a realistic nature bearing in mind the intended use of the edge styles for the encoding of strength of association in conditional independence graphs. The general form of the three types of questions was as follows:

1. Which two variables are most strongly associated with variable X (where X was one of the variables in the graph)?
2. Which of the following four pairs of variables are most strongly associated (where the most strongly associated pair in the list did not necessarily correspond to the most strongly associated pair in the graph)?
3. Which of the following is the strongest 3-cycle in the graph (where the strongest 3-cycle was to be assessed by some sort of *visual* summation or averaging by the subject of the three edges in the cycle, and did indeed corresponded to the strongest 3-cycle in the graph as assessed by summing or averaging the numerical values assigned to the edges)?

The three questions were always asked in the same order. For each question, four options were presented, only one of which did indeed correspond to the correct answer. The options were always presented in alphabetical order (for example, a possible answer ACD would be listed before a possible answer BCD). For a few of the graphs drawn on a small number of vertices with a large proportion of edges missing, one or more of the

options for one or more of the questions had to be a dummy option, which could not actually be found in the graph (either a listed variable was not linked to the variable named in the first question, or an edge did not exist between a pair of variables listed in the second question, or a three-cycle did not exist in the third question), but wherever possible, the options presented were to be found in the graph displayed.

Each graph was displayed prior to the display of the first question to allow the user to study the graph. The user could request the first question when ready by clicking the mouse on a button on the screen. Having answered the first question, the user could then request the display of the second question when ready in the same way. Having answered the second question, the user could then request the display of the third question when ready. Having answered the third question the user could then request the display of the next graph, again by clicking the mouse on the button. Having selected one of the four options as the answer to the displayed question, the user could change his/her answer provided they did so before going on to the next question. The user could not move on to the next question or graph until they had selected one of the four options displayed.

The user was initially shown two practice graphs, one with strengths encoded by widths, and one with strengths encoded by grey-tone. The same two practice graphs were used for each subject. These are shown in Figures 10-10 and 10-14 respectively. Using the first practice graph, the necessary basic concepts were explained to the subject — for example, the five different levels of width/grey-tone were pointed out, and their correspondence to different levels of strength was explained (ie. that the stronger the association, the wider or darker the edge). It was explained that missing edges correspond to zero or no association. It was also explained that 3-cycles correspond to sets of three non-zero edges between three variables. Then the use of the graph-mouse interface was explained to the subject, and the subject was talked through each of the three questions for the first graph. For the practice graphs only, the subject was told whether their chosen answer was correct or not. The three questions, and the correct responses, are shown in Figures 10-11, 10-12 and 10-13 for the widths practice graph. Having answered each question correctly for the first practice graph, the subject was then allowed to go through the questions for the second practice graph by themselves. The three questions, and the correct responses, are shown in Figures 10-15, 10-16 and 10-17 for the grey-tone practice graph. Prior to displaying the first experimental graph, the subject was reminded of the

most important features of the graphs and of the experimental set-up, and was given an opportunity to ask questions.

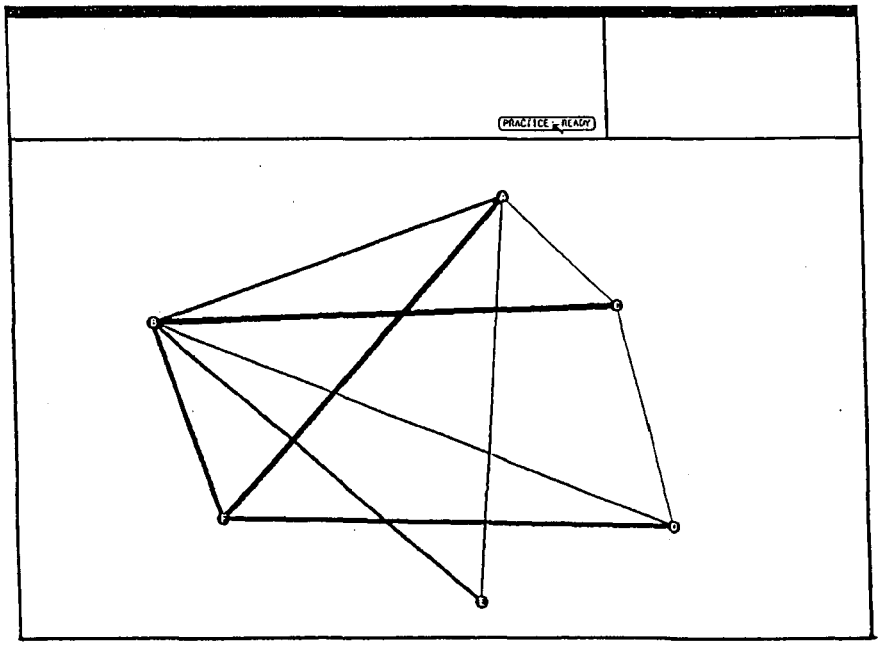

Figure 10-10: Practice graph with strength of association encoded by width of edges.

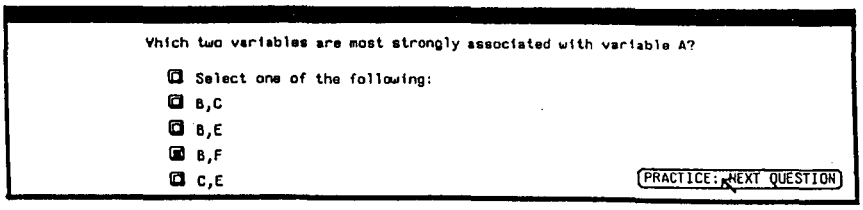

Figure 10-11: First question for practice graph with strength of association encoded by width

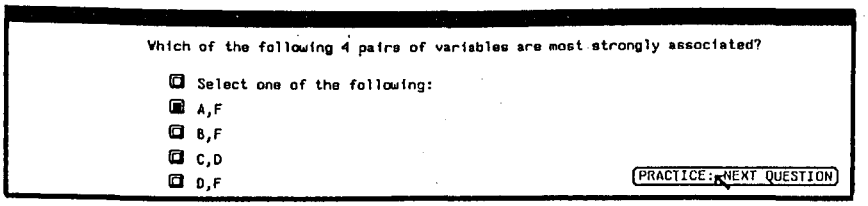

Figure 10-12: Second question for practice graph with strength of association encoded by width

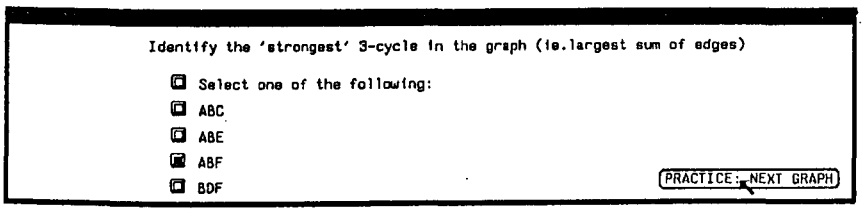

Figure 10-13: Third question for practice graph with strength of association encoded by width

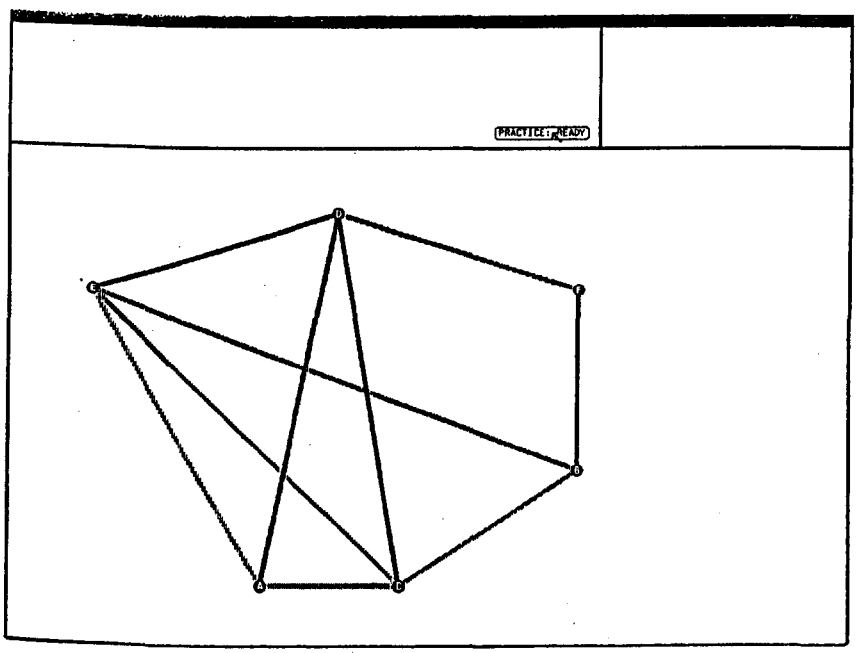

Figure 10-14: Practice graph with strength of association encoded by grey-tone

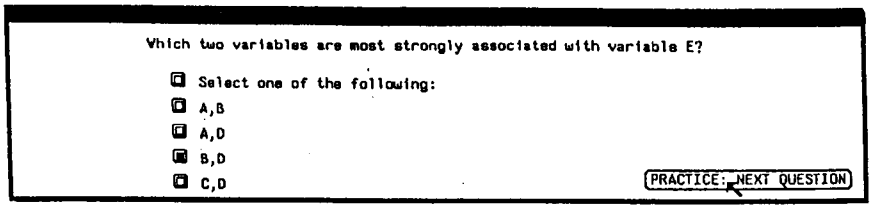

Figure 10-15: First question for practice graph with strength of association encoded by grey-tone

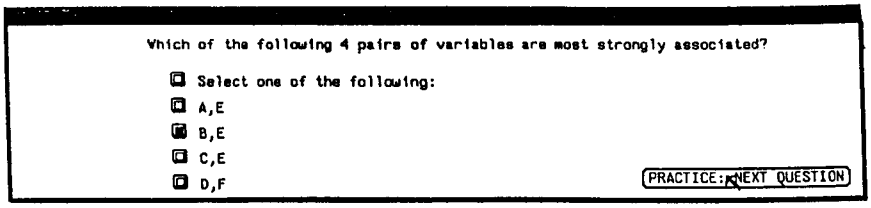

Figure 10-16: Second question for practice graph with strength of association encoded by grey-tone

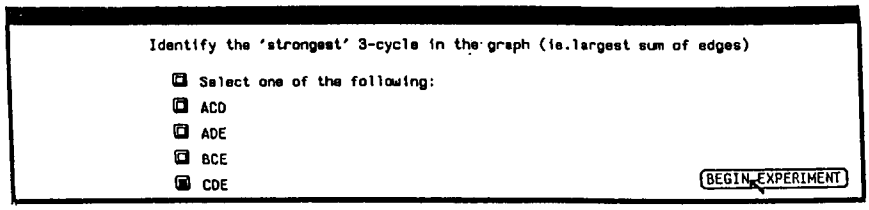

Figure 10-17: Third question for practice graph with strength of association encoded by grey-tone

Half the subjects (three from SG1 and three from SG2) received the widths practice graph first, and the other half received the grey-tone practice graph first, so that any effect resulting from which graph was explained most thoroughly could be tested for. These two groups of subjects and the practice graphs they received first will be referred to as WP and GP respectively.

For each experimental graph, the latency (ie. the time taken to answer each question) was measured using the computer's internal clock (which is accurate for time measured in seconds) from the moment the question was displayed until the user selected an option (or options, if they changed their mind), and the option(s) chosen were recorded. Whether the question was answered correctly or not was assessed subsequently. Subjects had been informed that both latency and accuracy were being recorded, without placing emphasis on either form of response, thus leaving it to the subjects to make any trade-off themselves. Assessment of both latency and accuracy is consistent with the experimental design used in a graphical perception experiment by Lewandowsky & Spence (1989b) who state that "When the performance of observers of statistical graphs is examined, it is desirable to measure not only accuracy, but also response latency".

The experimental design and allocation of subjects is summarised in Table 10-3.

		SG1	SG2
WP	Widths	A-2nd B-1st	A-1st B-2nd
	Grey-tone	A-1st B-2nd	A-2nd B-1st
	Subjects	S7, S9, S11	S8, S10, S12
GP	Widths	A-2nd B-1st	A-1st B-2nd
	Grey-tone	A-1st B-2nd	A-2nd B-1st
	Subjects	S1, S3, S5	S2, S4, S6

Table 10-3: Summary of experimental design, detailing the allocation of subjects to subject groups SG1 and SG2, and to initial practice graphs WP and GP, together with the order of presentation of the group A and group B graphs

Hypotheses

The following, main, hypothesis was to be tested using both accuracy and latency as measures of performance:

1. There is no difference in performance according to the edge style used to encode strength of association.

Two other, secondary, hypotheses which were also to be tested using the same measures were as follows:

2. Any carry-over effect affects performance between the two presentations of each graph in the same way, irrespective of the order of presentation of the edge styles.
3. The edge style of the practice graph which was presented first and therefore explained most thoroughly does not affect performance for the two different edge styles.

The main hypothesis cannot be answered satisfactorily without first investigating the two secondary hypotheses.

Initial Analyses

Tables 10-4 to 10–6 give, for each subject and each question, the mean time to a correct response, the number of correct responses (maximum N=12), and the standard deviation of the correct response times for the graphs encoded by the grey-tone edge style (G) and for the graphs encoded by the widths edge style (W). The latencies of incorrect responses are not considered since the experiment is concerned with the effectiveness of the two different edge styles for accurate, as well as quick, interpretation of encoded information. Furthermore, including the latencies of incorrect responses may distort the results if, for example, an inaccurate response is the result of a subject being more concerned with speed than accuracy, or if a subject is very poor at interpreting the graphs (whether fast or slow).

The same summary values are also given in Tables 10-4 to 10–6 for the differences in response times, G–W, computed for each pair of graphs. Where a subject gave an inaccurate response for one or both presentations of a graph, the G–W difference for that graph has not been included in the calculation of the mean G–W difference.

It can be seen from Tables 10-4 to 10–6 that the number of errors made was quite small relative to the total number of answers given. Of the 864 answers (12 subjects × 24 graphs × 3 questions), a total of 44 errors were made. The errors made, broken down by question and edge style of graph, are summarised in Table 10-7. Most of the errors made for Question 3 were made because some subjects chose options for which only two of the three possible edges between the vertices were present in the graph where these two edges 'summed' to more than the three edges in any of the 3-cycles in the graph. Even though it was stressed in the introduction to the experimental task that 3-cycles involve three edges

Question 1							
Subject	G	W	G – W	Subject	G	W	G – W
S1	12.09 N=11 (3.88)	11.82 N=11 (4.29)	0.90 N=10 (3.28)	S7	9.33 N=12 (2.87)	9.50 N=12 (5.02)	-0.17 N=12 (5.04)
S2	12.00 N=12 (5.12)	16.58 N=12 (10.86)	-4.58 N=12 (12.03)	S8	23.80 N=10 (10.30)	29.89 N=9 (17.16)	-9.75 N=8 (21.50)
S3	9.58 N=12 (3.45)	9.67 N=12 (2.67)	-0.08 N=12 (4.12)	S9	8.00 N=12 (2.13)	8.64 N=11 (2.25)	-0.55 N=11 (3.33)
S4	6.83 N=12 (3.30)	11.42 N=12 (10.62)	-4.58 N=12 (10.00)	S10	9.50 N=12 (7.63)	8.92 N=12 (5.79)	0.58 N=12 (9.40)
S5	8.50 N=12 (1.24)	10.25 N=12 (3.60)	-1.75 N=12 (3.42)	S11	10.42 N=12 (6.87)	10.64 N=11 (4.80)	0.18 N=11 (8.89)
S6	7.25 N=12 (1.29)	8.25 N=12 (2.14)	-1.00 N=12 (2.34)	S12	21.45 N=11 (17.07)	19.64 N=11 (4.80)	2.10 N=10 (25.96)

Table 10-4: Mean latencies, number of correct responses and standard deviations for each subject on Question 1

Question 2							
Subject	G	W	G – W	Subject	G	W	G – W
S1	20.73 N=11 (8.67)	18.83 N=12 (10.51)	2.18 N=11 (11.55)	S7	16.00 N=12 (6.84)	16.75 N=12 (7.68)	-0.75 N=12 (5.19)
S2	19.42 N=12 (5.87)	23.75 N=12 (10.63)	-4.33 N=12 (10.05)	S8	29.00 N=10 (9.51)	32.08 N=12 (14.64)	-5.50 N=10 (8.46)
S3	20.08 N=12 (5.76)	22.42 N=12 (10.10)	-2.33 N=12 (6.10)	S9	18.09 N=11 (9.19)	17.92 N=12 (9.89)	0.45 N=11 (5.68)
S4	17.33 N=12 (7.44)	16.42 N=12 (8.05)	0.92 N=12 (11.15)	S10	25.17 N=12 (8.13)	25.42 N=12 (9.12)	-0.25 N=12 (9.50)
S5	20.17 N=12 (4.88)	18.83 N=12 (6.25)	1.33 N=12 (4.68)	S11	20.83 N=12 (8.58)	19.09 N=11 (9.50)	1.18 N=11 (6.65)
S6	15.08 N=12 (5.52)	18.83 N=12 (6.73)	-3.75 N=12 (6.40)	S12	32.42 N=12 (16.33)	27.25 N=12 (5.82)	5.17 N=12 (16.02)

Table 10-5: Mean latencies, number of correct responses and standard deviations for each subject on Question 2

Question 3							
Subject	G	W	G – W	Subject	G	W	G – W
S1	22.78 N=9 (14.32)	23.12 N=8 (14.87)	-3.71 N=7 (13.34)	S7	34.27 N=11 (15.64)	35.82 N=11 (17.15)	-1.55 N=11 (14.96)
S2	40.42 N=12 (18.76)	42.55 N=11 (20.24)	-6.91 N=11 (15.00)	S8	37.00 N=9 (18.25)	29.70 N=10 (16.25)	2.38 N=8 (14.55)
S3	35.60 N=10 (26.49)	29.17 N=12 (12.60)	7.00 N=10 (29.07)	S9	7.25 N=12 (3.33)	10.45 N=11 (9.88)	-3.55 N=11 (11.36)
S4	28.18 N=11 (24.26)	29.36 N=11 (14.00)	-1.18 N=11 (16.63)	S10	41.50 N=12 (17.59)	37.60 N=10 (10.76)	0.20 N=10 (13.46)
S5	29.30 N=10 (12.15)	30.67 N=12 (16.18)	2.80 N=10 (12.59)	S11	42.50 N=11 (37.50)	31.27 N=11 (18.78)	1.50 N=10 (30.69)
S6	23.36 N=11 (16.82)	25.75 N=12 (19.96)	-3.36 N=11 (19.48)	S12	39.33 N=12 (17.90)	45.27 N=11 (17.68)	-4.82 N=11 (19.15)

Table 10-6: Mean latencies, number of correct responses and standard deviations for each subject on Question 3

	Question 1	Question 2	Question 3	Overall
Widths	7	1	14	22
Grey-tone	4	4	14	22
Overall	11	5	28	44

Table 10-7: Table giving numbers of errors made, broken down by question and edge style

between three vertices, some subjects interpreted missing edges as edges 'present' but having zero strength.

Because of the relatively small number of errors made, and because, overall, the same number of errors were made for each edge style, it was decided not to investigate accuracy any further. However, it is of interest to note that each subject made at least one error. Two subjects made one error only; five subjects made two errors; two subjects made three; one subject made four; one subject (S1) made ten; and one subject (S8) made twelve. Therefore, two subjects were responsible for half the total number of errors made. Because of the small number of subjects who took part in the experiment, it was decided not to drop these subjects from the analysis on the basis of their accuracy. For all the

subjects, the latency of the response to a question was not considered if the wrong answer had been given.

Looking at the data in Tables 10-4 to 10–6, it is apparent that there were two exceptionally slow subjects, S8 and S12 (both in subject group SG2 and both of whom received the widths practice graph first). It was decided not to drop these subjects from the analyses, on the assumption that their slow performance would affect both graph types such that they would not bias the results used to test Hypotheses 1 and 2, although they could affect the results used to test Hypothesis 3. Although it may be the case that S12, who made a total of 3 errors, was sacrificing speed for accuracy, S8 was both slow and inaccurate. With a larger sample size, it would be expected that a subject like S8 would not be so influential.

It can be seen from Tables 10-4 to 10–6 that the mean latencies and the standard deviations, for both widths and grey-tone graphs, tend to increase with question number, confirming that the 3 questions represent increasing complexity. However, if the G–W differences can be shown to be the same for each of the 3 questions, it may be possible to combine the results for the 3 questions to give a potentially more sensitive analysis. To test this, a repeated measures analysis was used. The analysis of repeated measures is described in detail in Crowder & Hand (1990) and Hand & Taylor (1987).

Given the simplicity of the design used here, having just one within-subjects factor with three levels, the analysis was carried out by hand using the multivariate approach described by Morrison (1976) and Myers (1979). The null hypothesis to be tested, based on the G–W differences for each of the three questions Q1, Q2 and Q3, was that $\mu_{q1} = \mu_{q2} = \mu_{q3}$. The analysis proceeded by consideration of the two sets of difference scores, $x_{q1} - x_{q2}$ and $x_{q2} - x_{q3}$, computed for each of the 12 subjects. It was found that $F=1.003$ with $\nu=2,10$df. Since $F_{critical}$ at $\alpha=0.05$ is 4.81, the null hypothesis can not be rejected and it was decided that the results for each of the three questions may be combined for each subject.

Table 10-8 gives the same summary values as presented in Tables 10-4 to 10–6, computed for the combined questions.

The three hypotheses were tested in accordance with the procedures for the analysis of a 2×2 cross-over design contained in Jones & Kenward (1989). Table 10-9

Combined Questions							
Subject	G	W	G – W	Subject	G	W	G – W
S1	18.26 N=31 (10.35)	17.45 N=31 (10.92)	0.25 N=28 (9.91)	S7	19.46 N=35 (14.18)	20.26 N=35 (15.40)	-0.80 N=35 (9.11)
S2	23.94 N=36 (16.69)	27.20 N=35 (17.76)	-5.23 N=35 (12.12)	S8	29.69 N=29 (13.71)	30.68 N=31 (15.42)	-4.38 N=26 (15.45)
S3	20.94 N=34 (17.85)	20.42 N=36 (12.31)	1.21 N=34 (16.24)	S9	10.91 N=35 (7.37)	12.50 N=34 (8.99)	-1.21 N=33 (7.54)
S4	17.14 N=35 (16.48)	18.77 N=35 (13.13)	-1.63 N=35 (12.62)	S10	25.39 N=36 (17.66)	23.18 N=34 (14.47)	0.18 N=34 (10.44)
S5	18.74 N=34 (11.02)	19.92 N=36 (13.06)	0.68 N=34 (7.62)	S11	24.09 N=35 (25.14)	20.33 N=33 (14.82)	0.94 N=32 (17.70)
S6	15.00 N=35 (11.72)	17.61 N=36 (13.93)	-2.69 N=35 (11.32)	S12	31.34 N=35 (18.17)	30.62 N=34 (16.65)	0.91 N=33 (20.27)

Table 10-8: Mean latencies, number of correct responses and standard deviations for each subject on the Combined Questions

shows the mean response times or latencies for each subject, for the 6 graphs which were first of all presented using grey-tone shading (G1) and then using widths of edges (W2), and for the other 6 graphs which were first of all presented using widths of edges (W1) and then using grey-tone shading (G2). Subject groups SG1 and SG2 were presented with the same two sets of 6 graphs, but in opposite ways. The means of various sums and differences used in the analyses are also presented. The overall means are displayed in Table 10-10 in relation to the design of the cross-over trial, and summarised visually in Figure 10-18.

The use of parametric tests requires assumptions of normality. With just 12 observations it is difficult to assess normality. Histograms constructed using MINITAB, but not presented here, suggest that the distributions of some measures were more normal in appearance than others. However, on the assumption that the t-test is generally robust, the decision was made to use parametric tests and to test each hypothesis at the usual 5% level of significance.

The results of the tests of the hypotheses are presented for each question in turn below. The results of all the analyses carried out are summarised in Appendix B. All analyses were carried out using MINITAB.

Table 10-9: Mean response times for each presentation style and period, together with sums and differences used in the analyses

Subject	Practice Graph	Subject Group	G1	W2	W1	G2	G1 + W2	W1 + G2	G1 – W2	W1 – G2
S1	GP	SG1	16.94	16.94	18.15	19.86	34.12	37.55	-0.24	-1.00
S2	GP	SG2	25.56	20.88	33.17	22.33	42.47	55.50	0.71	10.83
S3	GP	SG1	24.41	20.50	20.33	17.47	44.06	37.29	4.76	2.35
S4	GP	SG2	18.18	15.06	22.28	16.17	33.24	38.44	3.12	6.11
S5	GP	SG1	19.71	17.44	22.39	17.76	34.65	38.94	4.76	3.41
S6	GP	SG2	14.94	14.50	20.72	15.06	29.41	35.78	0.47	5.67
S7	WP	SG1	21.56	19.11	21.47	17.24	40.67	38.71	2.44	4.24
S8	WP	SG2	31.79	30.93	30.44	27.73	64.08	56.93	-3.08	5.50
S9	WP	SG1	11.06	9.29	15.71	10.78	19.62	26.71	2.50	4.71
S10	WP	SG2	28.67	21.47	24.88	22.11	46.88	46.18	3.94	3.59
S11	WP	SG1	33.44	18.31	22.24	14.18	45.88	36.75	9.25	7.37
S12	WP	SG2	35.61	27.65	33.59	26.82	64.76	61.81	9.47	8.19

Practice Graph	Subject Group	Graph Group	Presentation 1	Presentation 2
GP	SG1 (S1,S3,S5)	A	(G1) 20.35	(W2) 18.29
		B	(W1) 20.29	(G2) 18.36
	SG2 (S2,S4,S6)	A	(W1) 25.39	(G2) 17.85
		B	(G1) 19.56	(W2) 16.82
WP	SG1 (S7,S9,S11)	A	(G1) 22.02	(W2) 15.57
		B	(W1) 19.81	(G2) 14.07
	SG2 (S8,S10,S12)	A	(W1) 29.64	(G2) 25.55
		B	(G1) 32.02	(W2) 26.68

Table 10-10: Mean latencies presented in relation to cross-over design

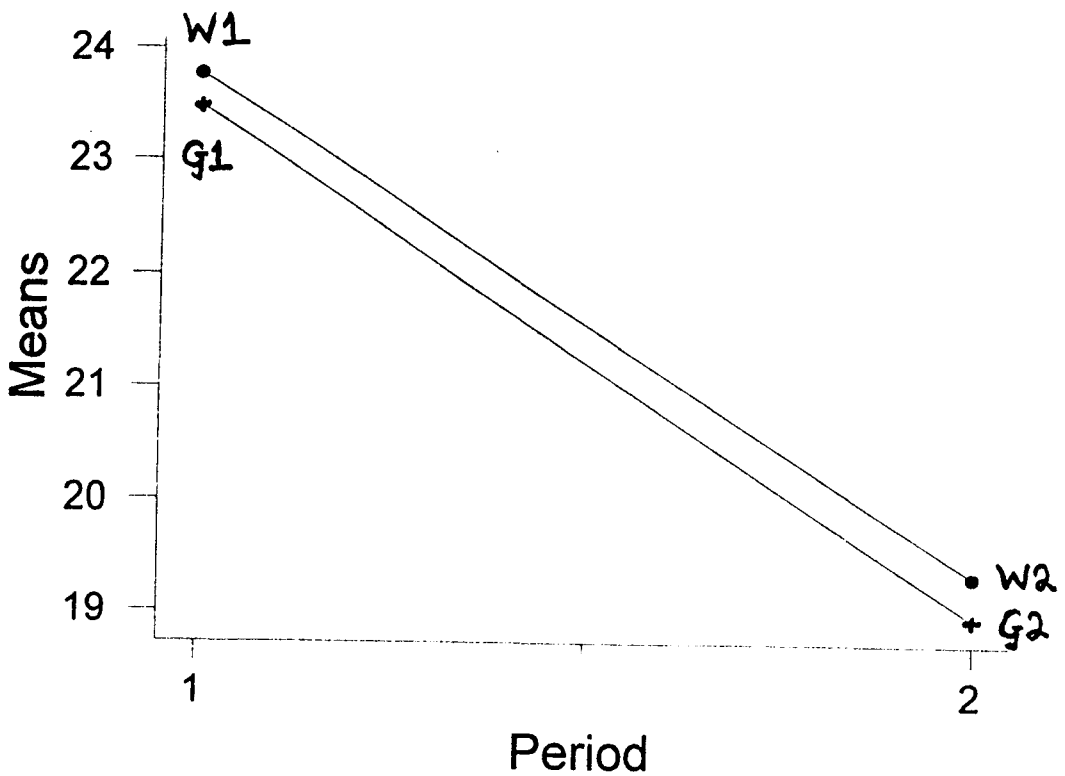

Figure 10-18: Visual display of overall means resulting from cross-over design

Analyses:

To test Hypothesis 2, that there is no carry over effect between the two presentations of each graph, a paired t-test was carried out to compare the two sets of (twelve) scores G1+W2 and G2+W1. This analysis is based on the standard procedure for a 2×2 cross-over design, although a (two-tailed) paired t-test is employed instead of the usual two-sample t-test because data from all of the subjects is used to obtain both sets of scores. Use of this procedure cannot show that there is no carry-over effect. Rather, it may be shown that any carry-over effect affects the two different orders of presentation (G followed by W and W followed by G) in the same way.

It was found that any carry-over effect between the two presentations of the graphs affects the two orders of presentation equally (mean of G1+W2=41.7; mean of G2+W1=42.5; mean difference =–0.90; t=–0.46; p=0.65; 11df; 95% CI (–5.19, 3.40)). I shall therefore consider all the graphs, irrespective of period of presentation, in testing Hypotheses 1 and 3.

It is possible to test whether there is any difference between the two periods of presentation by comparing the two sets of difference scores G1–W2 and G2–W1. Using a paired t-test it is found that there is indeed evidence of a significant difference between the two periods (mean of G1–W2=3.17; mean of G2–W1=–5.08; mean difference =8.26; t=5.44; p=0.0002; 11df; 95% CI (4.91, 11.60)). Although a two-tailed hypothesis test was employed, the values of the means suggest that the subjects are faster the second time they see a graph. However, it has already been shown that the carry-over effect is the same irrespective of which style of graph was seen first.

To test Hypothesis 3, that there is no effect of the edge style of the first practice graph upon performance, two two-sample t-tests were conducted, one to compare the six subjects receiving GP first with the six receiving WP first for G1+W2, and the other to compare these two groups for W1+G2. There was no evidence of any effect of practice graph upon performance based on these two measures (mean of G1+W2 for GP =36.33; mean of G1+W2 for WP =47.00; t=–1.48; p=0.17; 10df; 95% CI (–26.7, 5.4); mean of W1+G2 for GP =40.58; mean of W1+G2 for WP =44.50; t=–0.64; p=0.54; 10df; 95% CI (–17.7, 9.8)).

To test Hypothesis 1, that there is no evidence of a difference in performance according to the edge style used, a paired t-test was carried out to compare the two sets of

twelve difference scores G1–W2 and W1–G2. It was found that there is no evidence of a difference between the two edge styles, widths and grey-tone shading (mean of G1–W2=3.17; mean of W1–G2=5.08; mean difference =–1.91; t=–1.59; p=0.14; 11df; 95% CI (–4.55, 0.74)).

Analyses: Summary

All the above results for the three hypotheses are summarised in Appendix B. The main results are also summarised in Table 10-11.

Hypothesis	t	df	p
1 (treatment effect)	–1.59	11	0.14
2 (practice effect: G1+W2)	–1.48	10	0.17
2 (practice effect: W1+G2)	–0.64	10	0.54
3 (carry-over effect)	–0.46	11	0.65
3 (period effect)	5.44	11	0.00

Table 10-11: Table summarising the t-tests carried out for each hypothesis, for each question

In conclusion, this experiment provides no evidence to reject the null hypothesis that the edge style used to represent strength of association (ie. widths or grey-tone shading) has no effect upon the speed (or accuracy) of the interpretation of the displayed graph.

One possible explanation of the 'failure' of this experiment to distinguish between the two edge styles in terms of speed or accuracy is that there is no difference between the two styles. An alternative explanation which should be considered is that the power of the t-test used to test the main hypothesis was not sufficient to detect any real difference. This is a potentially plausible explanation given that just 12 subjects were used in the experiment, owing to difficulties in recruiting volunteers.

Power calculations have been carried out to determine the sample size which would have been required to detect a significant difference for the main hypothesis, assuming high power and some meaningful difference. Since this was a pragmatic experiment, there is no theoretically significant difference which I would wish to detect. Instead, therefore, an arbitrary difference of 3 seconds has been chosen — no doubt a different conclusion would be reached if a smaller value were chosen, but I would suggest that a larger difference would also be acceptable. No direction for the difference is

hypothesised. The value used as an estimate of the standard deviation of the differences is 3.66 (the larger, and therefore the more conservative, of the two sample standard deviations obtained for the G1–W2 and W1–G2 differences).

The power calculations were carried out using the nomogram presented on p.456 of Altman (1991) at the 5% level of significance and for power values of 70%, 80% and 90%, which should yield sample sizes appropriate to give a moderately high, high and very high (respectively) probability of detecting as significant a difference of the magnitude considered. The results of the power calculations were 36 subjects for 70% power, 47 for 80% power and 62 for 90% power. It can be seen that the sample size of 12 used in the experiment was considerably less than required to give power of 70%–90% to detect a difference of 3 seconds.

The number of subjects who, on average, actually took longer than the arbitrarily chosen difference of 3 seconds for one graph type compared with the other to select the correct response was considered. In the test of Hypothesis 1, 6 subjects were faster, on average, on widths graphs than the equivalent grey-tone graphs for G1–W2, whereas one subject was faster on the grey-tone graphs. For W1–G2, 10 subjects were faster on the grey-tone graphs than on the widths graphs, and none was faster on the widths graphs. In practice, these seemingly large differences are likely to reflect the significant difference previously found between the two presentation periods.

Another possible explanation for the apparent failure of the experiment to distinguish between the two edge styles is that it was inappropriate to use parametric tests without some transformation of the data. The analyses were subsequently repeated using equivalent non-parametric tests (the Wilcoxon (W) test in place of the paired t-tests, and the Mann-Whitney (U) test in place of the two-sample t-tests) and of the 5 hypothesis tests carried out (results presented in Appendix B), the conclusions reached using the non-parametric tests agreed with the conclusions reached using the parametric tests in every case, at the usual 5% level of significance.

Regression Results

Regression lines were fitted to the data for the twelve widths graphs and the twelve grey-tone graphs for each of the three questions separately, to investigate whether

the mean latency of the response could be related to the number of vertices and/or the number of edges in the graph, according to the edge style used. Simple (linear) regression was used to fit the number of vertices V and the number of edges E separately, and multiple (linear) regression was used to fit V and E together, with and without an additional term corresponding to $V*E$. The fitted regression lines and corresponding coefficients of determination R^2 (adjusted for degrees of freedom) obtained using MINITAB are presented in Appendix B and summarised below.

For Question 1, all of the fitted regression lines for both edge styles have an R^2_{adj} value between 0% and 15.8%. This indicates that the fitted regression lines explain none or very little of the variation in the observed response latencies. This finding may reflect the simplicity of this particular question, in which case we would hope for more interesting results for the more complex Questions 2 and 3.

For Question 2, all of the fitted regression lines for both edge styles have an R^2_{adj} value between 37.9% and 79.4%. Some of these values are large enough to indicate that the fitted regression lines explain a sizeable proportion of the variation in the observed response latencies.

For Question 3, all of the fitted regression lines for both edge styles have an R^2_{adj} value between 0.0% and 8.5%. Contrary to expectations, these values are no better than for Question 1 and again indicate that the fitted regression lines explain none or very little of the variation in the observed response latencies.

The fitted linear regression lines only succeed in explaining the observed response latencies in terms of the number of vertices and/or the number of edges in the graph for Question 2, and then for widths graphs better than for grey-tone graphs. Examination of plots of the mean latencies against E and against V (not presented here) suggest that many relationships may in fact be non-linear, although for some of the plots there are outliers which may be distorting the relationships; thus the appropriateness of the use of linear regression should be examined. It may, however, be necessary to incorporate other, uncontrolled, variables in order to explain the variation in the recorded latencies using the regression approach.

Because it was not the main aim of the experiment to be able to predict latency from the number of edges and/or the number of vertices in the graph, the significance of the regression coefficients in each equation has not been considered, and the fitting of regression lines has not been pursued any further.

It is therefore the conclusion of this experiment that there is no evidence that one edge style (width or grey-tone shading) has a greater influence on the speed or accuracy with which subjects can extract different types of information about strengths of association from independence graphs than the other edge style. Some subjects remarked that they found one edge style more pleasant to use than the other, and the majority of those who expressed a preference preferred the grey-tone style, but it was not always the case that a subject performed faster on the edge style he/she preferred!

It is therefore suggested that the edge style to be used could be left to the individual preference of the user. Alternatively, both width and grey-tone could be combined in a single representation, such that thin faint edges correspond to weak non-zero associations, and fat dark edges correspond to strong associations. However, although intuitively attractive, there is no experimental evidence available to indicate whether a combination of the two edge styles would be preferable (in terms of speed and/or accuracy) to a single edge style alone.

10.6 Encoding Sign of Association

This chapter has been concerned with the graphical representation of strengths of associations, primarily for continuous data. The strength of an association between two continuous variables may be indicated by the value of the negative partial correlation coefficient between the two variables, which may be positive or negative in sign, or by the edge exclusion deviance, which is monotonically related to the absolute value of the partial correlation coefficient and is always positive in sign. The strength of an association between two discrete variables may be indicated by the exclusion deviance, which is always positive in sign, or, in certain situations, by the cross-product ratio, which may be positive or negative in sign.

It may be of interest to represent graphically which pairs of variables are positively associated, and which are negatively associated, when this information is known. One edge style (eg. a thick edge) could be used to represent positive associations, and an alternative edge style (eg. a thin edge) to represent negative associations.

It is not suggested that information about the signs of association be combined with information about their strength, since this may lead to an excessive amount of information on a single graph. It may be preferable to have two graphs available, one of which communicates information about the strengths of associations, and the other of which communicates information about the signs of these associations. This idea will be pursued further in Chapter 11.

10.7 Summary

This chapter has been concerned with the encoding of strength of association in conditional independence graphs. This is applicable both to associations between continuous variables (represented by the negative partial correlation coefficient or the edge exclusion deviance) and to associations between discrete variables (represented by the exclusion deviance for graphical models or the cross-product ratio for 2×2 contingency tables).

One approach to the encoding of the strengths of the associations is to use the distance between vertices in the independence graph. By applying PCA or MDS to the matrix of associations, it is to be hoped that the vertices representing highly associated variables will be plotted close together. However, in practice the two-dimensional PCA or MDS solution may not produce an accurate representation of the multi-dimensional distances representing the relative strengths of the associations between variables, and the resultant graphs can also appear quite illegible.

It was therefore decided to encode the strengths of the associations in the edges of the graph, with no restrictions on the location of the vertices. Two edge coding schemes were proposed: one based on the widths of edges and the other based on the grey-tone shading of edges. A graphical perception experiment gave no evidence of any important differences between the use of these two coding styles in terms of accuracy or latency in the interpretation of information about the strength of associations from graphs constructed employing the two styles. A third style was also suggested, this being a combination of the width and grey-tone shading styles.

Finally, it was suggested that the sign of the association might also be incorporated in a conditional independence graph.

The use of edge styles and the encoding of the sign of the associations will be considered further in Chapter 11, together with the problem of representing the parameter values of the interactions contained within log-linear interaction models for discrete data.

BLANK PAGE IN ORIGINAL

11. Conditional Independence Graph Enhancer

11.1 Introduction to the CIGE Package

A computer package has been written to implement the work and ideas which have been presented in Chapters 7, 9 and 10, based on the conditional independence graph approach. Thus the package is intended as a tool for the construction of representations of fitted statistical models based on independence and interaction graphs. In addition to this, however, the package serves as a tool for the interpretation of the model represented, and may also be of use in assessing the appropriateness of the fitted model.

Because of the various features of the package, it has been called the "Conditional Independence Graph Enhancer", or CIGE (pronounced "Siggy"). The package has been written for monochrome Sun workstations with a three-button mouse pointer, and is written in "C" using SUNView graphics library routines. There are approximately 2310 lines of code. The decision to write the package for the SUN computer rather than the more ubiquitous IBM PC-compatible is due to the highly interactive and high-speed performance of this dedicated graphics workstation. The choice of monochrome, rather than colour, workstations is in keeping with the philosophy that representations should be readily reproducible.

In the following section (Section 11.2), the main features of the CIGE package are described, including the initial data input and graph construction, incorporation of the extensions to independence graphs described in earlier chapters, and additional modifications to the use of independence graphs as a model representation technique which have not been described previously. Some of these additional modifications have parallels in the work of Becker *et al* (1987, 1989, 1990), among others, on the use of dynamic graphics for the analysis of network data (see Chapter 2). All of the features described will be illustrated in the next chapter (Chapter 12), in which the use of CIGE applied to a number of example data sets is described. The reader is also referred to Cottee (1990). Full details of the use of CIGE are contained in a User's Guide presented in Appendix A.

In many respects it is appropriate to think of CIGE as an interface between the user and numerical information relating to the displayed fitted model. Through the graph drawn to represent the model, the user can access information relating to the model. In

some instances, this information may be encoded pictorially in the graph itself, or will be presented numerically, or in an additional pictorial form. The interface is intended to be self-explanatory and easy to use, guiding the user in their choice of information.

11.2 Features of the Package

11.2.1 Data Input

Before CIGE can be employed to represent a fitted model, it is necessary for the user to first fit the model to their data. It is not possible to do this within CIGE, since the package does not possess any model fitting capabilities. Instead, it is recommended that the user should use some existing model fitting package, such as GLIM or MIM (see Chapter 7). Having fitted a model, the user can then input the pertinent features of the model into the CIGE package.

For continuous data, the user should input the 'strengths' of the pair-wise associations in the fitted covariance selection model. These 'strengths' may take the form of the fitted partial correlation coefficients or of the edge exclusion deviances (the two are monotonically related). For discrete data, the user can input either the generating class of the fitted log-linear model which they wish to represent, or the parameters of each of the terms in the model (from which the generating class can be determined).

11.2.2 Graph Construction and Manipulation

By default, CIGE will construct the initial independence graph with one vertex per variable, whereby the vertices are located equidistant around the circumference of an imaginary circle. The vertices are labelled with a variable name specified by the user.

In the conventional manner for independence graphs, single continuous edges are drawn between the vertices to correspond to pair-wise associations or two-way interactions. In the case of continuous data, the inclusion of edges in the graph is dependent upon a default threshold value corresponding to the 5% significance level for the 'strengths' of the associations (negative partial correlations or edge exclusion deviances) or upon a threshold value specified by the user. If the strength of the association between a pair of variables exceeds the threshold value, then an edge is drawn in the graph between the two vertices corresponding to this pair of vertices. In the case of

discrete data, the inclusion of edges in the graph is dependent upon implication of the corresponding two-way interactions by the elements of the generating class, or upon the presence of non-zero model parameters for the two-way interactions. If the generating class of the model (discrete data only) is given as input, then this is sufficient to determine which edges are present in the independence graph, otherwise the generating class of the model must first be determined from the inputted model parameters.

If the initial, default, lay-out of the independence graph is not to the user's liking, it is a simple matter to click the mouse on any offending vertex and then, whilst holding the mouse button down, to 'drag' the vertex across the screen to a preferred position. The user will see the edges of the graph which are linked to this vertex being dragged along with the vertex as it moves.

Reasons why a user may wish to relocate the vertices may include aesthetic reasons, such as removing cross-overs which may the graph difficult to read, and interpretative reasons; for example, the user may wish to cluster variables together to highlight the associations within and between different groups of variables.

It could be argued that there are alternative methods of obtaining a default representation. For example, a PCA or MDS solution could be used to determine a lay-out of the points, but the usefulness of such a representation was questioned in Chapter 10. Another method might ideally represent the independence graph with the minimum possible number of cross-overs, so as to improve the readability of the graph. However, incorporation of a routine to execute Nicholson's algorithm in CIGE (described in Chapter 8) is rejected on the grounds that it does not necessarily give the best solution (although it is the best algorithm that I know of), and, in particular, because it cannot be readily implemented. The computation is straightforward, but the graph construction presents too many problems.

Hence the user is provided with an initial construction of the graph which is easy to construct and involves very little computation, even though the resultant representation is not necessarily the best, either in terms of cross-overs or on aesthetic grounds (the latter problem being one which almost certainly could not be solved by a programmable algorithm). If the user wishes to employ the result of Nicholson's algorithm applied to the graph under consideration, or to implement a PCA or MDS solution, the facility to drag the vertices and edges may be used. However, the desired lay-out of the points must be determined outside of CIGE.

11.2.3 Incorporation of Edge Codes for Interactions

This feature of CIGE is only appropriate for, and hence only available for, models for discrete data.

The edge codes described in Chapter 9, whereby each edge is drawn using a particular edge code according to the largest interaction involving that edge and which leads to construction of the conditional interaction graph, have been incorporated into CIGE. At any stage the user may display the interaction graph in preference to the independence graph. A key can be displayed together with the graph, to indicate which edge code corresponds to each order of interaction.

11.2.4 Display of Generating Class Elements

This feature of CIGE is only appropriate for, and hence only available for, models for discrete data.

In order to identify uniquely the model corresponding to a displayed independence or interaction graph, it is necessary to be able to determine the elements of the generating class of the model represented. As has already been seen (Chapters 7 and 9), it is not always possible to determine the elements of the generating class from the displayed graph, only to identify the corresponding graphical (independence or interaction) model, of which the model actually represented may be a sub-model.

Thus CIGE contains a facility whereby the elements of the generating class of the model represented may be displayed individually, as sub-graphs of the independence or interaction graph which is displayed. Having selected the option to do this, a second window is displayed (the original graph being contained in the first), and the sub-graphs corresponding to each element may be displayed either simultaneously, or sequentially. The edges of the sub-graphs will be edge codes appropriate to the order of the interaction of the element. The positions of the vertices will correspond to those in the main graph, and will be labelled with the initial letter of each variable label used in the main graph. However, if the user relocates the vertices of the main graph whilst the second window is open, the vertices of the sub-graphs will be not be relocated until the sub-graphs are re-displayed.

11.2.5 Incorporation of Edge Styles for Associations

This feature of CIGE is only appropriate for, and hence only available for, models for continuous data.

For continuous data, the original independence graph (the interaction graph not being available for this data type) can be modified to incorporate the edge styles described in Chapter 10. All three types of edge styling described in Chapter 10 are available — namely width, grey-tone shading, and width and grey-tone shading combined. Six levels of each (consisting of five 'non-zero' levels, and one 'zero' or 'sub-threshold' level) are available, since (as mentioned previously in Chapter 10) this is the number of different widths or shades which are readily distinguishable given the pixel resolution of the screen. The user must specify the edge style to be used, and must also specify which of three approaches should be used to find the boundary values between each of the six levels. For all of these options, only the absolute values of the associations are considered. CIGE can calculate and use boundary values based on equidistant values between the two extreme data values, or calculate and use boundary values based on significance levels. Alternatively, the user may specify boundary values between the levels, the lower and upper limits being zero and the maximum association respectively for this option. For further details, see Appendix A.

11.2.6 Use of Sliders

This feature of CIGE is only appropriate for, and hence only available for, models for continuous data.

Through the use of a slider, it is possible to obtain a dynamic display of the absolute strengths of the associations represented in a given independence graph. This is an alternative approach to the use of edge styles in the previous section (11.2.5), and is a useful tool for determining a threshold value below which associations will be regarded as being zero (ie. for determining which edges it is appropriate to drop from the fitted model).

In the initial display, the graph corresponding to the current threshold value is drawn, and above the graph a slider is displayed of which the lower limit is zero and the upper limit corresponds to the maximum absolute strength of association in the data. The

slider is initially at the current threshold value used in the construction of the displayed graph, but the mouse may be used to move the slider between the upper and lower limits. As the slider moves, edges are added to or dropped from the graph as appropriate for the threshold value of the strength of association corresponding to the current position of the slider. This threshold value is displayed above the slider, and is updated as the slider moves. Thus the slider can be used as a tool for determining a new threshold value for the strength of association in an *ad hoc* manner.

11.2.7 Numerical Information from Vertices, Edges and Cliques

Continuous Data

For continuous data, the user may obtain the numerical value of the strength of any pair-wise association within the data. When this option is chosen, the complete graph on the vertices is displayed, with edges set to 'zero' (ie. below the current threshold value) represented as thinner continuous lines than those used to represent 'non-zero' edges. Using this graph, the user may click the mouse on the edges which are missing from the independence graph, and thus obtain the true value (with sign) of the association (which is not necessarily zero) between any pair of unlinked variables. In the same way, the user may also obtain the true value (with sign) of the association between any pair of linked variables.

Discrete Data

For discrete data, if the data used as input is in the form of parameter estimates, the user may click the mouse on a single vertex (corresponding to a main effect), on two vertices (corresponding to a two-way interaction effect), or on three or more vertices (corresponding to a higher order interaction effect), in order to obtain the parameter estimates of the terms of the model represented by the displayed independence or interaction graph. The estimates are displayed in the form of a table with one entry per level of the variable (in the case of a main effect), or one entry per combination of the levels of the variables (in the case of an interaction effect).

If the data used as input is in the form of the generating class, this option can not be chosen since the requested numerical information is not available.

Sign of Association

This option is only appropriate for, and therefore only available for, models for continuous data for which the partial correlation coefficients are used as input. For models for discrete data, it is possible to determine the sign of the association by examination of the parameter values (see above).

When this option is selected, the independence graph is re-displayed with the usual continuous lines drawn between pairs of variables having positive values for their partial correlations, and thinner continuous lines drawn between pairs of variables having negative values for their partial correlation. Because the true values of the partial correlations are used, the signs represented will be the opposite of those contained in the matrix of negative partial correlations.

If the matrix of edge exclusion deviances was used as input, it is not possible to examine the sign of the association between pairs of variables, since the value of the edge exclusion deviance will be positive for all pairs of variables.

11.3 Summary

In this chapter, the Conditional Independence Graph Enhancer (CIGE) package was introduced. This package has been designed to implement the work and ideas concerned with the use of conditional independence graphs for the representation of fitted models presented in preceding chapters, but also incorporates a number of other ideas.

The main features of the CIGE package, including the data input and graph construction and manipulation, as well as the encoding and accessing of numerical information were outlined. Full details of these features and of the use of CIGE are contained in Appendix A.

The usefulness of the various features of CIGE when applied to models fitted to a number of actual data sets will be illustrated in the next chapter.

BLANK PAGE IN ORIGINAL

12. Examples Using CIGE

12.1 Introduction

In this chapter, I shall describe the use of CIGE for the representation of models fitted to various example data sets.

Covariance selection models have been fitted, using the naive covariance selection modelling technique described in Chapter 7, to two continuous data sets. The continuous data sets used are as follows:

1. The pit-prop data set considered by Jeffers (1967). This data set involves 13 variables, consisting of various measurements relating to the maximum compressive strength of N=180 pit-props manufactured from Corsican pine trees. This data set has already been examined in Chapter 10 within the context of the encoding of strength of association by distance, and has also been used by Whittaker (1988, 1990), in a graphical modelling context.
2. The kangaroo skeleton data set of Poole, presented in Andrews & Herzberg (1985). This data set involves 18 variables, consisting of various measurements made on N=148 kangaroo skeletons. Although measurements were made on three different species of kangaroo of both sexes, no distinction has been made between the species or the sexes for the purposes of illustrating CIGE. Missing values were replaced by the intra-species intra-variable mean prior to the calculation of the correlation matrix. This data set was also examined in Chapter 10 within the context of the encoding of strength of association by distance.

The use of CIGE for the enhancement of conditional independence graphs for covariance selection models is described by consideration of each of the available menu options in turn. This is done in some detail for the pit-prop data set but, for the other continuous data sets, only the more interesting aspects of the displayed models which can be determined using CIGE will be presented.

A discrete data set will also be used to illustrate CIGE, to which a log-linear interaction model has been fitted. This is the byssinosis incidence data analysed by Higgins & Koch (1977), and also contained in Andrews & Herzberg (1985). This data set

forms a 3×2×2×2×3×2 contingency table, and is one of very few large multidimensional contingency tables which have been published. Although one of the variables in the data set is, strictly speaking, a dichotomous response variable, and so it is perhaps more appropriate to fit a logistic model to the data, I have chosen to treat all the variables as independent variables for the purposes of illustrating CIGE. In addition to this data set, a number of contrived generating classes of models, first considered in Chapter 6, are reconsidered using CIGE, in order to demonstrate the usefulness of the features of CIGE for uniquely determining the represented models. This achievement should be contrasted with the limited successes of the simple two-dimensional combination of points used in Section 6.2, and of the other techniques considered in that chapter, for the representation of these generating classes.

In describing the use of CIGE to investigate these example data sets, my intention is to provide the reader with a detailed illustration of how CIGE can be used for the representation and interpretation of covariance selection models for continuous data and of log-linear interaction models for discrete data. In particular, I wish to show how the representation can be used as part of an interface formed by CIGE and the graph in order to obtain additional information about the model displayed, in either numerical or pictorial form, which cannot be determined directly from the independence graph. The description is illustrated with screen-dumps made during actual interactive use of the package, and some suggestions are made about what can be learnt about the data and the fitted models using CIGE.

Actual instructions on the use of CIGE and details of the design of the package are omitted from this chapter. The interested reader is referred to Appendix A for details of the design of the package and instructions on how to run CIGE, and for details of how to use each of the various menu options available to obtain the screen-dumps used to illustrate the Jeffers pit-prop data and Higgins & Koch byssinosis data examples.

12.2 Covariance Selection Models

12.2.1 Pit-Prop Data Set

The 13 variables in this data set are as listed in Table 12-1, and the lower-triangular matrix of correlations between these 13 variables is presented in Figure 12-1.

LABEL	VARIABLE
DIA	Top diameter of prop (inches)
LEN	Length of prop (inches)
MOIST	Moisture content of prop (% dry weight)
SPG1	Specific gravity of timber
SPG2	Oven-dry specific gravity of timber
RNGT	Number of annual rings at top
RNGB	Number of annual rings at bottom
BOW	Maximum bow (inches)
BDIST	Distance of point of maximum bow from top (inches)
WHORL	Number of knot whorls
CLEN	Length of clear prop from top
KNOT	Average number of knots per whorl
KDIA	Average diameter of knots (inches)

Table 12-1: Pit-prop data: Variable list

Given the correlation matrix, the naive approach to graphical modelling described in Chapter 7 was used in order to obtain the lower-triangular matrix of (negative) partial correlations presented in Figure 12-2, from which the lower-triangular matrix of edge exclusion deviances presented in Figure 12-3 was derived.

Data Input

Although the data is available in two forms (see Figures 12-2 and 12-3), the matrix of negative partial correlations will be used as input. Although the absolute values of the partial correlations are monotonically related to the edge exclusion deviances, the edge exclusion deviances do not preserve information about the sign of the association.

For the pit-prop data set, the variable labels used shall be as given in Table 12-1, and the number of observed units is N=180 pit-props.

Figure 12-1: Pit-prop data: Lower triangular matrix of correlations

	DIA	LEN	MOIST	SPG1	SPG2	RNGT	RNGB	BOW	BDIST	WHORL	CLEN	KNOT
LEN	0.954											
MOIST	0.364	0.297										
SPG1	0.342	0.284	0.882									
SPG2	-0.129	-0.118	-0.148	0.220								
RNGT	0.313	0.291	0.153	0.381	0.364							
RNGB	0.496	0.503	-0.029	0.174	0.296	0.813						
BOW	0.424	0.419	-0.054	-0.059	0.004	0.090	0.372					
BDIST	0.592	0.648	0.125	0.137	-0.039	0.211	0.465	0.482				
WHORL	0.545	0.569	-0.081	-0.014	0.037	0.274	0.679	0.557	0.526			
CLEN	0.084	0.076	0.162	0.097	-0.091	-0.036	-0.113	0.061	0.085	-0.319		
KNOT	-0.019	-0.036	0.220	0.169	-0.145	0.024	-0.232	-0.357	-0.127	-0.368	0.029	
KDIA	0.134	0.144	0.126	0.015	-0.208	-0.329	-0.424	-0.202	-0.076	-0.291	0.007	0.184

Figure 12-2: Pit-prop data: Lower triangular matrix of (negative) partial correlations

	DIA	LEN	MOIST	SPG1	SPG2	RNGT	RNGB	BOW	BDIST	WHORL	CLEN	KNOT
LEN	-0.863											
MOIST	-0.102	0.045										
SPG1	-0.007	-0.011	-0.929									
SPG2	0.062	0.001	0.663	-0.693								
RNGT	-0.015	0.047	0.037	-0.175	0.015							
RNGB	-0.063	-0.058	0.113	0.019	-0.059	-0.858						
BOW	-0.175	0.102	-0.069	0.118	-0.109	-0.005	0.062					
BDIST	0.166	-0.303	-0.009	-0.030	0.022	0.155	-0.176	-0.212				
WHORL	-0.043	-0.115	0.035	-0.051	0.097	0.461	-0.542	-0.244	-0.026			
CLEN	-0.044	-0.088	-0.034	0.021	0.038	0.092	-0.093	-0.171	-0.100	0.510		
KNOT	-0.058	-0.047	0.022	-0.021	0.112	-0.211	0.191	0.191	-0.077	0.156	0.136	
KDIA	-0.084	-0.185	0.046	0.002	-0.007	-0.038	0.241	0.116	0.017	0.191	0.208	0.079

Figure 12-3: Pit-prop data: Lower triangular matrix of edge exclusion deviances

	DIA	LEN	MOIST	SPG1	SPG2	RNGT	RNGB	BOW	BDIST	WHORL	CLEN	KNOT
LEN	246.08											
MOIST	1.88	0.36										
SPG1	0.01	0.02	357.70									
SPG2	0.70	0.00	104.44	117.96								
RNGT	0.04	0.41	0.25	5.62	0.04							
RNGB	0.72	0.61	2.30	0.07	0.62	239.76						
BOW	5.59	1.87	0.85	2.54	2.14	0.01	0.69					
BDIST	5.01	17.39	0.02	0.16	0.09	4.39	5.67	8.25				
WHORL	0.34	2.40	0.23	0.47	1.70	43.06	62.48	11.02	0.13			
CLEN	0.36	1.40	0.21	0.08	0.26	1.53	1.56	5.35	1.81	54.22		
KNOT	0.62	0.40	0.09	0.08	2.27	8.16	6.66	6.72	1.06	4.45	3.38	
KDIA	1.27	6.30	0.37	0.00	0.01	0.26	10.75	2.45	0.05	6.71	7.96	1.13

Basic Display

The initial display of the basic independence graph, using the default 5% threshold value for the absolute value of the negative partial correlation coefficient $-\rho_{ij.K}$ and the default lay-out of the vertices equi-distant around the circumference of a circle, is presented in Figure 12-4.

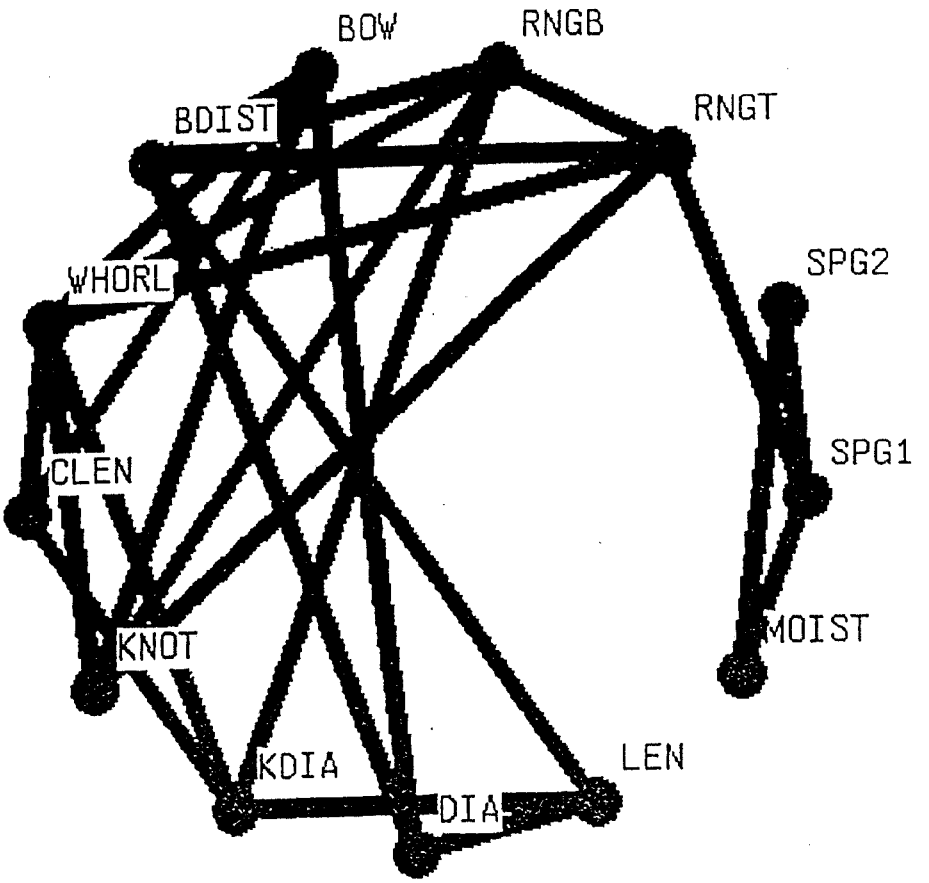

Figure 12-4: CIGE: Default lay-out of basic independence graph for pit-prop data model

It is already possible to discern some relationships between the variables. In particular there is a clique formed by MOIST, SPG1 and SPG2, which are variables related to the moisture content and wet and dry specific gravities of the timber respectively. However, there are a large number of very long edges (for example, the variables BDIST and LEN, although associated, are drawn far apart) and there are several cross-overs within the graph. The user might therefore begin by attempting to improve the

layout of the graph, so that the relationships between the variables can be discerned more easily.

By clicking and dragging the mouse on any of the vertices in this default display and relocating them, it is possible to obtain a much more aesthetically pleasing display of the basic independence graph, such as the display shown in Figure 12-5. The basic independence graph now has just one cross-over, and the associations between the variables can be determined more readily, although some of the edges in the graph are still very long. The conditional independence relationships between the variables in this data set may now be readily determined according to the missing edges in the graph (see Chapter 7). For example, it can be seen that BOW (the maximum bow of a pit-prop) and LEN (the length of the prop) are conditionally independent given DIA (the top diameter of the prop).

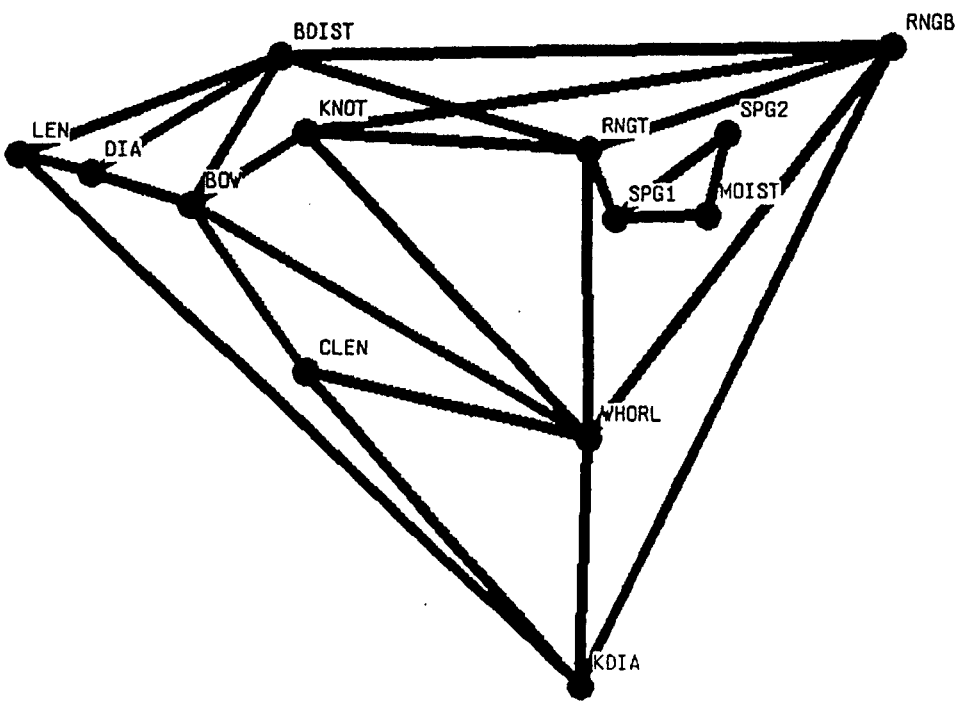

Figure 12-5: CIGE: Modified lay-out of basic independence graph for pit-prop data model

By clicking the left-hand mouse button anywhere within the window in which the graph is drawn, it is possible to display the menu of options listed in Table 12-2. The

function of each of these options will be described in turn in the remainder of this section with reference to the pit-prop data set.

Numerical Values
Encoded Strengths
Slider
Change Threshold
Signed Graph

Table 12-2: List of menu options for use with covariance selection models

Menu Option: Numerical Values

When this menu option is selected, the complete graph on all 13 vertices is drawn as shown in Figure 12-6. The thick edges correspond to the edges in the basic independence graph, and the thinner edges correspond to associations which fall below the current threshold level and are therefore not displayed in the basic independence graph. By clicking the mouse on any edge (thick or thin) between any pair of vertices, the value of the association between the corresponding pair of variables can be obtained. Through the use of this menu option it is therefore possible to determine the value of the associations between any or all pairs of variables.

In this example, the value of the association displayed will correspond to the partial correlation between the variables, the negative values of which have been presented in Figure 12-2 and were used as input. If the matrix of edge exclusion deviances had been used as input, then the value of the association displayed would correspond to the edge exclusion deviance, as presented in Figure 12-3. For example, if the user were to click on the edge between RNGB and KDIA, the strength of the association would be reported as –0.241 — thus there is a negative partial correlation between the number of rings at the base of the pitprop and the average diameter of the knots, and since the corresponding edge is thick, this association is significant at the 5% level.

Some idea of the strength and sign of the associations between every pair of variables can also be obtained, encoded within the graph in pictorial form, by the use of other menu options described below.

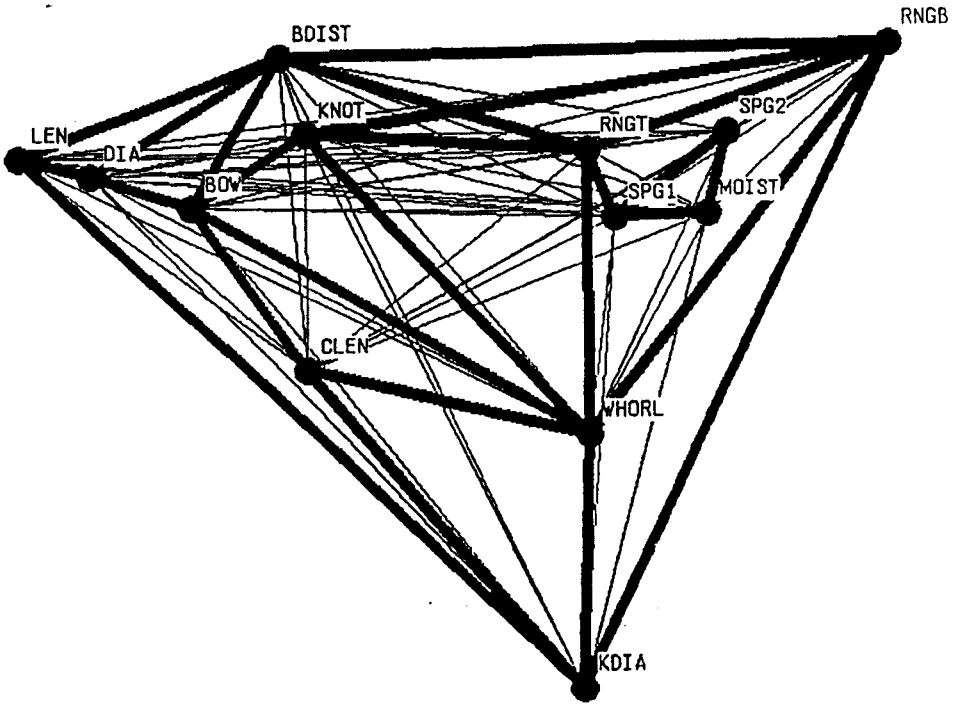

Figure 12-6: CIGE: Use of numerical value menu option for pit-prop data model

Menu Option: Encoded Strengths

This menu option has two sub-menus, one corresponding to the edge styles which can be chosen to encode the strength of association, and one corresponding to the method of determining the boundaries between the six different levels of strength. These sub-menus are presented in Table 12-3.

Widths
Grey Shades
Combination

Default 1: Significance Levels
Default 2: Equi-spaced
User-defined boundaries

Table 12-3 Sub-menus for use with Encoded Strengths menu option for covariance selection models

In Figure 12-7, the independence graph is presented for the pit-prop data set in which the strengths of the associations have been encoded by the widths of the edges, and the boundaries between the associations corresponding to each of the six levels of width of edge (including no edge) have been determined by the use of the default significance levels (see Section 11.2.5 of Chapter 11). Most of the edges appear quite thick, which implies that the associations represented by these edges are highly significant. There are some thin edges, for example between KNOT (the average number of knots per whorl) and WHORL (the number of knot whorls), and between KNOT and CLEN (the length of clear prop from the top), implying that these associations have low significance and that it may not be appropriate to attach too much importance to them. The precise value of these associations can be determined using the numerical values menu option already described.

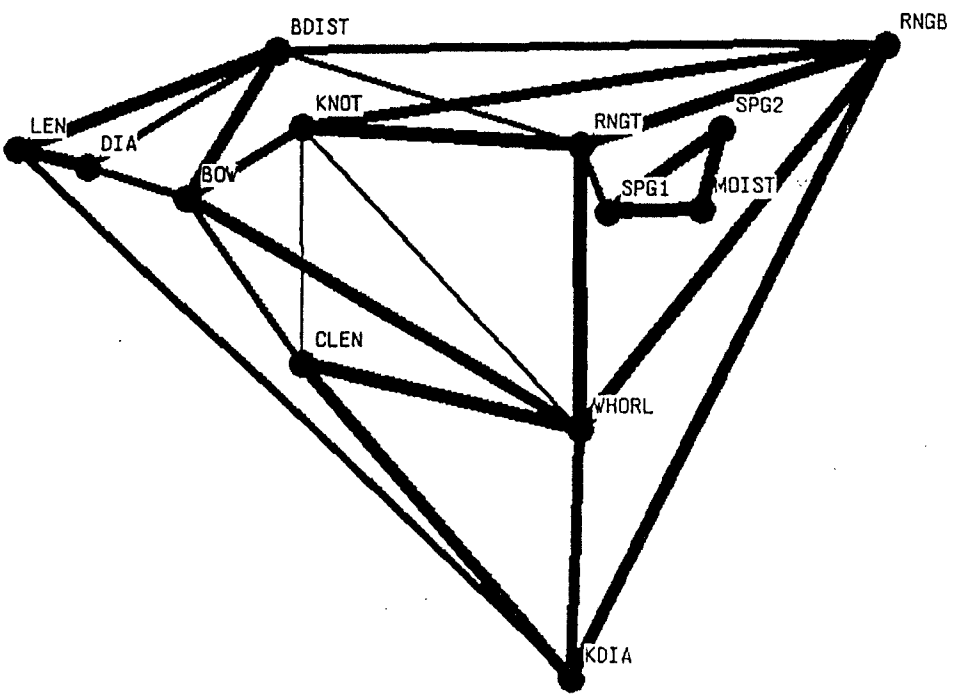

Figure 12-7: CIGE: Use of encoded strengths menu option, with default levels chosen by significance level and strength encoded by width for pit-prop data model

In Figure 12-8, the independence graph is presented for the pit-prop data set in which the strengths of the associations have been encoded by the grey-tone shading of the edges, and the boundaries between the associations corresponding to each of the six levels of grey-tone shading (including no edge) have been determined by the use of the default

equi-spacing (see Section 11.2.5 of Chapter 11). Use of the equi-spacing option means that the styles assigned to the edges correspond to the *relative* importance of the edges, whereas the use of the significance levels option, illustrated above, meant that the styles assigned to the edges corresponded to the importance of the edges in terms of their *significance*. Thus in Figure 12-8, it can be seen that only a few edges, represented by darker lines, have relatively strong associations, whereas the other edges have negligible associations (corresponding to no edge) or relatively small associations (corresponding to the faintest edges). The strongest associations are between RNGB and RNGT (both relating to the number of rings), between LEN and DIA (both relating to the size of the prop), and between SPG1 and MOIST (corresponding to the specific gravity of the pit-prop when wet and the moisture content of the pit-prop). Other strong associations involve SPG2 (the specific gravity of the pit-prop when dry) with SPG1 and MOIST.

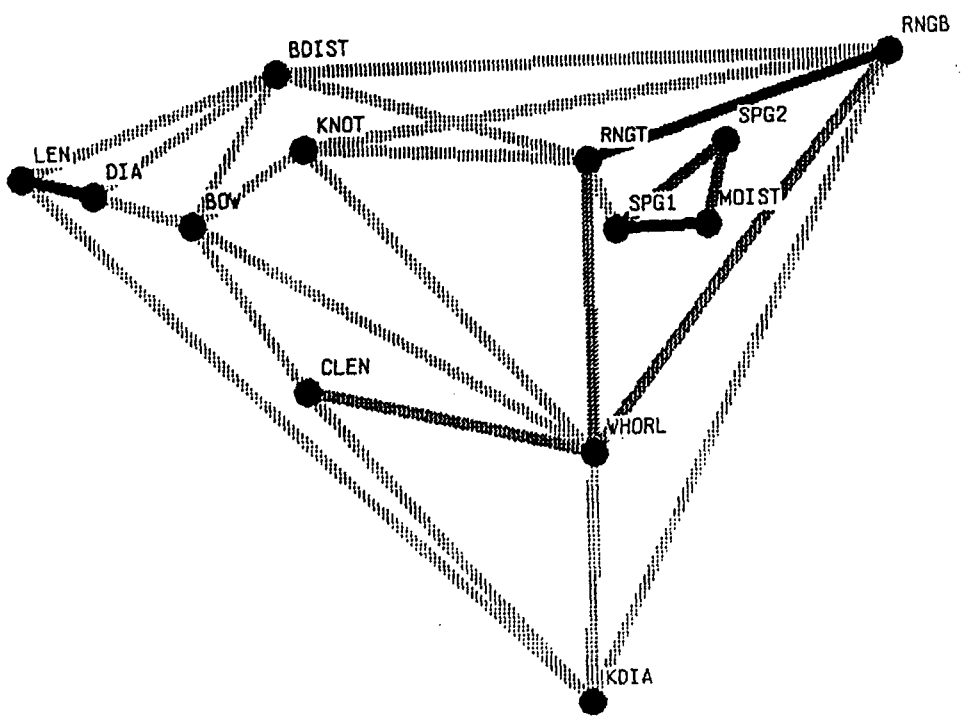

Figure 12-8: CIGE: Use of encoded strengths menu option, with default levels chosen by equi-distant spacing and strength encoded by grey-tone shading, for pit-prop data model

Use of the third option, for which the user must specify the boundaries between the associations corresponding to each of the six levels of encoding, has not been

illustrated here since it is not obvious what an appropriate choice of boundary values would be for this example.

Menu Option: Slider

In Figure 12-9, it can be seen how the slider initially appears when this menu option is chosen. The position of the slider and the appearance of the graph correspond to the current (default) threshold value of 0.1453. which is the (absolute) value of the partial correlation coefficient corresponding to the 5% significance level.

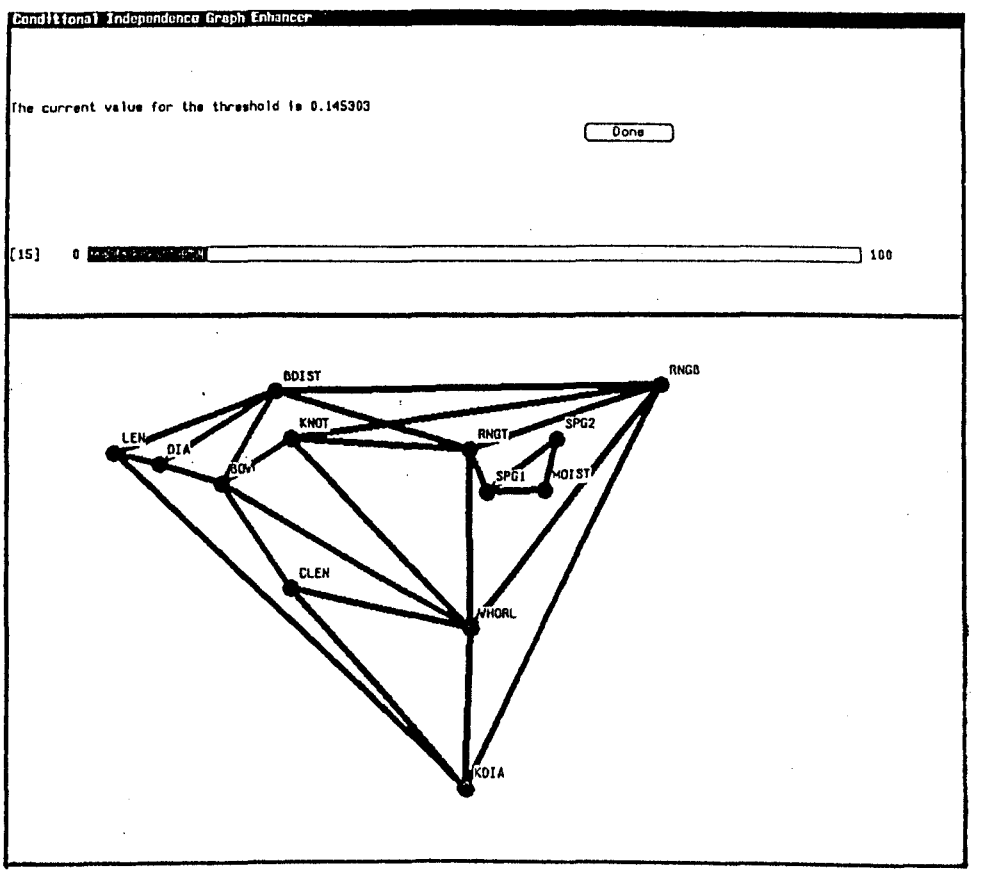

Figure 12-9: CIGE: Use of slider menu option, with threshold value=0.1453, corresponding to the 5% significance level, for pit-prop data model

In Figure 12-10, the slider has been moved to show how the independence graph would appear for a threshold value of 0.1951, which approximates, within the limits of the resolution of the slider, the 1% significance level of 0.1902. The edges remaining in the

graph will correspond to those having the thickest edges in the graph in Figure 12-7, since they represent the stronger, more significant, associations.

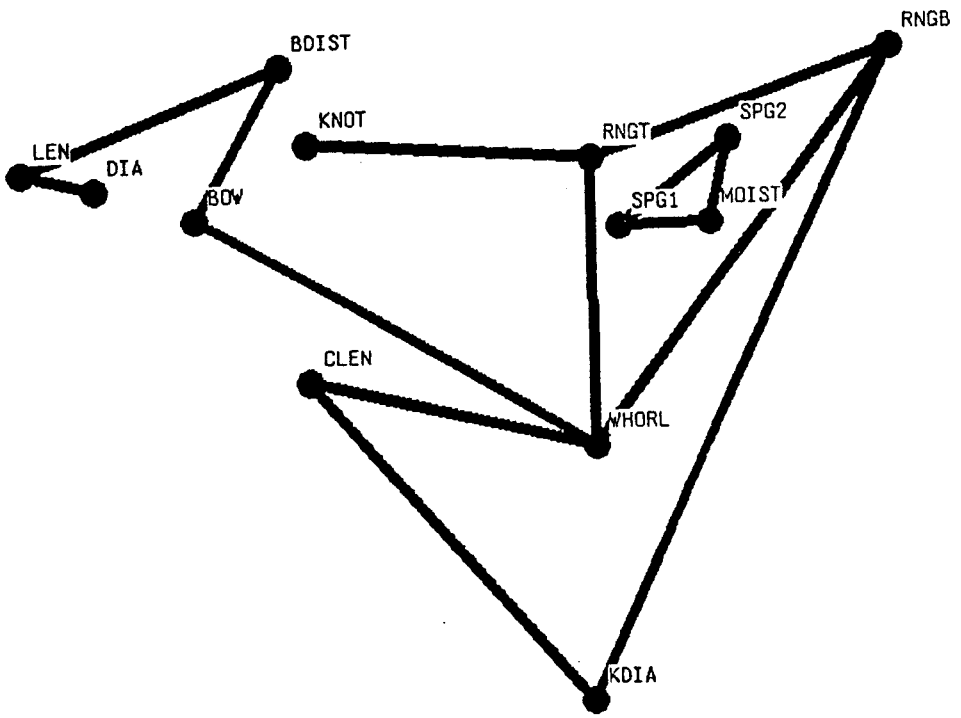

Figure 12-10: CIGE: Use of slider menu option, with threshold value=0.1951, corresponding to the 1% significance level, for pit-prop data model

Figure 12-11 shows how the independence graph appears for a much greater threshold value of 0.2508, which approximately corresponds to the much higher significance level of 0.1% (which has a true value of 0.2416). This indicates that the associations represented by the remaining edges are very strong indeed. By comparing this graph with the graph in Figure 12-8 it can be seen that some of the edges remaining must correspond to associations which are, relatively speaking, even more significant. In particular, there is a definite clique formed by SPG1, SPG2 and MOIST, which is independent of all other variables at this level of significance.

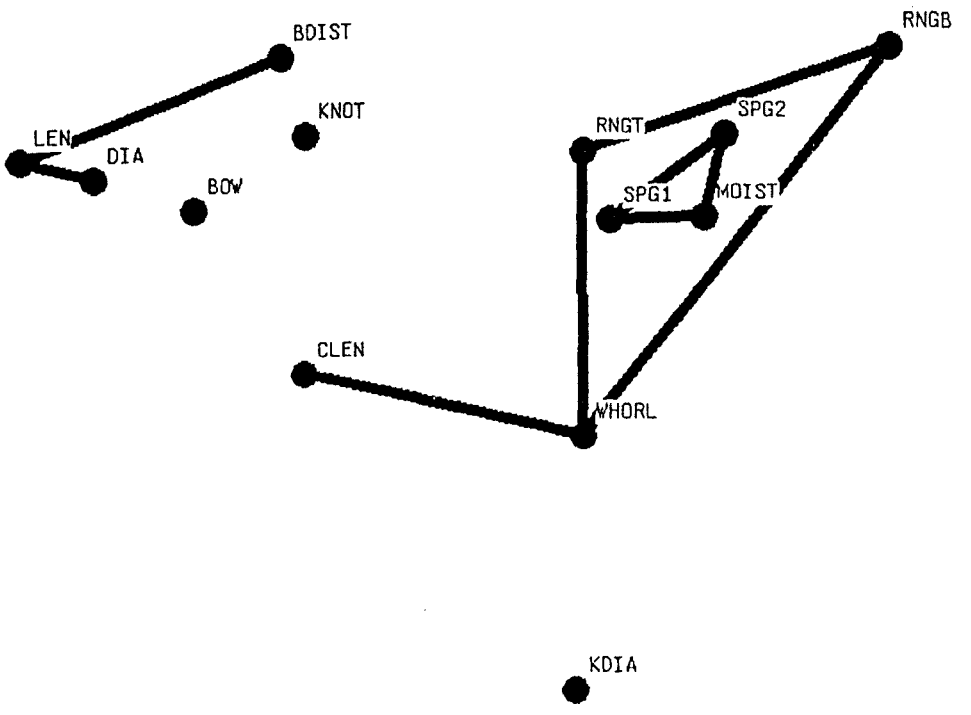

Figure 12-11: CIGE: Use of slider menu option, with threshold value=0.2508, corresponding to the 0.1% significance level, for pit-prop data model

Menu Option: Changing the Threshold

Use of the slider allows the user to examine the effect of changing the threshold in an exploratory, transient, manner. An alternative approach is to change the actual threshold value explicitly and permanently. In Figure 12-12, the threshold value has been changed to that corresponding exactly to the 1% significance level (ie. to 0.1902) and the independence graph automatically redrawn accordingly. It can be seen that the edges in this graph do not correspond exactly to those in the graph drawn using the slider option for a threshold value of 0.1951 presented in Figure 12-10. This suggests that if it is very important to determine which edges correspond to associations which are significant at a particular level, then it is better to change the threshold value explicitly, but for most purposes use of the slider will be adequate and a new threshold value need only be explicitly specified if a permanent change is required.

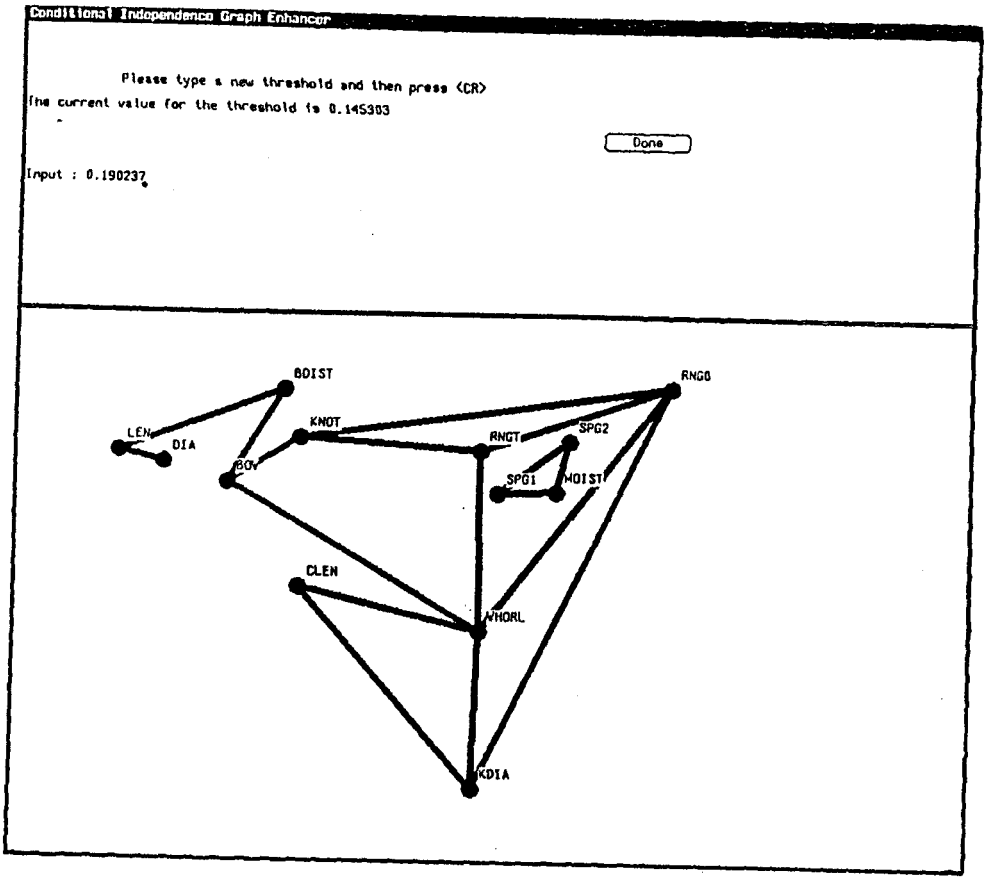

Figure 12-12: CIGE: Use of the change threshold menu option to change the threshold to 0.1902, corresponding to the 1% significance level, for pit-prop data model

Menu Option: Signed Graph

The signed graph corresponding to the independence graph constructed for the 5% significance level is presented in Figure 12-13. In this graph the signs of the partial correlations are represented (which will be the opposite of those contained in the matrix of negative partial correlations used as input). Thus the thin edge between MOIST and SPG2 corresponds to a negative partial correlation between the moisture content of the pit-prop and the oven dry specific gravity of the pit-prop, whereas the thick edge between MOIST and SPG1 corresponds to a positive partial correlation between the moisture content of the pit-prop and the specific gravity of the wet pit-prop.

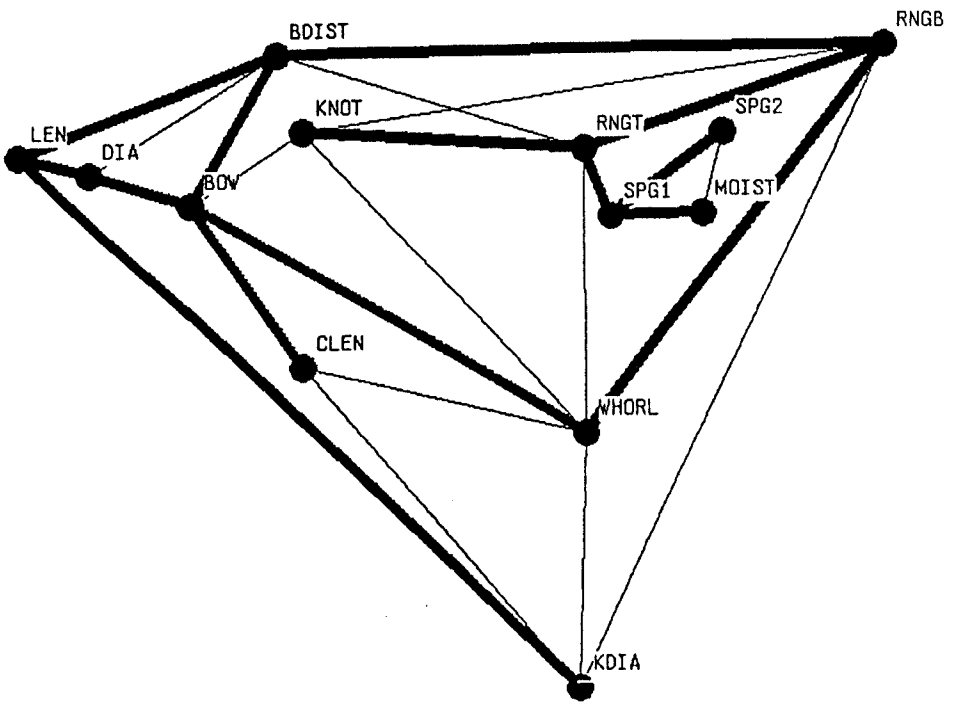

Figure 12-13: CIGE: Use of signed graph menu option, for pit-prop data model

12.2.2 Kangaroo Data Set

The 18 variables in this data set are as listed in Table 12-4, and the lower-triangular matrix of correlations between these 18 variables is presented in Figure 12-14.

Given the correlation matrix, the naive approach to graphical modelling described in Chapter 7 was used to obtain the lower-triangular matrix of (negative) partial correlations presented in Figure 12-15, from which the lower-triangular matrix of edge exclusion deviances presented in Figure 12-16 was derived.

LABEL	VARIABLE
BASIL	Basilar length
OCCIP	Occipitonasal length
PAL-L	Palatilar length
PAL-W	Palate width
NAS-L	Nasal length
NAS-W	Nasal width
SQUAM	Squamosal depth
LACRY	Inter-lacrymal width
ZYGOM	Zygomatic width
ORBIT	Post orbital width
ROST	Rostral width
SUPRA	Supra-occipital-paroccipital depth
CREST	Crest width
FORAM	Incisive foramina length
MAN-L	Mandible length
MAN-W	Mandible width
MAN-D	Mandible depth
RAMUS	Ascending ramus height

Table 12-4: Kangaroo skeleton data: Variable list

Data Input

Although the data is available in two forms (see Figures 12-15 and 12-16), the matrix of negative partial correlations will again be used as input.

For the kangaroo skeleton data set, the variable labels used shall be as given in Table 12-4, and the number of observed units is N=148 skeletons.

Basic Display

The default basic independence graph which is initially displayed is as presented in Figure 12-17. As can be seen from this graph, there are very many edges (47) corresponding to a large number of associations significant at the 5% level, and very many cross-overs. It is unlikely that a more pleasing representation of this model could be obtained easily by relocating the vertices.

Figure 12-14: Kangaroo skeleton data: Lower triangular matrix of correlations

	BASIL	OCCIP	PAL-L	PAL-W	NAS-L	NAS-W	SQUAM	LACRY	ZYGOM	ORBIT	ROST	SUPRA	CREST	FORAM	MAN-L	MANW	MAN-D
OCCIP	0.953																
PAL-L	0.985	0.947															
PAL-W	0.651	0.621	0.694														
NAS-L	0.836	0.934	0.847	0.569													
NAS-W	0.794	0.849	0.803	0.643	0.834												
SQUAM	0.727	0.670	0.717	0.517	0.540	0.639											
LACRY	0.923	0.904	0.919	0.661	0.787	0.851	0.737										
ZYGOM	0.898	0.807	0.891	0.666	0.637	0.663	0.770	0.879									
ORBIT	0.260	0.290	0.245	0.100	0.247	0.360	0.240	0.363	0.262								
ROST	0.866	0.821	0.853	0.613	0.720	0.692	0.724	0.838	0.827	0.145							
SUPRA	0.904	0.883	0.907	0.637	0.759	0.765	0.775	0.898	0.901	0.275	0.816						
CREST	-0.663	-0.731	-0.686	-0.465	-0.740	-0.618	-0.347	-0.602	-0.443	0.062	-0.580	-0.587					
FORAM	0.234	0.276	0.264	0.172	0.348	0.345	0.229	0.284	0.149	0.192	0.234	0.231	-0.203				
MAN-L	0.933	0.899	0.929	0.664	0.806	0.758	0.684	0.890	0.875	0.236	0.841	0.884	-0.671	0.221			
MANW	0.789	0.682	0.778	0.504	0.482	0.514	0.652	0.753	0.876	0.243	0.669	0.798	-0.341	0.093	0.745		
MAN-D	0.846	0.769	0.819	0.542	0.612	0.642	0.698	0.800	0.868	0.275	0.751	0.828	-0.417	0.091	0.819	0.816	
RAMUS	0.929	0.844	0.910	0.621	0.678	0.664	0.765	0.892	0.944	0.233	0.840	0.906	-0.532	0.162	0.897	0.866	0.886

Figure 12-15: Kangaroo skeleton data: Lower triangular matrix of (negative) partial correlations

	BASIL	OCCIP	PAL-L	PAL-W	NAS-L	NAS-W	SQUAM	LACRY	ZYGOM	ORBIT	ROST	SUPRA	CREST	FORAM	MAN-L	MANW	MAN-D
OCCIP	-0.475																
PAL-L	-0.749	0.161															
PAL-W	0.250	0.039	-0.338														
NAS-L	0.248	-0.725	-0.227	0.038													
NAS-W	-0.126	0.019	0.088	-0.311	-0.228												
SQUAM	0.007	-0.001	-0.011	0.079	0.102	-0.271											
LACRY	0.058	-0.176	-0.059	0.050	0.206	-0.533	0.105										
ZYGOM	-0.021	0.024	-0.113	-0.246	0.075	0.116	-0.101	-0.157									
ORBIT	0.038	-0.146	-0.001	0.132	0.061	-0.092	0.027	-0.237	0.024								
ROST	-0.162	0.029	0.090	-0.038	-0.061	0.111	-0.205	-0.182	-0.141	0.200							
SUPRA	0.292	-0.254	-0.173	0.075	0.051	-0.103	-0.229	-0.032	-0.231	0.011	0.017						
CREST	-0.136	0.238	0.222	-0.028	-0.100	0.068	-0.140	0.027	-0.333	-0.374	0.037	0.081					
FORAM	0.075	0.133	-0.144	0.069	-0.183	-0.052	-0.098	-0.094	0.101	-0.085	-0.077	-0.021	-0.065				
MAN-L	-0.130	0.130	0.013	-0.083	-0.206	0.044	0.100	-0.061	-0.188	-0.083	-0.075	-0.073	0.272	0.010			
MANW	0.038	-0.089	-0.153	0.134	0.225	0.021	0.112	0.024	-0.271	0.009	0.178	-0.103	-0.075	-0.041	0.064		
MAN-D	-0.092	-0.029	0.077	0.080	0.047	-0.220	0.029	0.170	-0.050	-0.040	-0.040	-0.054	-0.161	0.139	-0.136	-0.146	
RAMUS	-0.310	0.082	0.145	-0.060	-0.001	0.364	-0.184	-0.301	-0.191	0.065	-0.004	-0.176	0.100	-0.014	-0.106	-0.181	-0.265

Figure 12-16: Kangaroo skeleton data: Lower triangular matrix of edge exclusion deviances

	BASIL	OCCIP	PAL-L	PAL-W	NAS-L	NAS-W	SQUAM	LACRY	ZYGOM	ORBIT	ROST	SUPRA	CREST	FORAM	MAN-L	MANW	MAN-D
OCCIP	37.91																
PAL-L	121.94	3.89															
PAL-W	9.55	0.23	17.96														
NAS-L	9.42	110.57	7.83	0.22													
NAS-W	2.38	0.06	1.16	15.05	7.93												
SQUAM	0.01	0.00	0.02	0.92	1.54	11.32											
LACRY	0.49	4.65	0.52	0.37	6.44	49.48	1.64										
ZYGOM	0.07	0.08	1.89	9.22	0.82	2.00	1.52	3.70									
ORBIT	0.21	3.20	0.00	2.61	0.55	1.26	0.11	8.54	0.09								
ROST	3.95	0.12	1.21	0.21	0.56	1.84	6.33	4.97	2.96	6.04							
SUPRA	13.18	9.86	4.51	0.83	0.38	1.57	7.97	0.15	8.12	0.02	0.04						
CREST	2.76	8.60	7.50	0.12	1.49	0.68	2.93	0.10	17.45	22.24	0.21	0.98					
FORAM	0.84	2.64	3.12	0.70	5.06	0.41	1.44	1.32	1.50	1.08	0.87	0.07	0.64				
MAN-L	2.53	2.52	0.03	1.01	6.45	0.29	1.50	0.55	5.30	1.02	0.84	0.78	11.38	0.01			
MANW	0.21	1.18	3.52	2.68	7.67	0.07	1.86	0.08	11.31	0.01	4.76	1.57	0.83	0.25	0.60		
MAN-D	1.26	0.13	0.89	0.94	0.33	7.31	0.13	4.36	0.37	0.23	0.24	0.43	3.91	2.89	2.78	3.20	
RAMUS	15.00	1.01	3.16	0.53	0.00	21.09	5.12	14.02	5.49	0.63	0.00	4.65	1.49	0.03	1.68	4.92	10.77

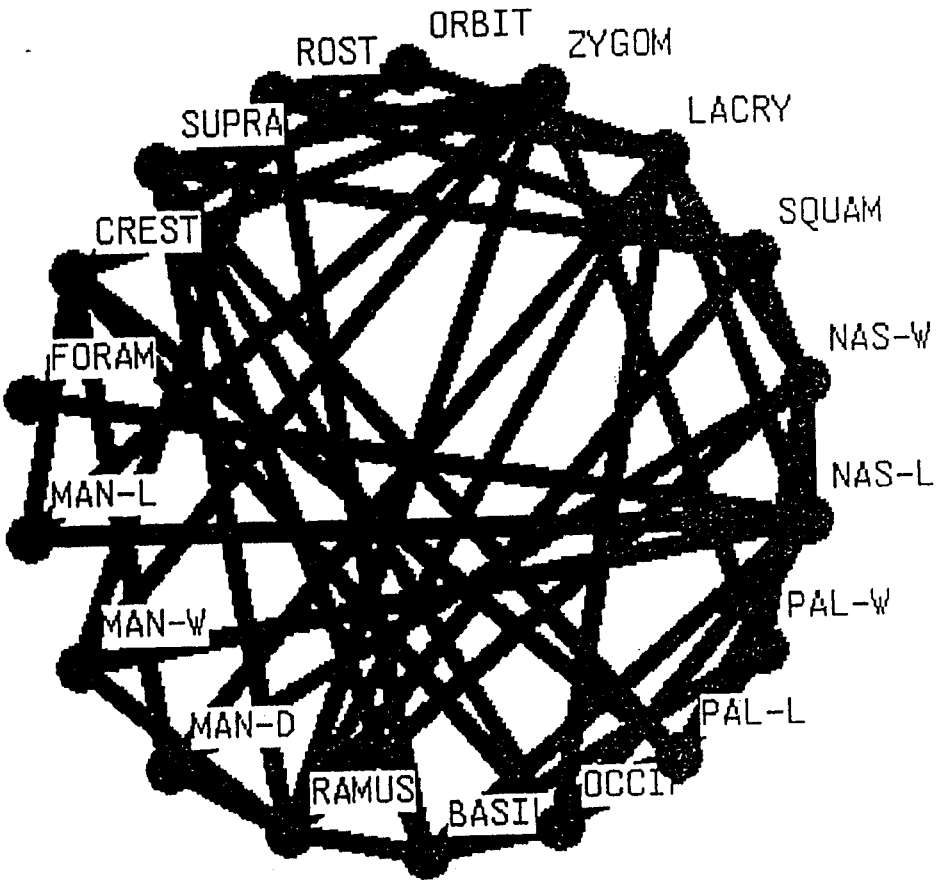

Figure 12-17: CIGE: Default lay-out of basic independence graph for kangaroo skeleton data model

To obtain a more pleasing display, the user could first simplify the model by reducing the number of significant associations by using a higher significance level. Using the appropriate menu option to change the threshold according to the new significance level would have the effect of reducing the number of edges in the graph. For example, Figure 12-18 shows the independence graph with the default lay-out of vertices constructed for the threshold value corresponding to the 1% significance level, and Figure 12-19 shows the independence graph constructed for the 0.5% significance level.

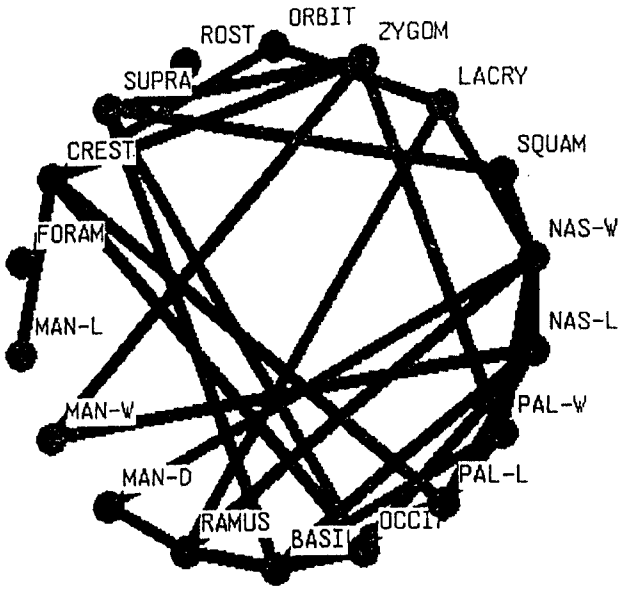

Figure 12-18: CIGE: Default lay-out of basic independence graph for kangaroo skeleton data model with threshold value changed to 1% significance level

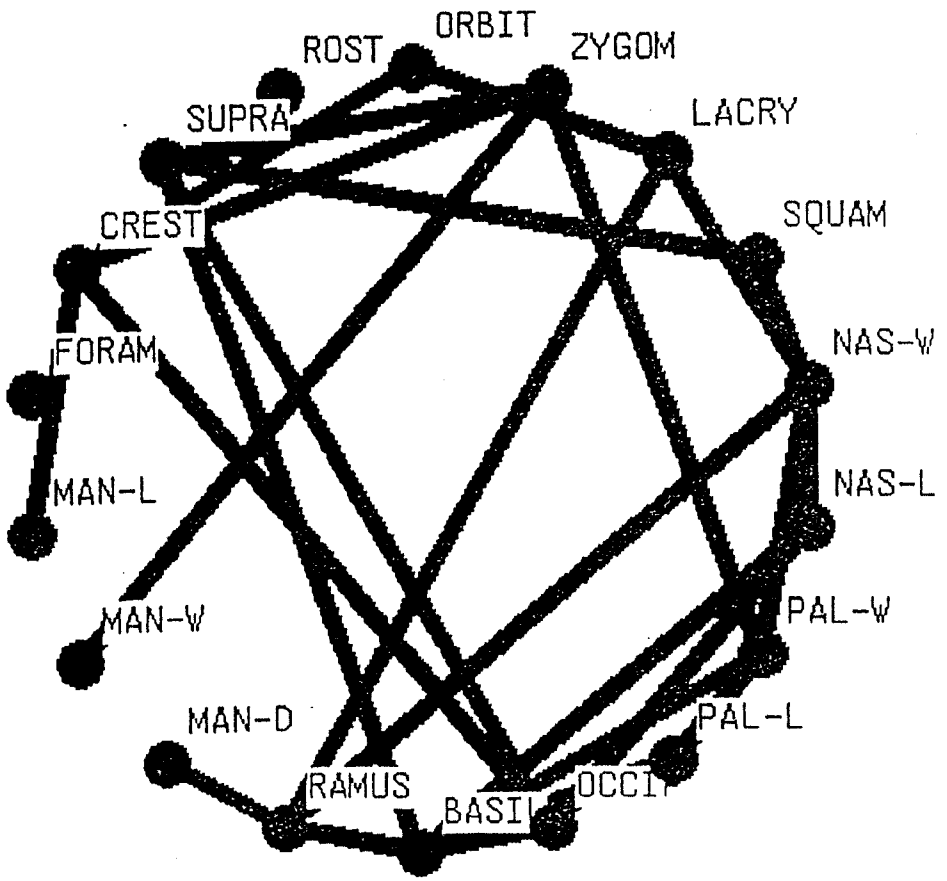

Figure 12-19: CIGE: Default lay-out of basic independence graph for kangaroo skeleton data model with threshold value changed to 0.5% significance level

The number of edges in the graph, corresponding to only the more significant associations, is now sufficiently small (having been reduced from 47 to 25) to enable the vertices to be relocated by hand and eye to give a more pleasing representation of this particular independence graph. One possible representation is as shown in Figure 12-20. It can be seen from this graph that the two variables FORAM (incisive foramina length) and ROST (rostral width) are not associated (at the 0.5% level) with any other variables, whereas the other 16 variables are all strongly associated with at least one other.

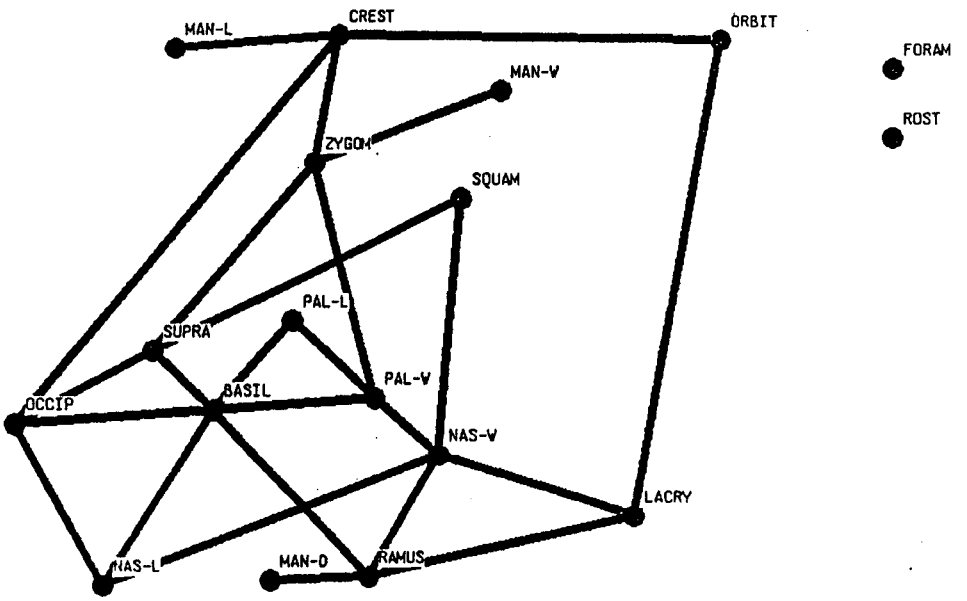

Figure 12-20: CIGE: Modified lay-out of basic independence graph for kangaroo skeleton data model (with threshold value corresponding to 0.5% significance level)

It is difficult to interpret the associations and cliques between the variables without specialised knowledge of kangaroo anatomy, but it can be seen, for example, that PAL-L and PAL-W (palatilar length and width) are associated, as are NAS-L and NAS-W (nasal length and width), whereas there are no associations between MAN-L, MAN-W and MAN-D (mandible length, width and depth) at this level of significance.

Menu Option: Numerical Values

The numerical values option was not employed with this data set because, with 18 variables in the data set, the corresponding complete graph would have 153 edges. The

representation of such a graph would be extremely crowded and it would be difficult to use the mouse pointer to select a particular edge in order to determine the precise numerical value of the association represented by that edge. If the user did wish to find the numerical value of the association corresponding to a particular edge, they should first rearrange the vertices so as to isolate the edge of interest before selecting this menu option.

Menu Option: Encoded Strengths

Using the (default) equi-distant spacing option in conjunction with the grey-tone shading edge style, the graph shown in Figure 12-21 is obtained. This graph indicates that, although there are a large number of associations in the data, only a handful of them are relatively strong. The two strongest associations can be seen to be between OCCIP (occipitonasal length) and NAS-L (nasal length), and between BASIL (basilar length) and PAL-L (palatilar length). Only one association occurs at the next level (between NAS-W (nasal width) and LACRY (inter-lacrymal width)), and one at the next (between OCCIP and BASIL). Relatively speaking, the majority of associations occur at the lowest 3 levels (including no edge).

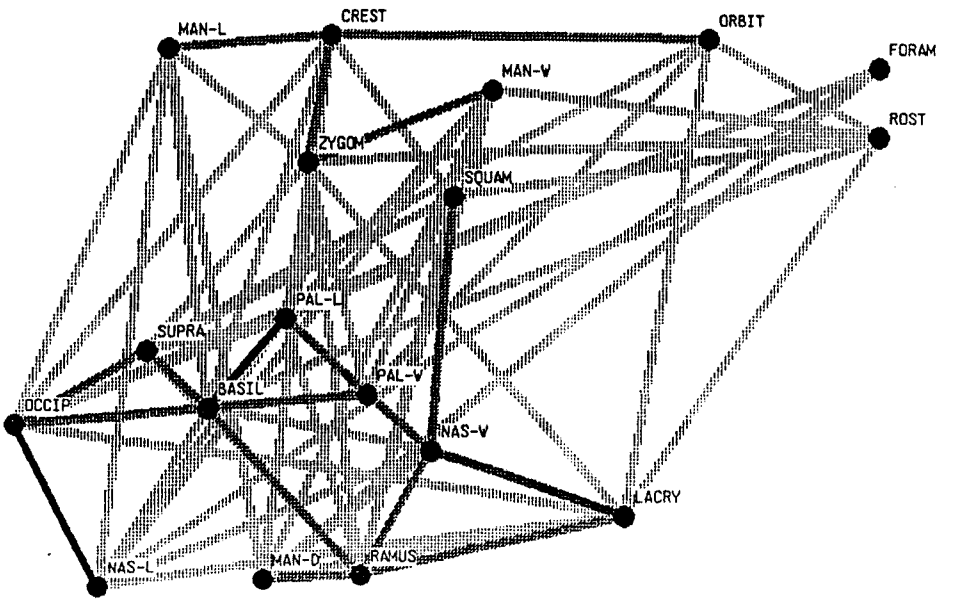

Figure 12-21: CIGE: Use of encoded strengths menu option, with default levels chosen by equi-distant spacing and strength encoded by grey-tone shading, for kangaroo data model

Using the (default) significance levels option in conjunction with the combination of width and grey-tone shading edge style, the graph displayed in Figure 12-22 is obtained. The thick black lines correspond to the basic graph obtained using the 0.5% significance level (see Figure 12-20). The whole set of edges contained in this graph are significant at the 10% level. Because of the lay-out of the vertices and because of the use of different widths and shades for the edges, this graph in fact seems more pleasing visually than the original default graph presented in Figure 12-17 (with edges significant at the 5% level), despite having more edges.

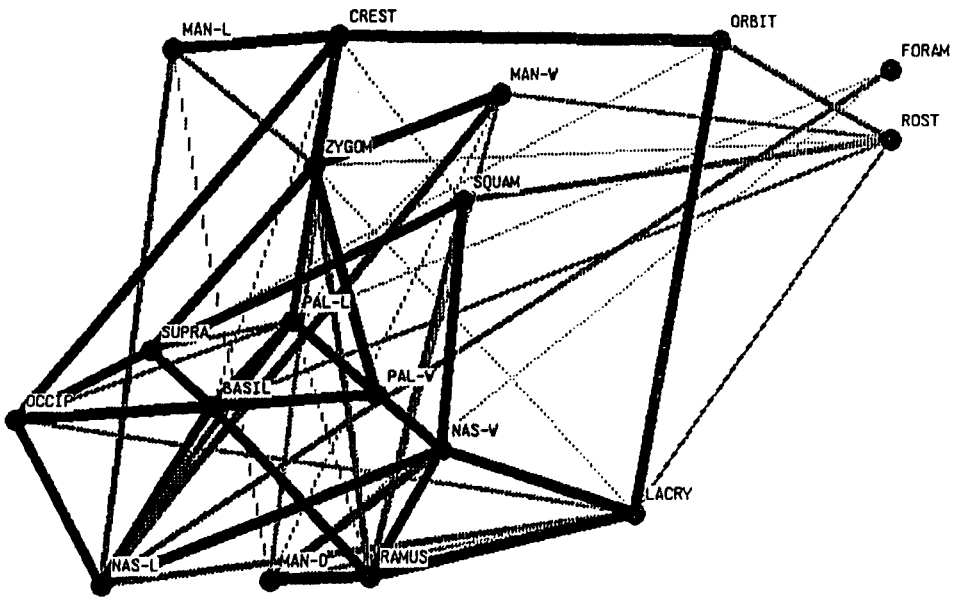

Figure 12-22: Use of encoded strengths menu option, with default levels chosen by significance levels and strength encoded by the combination of width and grey-tone shading, for kangaroo data model

Menu Option: Slider

The slider option has not been employed with this data set because a lot of information has already been acquired as to how the graph will change in appearance with an increase or decrease in significance level.

Menu Option: Changing the Threshold

The threshold has already been changed explicitly in order to obtain a simplified independence graph.

Menu Option: Signed Graph

The graph obtained when this option is chosen is presented in Figure 12-23. From this graph it can be seen that, of the strongest associations (see Figure 12-21), the association between BASIL and PAL-L is in fact negative in sign, whereas the association between NAS-L and OCCIP is positive, although again one would need a certain amount of knowledge of kangaroo anatomy to interpret the signs of the associations.

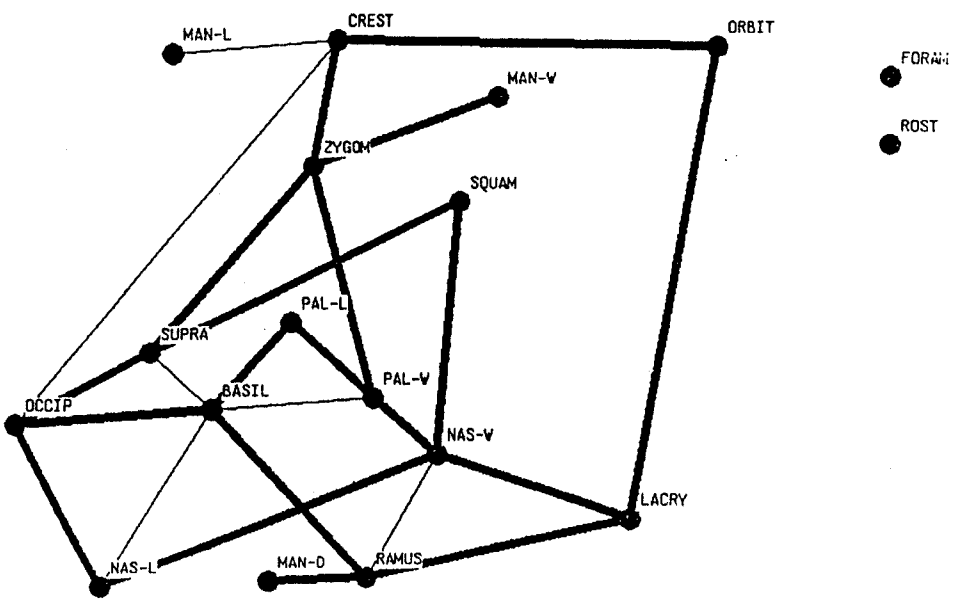

Figure 12-23: CIGE: Use of signed graph menu option, for kangaroo data model

12.3 Log-Linear Interaction Models

12.3.1 Byssinosis Data Set

The data take the form of a 3×2×2×2×3×2 contingency table. The six variables are as listed in Table 12-5, and the contingency table is reproduced in Table 12-6.

LABEL	VARIABLE	LEVELS
EMP	Length of employment	1: <10 years 2: 10–19 years 3: ≥20 years
SMOKE	Smoking habit	1: Smoker (in last 5 years) 2: Non-smoker (in last 5 years)
SEX	Sex	1: Male 2: Female
RACE	Race	1: White 2: Other
WPLACE	Dustiness of work place	1: Most dusty 2: Less dusty 3: Least dusty
BYSS	Incidence of byssinosis	1: Yes 2: No

Table 12-5: Byssinosis data: List of variables

GLIM was used to fit a log-linear interaction model to the data contained in the contingency table, ignoring potential problems arising from the sparsity of the data. A forwards stepwise selection procedure was used, in which the interaction making the largest significant improvement to the fit of the model was added at each stage and the model re-fitted until no further improvement in significance, at the 5% significance level, could be achieved by the addition of any remaining terms. This procedure may not have resulted in the 'best' or most appropriate model which can be fitted to this data set, and the resulting model is quite different from that fitted by Higgins & Koch (1977) using a different approach, but the model obtained is adequate for the purpose of illustrating the use of CIGE for the representation of a log-linear interaction model fitted to a discrete data set.

				WPLACE					
				1		2		3	
				BYSS		BYSS		BYSS	
EMP	SMOKE	SEX	RACE	1	2	1	2	1	2
1	1	1	1	3	37	0	74	2	258
1	1	1	2	25	139	0	88	3	242
1	1	2	1	0	5	1	93	3	180
1	1	2	2	2	22	2	145	3	260
1	2	1	1	0	16	0	35	0	134
1	2	1	2	6	75	1	47	1	122
1	2	2	1	0	4	1	54	2	169
1	2	2	2	1	24	3	142	4	301
2	1	1	1	8	21	1	50	1	187
2	1	1	2	8	30	0	5	0	33
2	1	2	1	0	0	1	33	2	94
2	1	2	2	0	0	0	4	0	3
2	2	1	1	2	8	1	16	0	58
2	2	1	2	1	9	0	0	0	7
2	2	2	1	0	0	0	30	1	90
2	2	2	2	0	0	0	4	0	4
3	1	1	1	31	77	1	141	12	495
3	1	1	2	10	31	0	1	0	45
3	1	2	1	0	1	3	91	3	176
3	1	2	2	0	1	0	0	0	2
3	2	1	1	5	47	0	39	3	182
3	2	1	2	3	15	0	1	0	23
3	2	2	1	0	2	3	187	2	340
3	2	2	2	0	0	0	2	0	3

Table 12-6: Byssinosis data: Contingency table data

The fitted model is as follows:

$$1 + EMP + SMOKE + SEX + RACE + WPLACE + BYSS$$
$$+ EMP.SMOKE + EMP.SEX + EMP.RACE$$
$$+ EMP.WPLACE + EMP.BYSS + SMOKE.SEX$$
$$+ SMOKE.RACE + SMOKE.BYSS + SEX.RACE$$
$$+ SEX.WPLACE + RACE.WPLACE + WPLACE.BYSS$$
$$+ EMP.SMOKE.SEX + EMP.SEX.RACE + EMP.SEX.WPLACE$$
$$+ EMP.RACE.WPLACE + SMOKE.SEX.RACE$$

The generating class of the fitted model is as follows:

[EMP.SMOKE.SEX] [EMP.SEX.RACE] [SMOKE.SEX.RACE]

[EMP.SEX.WPLACE] [EMP.RACE.WPLACE]

[WPLACE.BYSS] [EMP.BYSS] [SMOKE.BYSS]

The parameter value estimates obtained using GLIM are as presented in Table 12-7.

ESTIMATE	PARAMETER	ESTIMATE	PARAMETER
1.674	1	2.344	SEX(2).WPLACE(2)
0.169	EMP(2)	1.976	SEX(2).WPLACE(3)
1.637	EMP(3)	–1.251	RACE(2).WPLACE(2)
–1.260	SMOKE(2)	–1.488	RACE(2).WPLACE(3)
–2.264	SEX(2)	–0.507	EMP(2).BYSS(2)
1.420	RACE(2)	–0.655	EMP(3).BYSS(2)
–1.762	WPLACE(2)	0.555	SMOKE(2).BYSS(2)
–0.680	WPLACE(3)	2.531	WPLACE(2).BYSS(2)
1.854	BYSS(2)	2.709	WPLACE(3).BYSS(2)
–0.498	EMP(2).SMOKE(2)	0.620	EMP(2).SMOKE(2).SEX(2)
–0.251	EMP(3).SMOKE(2)	1.113	EMP(3).SMOKE(2).SEX(2)
–8.536	EMP(2).SEX(2)	–1.392	EMP(2).SEX(2).RACE(2)
–2.192	EMP(3).SEX(2)	–2.510	EMP(3).SEX(2).RACE(2)
0.515	SMOKE(2).SEX(2)	0.226	SMOKE(2).SEX(2).RACE(2)
–1.225	EMP(2).RACE(2)	8.122	EMP(2).SEX(2).WPLACE(2)
–2.414	EMP(3).RACE(2)	8.117	EMP(2).SEX(2).WPLACE(3)
0.058	SMOKE(2).RACE(2)	1.818	EMP(3).SEX(2).WPLACE(2)
0.386	SEX(2).RACE(2)	1.433	EMP(3).SEX(2).WPLACE(3)
–0.082	EMP(2).WPLACE(2)	–0.919	EMP(2).RACE(2).WPLACE(2)
0.024	EMP(2).WPLACE(3)	–0.657	EMP(2).RACE(2).WPLACE(3)
–0.433	EMP(3).WPLACE(2)	–1.781	EMP(3).RACE(2).WPLACE(2)
–0.338	EMP(3).WPLACE(3)	0.113	EMP(3).RACE(2).WPLACE (3)

Table 12-7: Byssinosis data: Parameter value estimates obtained using GLIM

Data Input

For the purposes of this example, the parameter values presented in Table 12-7 will be used as input in preference to the generating class, which can be determined from the parameter values.

For the byssinosis data set, the number of variables is 6 and the variable labels used are as presented previously in Table 12-5.

Basic Display

The basic independence graph, having the initial default lay-out, is as displayed in Figure 12-24. There are only a few cross-overs in the graph but, nevertheless, it is not as clear as it could be. By relocating the vertices it is possible to obtain a more aesthetically pleasing display of the basic independence graph, such as the one presented in Figure 12-25.

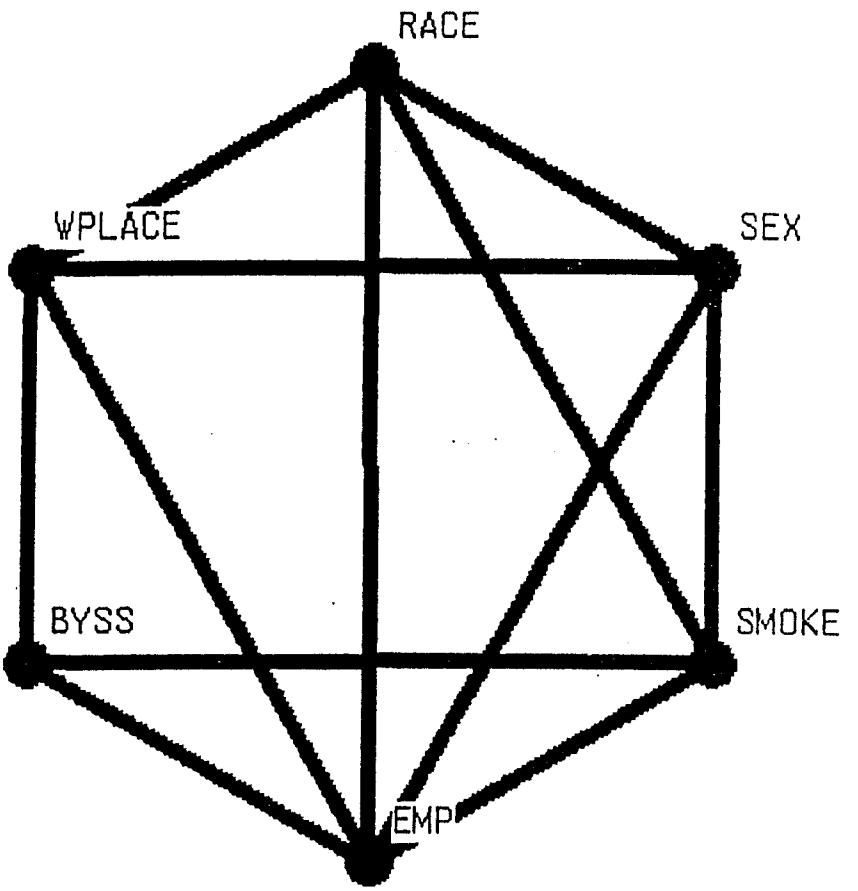

Figure 12-24: CIGE: Default lay-out of independence graph for byssinosis incidence data model

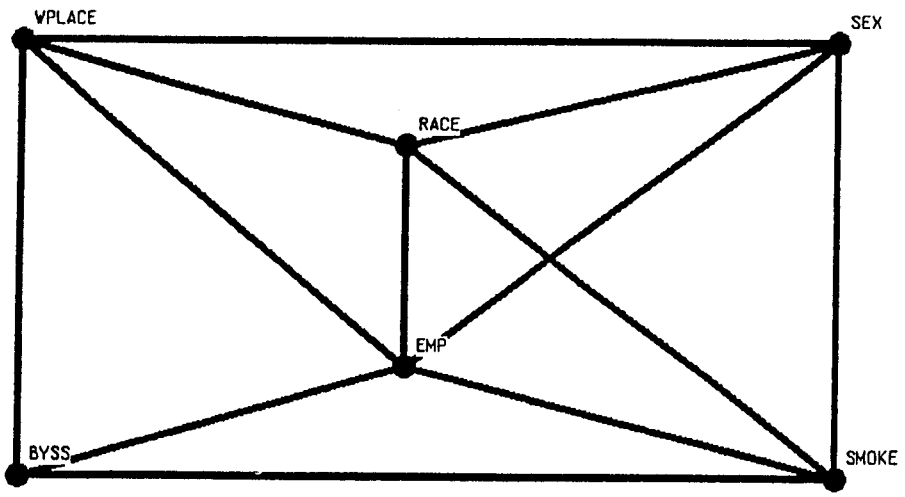

Figure 12-25: CIGE: Modified lay-out of independence graph for byssinosis incidence data model

Menu Option: Independence Graph

The independence graph for the byssinosis example data set is the same as the initial graph which has already been presented in Figure 12-25.

From this graph, it is possible to determine a number of conditional independence relationships (see Chapter 7). For example, SEX and BYSS (the incidence of byssinosis) are conditionally independent given EMP (length of employment), WPLACE (dustiness of work-place) and SMOKE (smoker or non-smoker). This immediately leads one to hypothesise that the incidence of byssinosis may differ for men and women because men and women tend to work in different environments, albeit within the same factory, because men and women had different smoking habits at the time of the study (1973), and because men tend to stay in the same job for longer than women, women tending to leave to raise a family.

Menu Option: Interaction Graph

The interaction graph for the byssinosis data example is presented in Figure 12-26. It can be seen that there are only two edge styles used, corresponding to two-way and

three-way interactions, since these are the only orders of interactions contained within the generating class.

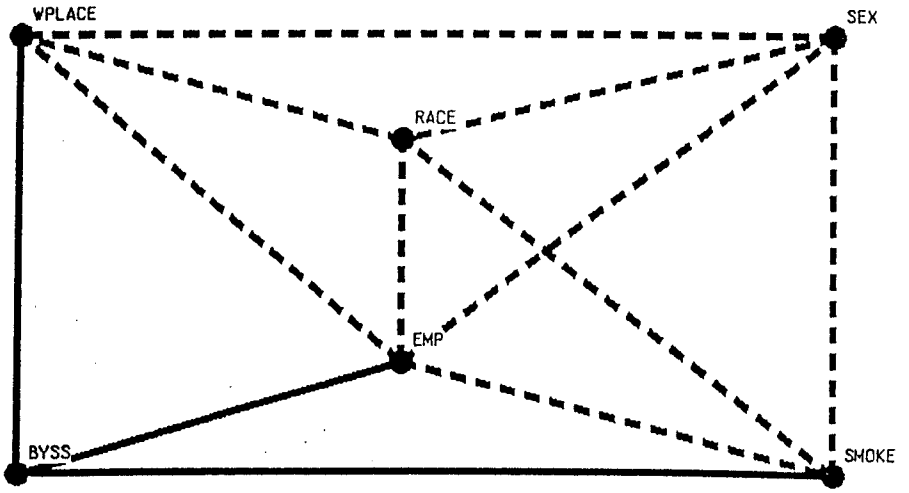

Figure 12-26: CIGE: Interaction graph for byssinosis incidence data model

From the interaction graph it would appear that the three-way interaction [EMP.SMOKE.RACE] is also contained in the generating class of the model represented. To be able to determine the represented model correctly, it is necessary, in this case, to display the elements of the generating class separately, either simultaneously or sequentially.

Menu Option: Elements of Generating Class

The simultaneous display of the 8 elements of the generating class for the model fitted to the byssinosis data is as shown in Figure 12-27. From such a display it is possible to determine the generating class of the model represented without ambiguity.

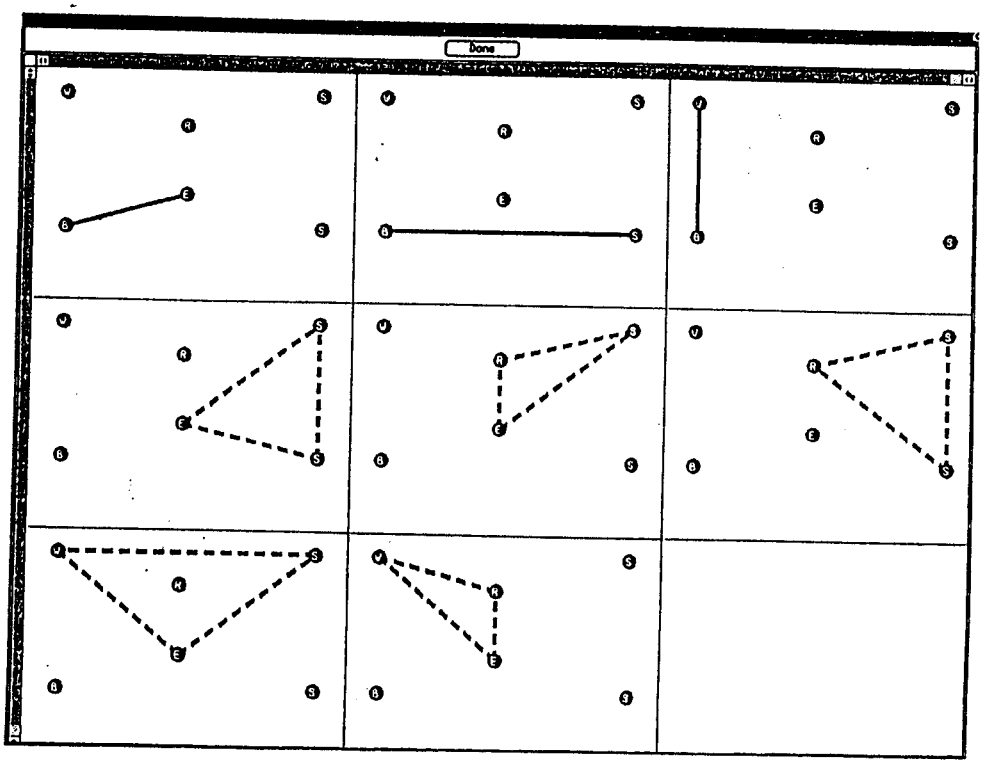

Figure 12-27: CIGE: Use of menu option to display elements of generating class simultaneously for byssinosis incidence data model

An alternative to the simultaneous display of the generating class elements is the use of the sequential display. The elements of the generating class displayed sequentially for the model fitted to the byssinosis data would, of course, be identical to those already shown in Figure 12-27, except they would be displayed one at a time.

Menu Option: Parameter Values

The user can access the actual values of the parameters in the model for a particular main effect or interaction by using the mouse pointer to highlight the vertex or vertices corresponding to the variable(s) in the main effect or interaction effect of interest. In Figure 12-28 the vertices corresponding to EMP, SEX and WPLACE can be seen to have been highlighted.

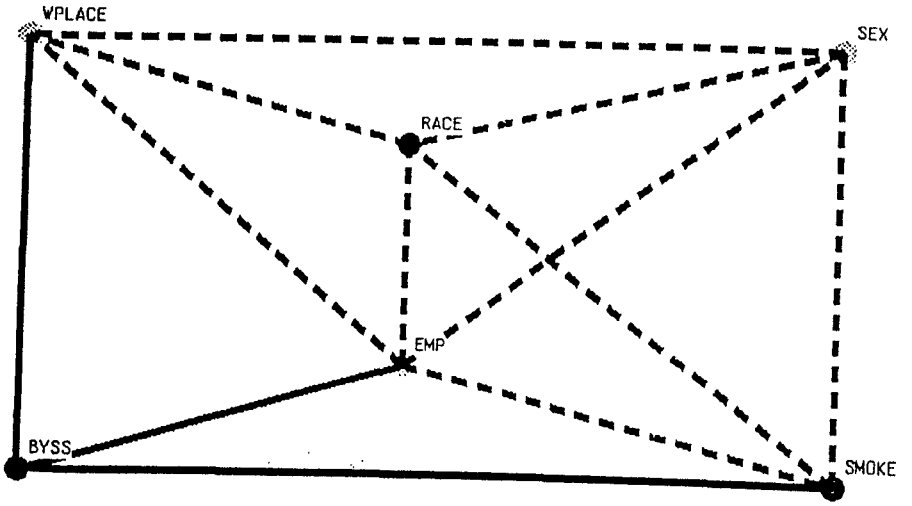

Figure 12-28: CIGE: Use of menu option to display parameter values: highlighted interaction for byssinosis incidence data model

The table of parameter values corresponding to this interaction will be displayed in a separate window on the screen. For the interaction highlighted in Figure 12-28, the parameter values are as given in Figure 12-29. The values are displayed for each combination of the levels of the three variables. Even though not all these values were contained in the parameter values used as input, CIGE carries out the necessary calculations in order to determine the missing values.

If vertices are highlighted which correspond to an interaction which is not contained in the model, the user is told that there is no such interaction. Thus if the user had highlighted the three vertices corresponding to EMP, SMOKE and RACE, they would have been informed that there is no such interaction, even though it would appear from the interaction graph in Figure 12-26, as has already been mentioned, that such an interaction exists in the model. If the user had highlighted just two of these three vertices, then the parameter values of the corresponding two-way interaction would have been displayed, because all of the two-way interactions involving these three variables are implied by the other three way-interactions actually contained in the represented model.

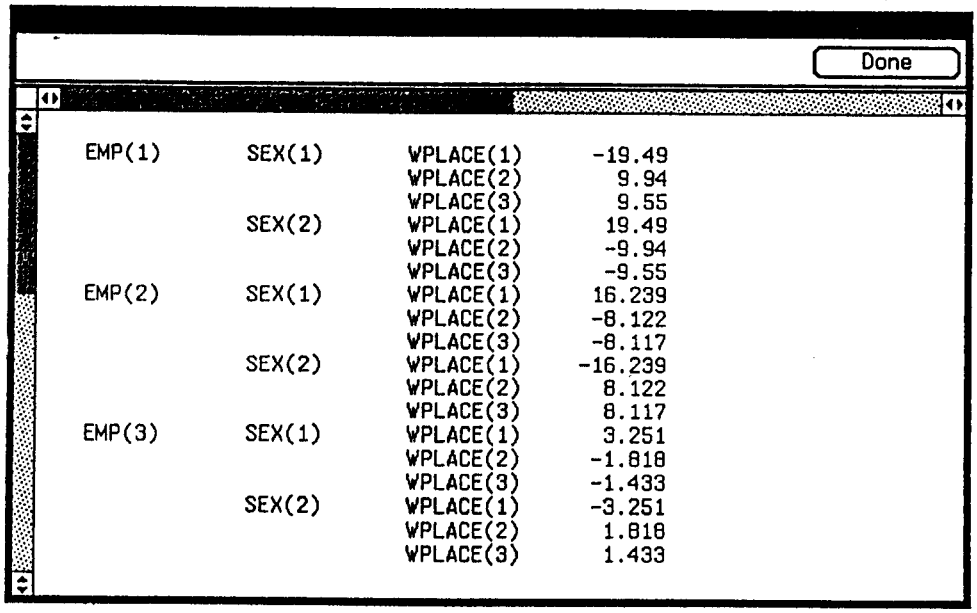

Figure 12-29: CIGE: Use of menu option to display parameter values: table of parameter values corresponding to interaction for byssinosis incidence data model

12.3.2 Example Generating Classes

In Chapter 6, some simple pen-and-paper approaches to the representation of the generating class of a fitted model were considered involving the two-dimensional combination of points (Section 6.2). A number of example generating classes were used in order to illustrate and assess the usefulness of each of these technique. In particular, it proved difficult to distinguish between the three variable generating classes {[*AB*] [*AC*] [*BC*]} and {[*ABC*]}, and between the four variable generating classes {[*AB*] [*AC*] [*AD*] [*BC*] [*BD*] [*CD*]}, {[*ABC*] [*BCD*] [*AD*]}, {[*ABC*] [*ABD*] [*BCD*]}, {[*ABC*] [*ABD*] [*ACD*] [*BCD*]} and {[*ABCD*]}.

The techniques which were developed and considered in Section 6.2 included the use of links between vertices (6.2.2), the representation of interaction type by line styles (6.2.3) and the separate display of generating class elements (6.2.8), among others. These three techniques have since been incorporated into the CIGE approach.

In this section, therefore, I wish to show how CIGE can be applied to the generating classes presented above, and the use of the different features in order to

determine, without ambiguity, the generating class which has been used as input in each case.

Example Generating Classes Involving Three Variables

If either of the generating classes {[*AB*] [*AC*] [*BC*]} or {[*ABC*]} are used as input to CIGE, the corresponding independence graph will be as shown in Figure 12-30. As discussed in Chapter 6, it is impossible to determine for certain the generating class of the model this independence graph is intended to represent.

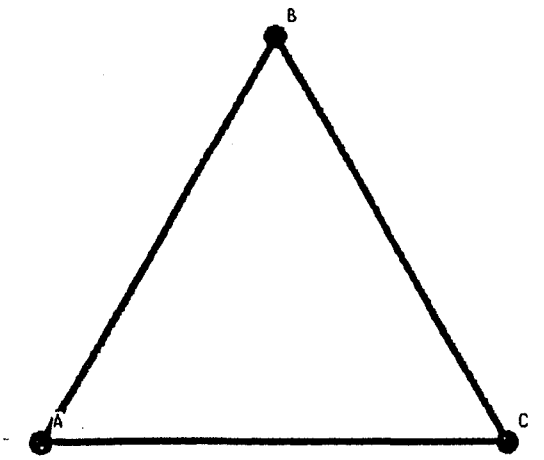

Figure 12-30: CIGE: Use of menu option to display independence graph corresponding to example generating classes involving three variables

However, if the interaction graph menu option is selected, then if the interaction graph is identical to the independence graph in Figure 12-30, it must represent the model with generating class {[*AB*] [*AC*] [*BC*]}, whereas if the interaction graph appears as in Figure 12-31, it must represent the model with generating class {[*ABC*]}.

Alternatively, the menu option to display the elements of the generating class (simultaneously or sequentially) may be selected. In the case of the model with generating class {[*AB*] [*AC*] [*BC*]}, the elements will be displayed as in Figure 12-32, whereas in the case of the model with generating class {[*ABC*]}, the (single) element will be displayed as in Figure 12-33. This element is identical to the interaction graph in Figure 12-31.

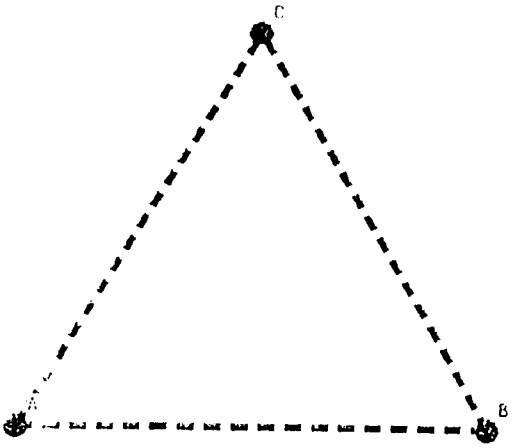

Figure 12-31: CIGE: Use of menu option to display interaction graph corresponding to example generating class {[ABC]}

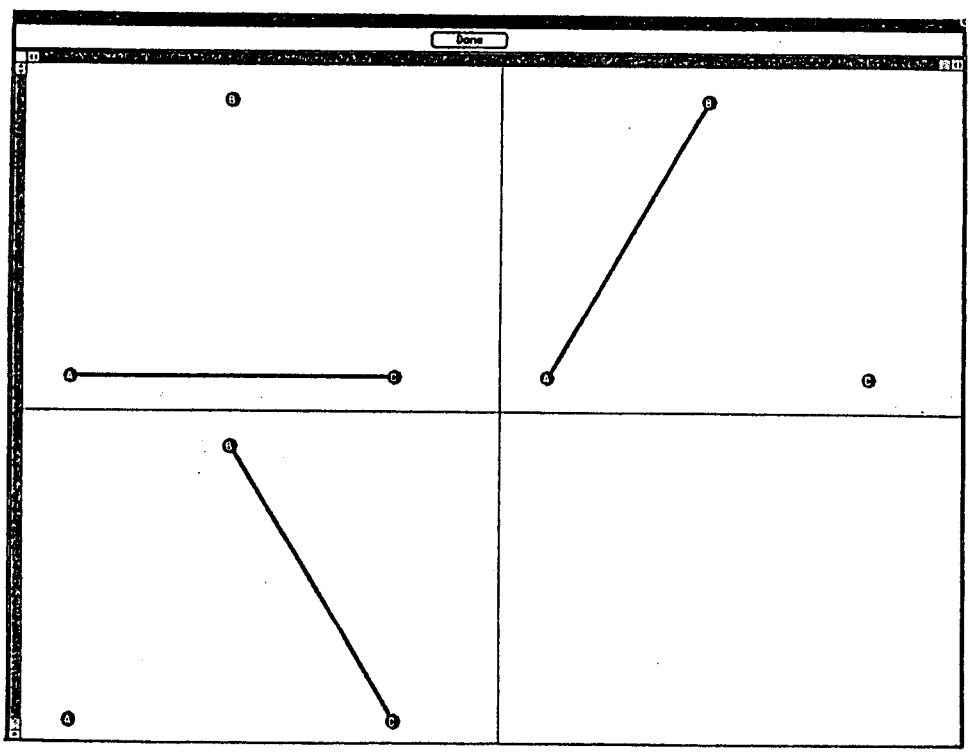

Figure 12-32: CIGE: Use of menu option to display elements of generating class {[AB] [AC] [BC]}

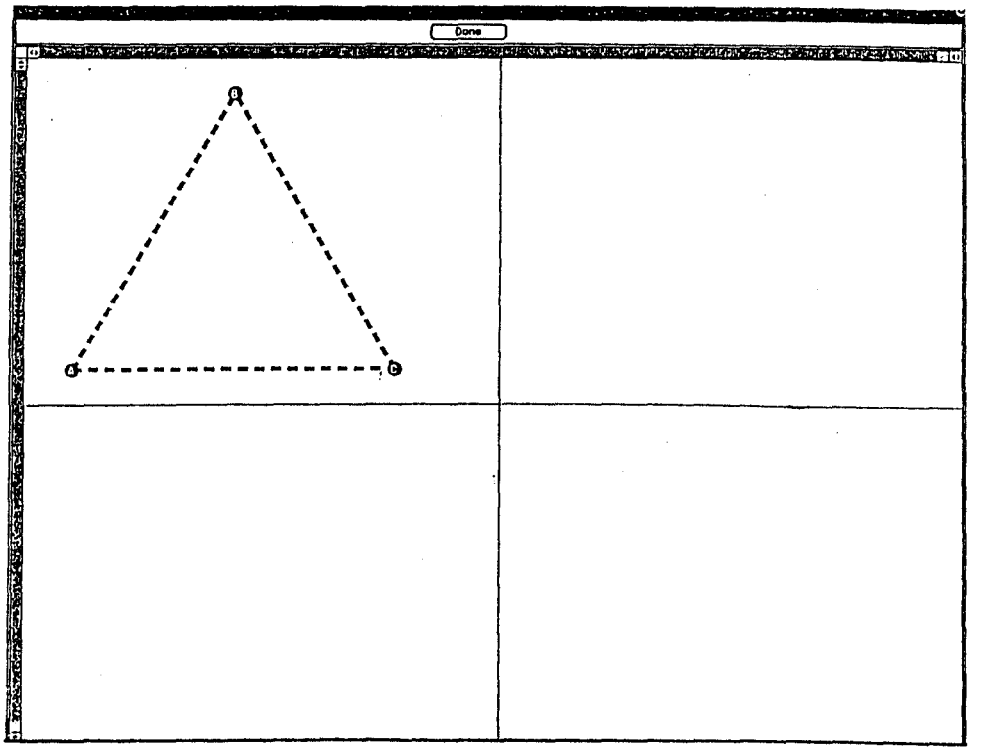

Figure 12-33: CIGE: Use of menu option to display element of generating class {[ABC]}

Example Generating Classes Involving Four Variables

If any of the five generating classes {[*AB*] [*AC*] [*AD*] [*BC*] [*BD*] [*CD*]}, {[*ABC*] [*BCD*] [*AD*]}, {[*ABC*] [*ABD*] [*BCD*]}, {[*ABC*] [*ABD*] [*ACD*] [*BCD*]}, or {[*ABCD*]} is used as input to CIGE, the independence graph displayed will be as in Figure 12-34 (vertices relocated from initial default display). As discussed in Chapter 6, it is impossible to determine for certain the generating class of the model this independence graph is intended to represent.

If the interaction graph menu option is selected, and the interaction graph is identical to the independence graph in Figure 12-34, then it must represent the model with generating class {[*AB*] [*AC*] [*AD*] [*BC*] [*BD*] [*CD*]}. If the interaction graph appears as in Figure 12-35, then it must represent the model with generating class {[*ABC*] [*BCD*] [*AD*]}, or if the interaction graph appears as in Figure 12-36, then it must represent the

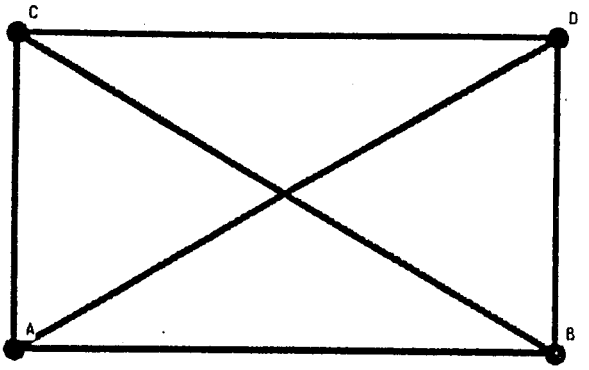

Figure 12-34: CIGE: Use of menu option to display independence graph corresponding to example generating classes involving four variables

model with generating class {[*ABCD*]}. However, if the interaction graph appears as in Figure 12-37, it may represent the model with generating class {[*ABC*] [*ABD*] [*BCD*]} or the model with generating class {[*ABC*] [*ABD*] [*ACD*] [*BCD*]} (or even one of the other three models involving three three-way interactions). Only by selecting the menu option to display the elements of the generating class separately is it possible to distinguish between the model with generating class {[*ABC*] [*ABD*] [*BCD*]} (see Figure 12-38) and the model with generating class {[*ABC*] [*ABD*] [*ACD*] [*BCD*]} (see Figure 12-39).

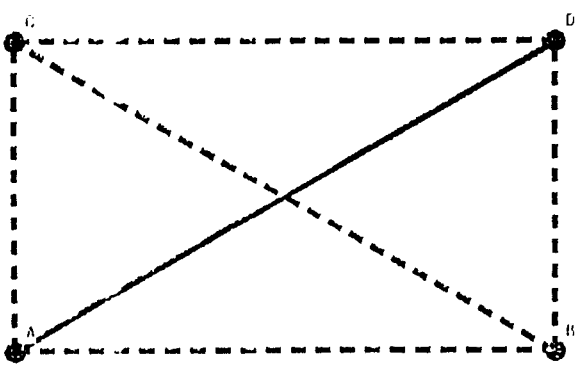

Figure 12-35: CIGE: Use of menu option to display interaction graph corresponding to example generating class {[ABC] [BCD] [AD]}

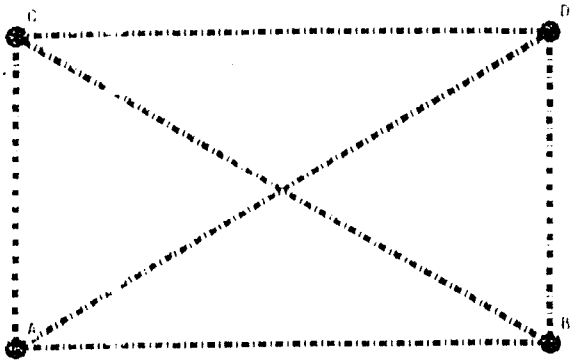

Figure 12-36: CIGE: Use of menu option to display interaction graph corresponding to example generating class {[ABCD]}

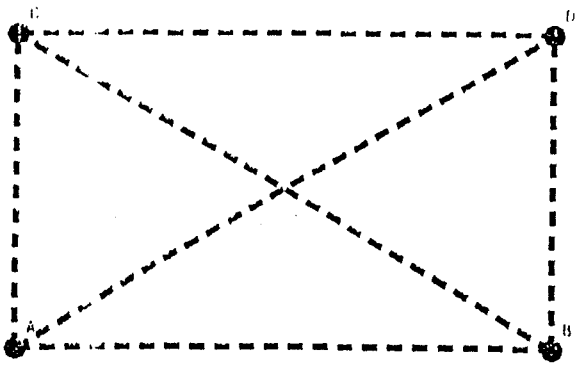

Figure 12-37: CIGE: Use of menu option to display interaction graph corresponding to example generating classes {[ABC] [ACD] [BCD]} and {[ABC] [ABD] [ACD] [BCD]}

[N.B. The figures produced using CIGE to illustrate this section are identical in appearance to some of the figures used to illustrate the pen-and-paper techniques developed in Chapter 6. This is because CIGE was in fact used to produce the figures in Chapter 6 wherever possible.]

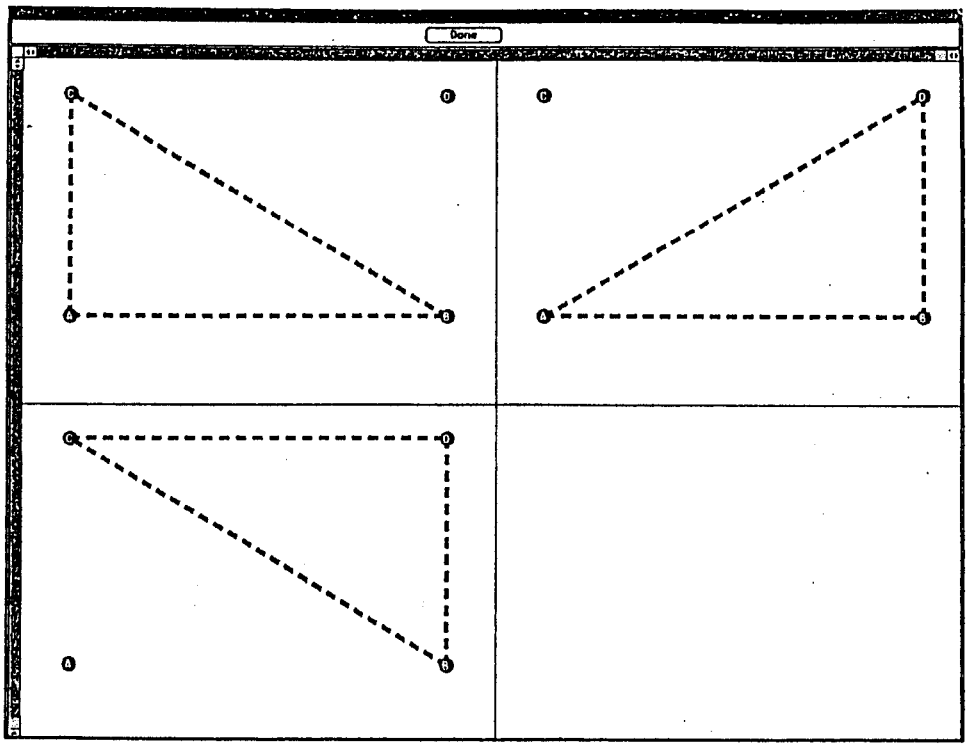

Figure 12-38: CIGE: Use of menu option to display elements of generating class {[ABC] [ABD] [BCD]}

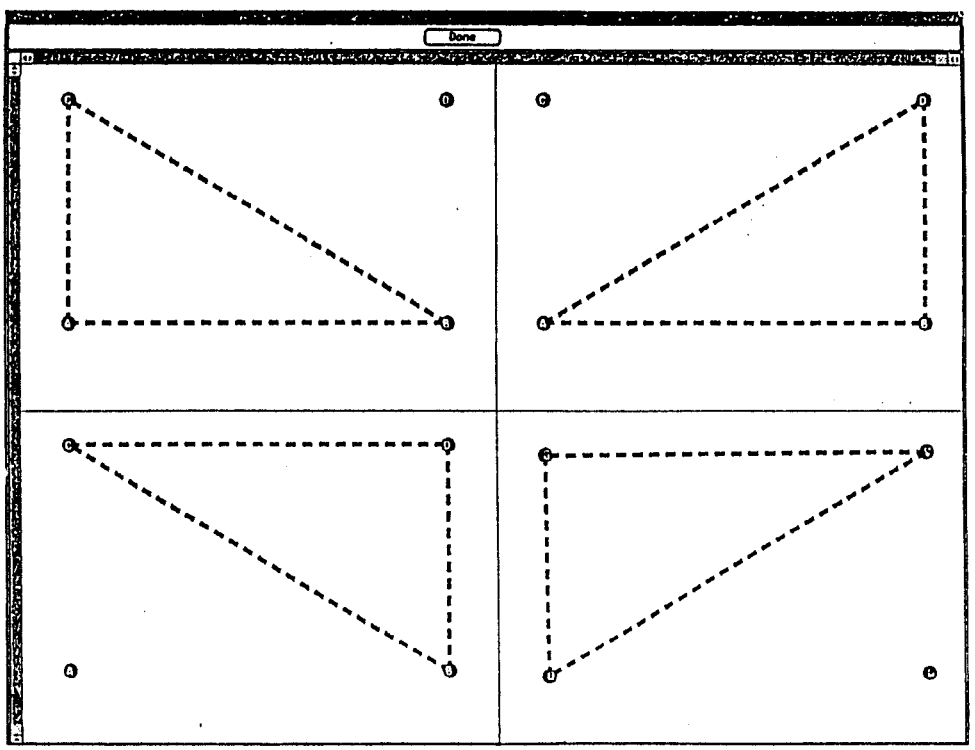

Figure 12-39: CIGE: Use of menu option to display elements of generating class {[ABC] [ABD] [ACD] [BCD]}

12.4 Summary

In the preceding chapter (Chapter 11) the main features of the CIGE computer package, which incorporates the work in earlier chapters, were described in general terms. This chapter was concerned with the application of these features of CIGE to real data sets, both to illustrate the various features and (hopefully) to demonstrate the usefulness of CIGE as a technique for the representation of structured multivariate data. Although this chapter was illustrated with graphs obtained during the actual interactive use of CIGE, the reader was spared too many details on how to use CIGE in practice — such details are provided in the form of a User's Guide in Appendix A, together with numerous other screen dumps.

Covariance selection models fitted to two continuous data sets were used to illustrate CIGE. The first, concerned with measurements relating to pitprops, was a moderate sized data set involving 13 variables used primarily to illustrate the features of CIGE for continuous data, although some consideration was given to the relationships found in the data. The second set of data, concerned with measurements relating to kangaroo skeletons, was a fairly large data set involving 18 variables. Some problems were encountered with the display and interpretation of such a large amount of information in graphical form, but these were tackled within CIGE and a lot was eventually learnt about the structure within the data.

A log-linear interaction model fitted to a multidimensional contingency table was then considered, involving 5 variables relating to the incidence of byssinosis in the textile industry. This model was primarily used to illustrate the features of CIGE for discrete data.

The discrete data features of CIGE were also applied to a number of example generating classes which had previously been considered in Chapter 6 where they had been used to illustrate a number of simple two-dimensional pen-and-paper approaches to the representation of fitted models which had not always been successful. Consideration of these generating classes using CIGE made it possible to determine the represented model quickly and correctly in each case, with the added advantage that the computer could draw the required graphical display instantly.

Having considered each of these data sets using CIGE, CIGE can be seen to be a useful tool for the representation of certain types of fitted model. CIGE facilitates the

display, in graph form, of numerical information relating to the fitted model, and the user is able to interact with the graph, and therefore with the data, in a straightforward manner which enables the user to develop an understanding of the structure within the data. Furthermore, as has been seen, the graphical display obtained using CIGE can be printed out at any stage of the analysis for inclusion in reports, papers, etc.

Although CIGE is already a useful package for the display and consideration of certain types of structured multivariate data, in the following chapter I wish to consider a number of ways in which CIGE may be extended and improved, so as to increase its usefulness.

13. Extensions and Improvements to CIGE

13.1 Introduction

The preceding chapters (Chapters 11 & 12) have been concerned with the development and use of CIGE ("Conditional Independence Graph Enhancer"). CIGE is a computer program which has been written to implement the work and ideas contained in the second part of this thesis (ie. Chapter 6 onwards).

CIGE has been written for the construction and interpretation of representations for fitted statistical models obtained by the application of covariance selection modelling and log-linear modelling, and implemented on the SUN workstation.

This chapter is concerned with possible extensions and improvements to CIGE. In Section 13.2, the use of CIGE to represent other models, such as regression or ANOVA models, will be discussed. If the actual model fitting could be carried out in CIGE, this would make CIGE a very versatile package for both the fitting and display of statistical models for multivariate data. This possibility is considered in Section 13.3.

The graphical representations used in CIGE, whilst developed for the display of the terms in a covariance selection model or log-linear model, are potentially of use for the display of other information. The use of CIGE for the representation of correlation and covariance matrices is discussed in Section 13.4.

Finally, the possibility of writing a PC version of CIGE is considered in Section 13.5.

13.2 Representation of Other Models

If one has any model which contains the equivalent of main effects and interaction terms, for example, regression models or ANOVA models, then this model may be expressed in terms of a generating class and the generating class may be used as input to CIGE in the same way as the generating class of a log-linear interaction model. In this way, a graphical representation of the model can be obtained. Features of CIGE which would be of use with such representations include:

- The facility to display an 'interaction' graph.

- The facility to display the elements of the generating class of the model either simultaneously or sequentially, to resolve any ambiguities in the interaction graph.

If it is possible to input parameter values as well, then it may also be possible to use CIGE to obtain the values of the parameters for various chosen terms in the model.

13.3 Incorporation With Model-Fitting Routines

One particular way in which CIGE could be improved would be if it could be used as a tool for modelling as well as for model interpretation and communication. It was not the aim of this thesis to develop a tool for modelling, but it would seem to be a logical extension. For example, if the user could specify edges to add to or delete from the displayed graph (perhaps for theoretical rather than statistical reasons), a model-fitting version of CIGE (M-CIGE?) could then provide details of the fit of the new model. Moreover, at present the user is required to fit a model to his/her data outside of the CIGE package (using GLIM, SPSS, SAS, or whatever) and then to input the details of the fitted model to CIGE before the corresponding graph can be displayed. M-CIGE would conceivably be a unified package which requires only the raw data as input, and which can then be used to fit a model prior to its display. Alternatively, the display may actually form part of the model fitting procedure.

CIGE could therefore be improved by the incorporation of model fitting routines, which may or may not make use of the graphical display of the current model. However, there already exists a modelling package called "Mixed Interaction Modelling", written by Edwards (1995) and described in Section 7.5.1 of Chapter 7, which is designed for use in the fitting of covariance selection models (graphical Gaussian models) and graphical log-linear models, as well as conditional Gaussian models for mixed data. Moreover, MIM provides a visual display of the current *graphical* model, which can be used as an aid to the fitting of models (eg. by the deletion of one edge at a time from the graph, or by the addition of one edge at a time). It may therefore be appropriate to combine CIGE with a package such as MIM. Then not only would it be possible to fit the initial model in M-CIGE, but it would be possible at any stage in the examination of the fitted model through its graphical representation, to investigate the effect upon the fit of the model of

modifications or alternatives to this model. Because CIGE can be used to represent non-graphical models too, it could be used to extend the usefulness of MIM.

13.4 Representation of Correlation and Covariance Matrices

CIGE could be readily adapted as a tool for the exploration of correlation and covariance matrices. There are obvious parallels between the current use of CIGE to explore the contents of matrices of negative partial correlations and edge exclusion deviances, and its potential use to explore the contents of covariance and correlation matrices. Features of the package which would be of relevance to the exploration of such matrices include:

- The display of those edges corresponding to all associations greater than a specified threshold value.
- The facility to use the slider to investigate the strengths of the associations.
- The facility to obtain the value of the association between any pair of variables by clicking on the appropriate edge in the complete graph.
- The facility to display the strengths of associations by the use of edge codes (width and/or grey-shading).
- The signed graph display which makes the sign of the association between pairs of variables apparent.

Indeed, the only distinction between the use of CIGE for the representation of covariance selection models and the use of CIGE for the representation of correlation and covariance matrices is that no interpretation can be made in terms of conditional independence of the absence of edges below a particular threshold from a graph drawn for a correlation or covariance matrix. However, CIGE would still be invaluable as an exploratory tool to investigate the strongest associations (correlations or covariances) between variables.

13.5 PC Version of CIGE

Although the SUN version of CIGE has the benefits of speed and excellent graphics facilities as a result of the quality of the hardware (eg. a high resolution screen and a three-button mouse) and of the powerful standard software (SUNView graphics library routines), it is not readily portable, except to other SUN users.

It is therefore my intention to write a PC version of CIGE (PC-CIGE). This will increase the number of potential users of CIGE. However, there are different software and hardware specifications available on supposedly compatible PCs, and programming decisions will need to be made taking into account the choice of graphics board (eg. VGA or SuperVGA?), the fact that most users will have a mouse with just one or at most two buttons, what operating system to write for (eg. Windows 3.1, Windows 95 or Windows NT?), what commercial graphics software packages the user may have, what dialect of "C" would be most appropriate (eg. Microsoft C, or an Object Oriented version such as Borland C++?), how much RAM the user is likely to have (8Mb or more?), and what speed of processing could be achieved (eg. 386, 486 or Pentium?).

It would not be a straightforward matter to rewrite the SUN version of CIGE to run on a PC. It is likely that the program would not only need to be rewritten, but also redesigned.

13.6 Summary

In this chapter, a number of possible extensions and improvements to CIGE have been considered which would further increase the usefulness of the program. It is hoped that these improvements may be implemented at some stage in the future.

14. Conclusions

This thesis has been concerned with the development of graphical representations for structured multivariate data.

Graphical representations have been shown (in Chapter 1) to have played an important role in the application and development of statistics, particularly within the past 20 years or so. This surge of interest in the development and use of graphical techniques in statistics has been coupled with the surge in the availability and capabilities of computing power and computer graphics software, and shows every sign of continuing.

However, while very many techniques have been developed for exploratory data analysis (ie. for determining which model it is most appropriate to fit to the data), and whilst several techniques have been developed for assessing the fit of the model to the data, it was the concern of this thesis that very few techniques exist for the representation of the fitted model itself.

In Chapter 2, some of the most common graphical techniques for data exploration were described, followed in Chapter 3 by a review of the most common techniques for model diagnostics. Some mention was also made in Chapter 3 of the techniques available for experimental design symbolisation in ANOVA, since the design of the experiment has a direct influence on the models which may be fitted to the data. In Chapter 4, the few techniques existing for the representation of fitted models were described in some detail, with the exception of conditional independence graphs, which were considered in considerable detail later on. An attempt was made to extend the usefulness of two of the techniques described: namely the extension of interaction plots made by Monlezun (1979), and the technique for the representation of an ANOVA summary table by Bond (1988). It was concluded that the Monlezun plots were not easily interpretable, and that extensions to them were not worth pursuing. The alternative approaches to Bond's representation of an ANOVA summary table were successful, but obviously only of use for the purpose of representing an ANOVA summary table. Finally, in Chapter 5, some mention was made of issues in graphical perception and presentation which should be borne in mind when considering the effectiveness of any graphical representation.

In Chapter 6 an attempt was made to develop some novel pen-and-paper based techniques for the representation of fitted statistical models. Three approaches were pursued. Firstly, the two-dimensional combination of points was considered in Section

6.2. Most of the ideas developed were found to be of use for the unique representation of the generating class of some models, but not of all models. The models which could be successfully represented differed according to the technique used. A few ideas were found to be of use for the representation of the generating class of all models but, with the exception of the separate display of the generating class elements, these could lead to quite messy representations for models involving many variables. Secondly, in Section 6.3 an alternative representation of Venn diagrams developed by Edwards (1989) was developed for the representation of the generating class of a fitted model. Although this technique, unlike the previous techniques, did not lead to problems of ambiguity, it was felt to be unsatisfactory because of the differing emphasis made of different elements of the generating class. Thirdly, in Section 6.4 a completely new technique was developed: the Topological–Magnitude Graph. This technique did not aim to represent the elements of the generating class of the fitted model, but to show every variable in the data and the nature of any interactions between these variables. Use of the Topological Graph allowed the terms within the model to be determined by means of an algebraic interpretation of the links between blocks. Use of the Magnitude Graph conveyed this same information, but the actual size of the effects corresponding to the terms on the model could be represented as well, with only a slight increase in complexity. However, in practice it was found that there were certain types of models which could not be successfully represented by these graphs, due to limitations of the algebraic interpretation of the links.

Having failed to develop a novel pen-and-paper technique for the representation of fitted models which is of use for the representation of all fitted models, attention was switched to the conditional independence graph used in graphical modelling. This is similar to the graph constructed using links between vertices in Section 6.2.2 of Chapter 6. Since graphical modelling is a relatively new and little used, but nevertheless useful and important, statistical technique, a detailed description of both graphical modelling and conditional independence graphs was presented in Chapter 7. However, as was found for the techniques I developed for the two-dimensional combination of points, use of the conditional independence graph as a way of representing statistical models is only successful for some, but not all, models, and the technique has other limitations due to crossovers and the amount of information conveyed. These limitations were considered in the next three chapters. In Chapter 8, the problem of crossing numbers in graphs was considered in general, and I was forced to conclude that no theoretical or algorithmic solution exists at present to the problem of constructing a graph with the minimum

number of cross-overs. In Chapter 9, the usefulness of the conditional independence graph for the representation of log-linear interaction models for discrete data was extended by the incorporation of edge codes corresponding to interactions, thus leading to the development of the "conditional interaction graph". Then, in Chapter 10, the usefulness of the conditional independence graph for the representation of covariance selection models for continuous data was extended by the incorporation of the strength of the (pair-wise) associations into the graph by the use of varying edge styles. A graphical perception experiment provided no evidence of any difference in speed of interpretation of the graphs for the encoding of strength by the width of edges compared with the encoding of strength by the grey-tone shading of edges. Some consideration was also made of the possibility, first suggested by Whittaker (1988), of encoding strength of association by distance. PCA and MDS were applied to two data sets, but did not necessarily lead to a successful or aesthetically pleasing graph. Mention was also made in this chapter of a simple technique to encode the sign of the association in the graph.

Having initially failed to develop a successful pen-and-paper technique, it was felt that the way ahead could be with computer-implemented interactive graphical techniques. Computer-implemented techniques not only have the advantage of allowing easy construction and manipulation of the graphical representation, but are in keeping with recent trends that only those statistical techniques which for which software is available are commonly used. Having had some success in extending the usefulness of the conditional independence graph, it was decided to implement these extensions in an interactive computer package called the "Conditional Independence Graph Enhancer" or CIGE, specifically for covariance selection models and log-linear interaction models.

The many features of CIGE were described in Chapter 11. Not all features are applicable to both model types considered. For continuous data, the most important features are: a facility for obtaining the actual value of the partial correlation or edge exclusion deviance for any pair of variables; the incorporation of edge styles corresponding to strength of association (where the level of the strength of association can be encoded in three different ways); a slider facility for the examination of the effect of changing the threshold upon the edges in the graph; a facility for changing the threshold permanently; and a facility for showing the sign of the partial correlations. For discrete data, the most important features are: the incorporation of edge codes corresponding to order of interaction; a facility to display the elements of the generating class separately,

either simultaneously or sequentially; and a facility to display the parameter values corresponding to highlighted interactions. In general, the graphs obtained can be manipulated in order to obtain a more aesthetically pleasing representation or one with fewer cross-overs and/or a more meaningful grouping of the vertices/variables. These features were illustrated on a number of example data sets for both continuous and discrete data in Chapter 12.

Thus, through the use of CIGE, covariance selection models for continuous data and log-linear interaction models for discrete data can now be represented in such a way that, through the use of a combination of the available facilities, it is possible to determine the represented model uniquely. Since CIGE has been written for use with monochrome graphics workstations, and does not rely on three-dimensional graphics, at any stage the graph displayed can be printed out for inclusion in a report or journal.

With CIGE, therefore, I would claim to have been successful in meeting the research aim of developing a graphical technique for the representation of structured multivariate data in the form of (certain) fitted statistical models. Up until now, this has been a much neglected area. By keeping the mathematics as simple as possible, I have hopefully developed an approach which will appeal to non-statisticians who need to interpret and communicate fitted statistical models, as well as to statistical consultants who may have to assist non-statisticians in the interpretation and communication of fitted models, and hopefully to other statisticians too. Where it is not possible to use CIGE interactively, it may be possible to use some of the ideas contained therein, or to use one of the pen-and-paper techniques I developed in Chapter 6, for a fitted model where the chosen technique will be successful.

All of the techniques I have developed appear, with hindsight, to be simplistic and perhaps even 'obvious', yet I would argue that these features should be regarded in favour of the use of the graphical representations for fitted models developed in this thesis.

As for future developments, it is unlikely that a straightforward universally successful pen-and-paper technique could be developed — it is more likely that the future will lie in the development of interactive computer implemented techniques, both due to the complexity of constructing a representation, and also because of the importance and popularity of computer implemented techniques in statistics in general. In Chapter 13, a number of ideas for the future development of CIGE were presented. The extension of the package to enable the representation of models for mixed data will make CIGE more widely applicable, and the writing of a PC version of the software will make it more

widely useable. Incorporation of CIGE with model-fitting routines, whilst not directly relevant to my aims of developing a technique for the representation of *fitted* models, will extend the usefulness of the package. In particular, incorporation of CIGE with a package such as MIM for the fitting of graphical models could be of use in making graphical modelling known and available to a larger number of users than at present.

BLANK PAGE IN ORIGINAL

15. Appendix A: CIGE User's Guide

15.1 Introduction

The contents of this appendix are complementary to the contents of Chapters 11 and 12 in that this appendix describes in detail the implementation of the features described in Chapter 11 and illustrates the use of CIGE to obtain the graphs used to display the models fitted to the analyses of the pit-prop data and byssinosis data sets presented in Chapter 12.

CIGE ("Conditional Independence Graph Enhancer") is an interactive computer package for the display, communication and interpretation of covariance selection models for continuous data and log-linear interaction models for discrete data. CIGE has been written in the "C" programming language (Kernighan & Ritchie (1988); Schildt (1987)), for use with monochrome SUN workstations (eg. 3/50 and 3/60 models), making use of SUNView library routines which facilitate the use of windows, icons, menus and pointers (WIMPs). Although this section forms a User's Guide in that it details the various features of CIGE and their use, it is not intended to be a technical section, so details of the program itself have been kept to a minimum. In this way, it is not necessary to have any specialised knowledge of the hardware or software in order to understand how to use CIGE, beyond a basic familiarity with the use of windows, menus, and a three-button mouse pointer.

CIGE is formed by three main modules called CIGE, CCIGE, and DCIGE. The CIGE module (approximately 550 lines of code) handles the input of the data, which is either continuous or discrete, and passes control to the CCIGE module (approximately 880 lines of code) if the data is continuous, or to the DCIGE module (approximately 880 lines of code) if the data is discrete. This modular structure was chosen because the features of CIGE for continuous data and for discrete data are quite different.

The relationship between the three main modules is as illustrated in Figure 15-1. In Figure 15-2, the input and output files used with CIGE, together with the various temporary files and a supplementary program are illustrated in relation to these three main modules. Figure 15-2 is intended for reference, and the function of these files and the supplementary program will hopefully become apparent in subsequent sections of this appendix.

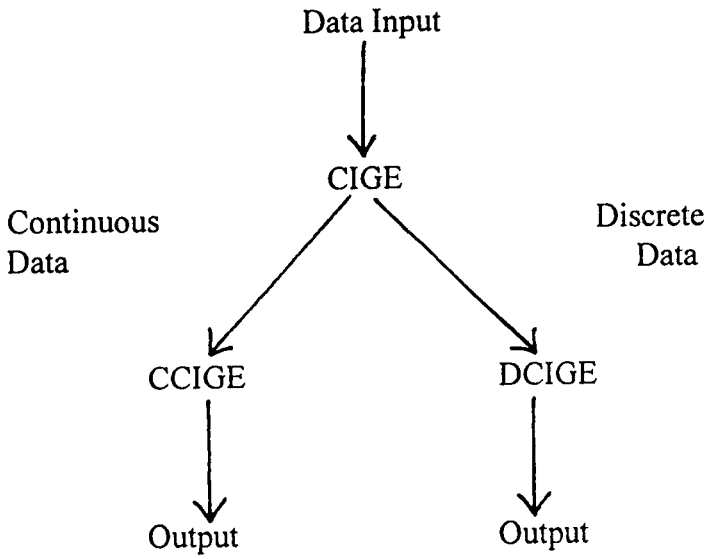

Figure 15-1: Diagram showing the relationship between the three main modules of CIGE: CIGE, CCIGE and DCIGE

The structure of this appendix is such that I shall begin by describing the use of the features of the data entry module CIGE. I shall then go on to describe the use of the continuous data module CCIGE, and conclude by describing the use of the discrete data module DCIGE. To avoid confusion, the name CIGE will, in general, be used throughout the text to refer to the package as a whole. The modules CIGE, CCIGE and DCIGE will only be referred to where it is relevant to the description of the functioning of the package. From the user's perspective, the modules form a coherent whole package, which is CIGE.

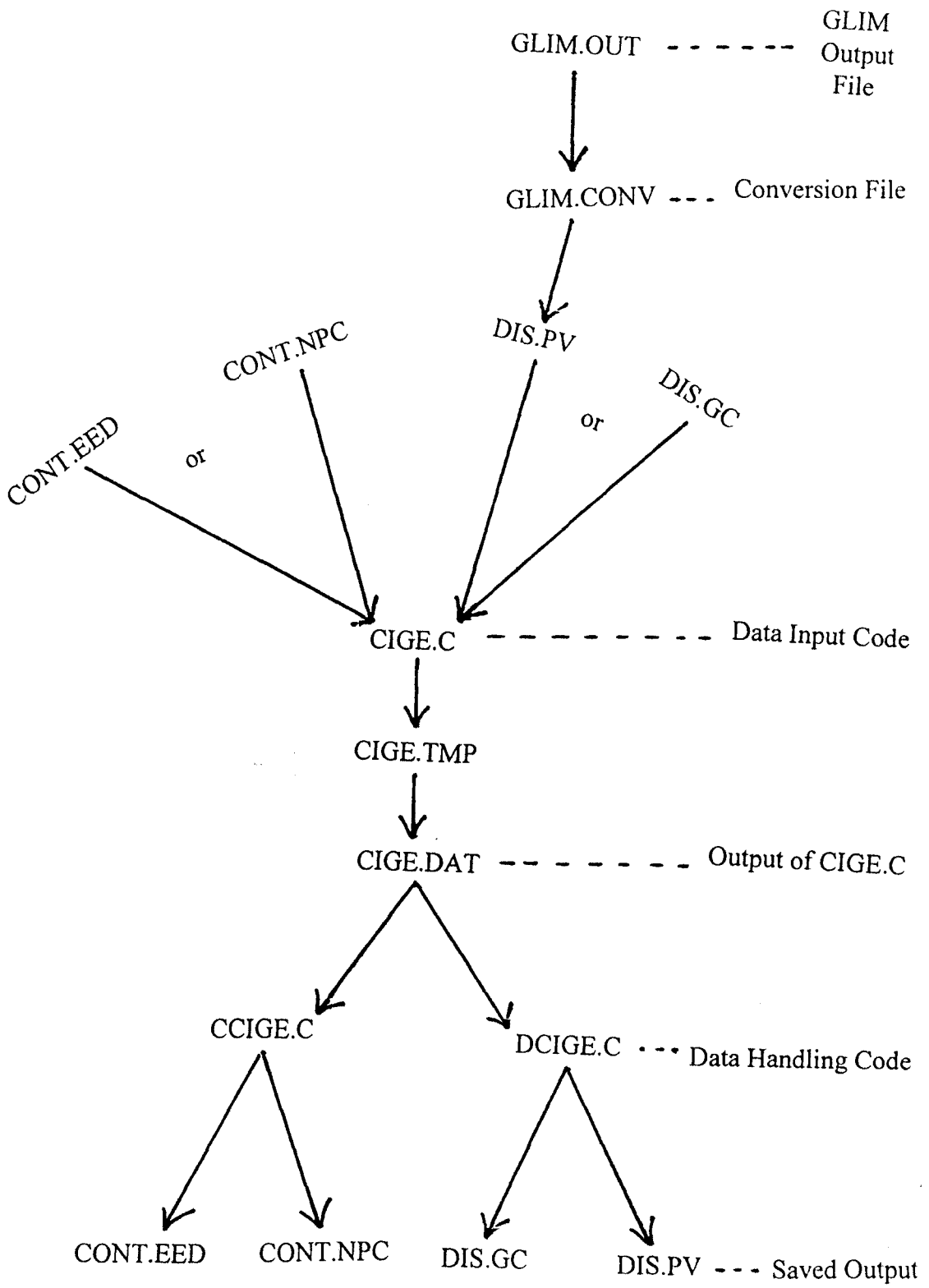

Figure 15-2: Diagram showing the relationship between the three main modules of CIGE, together with the temporary files and supplementary program

15.2 CIGE: Data Entry Module

Assuming that the three main modules and necessary icons, together with the supplementary program and any data files are resident in the current directory, CIGE is invoked simply by typing "cige" at the SUN shell prompt.

The first screen which the user encounters is part of the CIGE data entry module. This requires the user to choose between the following options:

1. Input new data
2. Restore previously saved graph.

If it is a new data set, the user would choose option 1. However, at any stage during the use of CIGE it is possible to save the current data, variable labels, variable positions, and threshold value for the basic graph so that it may be examined again at a later occasion. Were it the case that a graph had already been saved in this way, the user could select option 2, and then the name of the file containing the previously saved graph would be requested and the graph would be re-displayed exactly as it had appeared when saved. However, let us suppose that the user wishes to input new data and so selects option 1.

Having specified that he/she wishes to input new data, it is now necessary for the user to specify one of the following two options:

1. Continuous data
2. Discrete data

If a continuous data set is to be considered, the user would select option 1. What happens thereafter is described in Section 15.2.1. However, if a discrete data set is to be considered, the user would select option 2. What happens thereafter is described in Section 15.2.2.

15.2.1 Data Input for Continuous Data

Having indicated that continuous data is to be analysed, it is necessary to specify the form in which the continuous data has been prepared, by selecting one of the following two options:

1. Edge exclusion deviances
2. Negative partial correlations

If the user selects option 1, they are asked to specify the name of the data file in which the matrix of edge exclusion deviances is contained in upper triangular matrix form. In Figure 15-2, this corresponds to *cont.eed*. If the user selects option 2, they are first asked to specify the number of observed units N, and then to specify the name of the data file in which the matrix of negative partial correlation coefficients is contained in upper triangular matrix form. In Figure 15-2, this corresponds to *cont.npc*.

The upper triangular matrix must consist of one line per variable, and there must be at least one space between each data value. Otherwise, the data can be in free-format form. It is not necessary to explicitly give the number of variables in the data set (for which there is no maximum limit), since this can be determined from the size of the matrix used as input. It is, however, necessary for the user to have included the variable names with which they wish to label the vertices within the data file.

In Figure 15-3, it can be seen how the above options are displayed on the screen. The options are displayed one pair at a time, since the next pair of options to be displayed depends on the choice made between the previous options. However, the previous options continue to be displayed and it is possible to change an earlier response, in which case the subsequent options will change and be re-displayed as appropriate.

The data file specified is called *pitprop.npc*. This file contains the matrix of negative partial correlations corresponding to the covariance selection model examined using CIGE in Section 12.2.1 of Chapter 12. This data file is displayed, with annotations, in Figure 15-4. The file containing the matrix of edge exclusion deviances for the same data set, *pitprop.eed*, is identical, except that the upper triangle of the matrix of negative partial correlations is replaced by the upper triangle of the matrix of edge exclusion deviances.

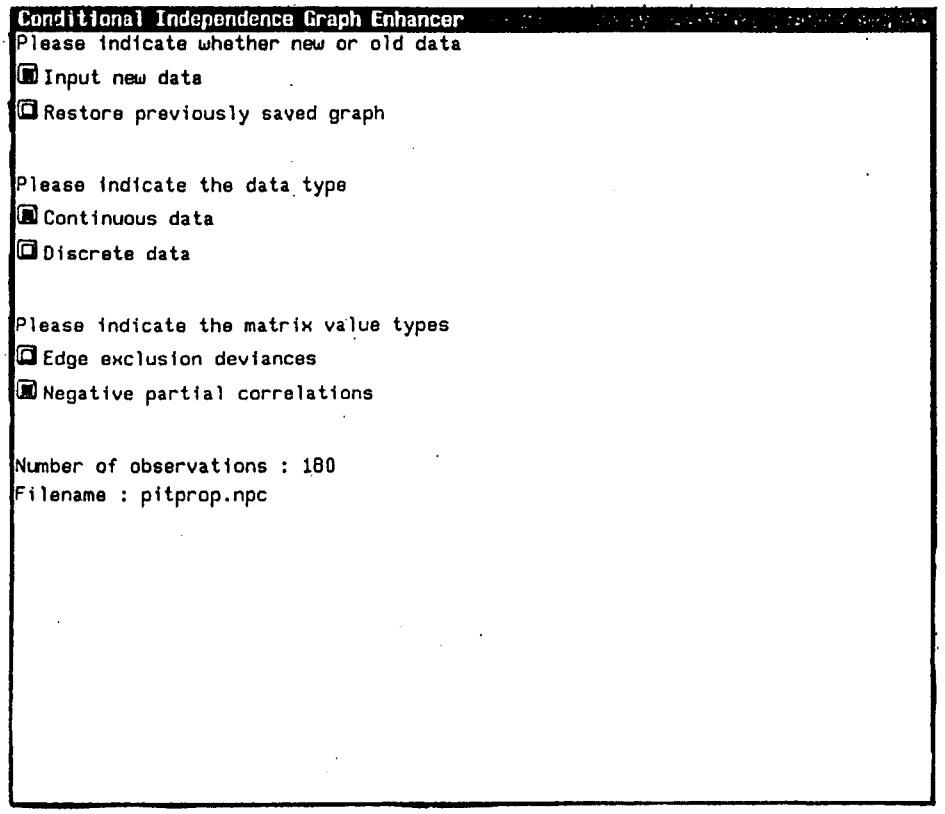

Figure 15-3: CIGE: Data entry screen for example continuous data model

```
-0.8632 -0.1018 -0.0066  0.0621 -0.0147 -0.0634 -0.1748  0.1
 0.0445 -0.0115  0.0010  0.0474 -0.0581  0.1015 -0.3035 -0.1
-0.9289  0.6635  0.0370  0.1127 -0.0688 -0.0092  0.0353 -0.0
-0.6933 -0.1753  0.0194  0.1184 -0.0299 -0.0512  0.0207 -0.0
 0.0153 -0.0586 -0.1086  0.0222  0.0970  0.0379  0.1119 -0.0
-0.8579 -0.0055  0.1553  0.4613  0.0921 -0.2106 -0.0379
 0.0619 -0.1760 -0.5416 -0.0930  0.1906  0.2407
-0.2117 -0.2437 -0.1711  0.1915  0.1163
-0.0263 -0.1001 -0.0767  0.0167
 0.5100  0.1563  0.1913
 0.1363  0.2080
 0.0792
```

negative partial correlation matrix

Figure 15-4: CIGE: Input data file (pitprop.npc) containing matrix of negative partial correlations, with annotations, for example continuous data model

The CIGE module modifies the data file used as input according to the choice made at each pair of options displayed and, in the case of the matrix of negative partial correlations only, the number of observations, to give *cige.tmp*. The default threshold value and the default lay-out of the vertices are calculated, and then the file is passed on to the CCIGE module as *cige.dat*. The lay-out of *cige.dat* is identical to the saved data file corresponding to a saved graph, which will be illustrated later.

15.2.2 Data Input for Discrete Data

Having specified that the data to be analysed is discrete, it is necessary to specify the form in which the discrete data has been prepared by selecting one of the following two options:

1. Parameter values
2. Generating class

The appearance of the above options on the screen is as appears in Figure 15-5

The user is then asked to specify the name of the data file in which the parameter values or the generating class elements are contained. In this example, the data file specified contains parameter values and is called *byss.pv*. It is again necessary to specify any names with which the user wishes to label the vertices corresponding to the variables, and also to specify the number of variables, within the file used as input.

Rather than type each of the parameter values into the data file, a special program has been written, to run independently of CIGE, called *glim.conv*. This program will take a GLIM output file (eg. *glim.out*), and strip it of all irrelevant text *except* for the list of parameter labels and values which are displayed by the GLIM command **$ disp e $**. This stripped output file can be used, with a few other additions, as the data file corresponding to *dis.pv* in Figure 15-2 for use with CIGE. CIGE is able to calculate the parameter values not explicitly contained in the GLIM output file.

The annotated input data file containing the parameter values for the byssinosis data considered in Chapter 12 is presented in Figure 15-7. In Figure 15-6, the same model is represented by its generating class.

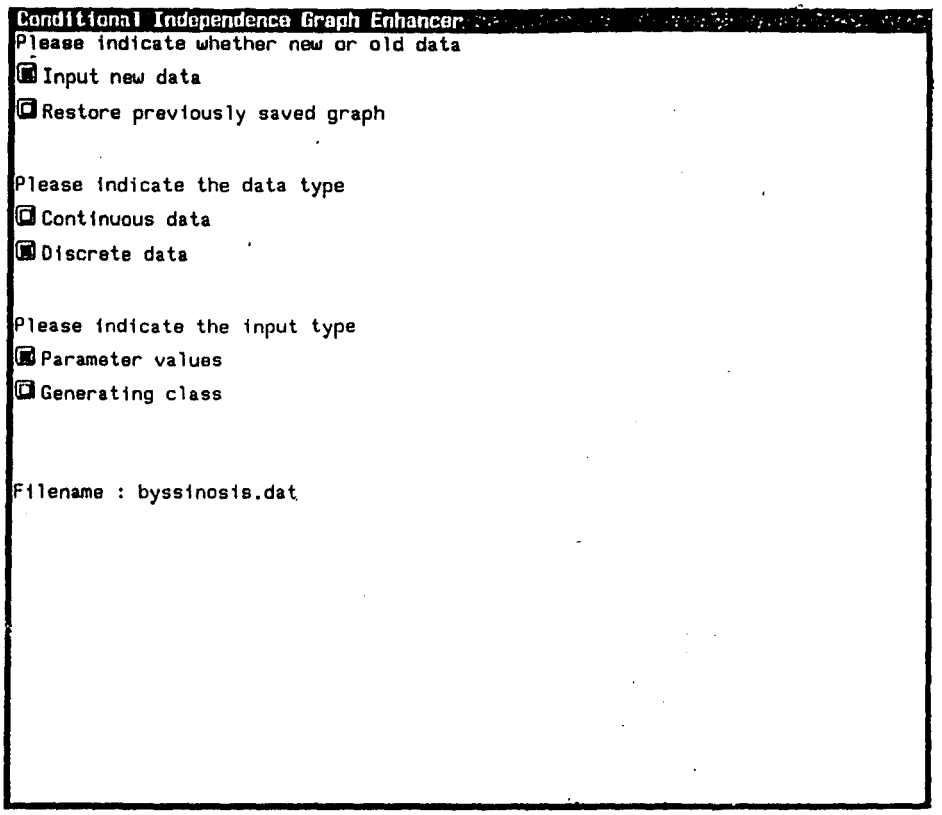

Figure 15-5: CIGE: Data entry screen for example discrete data model

```
6          <-- no. of variables
EMP
SMOKE
SEX        } variable labels
RACE
WPLACE
BYSS
EMP SMOKE SEX
EMP SEX RACE
SMOKE SEX RACE
EMP SEX WPLACE     } elements of generating class
EMP RACE WPLACE
WPLACE BYSS
EMP BYSS
SMOKE BYSS
```

Figure 15-6: CIGE: Input data file containing the generating class of the model, with annotations, for example discrete data model

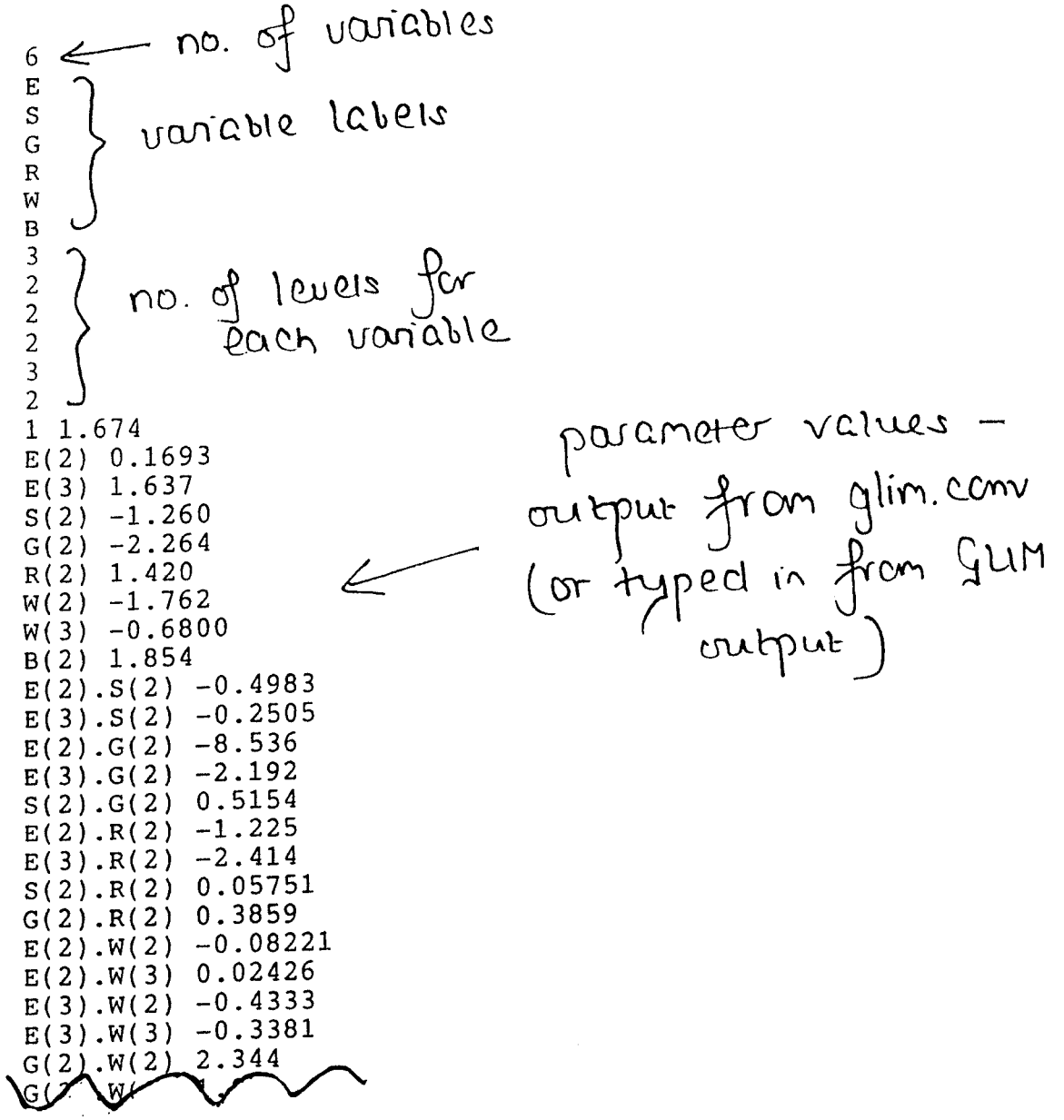

Figure 15-7: CIGE: Input data file containing the parameter values of the model, with annotations, for example discrete data model

Again, the CIGE module modifies the data file used as input according to the choices made at each pair of options displayed to give *cige.tmp* which is passed as *cige.dat* to the DCIGE module.

15.3 CCIGE: Continuous Data Module

The CCIGE module is automatically called by the CIGE module for use with continuous data in the form of the matrix of negative partial correlations or the matrix of edge exclusion deviances, corresponding to a covariance selection model.

15.3.1 Basic Display

CIGE automatically calculates a default threshold level for the display of edges, which corresponds to the 5% significance level. Thus if the matrix of edge exclusion deviances had been used as input, the default threshold would be $\chi^2_{[1]}$ =3.841. Since the matrix of negative partial correlations has been used as input instead, the default threshold will be the absolute value of $\rho_{ij.K}$ for which

$$\rho_{ij.K} = \sqrt{1 - e^{-\frac{3.841}{N}}}$$

where N is the number of observed units. This equation is derived from the formula described in Chapter 11 for the calculation of the matrix of edge exclusion deviances from the matrix of negative partial correlations.

The basic independence graph is drawn using the default lay-out obtained by locating the vertices corresponding to the variables equidistant around the circumference of an imaginary circle. These vertices are labelled with the variable names given as input. For those pairs of variables for which the absolute value of their (pair-wise) partial correlation exceeds the 5% (default) threshold value of $\rho_{ij.K}$, an edge is drawn between the corresponding vertices of the graph. The basic independence graph for the pit-prop data example is presented again in Figure 15-8.

By clicking and dragging the middle mouse button on any of the vertices in this default display and relocating them, it is possible to obtain a much more aesthetically pleasing display of the basic independence graph. An intermediate stage in the derivation of the modified lay-out is presented in Figure 15-9. This shows the appearance of the graph whilst the user is actually 'moving' one of the vertices. Because this much simplified version of the graph is drawn whenever the mouse button is depressed on a vertex, the graph can be continually updated, giving the impression that the edges are being dragged along by a moving vertex. This is very useful when attempting to

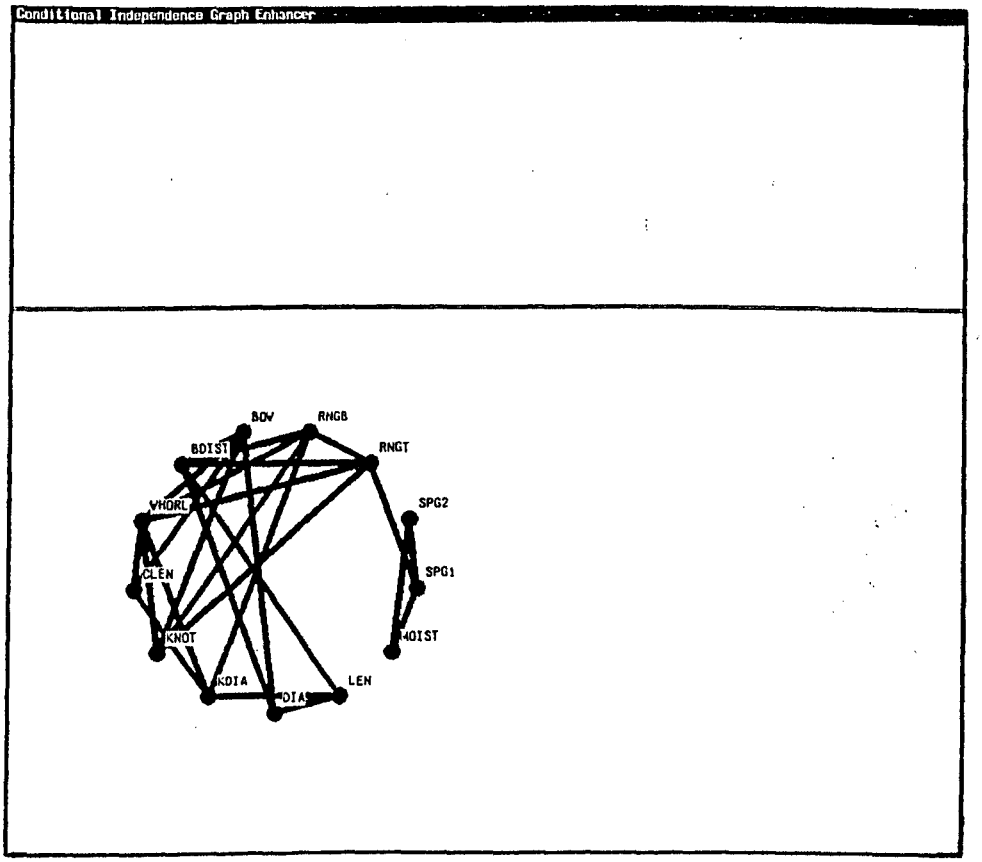

Figure 15-8: CCIGE: Default display of basic independence graph for continuous data example

determine the best (in some sense) location for each vertex, since the user can see, in real time, the effect of every position the vertex passes through upon the overall appearance of the graph. Only when the mouse button is released again is the graph redrawn with the thicker edges and variable labels, such as the graph presented again in Figure 15-10.

15.3.2 Menu Options

By clicking the right-hand mouse button anywhere in the window in which the basic graph is displayed, the menu of options is obtained. The appearance of the menu options is as shown in Figure 15-11. If the matrix of edge exclusion deviances had been used as input, the signed graph option would be automatically deselected, since it is not appropriate for this form of data.

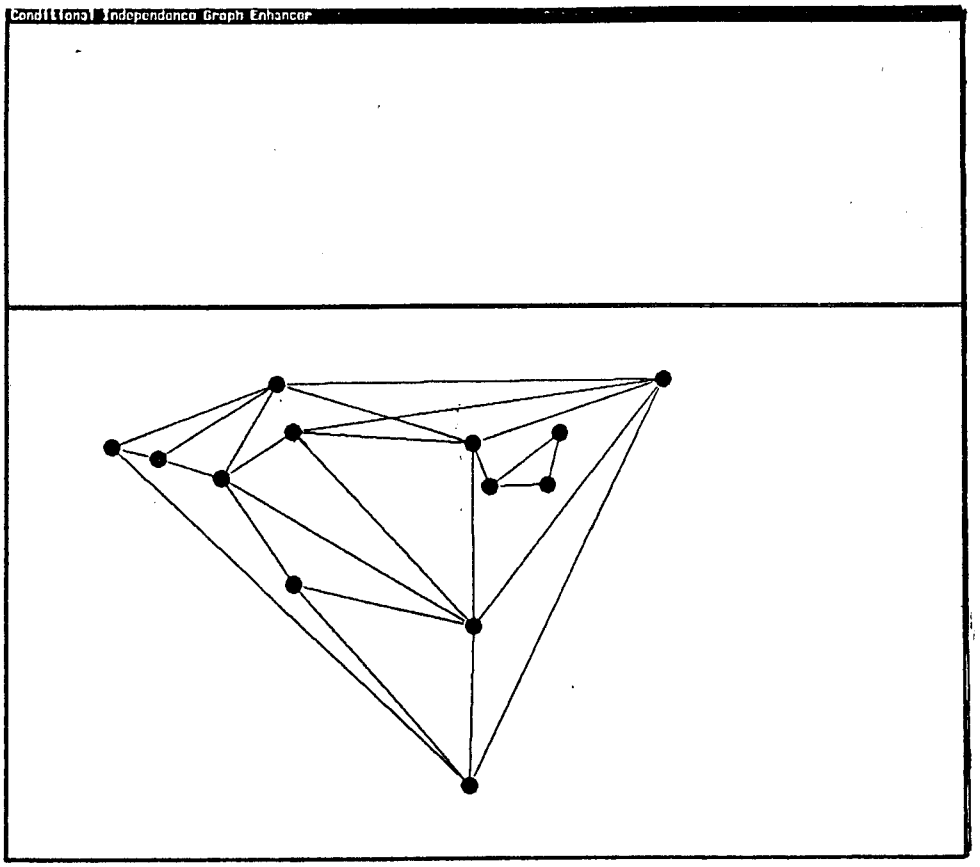

Figure 15-9: CCIGE: Intermediate stage in the modification of the display of the basic independence graph for continuous data example

In the remainder of this section, the effects of each of the menu options will be described in turn. Whenever no menu option is active, the basic independence graph is displayed by default. A menu option only ceases to be active when the user clicks the mouse on the DONE button. Whilst a menu option is active, it is still possible to relocate the vertices of the displayed graph.

Numerical Values

When this menu option is selected, the complete graph on all n vertices (involving $n(n-1)/2$ edges) is displayed. The edges which were included in the basic graph are still represented as continuous lines, but those edges in the complete graph which are not included in the basic graph are drawn narrower and so appear fainter.

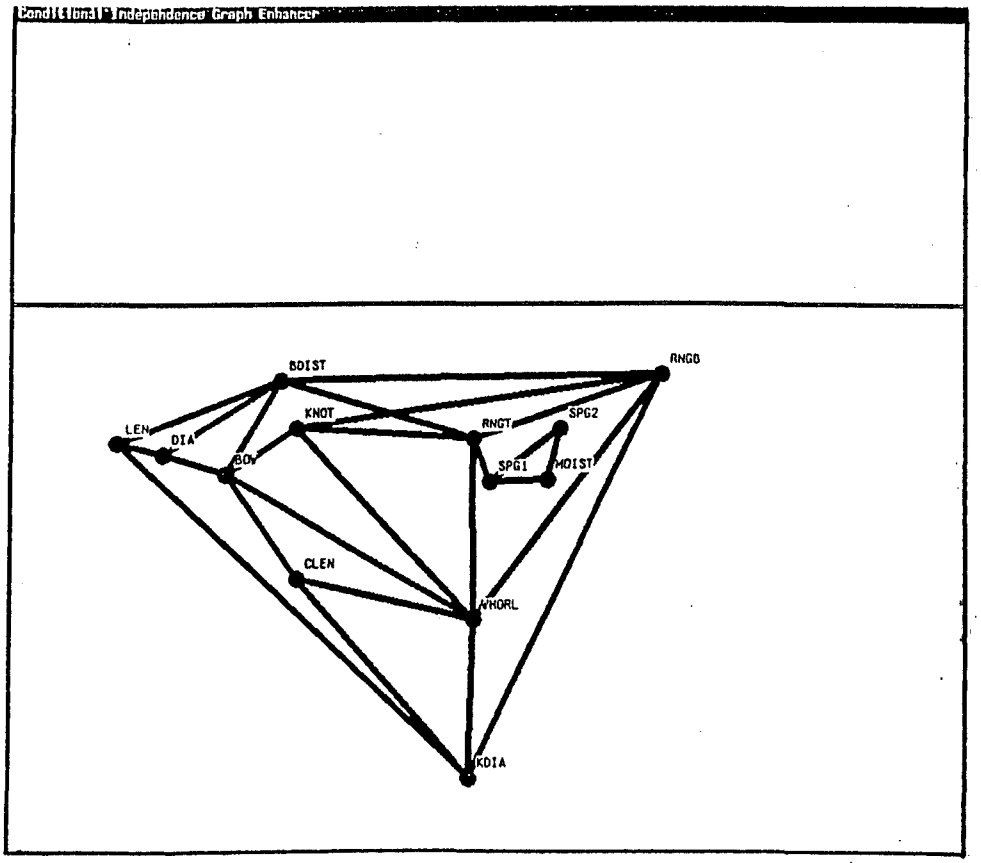

Figure 15-10: CCIGE: Modified display of basic independence graph for continuous data example

By clicking the left mouse button on any edge in the complete graph, it is possible to obtain the (signed) value of the partial correlation between the two variables joined by that edge. See, for example, the graph displayed in Figure 15-12 — having just clicked on the edge joining the vertices labelled RNGB and KDIA, the user is infromed that the strength of the association represented by this edge is –0.241. Note that the sign given is that which corresponds to the partial correlation, which is the opposite to that contained in the matrix of negative partial correlations used as input.

If the matrix of edge exclusion deviances had been used as input instead of the matrix of negative partial correlations, then the (always positive) value of the edge exclusion deviance between pairs of variables would be displayed.

If two edges are located close together, and the user clicks on an ambiguous area (ie. where the edges are very close), CIGE will display the partial correlation corresponding to one of the edges. Which edge this is is dependent on the way in which

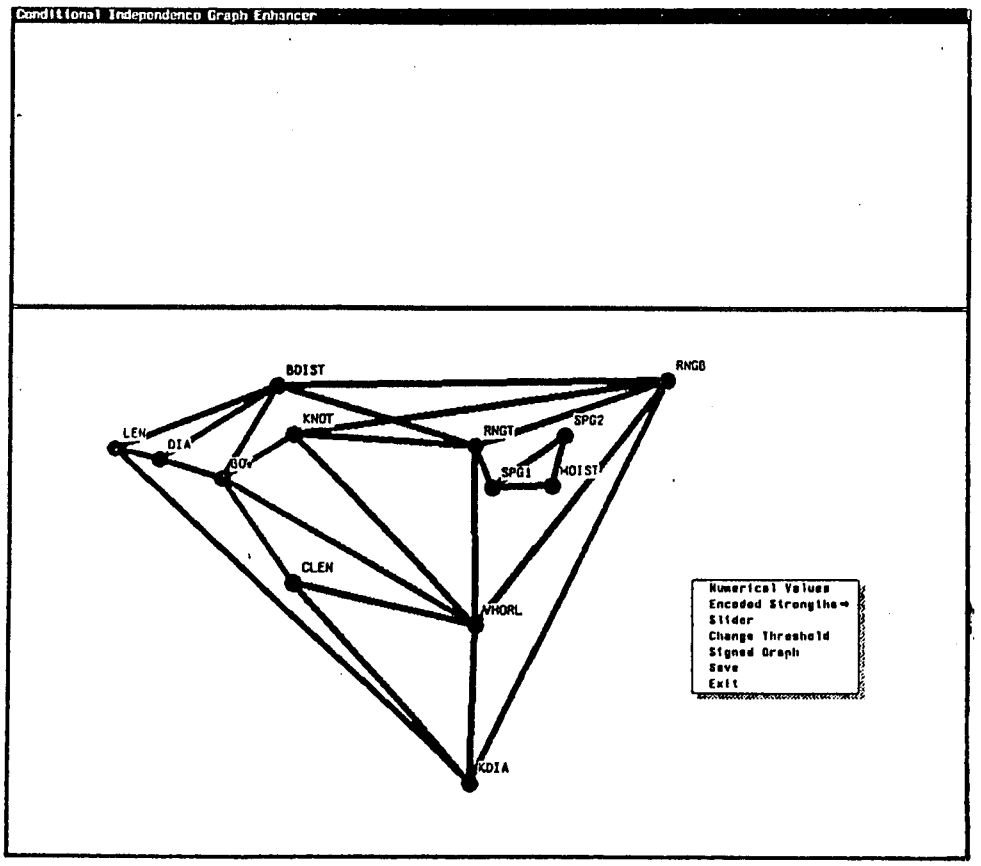

Figure 15-11: CCIGE: Menu options for use with continuous data example

the program tests each edge in turn until it finds one within a certain distance of where the mouse was clicked, but will be apparent from the labelling of the displayed value. If it is not possible to find an unambiguous part of an edge to click on in order to obtain the desired partial correlation it is possible to relocate the vertices using the middle mouse button as before.

Using this menu option it is therefore possible to obtain precise information about the strength of the association between pairs of variables where the strength of the association is below the threshold value and the edges are therefore not displayed in the basic graph, even though their value may be non-zero. It is also possible to obtain precise information about the strength of associations which are greater than the threshold value and are therefore represented by edges in the basic graph.

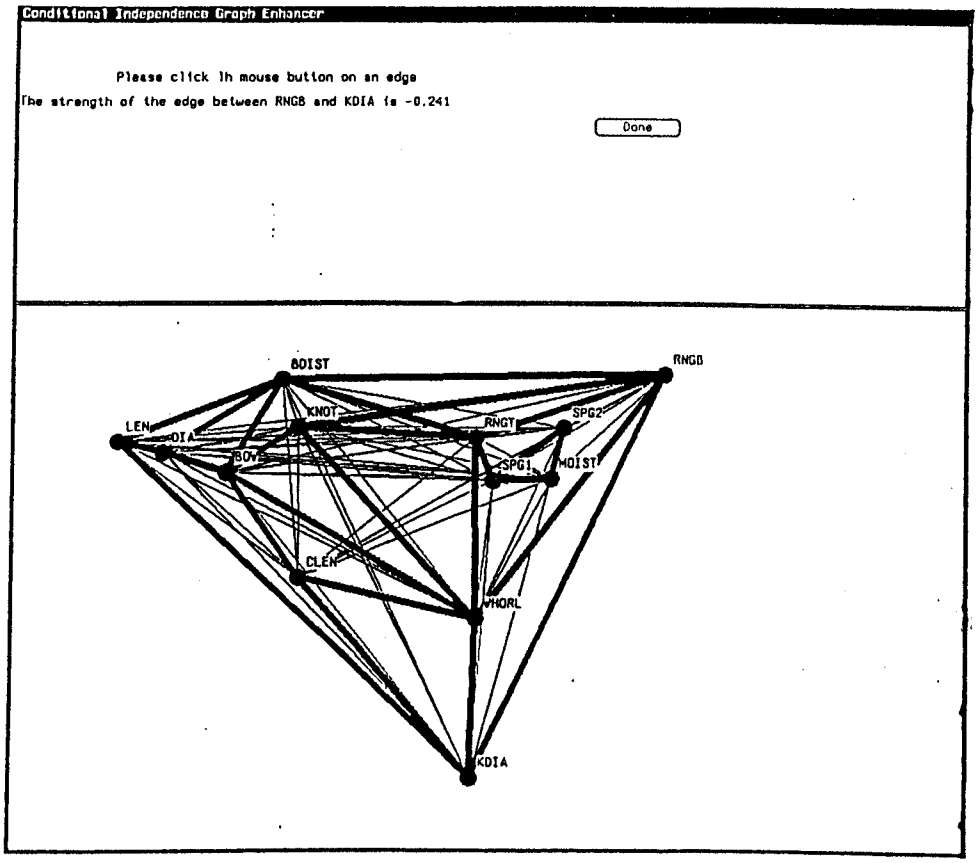

Figure 15-12: CCIGE: Numerical values menu option applied to continuous data example

Encoded Strengths

When this menu option is chosen, another, pull-right, menu is displayed listing the three different ways in which the strength of the associations represented by the edges in the basic graph can be incorporated in the graph. Having chosen one of these options, yet another pull-right menu is displayed listing the three different ways of specifying the boundaries between each of the levels of encoding used. The appearance of these menus and menu options is as shown in Figure 15-13.

For the first of the options for specifying the boundaries between each of the levels of encoding, which is based on significance levels, default significance levels are used which correspond to $p>0.1$, $0.1\geq p>0.05$, $0.05\geq p>0.025$, $0.025\geq p>0.01$, $0.01\geq p>0.005$ and $p\leq 0.005$. The values of the edge exclusion deviances or (absolute) partial correlations corresponding to these values of p can be calculated automatically by

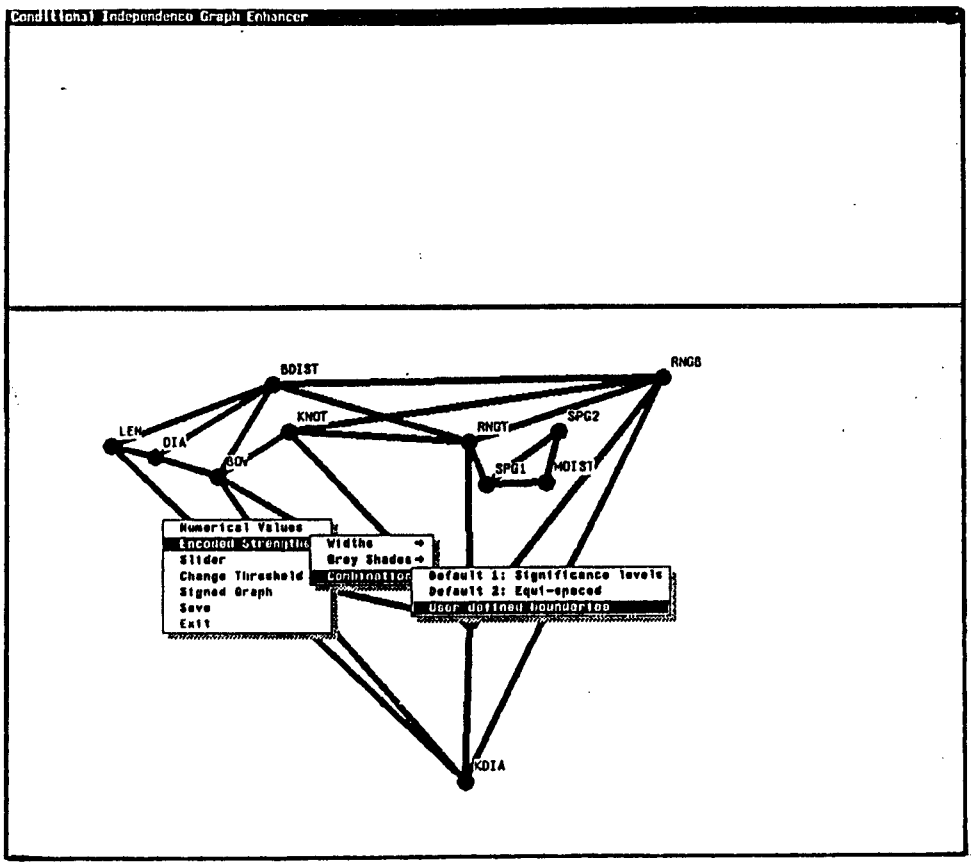

Figure 15-13: Additional menu options for use with Encoded Strengths option for continuous data example

CIGE since the values of the edge exclusion deviances follow a $\chi^2_{[1]}$ distribution, and are therefore tabulated as the critical values of $\chi^2_{[1]}$ corresponding to different values of p. The corresponding values of the partial correlations can be found using the equation presented in Section 15.3.1, substituting the appropriate value from the $\chi^2_{[1]}$ distribution for the default value of $\chi^2_{[1]}$ =3.841 (which corresponds to p=0.05). The smaller the value of p (in other words, the greater the significance of the association represented by a particular edge), then the wider and/or darker the edge. For example, Figure 15-14 shows the use of the default significance levels option in conjunction with the grey shades option for encoding the strength of association.

For the second option, where the levels are equi-spaced, the boundaries between levels are found by default by considering the maximum absolute partial correlation (or maximum edge exclusion deviance) and dividing this value by 6. Then the boundaries of

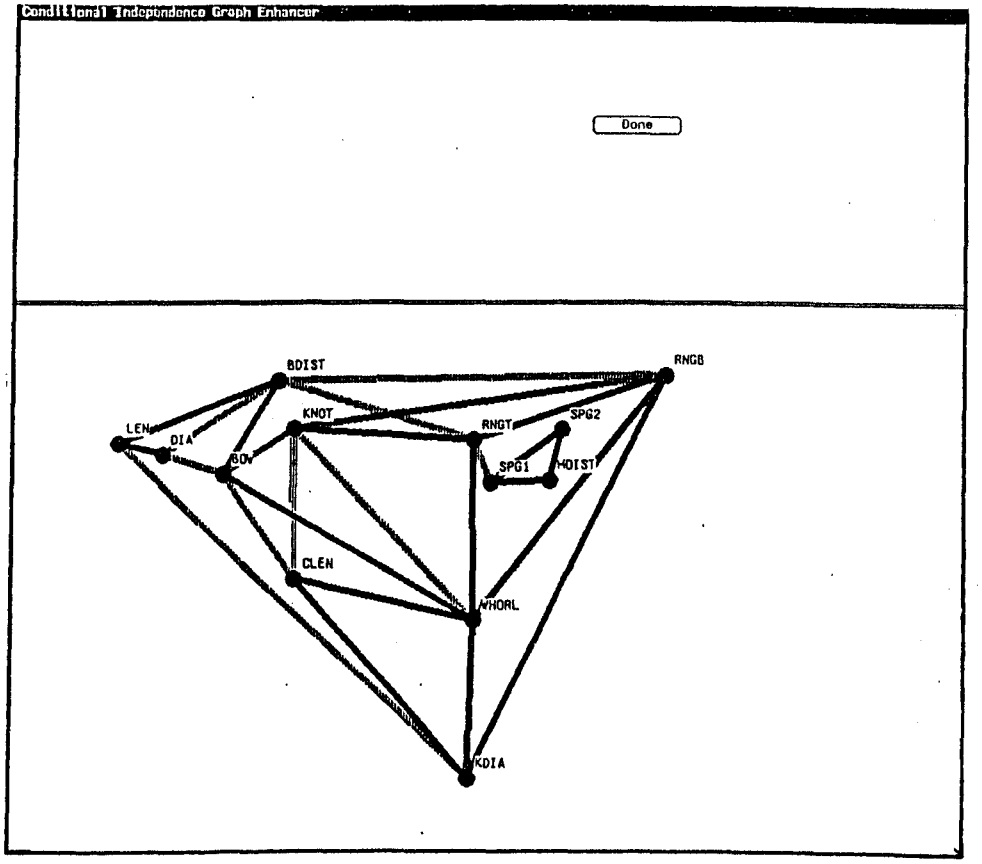

Figure 15-14: CCIGE: Default significance levels option used in conjunction with the grey shades option for encoding strength of association, for continuous data example

the six levels are calculated as max, $\frac{5}{6}$max, $\frac{4}{6}$max, $\frac{3}{6}$max, $\frac{2}{6}$max, $\frac{1}{6}$max, 0. The closer the level to the maximum value (in other words, the stronger the association), then the wider and/or darker the edge. For example, Figure 15-15 shows the use of the equi-spaced default option in conjunction with the widths option for encoding strength of association.

For the third option, the user is free to specify his/her own choice of values for the strengths of the associations (absolute partial correlations or edge exclusion deviances, as appropriate) corresponding to the boundaries of the six levels. The user's choice of values is subject to the logical requirement that the boundary values must increase monotonically. Again, the stronger the association, then the wider and/or darker the edge. Figure 15-16 shows the initial appearance of the graph after the user-defined boundaries option has been selected. Irrespective of the option which would have already been

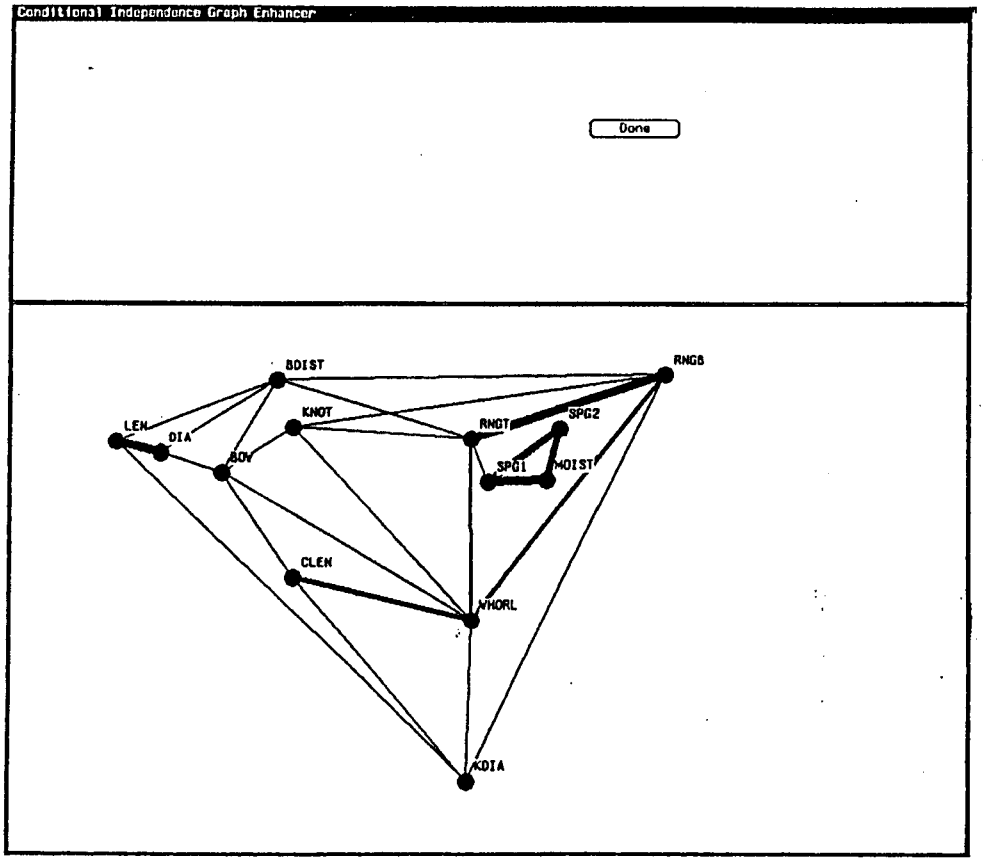

Figure 15-15: CCIGE: Default equi-spaced option used in conjunction with the widths option for encoding strength of association, for continuous data example

chosen for encoding the strengths of the associations, the basic independence graph is redisplayed and the 5 thresholds between the 6 levels are all initialised at the current threshold level. The user can then specify his/her own boundary values. For example, Figure 15-17 shows the use of the user-defined boundaries option for the trivial case where the user has defined the boundary values to correspond to the significance levels available automatically with the default significance levels option, in conjunction with the combination (widths and grey-shades) option for encoding the strengths.

By comparing Figure 15-15 with Figure 15-14 (or Figure 15-17), it can be seen that the two default options for calculating the boundaries between the levels can result in quite different encodings of the edges.

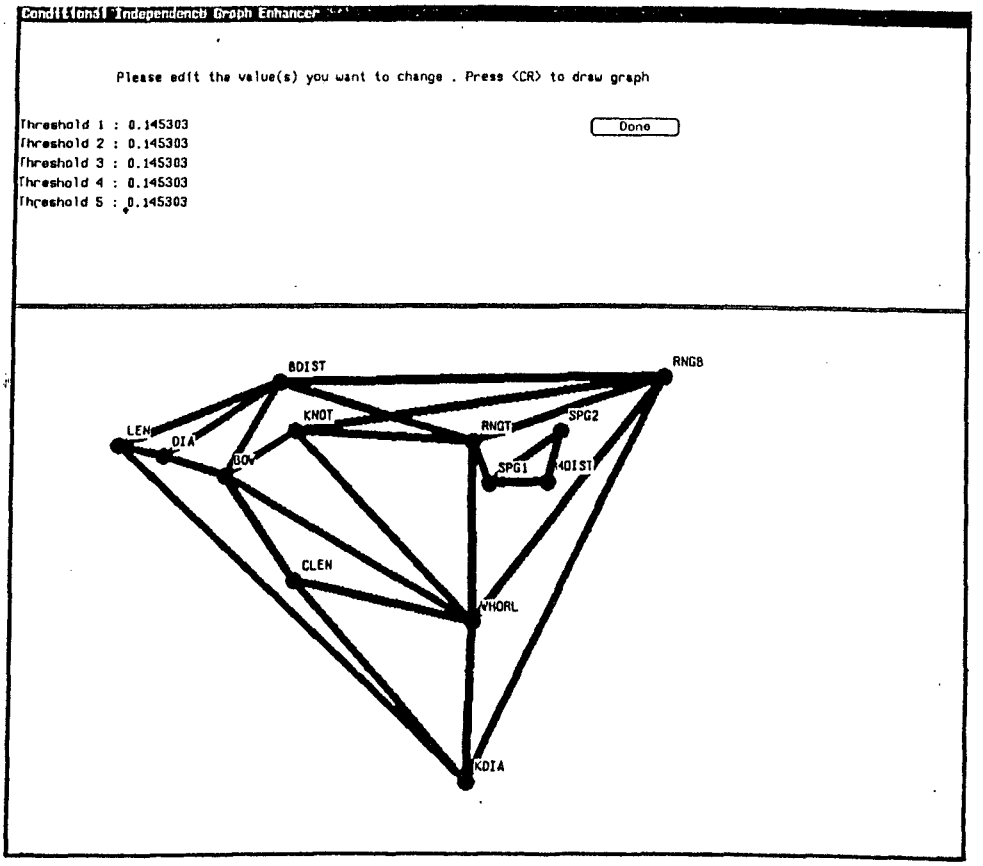

Figure 15-16: CCIGE: User-defined boundaries option applied to continuous data example

Slider

Use of this menu option makes it possible for the user to investigate the effect of changing the threshold value, above which edges are included in the graph, and below which edges are excluded.

When this option is selected, a slider is displayed above the graph. The left-hand end of the slider corresponds to a threshold value of zero, for which all edges would be included in the graph (since absolute values of the partial correlations are considered). The right-hand end of the slider corresponds to a threshold value equivalent to the maximum absolute partial correlation or maximum edge exclusion deviance, for which no edges would be included in the graph. Initially, the slider is located at the position corresponding to the current threshold value, as shown in Figure 15-18.

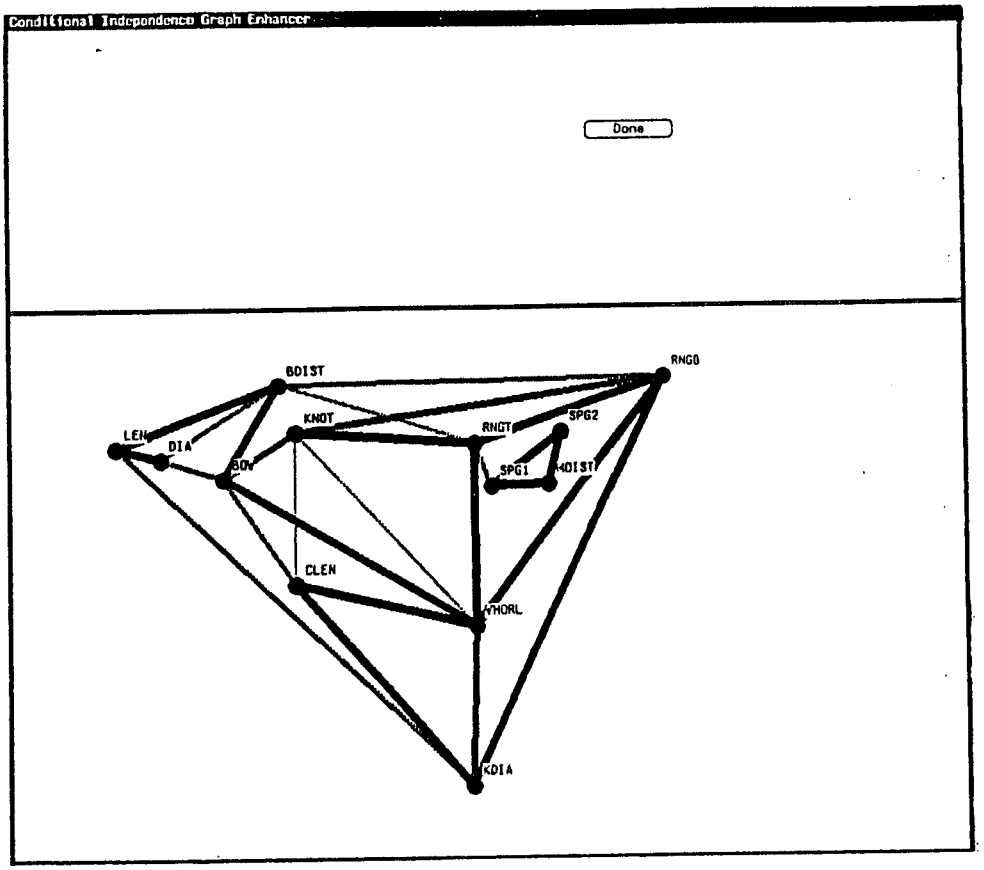

Figure 15-17: CCIGE: User-defined boundaries option used in conjunction with the combination option for encoding strength of association, for continuous data example

The user can click the left-hand mouse button on the slider and drag it in either direction, left or right. As they do so, edges are added to or dropped from the graph according to the change in threshold value, and the changing threshold value is displayed above the slider. For reasons of speed, the graph drawn whilst the mouse button is depressed on the slider has thin edges as shown in Figure 15-19. Only when the button is released is the graph redrawn with the usual, thicker, edges according to the prevailing threshold value, as shown in Figure 15-20. In Figure 15-20, the threshold value displayed above the slider is greater than the original threshold displayed in Figure 15-18 and so there are fewer edges in the redrawn graph.

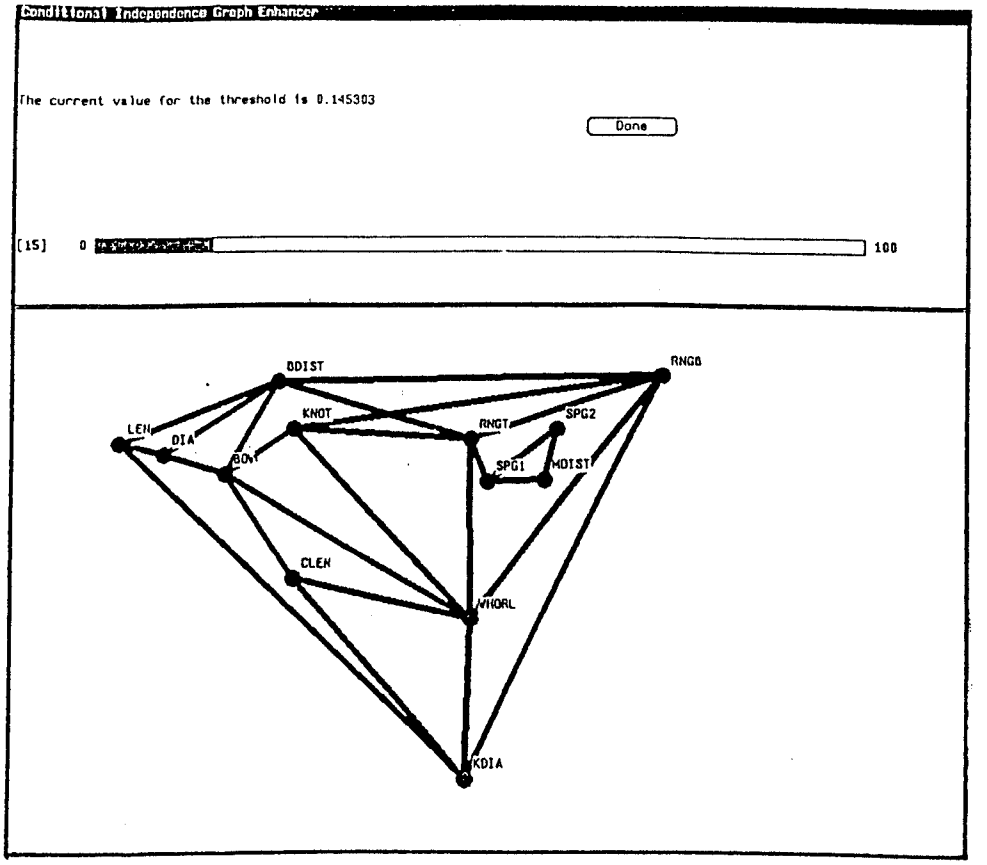

Figure 15-18: CCIGE: Initial display of slider for continuous data example

Note that this menu option is for exploring the effect of a change in threshold value and that the current value of the threshold is not permanently altered — as soon as this option is exited, the original basic independence graph is displayed. To change the default threshold permanently, the following menu option must be used.

Changing the Threshold

Having decided to permanently change the current threshold value, perhaps by calculation (outside of CIGE) of the default threshold corresponding to a significance level other than 5%, or by exploratory use of the slider menu option, this menu option should be used. Once a new value for the threshold is typed in, the independence graph is redrawn according to this new value. However, the new value does not become the new permanent threshold until the user clicks the left-hand mouse button on the button marked

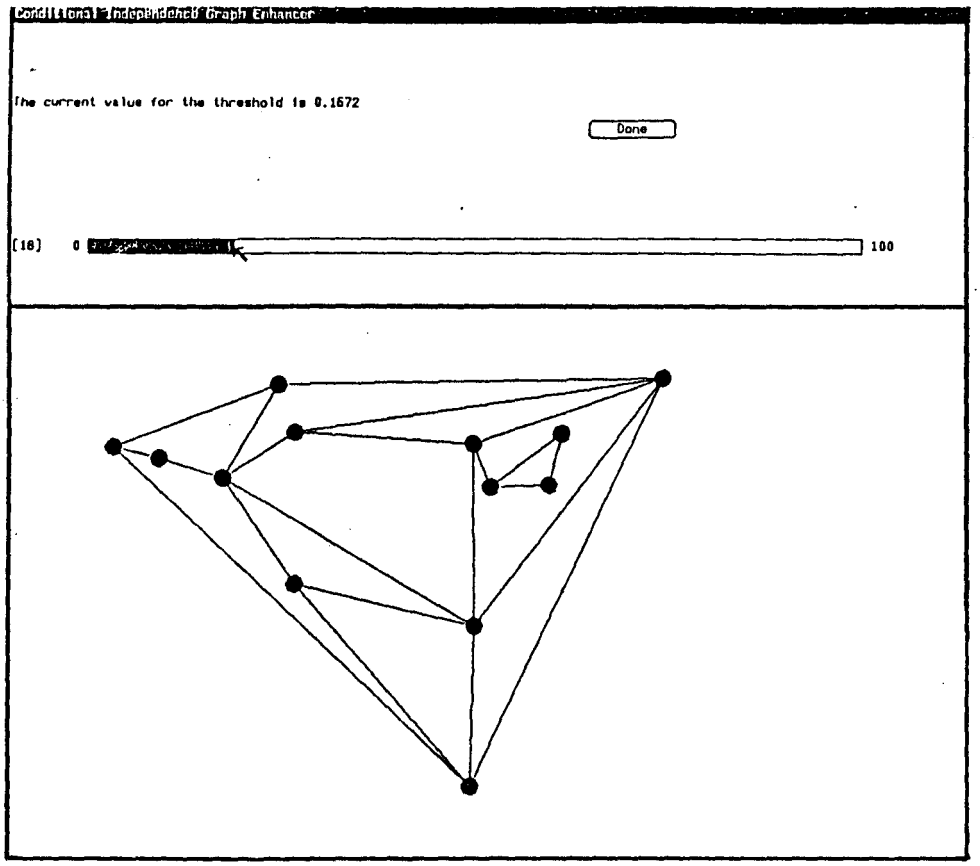

Figure 15-19: CCIGE: Intermediate stage in the use of the slider for continuous data example

'DONE'. This allows the user to experiment with different values before making a permanent change. After making a permanent change, the basic independence graph will thereafter be displayed with the edges corresponding to absolute partial correlations (or edge exclusion deviances) greater than this new value, until the value is again changed using this option. Use of this option is illustrated in Figure 15-21.

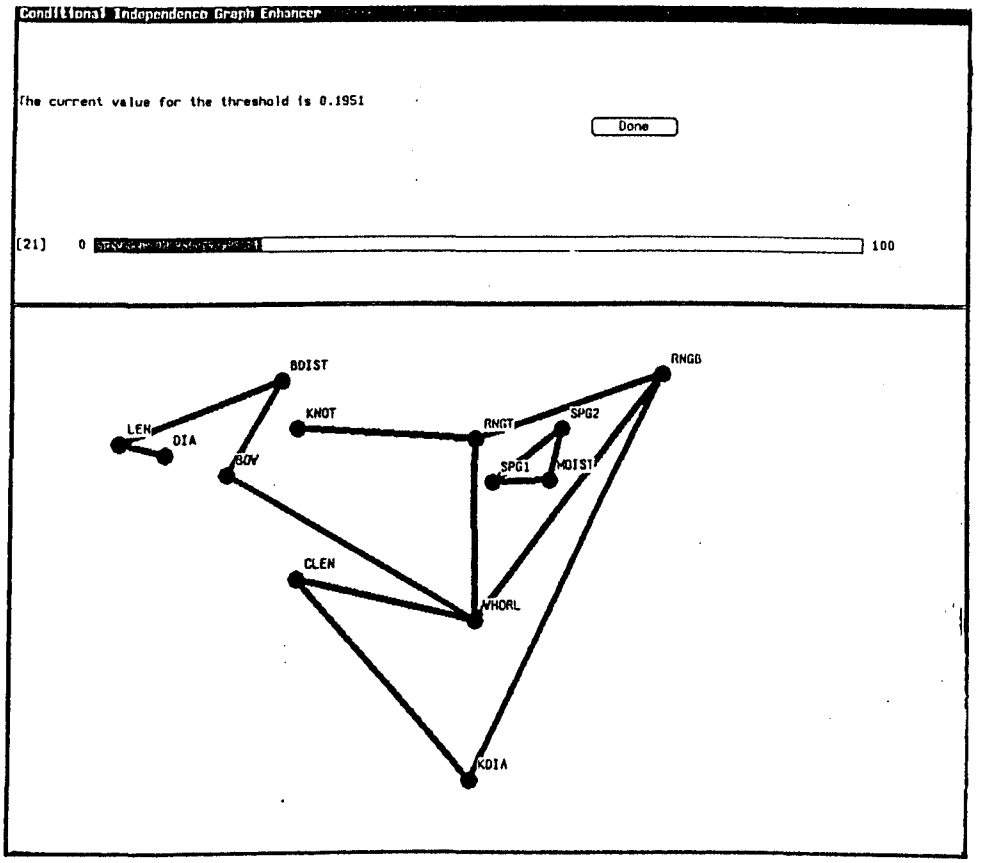

Figure 15-20: CCIGE: Display of independence graph for new value of threshold obtained using the slider for continuous data example

Signed Graph

In the basic independence graph, the edges displayed correspond to pairs of variables with an *absolute* partial correlation greater than the current threshold value. However, the user may wish to ascertain which pairs of variables are positively associated, and which are negatively associated. In choosing this menu option, the basic graph is re-displayed, but those edges corresponding to a positive partial correlation are drawn with the usual thick dark edge, and those edges corresponding to a negative partial correlation are drawn with a thinner and therefore fainter edge. Use of this option is illustrated in Figure 15-22.

As has already been mentioned, this menu option is not available if the edge exclusion deviance matrix is used as input, since there are no signs attached to their values.

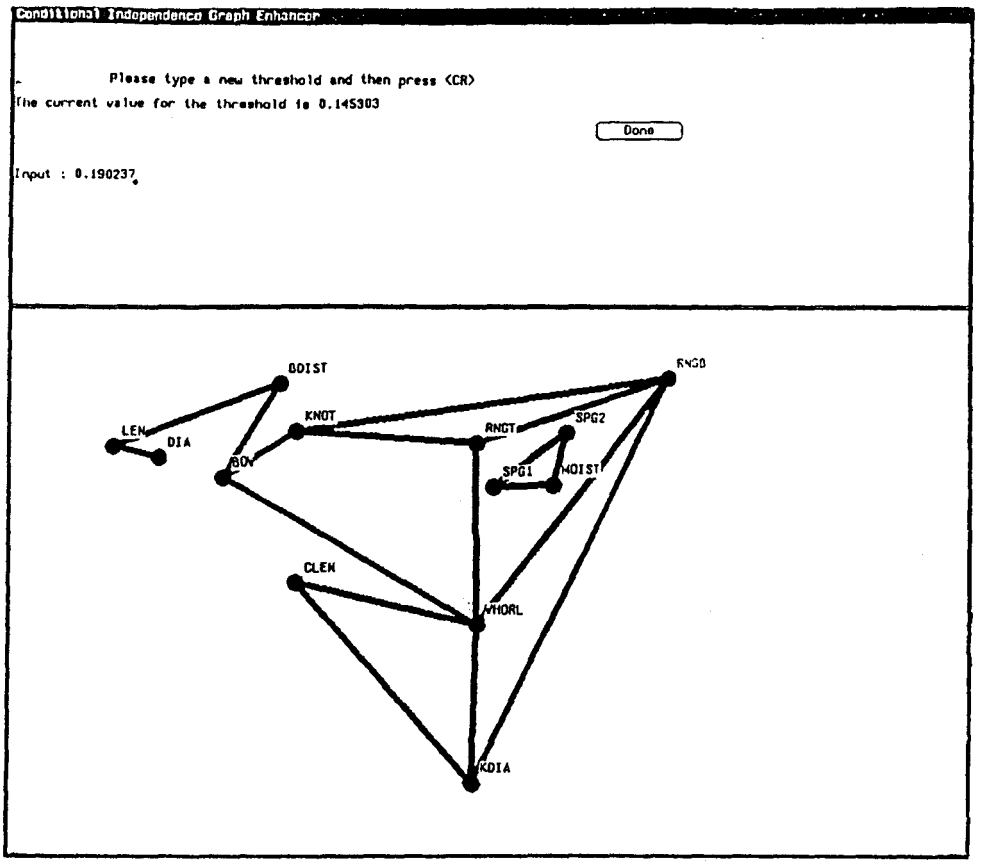

Figure 15-21: CCIGE: Change threshold menu option applied to continuous data example

Saving the Results

As mentioned above, the basic independence graph displayed can be saved at any time for further examination on another occasion. This is achieved by selecting this menu option and specifying the filename by which the data is to be saved (see Figure 15-23).

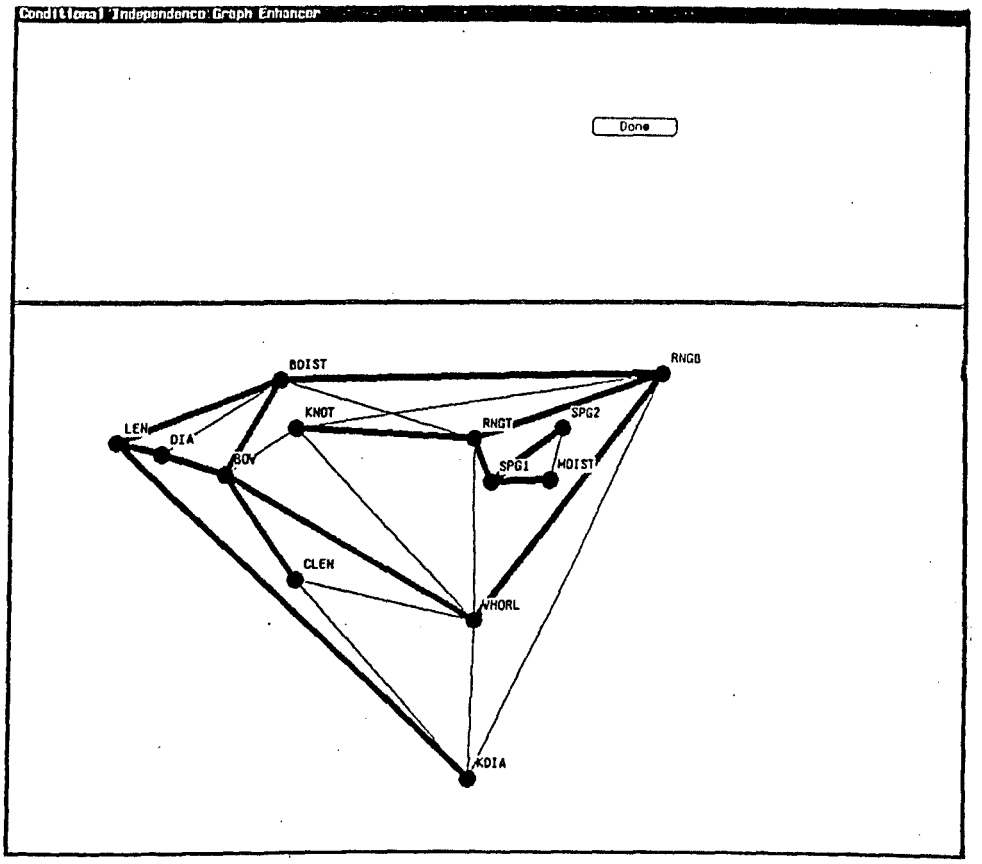

Figure 15-22: CCIGE: Use of signed graph menu option for continuous data example

The saved file, which can be specified in the data entry module any time CIGE is run (see Section 15.2 above), will contain the data matrix used as input, the locations of the vertices, the variable labels, and the default threshold level, and other information which was specified when the data was originally entered, such as whether the data was continuous or discrete, the form of the data matrix used as input, and the number of observed units. An annotated example of the saved data file for the pitprop data example is contained in Figure 15-24.

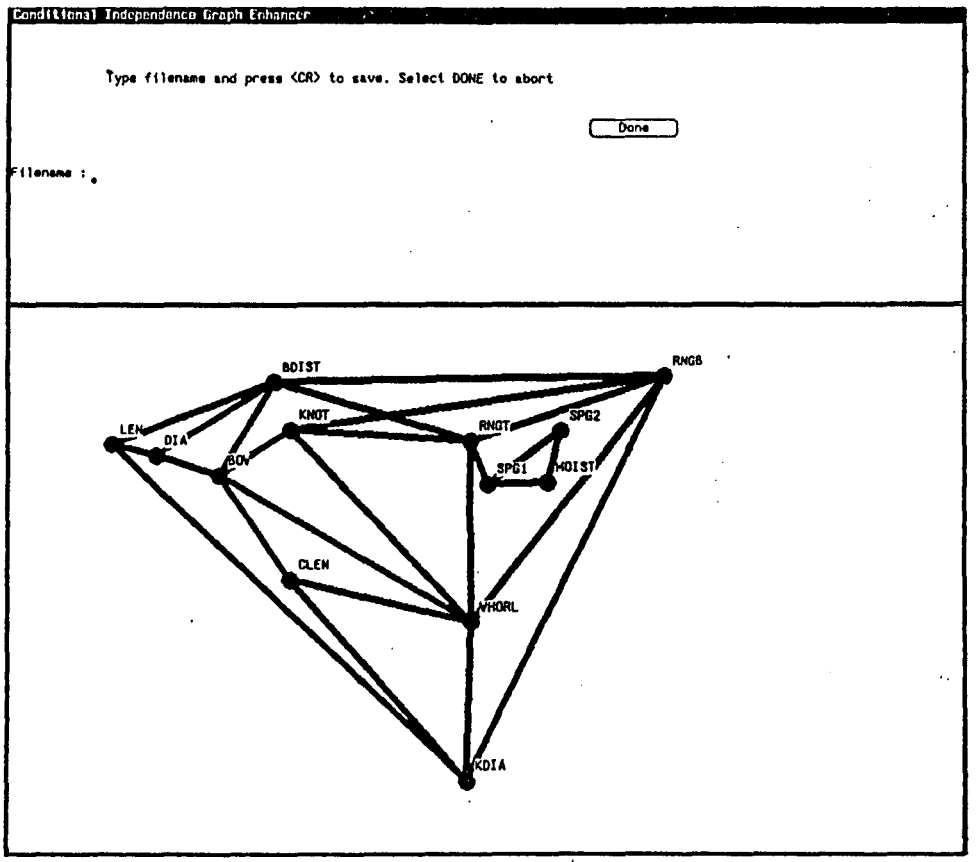

Figure 15-23: CCIGE: Save menu option for continuous data example

Exiting from CIGE

By selecting the menu option '**exit**', the CIGE program is exited, after seeking confirmation of the user's intention. If the graph has not been saved by explicit use of the save menu option, then all changes will be lost.

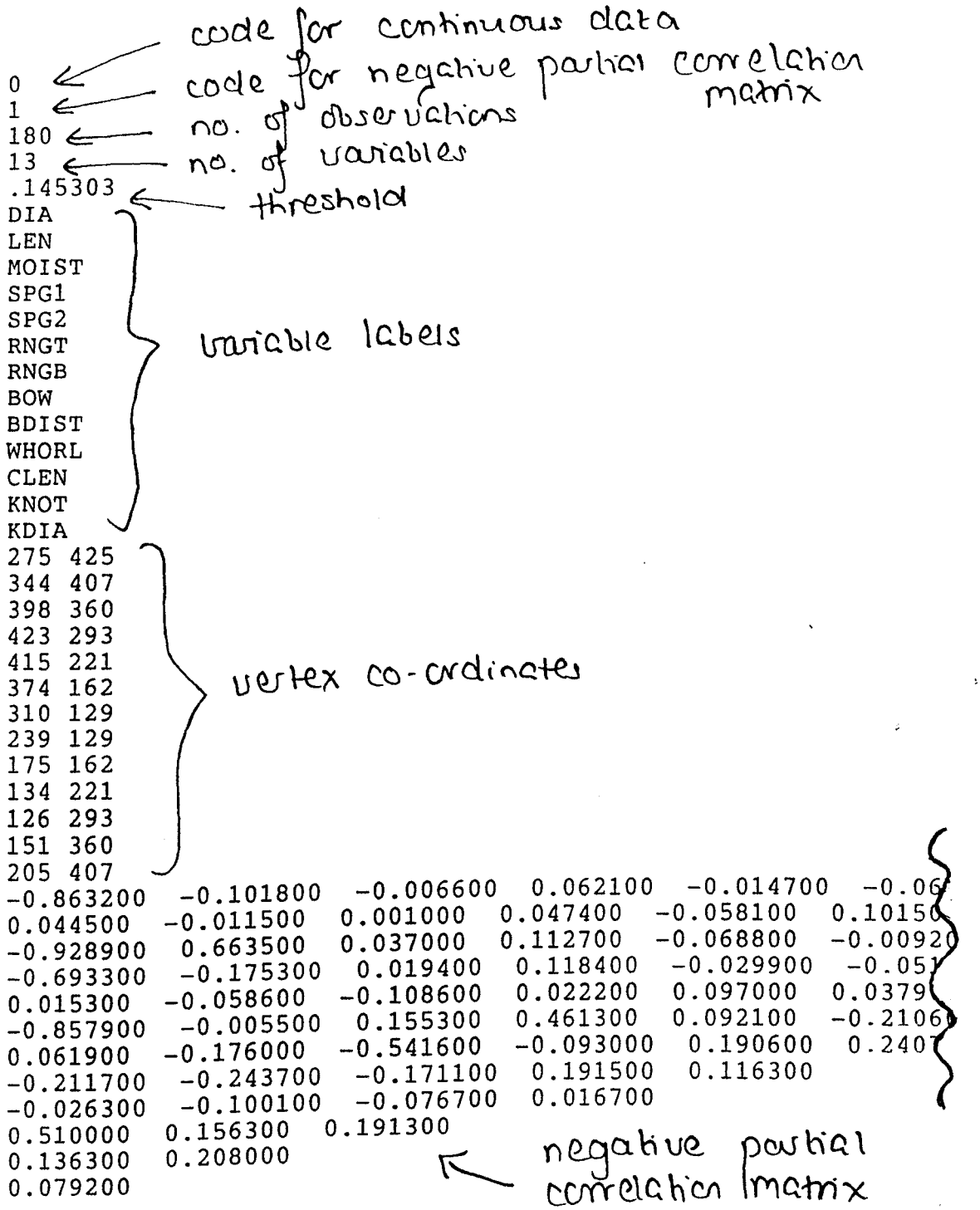

Figure 15-24: CIGE: Saved data file with annotations for example continuous data model

15.4 DCIGE: Discrete Data Module

The DCIGE module is automatically called by the CIGE module for use with discrete data, in the form of the generating class of a fitted log-linear interaction model, or of the actual parameter values of the terms in the fitted model. Given the parameter values as input, CIGE is able to derive the generating class corresponding to the fitted model.

15.4.1 Basic Display

The basic independence graph is drawn using the default lay-out obtained, as in the CCIGE module described above, by locating the vertices corresponding to the variables equidistant around the circumference of an imaginary circle. The vertices are labelled with the variable names which were inputted. For those pairs of variables for which a two-way interaction between the variables is either contained in the generating class of the model, or is implied by an element of the generating class of the model, an edge is drawn between the corresponding pair of vertices in the graph. Figure 15-25 again shows the default layout of the independence graph for the discrete data example, together with the menu of options (see below).

By clicking and dragging the middle mouse button on the vertices in this default display, it is possible to obtain a much more aesthetically pleasing display of the basic independence graph, such as the graph presented again in Figure 15-26.

15.4.2 Menu Options

By clicking the right-hand mouse button anywhere in the window in which the basic graph is displayed, the menu of options is obtained as already shown in Figure 15-25. In the remainder of this section, the effects of each of these options is considered in turn. Whilst a menu option is active, it is still possible to relocate the vertices of the displayed graph. When displaying the elements of the generating class, although the positions of the vertices when this option was called will be mirrored in the second window, any relocation of the vertices of the graph in the main window will not be mirrored in the second window until the option is re-selected.

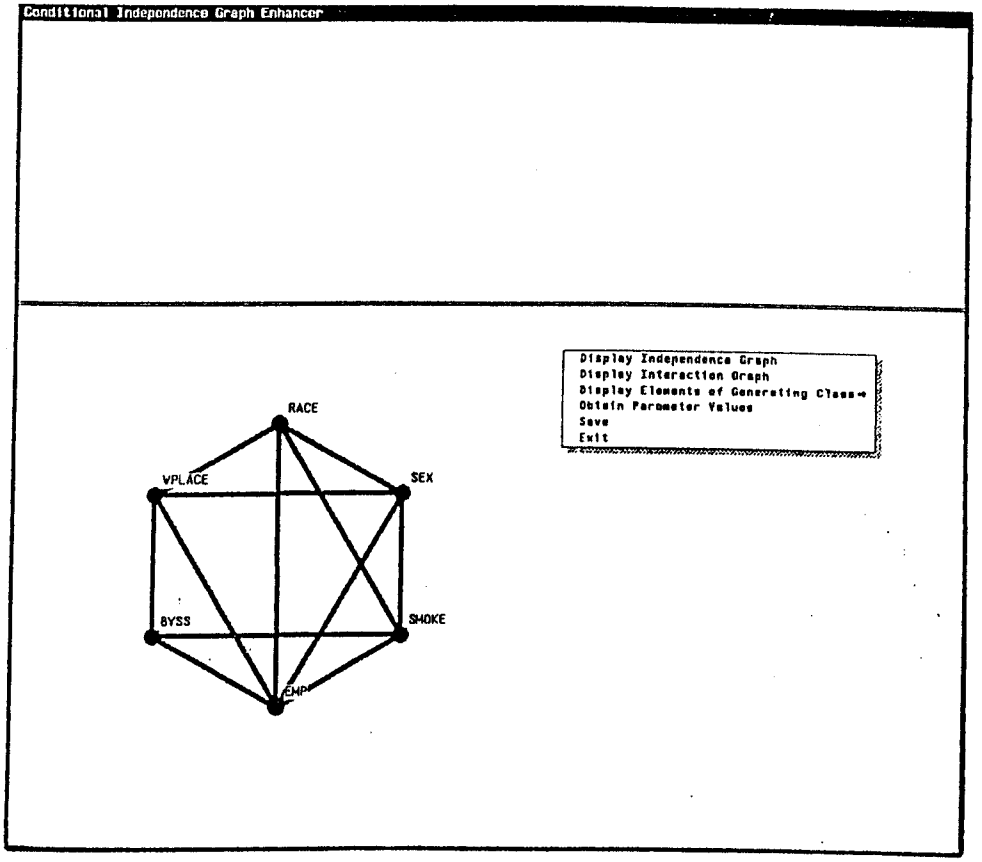

Figure 15-25: DCIGE: Default display of basic independence graph for discrete data example

As for the CCIGE module, a menu option remains active until the user clicks the mouse on the '**DONE**' button, with the exceptions of the independence graph and interaction graph options, one or other of which is active at all times (see below).

Independence Graph

This corresponds to the basic independence graph described in the previous section and displayed in Figure 15-26. Unlike continuous data, it is not the case for discrete data that this graph is displayed by default whenever no other menu option is active. Instead, either the independence graph or the interaction graph (described in the next section) may be displayed, depending on which was the most recently chosen menu option.

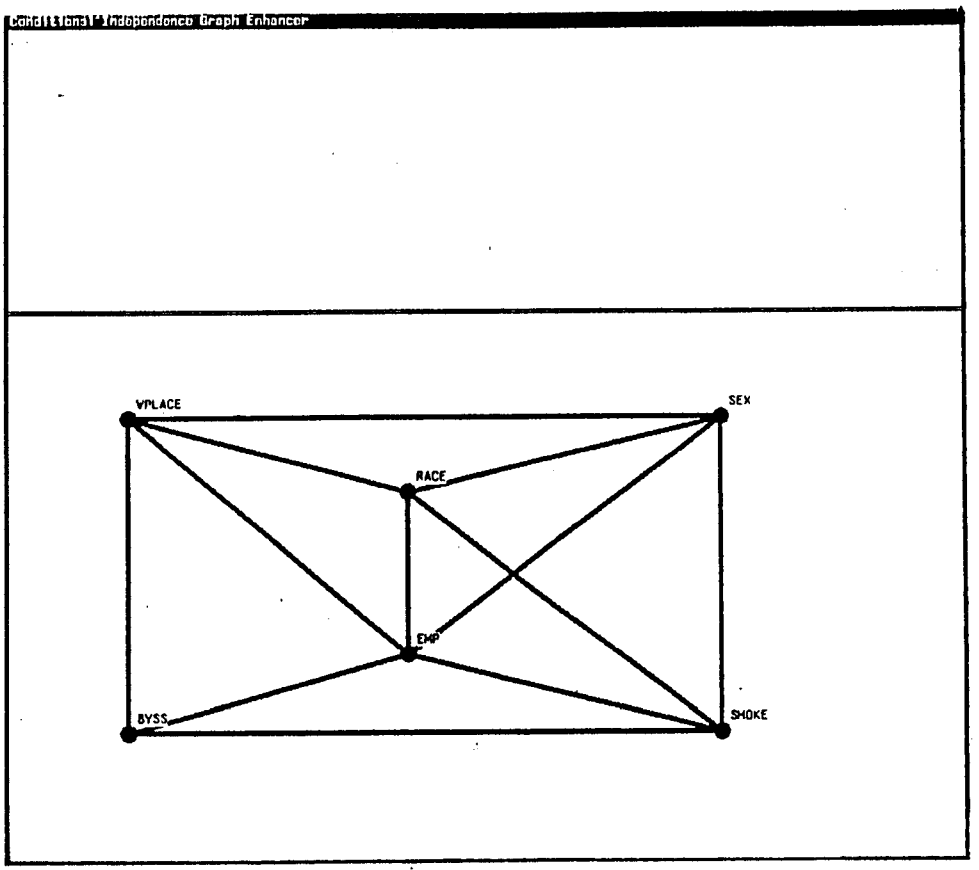

Figure 15-26: DCIGE: Modified display of basic independence graph for discrete data example

Interaction Graph

When this menu option is chosen, the interaction graph is displayed. This will have the same edges present and lay-out of the vertices as the independence graph in Figure 15-26, but the style of the edges will be coded according to the order of the largest interaction involving that pair of variables, as shown in Figure 15-27.

A key listing all the possible edge codes and the corresponding order of interaction is available for use with this menu option, and also for use with the display of the elements of the generating class described below, by clicking on the displayed key icon. This key is presented in Figure 15-28.

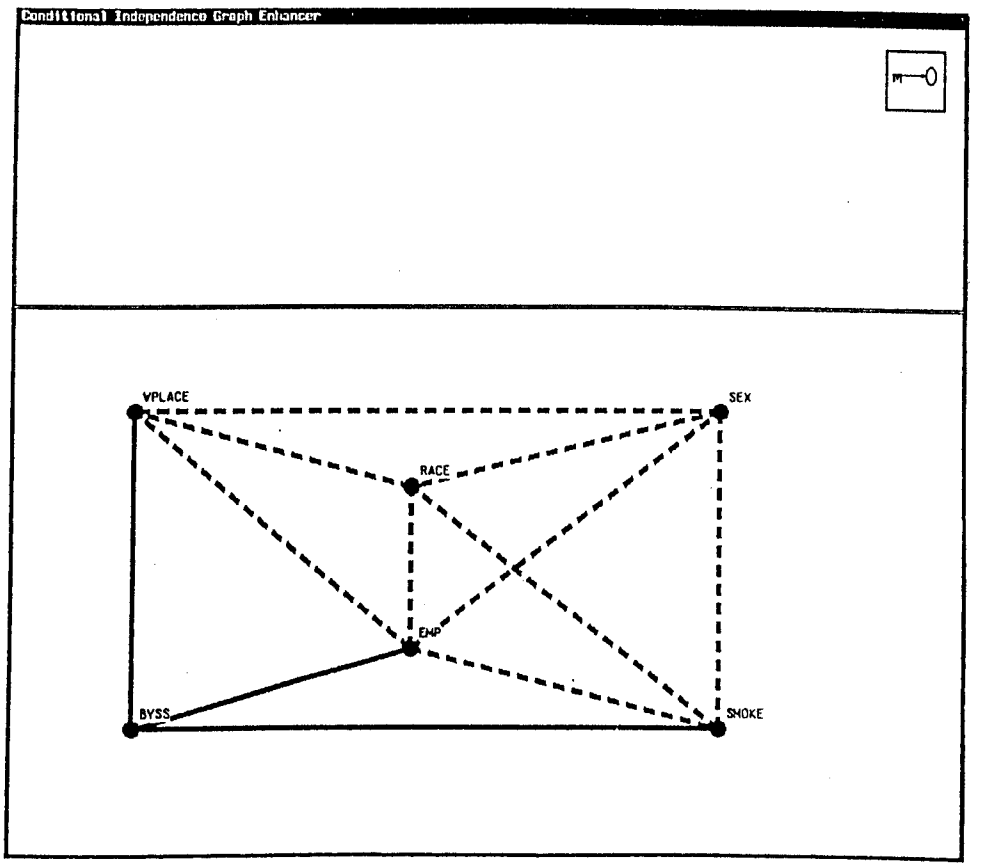

Figure 15-27: DCIGE: Display of interaction graph for discrete data example

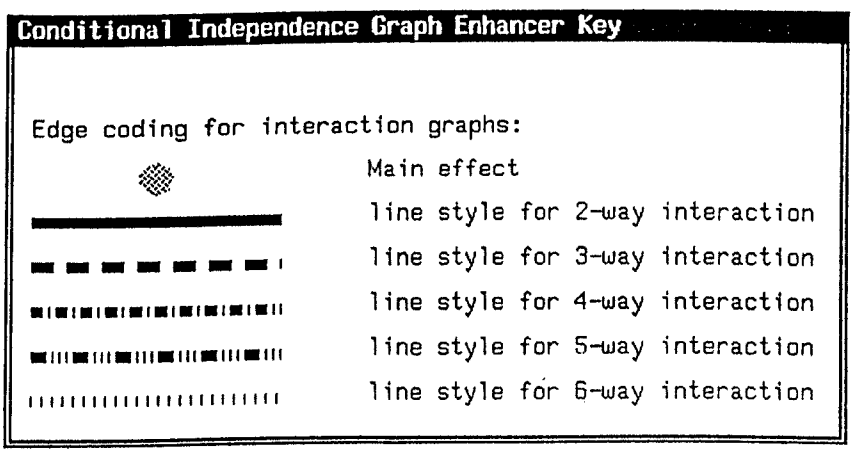

Figure 15-28: Key listing all possible edge codes and corresponding orders of interaction for interaction graphs for discrete data example

The appearance of the key when the key icon in Figure 15-27 is clicked, is as in Figure 15-29.

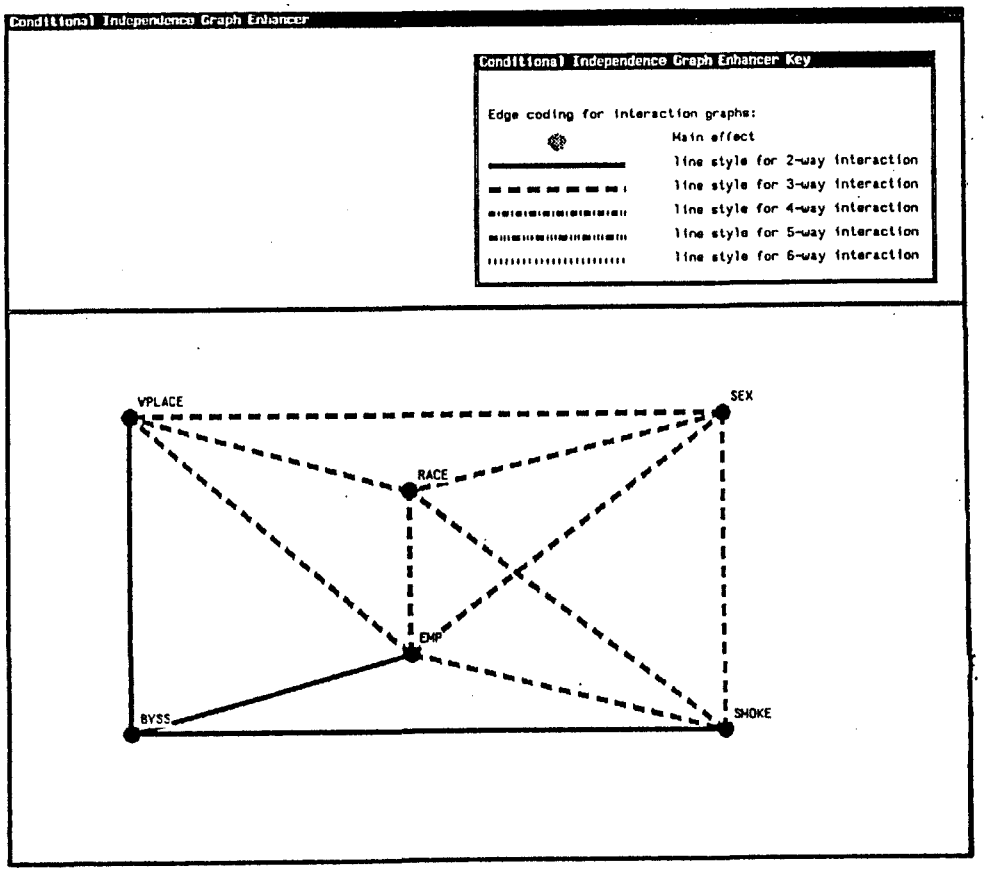

Figure 15-29: DCIGE: Display of interaction graph with key for discrete data example

Elements of Generating Class

If this menu option is chosen, an additional, pull-right, menu appears as shown in Figure 15-30. This allows the user to specify whether the elements of the generating class are to be displayed simultaneously or sequentially. For either option, the independence or interaction graph continues to be displayed in the current window, and a second window appears in which the elements of the generating class will be displayed.

If the elements of the generating class are to be displayed simultaneously, a full-screen sized 'canvas' is used to display the elements in a square grid of dimensions most appropriate to the number of elements in the generating class (eg. 2×2, 3×3, 4×4, etc.). In this example, the generating class has 8 elements, so a 3×3 grid is used. The second

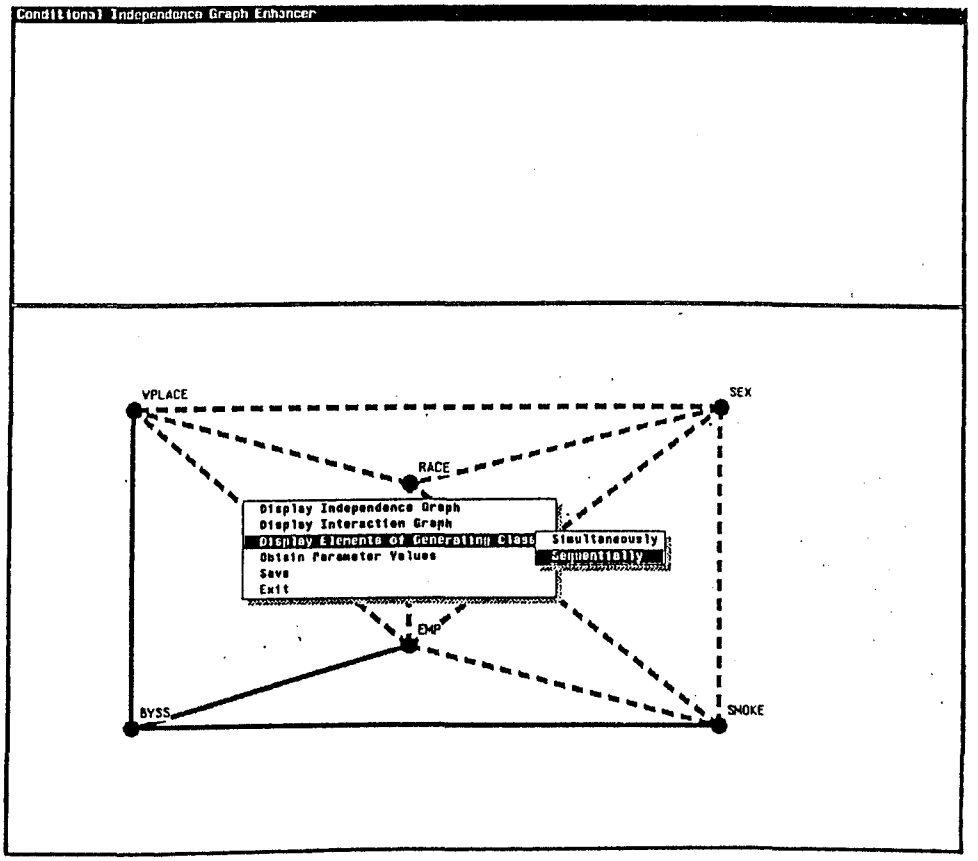

Figure 15-30: DCIGE: Appearance of pull-right menu for display of elements of generating class for discrete data example

window which initially appears is, by default, the same size as this grid, but the user can change the window size in order to access the main window, although re-sizing the second window will obscure part of the grid. In each square of the grid, a sub-graph is drawn corresponding to one of the elements of the generating class of the model, with the edges encoded or a single vertex highlighted, according to the size of the interaction corresponding to the displayed element.

See Figure 15-31 for the simultaneous display of the elements of the generating class corresponding to the discrete data example. The generating class elements were determined by CIGE from the parameter values in the data file used as input.

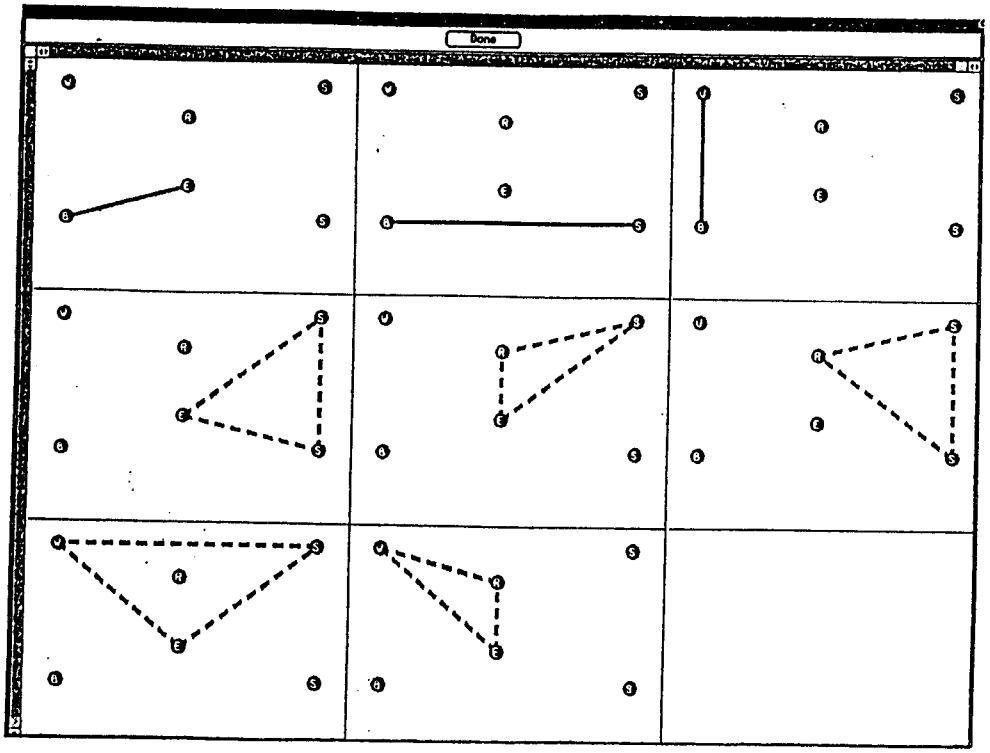

Figure 15-31: DCIGE: Simultaneous display of generating class elements for discrete data example

If the elements of the generating class are to be displayed sequentially, the first of the elements appears in a second window. Again, by default, this window is initially the same size as the screen, but if the user changes the size of the window the graph will automatically be re-drawn to scale. By clicking on the button marked '**NEXT**' each time, the user can repeatedly flip through each of the elements of the generating class in turn. A message tells the user that "**Graph *X* of *Y***" is currently being displayed, where Y is the number of elements in the generating class of the model, and X is the number of the element $(1,2,\ldots,Y)$ currently displayed. Again the edges of the sub-graphs displayed are encoded according to the size of the interaction.

Figure 15-32 shows graph 1 of 8 displayed in a re-sized second window for the discrete data example. The other 7 graphs to be displayed are as in Figure 15-31.

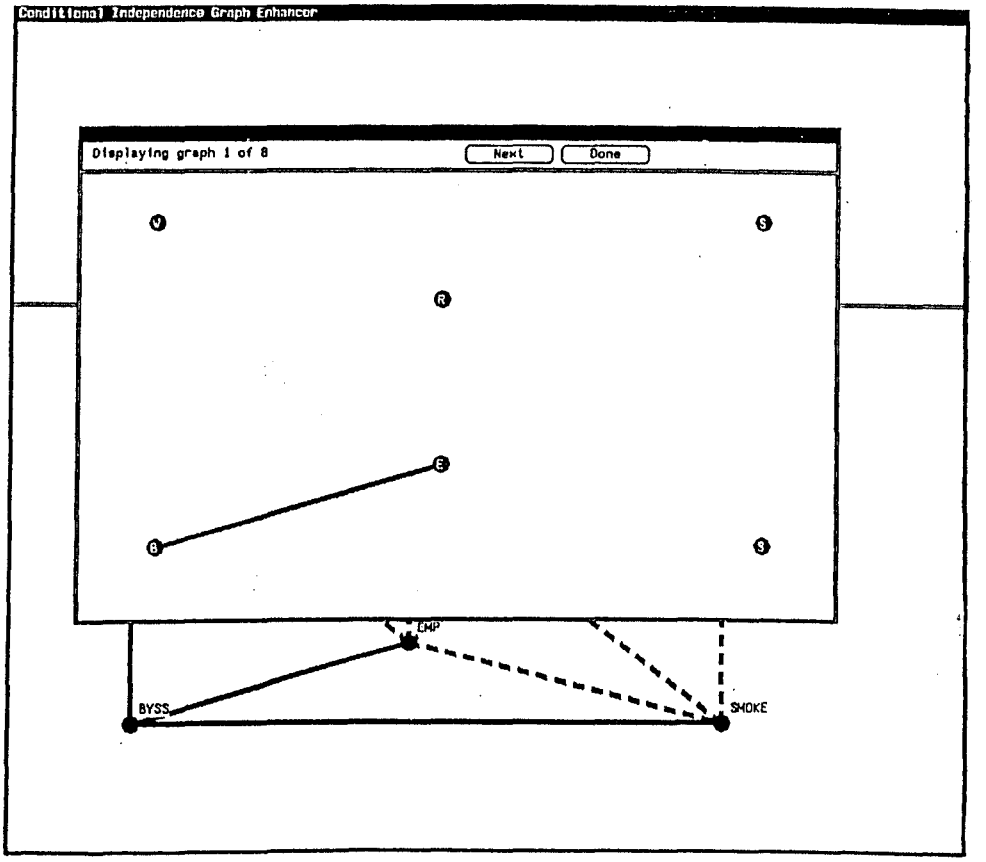

Figure 15-32: DCIGE: Graph 1 of 8 in the sequential display of generating class elements for discrete data example

Parameter Values

When this option is selected, the user can select the variable or variables which form the main effect or interaction effect of interest by clicking the left mouse button on the corresponding vertex or vertices in the graph. This acts as a toggle, such that one click will highlight the vertex (which is re-displayed with a shaded as opposed to a solid vertex), whilst a second click will cancel the selection of the vertex (which is then displayed as a solid vertex again). When all the variables in the interaction of interest have been selected, as shown in Figure 15-33, the user can click the mouse on the button marked '**SHOW**'. This will cause a second window to be opened in which the parameter values are displayed, as shown in Figure 15-34. However, if the user has selected a group of variables which form an interaction which is not contained in the generating class of

the model, then a message will be displayed in this second window saying "**No such interaction**".

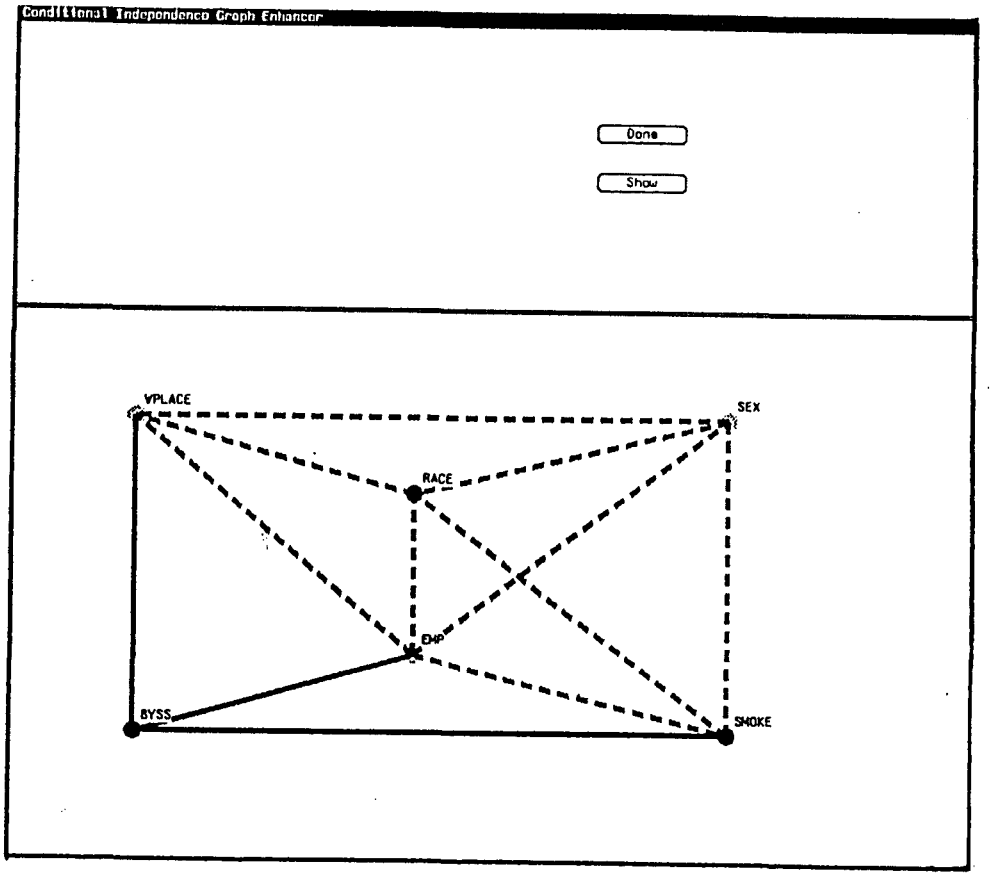

Figure 15-33: DCIGE: Highlighting of vertices corresponding to interaction of interest for discrete data example

For a main effect, the parameter values are listed in a single row; for a two-way interaction effect, the parameter values are listed in a two-way table; and for a three-way or higher order interaction effect, the parameter values are listed in a cross-referenced table as shown in Figure 15-34. The rows and columns of the tables are labelled with the vertex labels so that the parameter values can be determined for any combination of the levels of the selected variables.

The second window is initially displayed the same size as the screen, but can be re-sized to allow the user to select or deselect other variables/vertices in the graph in the main window. In re-sizing the second window, some of the parameter values may be obscured, but the user can enlarge the window if he/she wishes to inspect all of the values again. If the user selects or deselects other vertices in the main window whilst the second

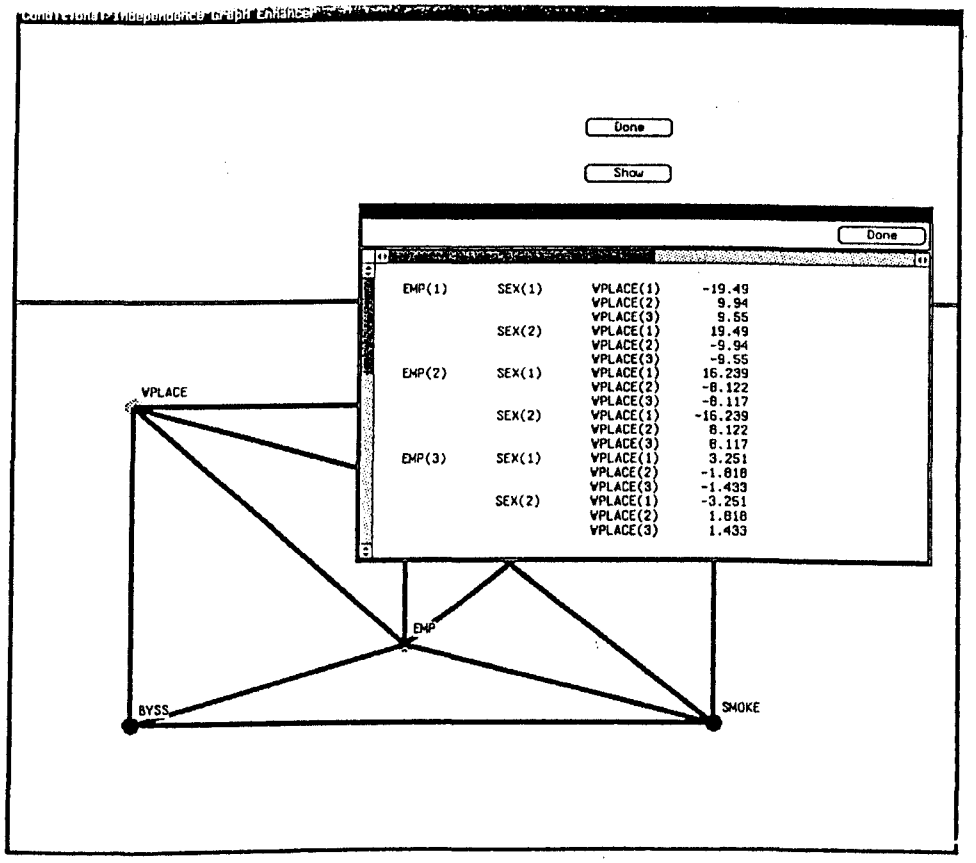

Figure 15-34: DCIGE: Parameter values corresponding to interaction of interest for discrete data example

window is open, and then clicks on the '**SHOW**' button, the display in the second window will change according to the interaction which is now highlighted.

Saving the Results

As indicated above, the independence graph or the interaction graph displayed can be saved at any time for further examination on another occasion (although a saved interaction graph will initially be re-displayed in the form of the corresponding independence graph). This is achieved by selecting this menu option and specifying the filename by which the data is to be saved. This file, which can be specified in the data entry module any time CIGE is run (see Section 15.2 above), will contain the parameter values given as input (if these were given), details of the interactions contained in the generating class of the model (for either form of data input), the locations of the vertices,

the variable labels, and other information which was specified when the data was originally entered, such as whether the data was continuous or discrete, the form of the data used as input and the number of variables. An annotated example of the saved data file for the byssinosis data example is contained in Figure 15-35 for the inputted parameter values, and in Figure 15-36 for generating class input.

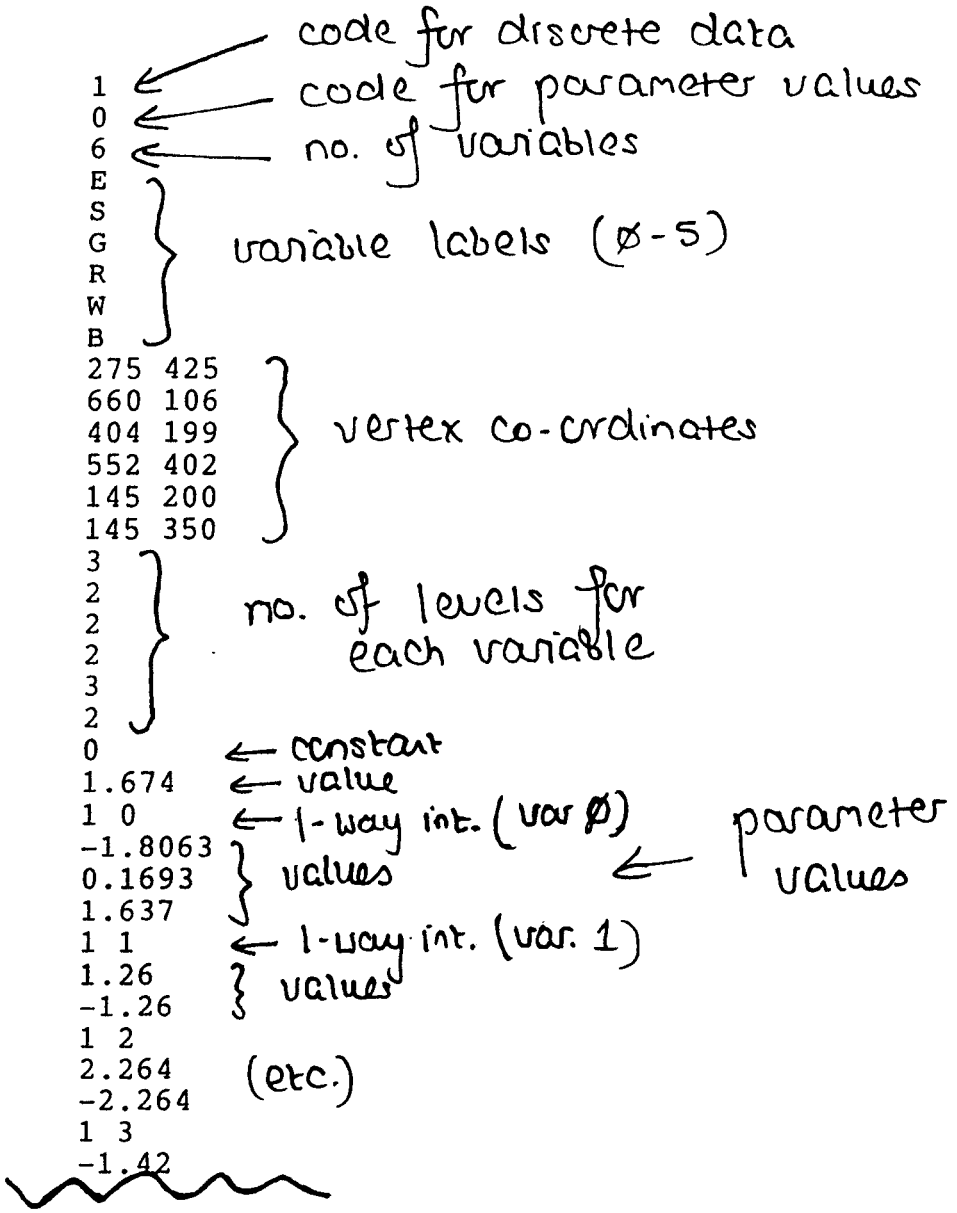

Figure 15-35: CIGE: Saved data file with annotations for example discrete data model with parameter value input

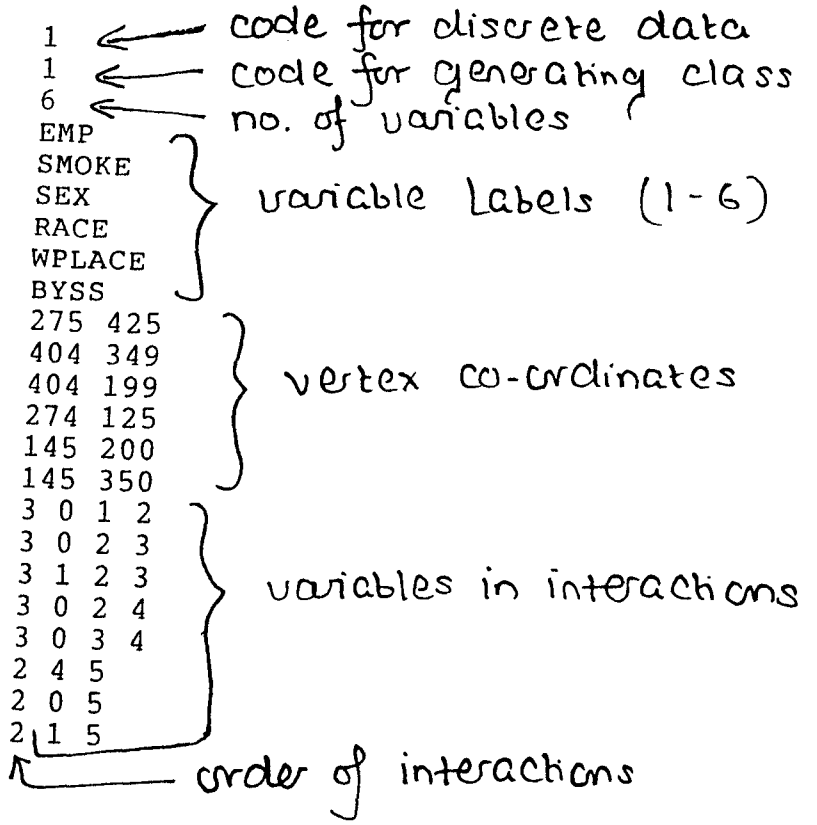

Figure 15-36: CIGE: Saved data file with annotations for example discrete data model with generating class input

Exiting from CIGE

As in the continuous data example, by selecting the menu option '**Exit**', the CIGE program is exited, after seeking confirmation of the user's intention to exit. If the graph has not been saved by explicit use of the save menu option, then all changes will be lost.

BLANK PAGE IN ORIGINAL

16. Appendix B: Graphical Perception Experiment – Summary of Results

16.1 Introduction

In this appendix, the results of all of the parametric and non-parametric hypothesis tests carried out for the experiment described in Section 10.5.1 of Chapter 10 are summarised. As mentioned in Section 10.5.1, both parametric and non-parametric tests were applied to the same data from the graphical perception experiment since the assumptions underlying the use of the parametric tests were not fulfilled for every test carried out. However, the results obtained by applying parametric and non-parametric tests to the same data were consistent for every test, suggesting that the reporting of only the parametric tests within Section 10.5.1 was sufficient.

The various parametric tests (paired t-tests for Hypotheses 1 and 3, and two-sample t-tests for Hypothesis 2) are summarised in Section 16.2, and the various non-parametric tests (Wilcoxon W tests for Hypotheses 1 and 3, and Mann-Whitney U tests for Hypothesis 2) are summarised in Section 16.3.

The main hypothesis tested was as follows:

1. There is no difference in performance according to the edge style used to encode strength of association.

The two secondary hypotheses tested were as follows:

2. Any carry-over effect affects performance between the two presentations of each graph in the same way, irrespective of the edge style which is presented first.
3. The edge style of the practice graph which was presented first and therefore explained most thoroughly does not affect performance for the two different edge styles.

The short-hand notation presented in Table 16-1 is used throughout. This short-hand is consistent with that adopted in Chapter 10.

W	Widths graphs
G	Grey-tone graphs
W1	Widths graphs presented at Time 1
W1	Widths graphs presented at Time 2
G1	Grey-tone graphs presented at Time 1
G2	Grey-tone graphs presented at Time 2
WP	Widths practice graphs
GP	Grey-tone practice graphs

Table 16-1: Short-hand notation adopted in the presentation of the results of the hypothesis tests

In the final section (16.4) the results of the attempts to fit regression lines to the data for the widths and grey-tone graphs, according to the number of vertices and the number of edges in the graph, are presented.

16.2 Parametric Tests

Hypothesis	Sample 1	Mean	S.D.	Sample 2	Mean	S.D.	t	df	p
1. Treatment effect	G1–W2	3.17	3.66	W1–G2	5.08	3.01	-1.59	11	0.14
2. Practice graph effect	GP:G1+W2	36.33	5.70	WP:G1+W2	47.00	16.70	-1.48	10	0.17
2. Practice graph effect	GP:W1+G2	40.58	7.39	WP:W1+G2	44.50	13.20	-0.64	10	0.54
3. Carry-over effect	G1+W2	41.70	13.10	W1+G2	42.50	10.40	-0.46	11	0.65
3. Period effect	G1–W2	3.17	3.66	G2–W1	-5.08	3.01	5.44	11	0.0002

16.3 Non-Parametric Tests

Hypothesis	Sample 1	Median	N	Sample 2	Median	N	W / U	p
1. Treatment effect	G1–W2	2.81	12	W1–G2	5.11	12	24	0.255
2. Practice graph effect	GP:G1+W2	34.39	6	WP:G1+W2	46.38	6	29	0.128
2. Practice graph effect	GP:W1+G2	37.99	6	WP:W1+G2	42.44	6	35	0.575
3. Carry-over effect	G1+W2	41.57	12	W1+G2	38.57	12	35	0.784
3. Period effect	G1–W2	2.81	12	G2–W1	-2.74	12	77	0.003

16.4 Regression Results

Dependent variable:

Y: Mean response latency (seconds)

Independent variables:

V: Number of vertices (3, 4, 5 or 6)

E: Number of edges (100%, 75% or 50% of $V(V-1)/2$)

V and *E* were fitted separately (simple linear regression) and together (multiple (linear) regression) for the widths graphs and for the grey-tone graphs separately for each question (1–3).

16.4.1 Question 1

Widths graphs

Regression Equation	R^2_{adj} (%)
$Y = 7.40 + 0.791V$	2.8
$Y = 12.6 - 0.005E$	0.0
$Y = 3.00 + 2.05V - 0.276E$	15.8
$Y = -1.7 + 2.72V + 0.189E - 0.064V*E$	8.0

Grey-tone graphs

Regression Equation	R^2_{adj} (%)
$Y = 6.66 + 0.705V$	6.0
$Y = 9.60 + 0.120E$	5.9
$Y = 7.72 + 0.401V + 0.067E$	0.0
$Y = 18.1 - 1.07V - 0.968E + 0.142V*E$	5.8

16.4.2 Question 2

Widths graphs

Regression Equation	R^2_{adj} (%)
$Y = -11.5 + 5.07V$	76.9
$Y = 12.7 + 0.641E$	37.9
$Y = -12.7 + 5.42V - 0.075E$	74.6
$Y = -32.4 + 8.23V + 1.90E - 0.270V*E$	79.4

Grey-tone graphs

Regression Equation	R^2_{adj} (%)
$Y = -4.16 + 3.88V$	53.1
$Y = 12.4 + 0.629E$	47.3
$Y = 0.52 + 2.54V + 0.293E$	53.3
$Y = -13.4 + 4.52V + 1.68E - 0.190V*E$	52.4

16.4.3 Question 3

Widths graphs

Regression Equation	R^2_{adj} (%)
$Y = 10.0 + 3.36V$	8.5
$Y = 25.2 + 0.485E$	3.3
$Y = 11.7 + 2.89V + 0.103E$	0.0
$Y = 30.7 + 0.17V - 1.80E + 0.261V*E$	0.0

Grey-tone graphs

Regression Equation	R^2_{adj} (%)
$Y = 10.2 + 3.44V$	1.9
$Y = 28.4 + 0.300E$	0.0
$Y = 4.0 + 5.23V - 0.391E$	0.0
$Y = 21.2 + 2.77V - 2.11E + 0.236V*E$	0.0

17. Bibliography

Aitkin M. (1980). A note on the selection of log-linear models. *Biometrics,* 36: 173–178.

Aitkin M., Anderson D., Francis B. & Hinde J. (1989). *Statistical Modelling in GLIM.* Clarendon Press, Oxford.

Allison D.B., Gorman B.S. & Primavera L.H. (1993). Some of the most common questions asked of statistical consultants: Our favourite responses and recommended readings. *Genetic, Social, and General Psychology Monographs,* 119(2): 155–185.

Altman D.G. (1991). *Practical Statistics for Medical Research.* Chapman & Hall, London.

Anderson E. (1960). A semigraphical method for the analysis of complex problems. *Technometrics,* 2(3): 387–391.

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

Andrews D.F. (1981). Dynamic displays for interactive statistical analysis. *Bulletin of the International Statistical Institute,* 43(2): 979–986.

Andrews D.F., Fowlkes E.B. & Tukey P.A. (1988). Some approaches to interactive statistical graphics. In: W.S. Cleveland & M.E. McGill (Eds.), *Dynamic Graphics for Statistics,* Wadsworth & Brooks/Cole.

Andrews D.F. & Herzberg A.M. (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker.* Springer-Verlag, New York.

Andrews H.P., Snee R.D. & Sarner M.H. (1980). Graphical display of means. *American Statistician,* 34(4): 195–199.

Anscombe F.J. (1973). Graphs in statistical analysis. *American Statistician,* 27(1): 17–21.

Anscombe F.J. & Tukey J.W. (1963). The examination and analysis of residuals. *Technometrics,* 5(2): 141–160.

Asimov D. (1985). The grand tour: A tool for viewing multidimensional data. *Siam Journal of Scientific and Statistical Computing,* 6(1): 128–143.

Asmussen S. & Edwards D. (1983). Collapsibility and response variables in contingency tables. *Biometrika,* 70(3): 567–578.

Atkinson A.C. (1985). *Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis.* Clarendon Press, Oxford.

Bachi R. (1975). Graphical methods: Achievements and challenges for the future. *Bulletin of the International Statistical Institute,* 40: 441–466.

Ball G.H. & Hall D.J. (1970). Some implications of interactive graphic computer systems for data analysis and statistics. *Technometrics,* 12(1): 17–31.

Barnett V. (1981). (Ed.) *Interpreting Multivariate Data. Proceedings of the Conference Entitled "Looking at Multivariate Data", University of Sheffield, 24–27 March 1980.* John Wiley & Sons, Chichester

Becker R.A., Chambers J.M. & Wilks A.R. (1986). Statistical software, graphics and future workstations for data analysis. In: D.M. Allen (Ed.), *Computer Science and Statistics: Proceedings of the 17th Symposium on the Interface, 1985,* pp.177–181. Elsevier Science Publishers, Amsterdam.

Becker R.A. & Cleveland W.S. (1987). Brushing scatterplots. *Technometrics,* 29(2): 127–142

Becker R.A., Cleveland W.S. & Wilks A.R. (1987). Dynamic graphics for data analysis. *Statistical Science,* 2(4): 355–395.

Becker R.A., Cleveland W.S. & Wilks A.R. (1988). Dynamic graphics for data analysis. In: W.S. Cleveland & M.E. McGill (Eds.), *Dynamic Graphics for Statistics.* Wadsworth & Brooks/Cole.

Becker R.A., Eick S.G., Miller E.O. & Wilks A.R. (1989). Dynamic graphical analysis of network data. *Statistical Research Reports, No. 78, November 1989.* AT&T Bell Laboratories, Murray Hill, New Jersey.

Becker R.A., Eick S.G. & Wilks A.R. (1990). Aspects of network visualization. *Statistical Research Reports, No. 92, April 1990.* AT&T Bell Laboratories, Murray Hill, New Jersey.

Becketti S. & Gould W. (1987). Rangefinder box plots: A note. *American Statistician,* 41(2): 149.

Benedetti J.K. & Brown M.B. (1978). Strategies for the selection of log-linear models. *Biometrics,* 34: 680–686.

Beniger J.R. & Robyn D.L. (1978). Quantitative graphics in statistics: A brief history. *American Statistician,* 32(1): 1–11.

Bertin J. (trans. Berg W.J.) (1983). *Semiology of Graphics: Diagrams, Networks, Maps.* The University of Wisconsin Press, Madison.

Bishop Y.M.M., Fienberg S.E. & Holland P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* The MIT Press, Cambridge MA.

Boardman T.J. (1977). Graphical contributions to the χ^2 -statistic for two-way contingency tables. *Communications in Statistics, A,* 6(15): 1437–1451.

Bond B.C. (1988). Personal communication.

Bradu D. & Gabriel K.R. (1978). The biplot as a diagnostic tool for models of two-way tables. *Technometrics,* 20(1): 47–68.

Broesma H.J. & Molenaar I.W. (1985). Graphical perception of distributional aspects of data. *CSQ–Computational Statistics Quarterly,* 2(1): 53–72

Buja A. & Asimov D. (1986). Grand tour methods: An outline. In: D.M. Allen (Ed.), *Computer Science and Statistics: Proceedings of the 17th Symposium on the Interface, 1985,* Elsevier Science Publishers, Amsterdam.

Buja A., Hurley C. & McDonald J.A. (1986). A data viewer for multivariate data. In: T.J. Boardman (Ed.), *Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface, March 19–21, 1986,* American Statistical Association, Washington D.C.

Boniface D.R. (1995). *Experiment Design and Statistical Methods for Behavioural and Social Research.* Chapman and Hall, London.

Calinski T. & Corsten L.C.A. (1985). Clustering means in ANOVA by simultaneous testing. *Biometrics,* 41: 39–48.

Carr D.B., Littlefield R.J. & Nicholson W.L. (1986). Scatterplot matrix techniques for large *N.* In: D.M. Allen (Ed.), *Computer Science and Statistics. Proceedings of the 17th Symposium on the Interface, 1985,* Elsevier Science Publications, Amsterdam.

Carr D.B., Littlefield R.J., Nicholson W.L. & Littlefield J.S. (1987). Scatterplot matrix techniques for large *N. Journal of the American Statistical Association,* 82(398): 424–436.

Chambers J.M., Cleveland W.S., Kleiner B. & Tukey P.A. (1983). *Graphical Methods for Data Analysis.* Wadsworth International Group; Belmont CA

Chatfield C. & Collins A.J. (1980). *Introduction to Multivariate Analysis.* Chapman and Hall, London

Chatterjee S. & Price B. (1991). *Regression Analysis by Example (2nd Edition).* John Wiley & Sons, New York.

Chernoff H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association,* 68(342): 361–368.

Chernoff H. & Rizvi M.H. (1975) Effect on classification error of random permutations of features in representing multivariate data by faces. *Journal of the American Statistical Association,* 70(351): 548–554.

Child D. (1970) *The Essentials of Factor Analysis.* Holt, Rinehart and Winston, London.

Clarkson D.B. & Gentle J.E. (1986). Methods for multidimensional scaling. In: D.M. Allen (Ed.), *Computer Science and Statistics. Proceedings of the 17th Symposium on the Interface, 1985,* Elsevier Science Publications, Amsterdam.

Clayton D. & Hills M. (1987). A two-period crossover trial. In: D.J. Hand & B.S. Everitt (Eds.), *The Statistical Consultant in Action*, pp.42–57. Cambridge University Press, Cambridge.

Cleveland W.S. (1984a). Graphs in scientific publications. *American Statistician*, 38(4): 261–269.

Cleveland W.S. (1984b). Graphical methods for data presentation: Full-scale breaks, dot charts, and multibased logging. *American Statistician*, 38(4): 270–280.

Cleveland W.S. (1987). Research in statistical graphics. *Journal of the American Statistical Association*, 82(398): 419–423.

Cleveland W.S., Diaconis P. & McGill R. (1982). Variables on scatterplots look more highly correlated when the scales are increased. *Science,* 216: 1138–1141.

Cleveland W.S., Harris C.S. & McGill R. (1983). Experiments on quantitative judgements of graphs and maps. *The Bell System Technical Journal,* 62(6): 1659–1674.

Cleveland W.S. & Kleiner D. (1975). A graphical technique for enhancing scatterplots with moving statistics. *Technometrics,* 17(4): 447–454.

Cleveland W.S. & McGill M.E. (1988). (Eds.) *Dynamic Graphics for Statistics,* Wadsworth & Brooks/Cole, Pacific Grove CA.

Cleveland W.S. & McGill R. (1983). A color-caused optical illusion on a statistical graph. *American Statistician,* 37(2): 101–105.

Cleveland W.S. & McGill R. (1984a). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387): 531–554.

Cleveland W.S. & McGill R. (1984b). The many faces of a scatterplot. *Journal of the American Statistical Association,* 79(388): 807–822.

Cleveland W.S. & McGill R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, 229: 828–833.

Cleveland W.S. & McGill R. (1987). Graphical perception: The visual decoding of quantitative information on graphical displays of data. *Journal of the Royal Statistical Society, A,* 150(3): 192–229 (incl. Discussion).

Cohen A. (1980). On the graphical display of the significant components in two-way contingency tables. *Communications in Statistics – Theoretical Methods,* A9(10): 1025–1041.

Cooper J.C.B. (1983). Factor analysis: An overview. *American Statistician,* 37(2): 141–147.

Costigan-Eaves P. & MacDonald-Ross M. (1990). William Playfair (1759–1823). *Statistical Science,* 5(3): 318–326.

Cottee M.J. (1987). The development of a fast algorithm for applying Gabriel's simultaneous test procedure. *M.Sc. thesis*, Institute of Psychiatry, University of London.

Cottee M.J. (1990). Computer assisted interpretation of conditional independence graphs. In: K. Momirovic & V. Mildner (Eds.), *COMPSTAT 1990: Proceedings in Computational Statistics*, pp.87–92. Physica-Verlag, Heidelberg.

Cottee M.J. & Hand D.J. (1989). Edge coding in conditional independence graphs. *Technical Report STAT–89–1*, Department of Statistics, The Open University, Milton Keynes.

Cox D.R. (1978). Some remarks on the role in statistics of graphical methods. *Applied Statistics*, 27(1): 4–9.

Cox D.R. & Lauh E. (1967). A note on the graphical analysis of multidimensional contingency tables. *Technometrics*, 9(3): 481–488.

Crowder M.J. & Hand D.J. (1990). *Analysis of Repeated Measures*. Chapman & Hall, London.

Dallal G. & Finseth K. (1977). Double dual histograms. *American Statistician*, 31(1): 39–41.

Daly F., Hand D.J., Jones M.C., Lunn A.D. & McConway K.J. (1995). *Elements of Statistics.* The Open University/Addison-Wesley Publishing Company, Wokingham.

Daniel C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics,* 1(4): 311–341.

Darroch J.N., Lauritzen S.L. & Speed T.P. (1980). Markov fields and log-linear interaction models for contingency tables. *Annals of Statistics*, 8(3): 522–539.

Dawid A.P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society, B*, 41(1): 1–31.

Dempster A.P. (1972). Covariance selection. *Biometrics*, 28: 157–175.

Denby L. & Pregibon D. (1987). An example of the use of graphics in regression. *American Statistician*, 41(1): 33–38.

Digby P.G.N. & Kempton R.A. (1987). *Multivariate Analysis of Ecological Communities*. Chapman & Hall, London.

Donoho A.W., Donoho D.L. & Gasko M. (1986). MacSpin: Graphical data analysis. In: T.J. Boardman (Ed.), *Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface, March 19–21, 1986,* pp.59–62. American Statistical Association, Washington D.C.

DuToit S.H.C., Steyn A.G.W. & Stumpf R.H. (1986). *Graphical Exploratory Data Analysis*. Springer-Verlag, New York.

Edwards A.W.F. (1989). Venn diagrams for many sets. *New Scientist*, 1646: 51–56.

Edwards D. (1990). Hierarchical interaction models. *Journal of the Royal Statistical Society, B*, 52(1): 3–20, (Discussion: 51–72).

Edwards D. (1995). *Introduction to Graphical Modelling.* Springer-Verlag, New York.

Edwards D. & Havranek T. (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrics,* 72(2): 339–351.

Edwards D. & Kreiner S. (1983). The analysis of contingency tables by graphical models. *Biometrika*, 70(3): 553–565.

Erdos P. & Guy R.K. (1973). Crossing number problems. *American Mathematical Monthly*, 80: 52–58.

Everitt B.S. (1978). *Graphical Techniques for Multivariate Data.* North-Holland, New York.

Everitt B.S. (1992). *The Analysis of Contingency Tables (2nd Ed.).* Chapman & Hall, London.

Everitt B. (1993). *Cluster Analysis (3rd Ed.).* Edward Arnold, London.

Everitt B.S. & Dunn G. (1991). *Applied Multivariate Data Analysis.* Edward Arnold, London.

Everitt B.S. & Nicholls P. (1975). Visual techniques for representing multivariate data. *The Statistician*, 24(1): 37–49.

Fary I. (1948). On straight line representation of planar graphs. *Acta Sci. Math. (Szeged)*, 11: 229–233.

Fienberg S.E. (1969). Preliminary graphical analysis and quasi-independence for two-way contingency tables. *Applied Statistics*, 18: 153–168.

Fienberg S.E. (1979). Graphical methods in statistics. *American Statistician*, 33(4): 165–178.

Flury B. & Riedwyl H. (1981). Graphical representation of multivariate data by means of asymmetrical faces. *Journal of the American Statistical Association*, 76(376): 757–765.

Freeman D.H. Jr. (1983). Graphical representation of multiway contingency tables: Alternative measures of association. *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 544–548

Friedman J.H., McDonald J.A. & Stuetzle W. (1982). Real time graphical techniques for analyzing multivariate data. In: *Proceedings of the Statistical Computing Section, American Statistical Association*, pp.49–54

Gabriel K.R. (1971). The biplot graphic display of matrices with application to principal components analysis. *Biometrika*, 58(3): 453–467.

Gabriel K.R. (1978). A simple method of multiple comparisons of means. *Journal of the American Statistical Association*, 73(364): 724–729.

Gabriel K.R. (1981). Biplot display of multivariate matrices for inspection of data and diagnosis. In: V. Barnett (Ed.), *Interpreting Multivariate Data*, pp.147–173. John Wiley & Sons, Chichester.

Gabriel K.R., Basu A., Odoroff C.L. & Therneau T.M. (1986). Interactive color graphic display of data by 3-D biplots. In: T.J. Boardman (Ed.), *Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface, March 19–21, 1986*, American Statistical Association, Washington D.C.

Gan F.F., Koehler K.J. & Thompson J.C. (1991). Probability plots and distribution curves for assessing the fit of probability models. *American Statistician*, 45(1): 14–21.

Gentleman J.F. (1977). It's all in a plot (Using interactive computer graphics in teaching statistics). *American Statistician*, 31(4): 166–175.

Gerson M. (1975). The techniques and uses of probability plotting. *The Statistician*, 24(4): 235–257.

Gibbons A. (1985). *Algorithmic Graph Theory*. Cambridge University Press, Cambridge.

Gnanadesikan R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. John Wiley & Sons, New York.

Gnanadesikan R. (1981). Statistical graphics: Capabilities, criteria and calibration. *Bulletin of the International Statistical Institute,* 43(2): 959–978.

Gnanadesikan R. & Wilk M.B. (1970). A probability plotting procedure for general analysis of variance. *Journal of the Royal Statistical Society, B,* 32(1): 88–101.

Goodchild N.A. & Vijayan K. (1974). Significance tests in plots of multi-dimensional data in two dimensions. *Biometrics,* 30: 209–210.

Gower J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika,* 53(3&4): 325–338.

Gower J.C. (1967). Multivariate analysis and multidimensional geometry. *The Statistician,* 17(1): 13–28.

Gower J.C. (1985). Multivariate analysis: Ordination, multidimensional scaling and allied topics. In: Leiderman & Williams (Eds.), *Encyclopaedia of Applicable Mathematics; Volume 6: Statistics,* pp.727–781. John Wiley & Sons.

Gower J.C. (1990). Three-dimensional biplots. *Biometrika,* 77(4): 773–785.

Gower J.C. & Digby P.G.N. (1981). Expressing complex relationships in two dimensions. In: V. Barnett (Ed.), *Interpreting Multivariate Data,* pp.83–118. Wiley.

Gower J.C. & Hand D.J. (1995). *Biplots.* Chapman & Hall, London.

Gower J.C. & Harding S.A. (1988). Nonlinear biplots. *Biometrika,* 75(3): 445–455.

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

Greenacre M.J. (1993). *Correspondence Analysis in Practice.* Academic Press, London.

Greenacre M. & Hastie T. (1987). The Geometric interpretation of correspondence analysis. *Journal of the American Statistical Association,* 82(398): 437–447.

Greenacre M.J. & Vrba E.S. (1984). Graphical display and interpretation of antelope census data in African wildlife areas, using correspondence analysis. *Ecology,* 65(3): 984–997.

Guy R.K. (1971). Latest results on crossing numbers. In: M. Capobianco, J.B. Frechen & M. Krolik (Eds.), *Recent Trends in Graph Theory: Proceedings of the First New York City Graph Theory Conference, 1970,* pp.143–156. Springer-Verlag: Lecture Notes in Mathematics, Volume 186.

Guy R.K. (1972). Crossing numbers of graphs. In: Y. Alavi, D.R. Lick & A.T. White (Eds.), *Graph Theory and Applications: Proceedings of the Conference at Western Michigan University,* pp.111–124. Springer-Verlag: Lecture Notes in Mathematics, Volume 303.

Guy R.K. (1988). Personal communication.

Hand D.J. (1979). Psychiatric examples of Simpson's Paradox. *British Journal of Psychiatry,* 135: 90–91.

Hand D.J. & Taylor C.C. (1987). *Multivariate Analysis of Variance and Repeated Measures: A Practical Approach for Behavioural Scientists.* Chapman & Hall, London.

Harary F. (1969). *Graph Theory.* Addison-Wesley.

Harary F. & Hill A. (1962/63). On the number of crossings in a complete graph. *Proceedings of the Edinburgh Mathematical Society,* 13(2): 333–338.

Hashimoto A. & Noshita K. (1971). A property of *N*-graphs. *IEEE Transactions on Computers,* C20(1): 95–97.

Haslett J., Bradley R., Craig P., Unwin A. & Wills G. (1991). Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *American Statistician,* 45(3): 234–242.

Havranek T. (1984). A procedure for model search in multidimensional contingency tables. *Biometrics,* 40: 95–100.

Healy M.J.R. (1968). Multivariate normal plotting. *Applied Statistics,* 17: 157–161.

Higgins J.E. & Koch G.G. (1977). Variable selection and generalized chi-square analysis of categorical data applied to a large cross-sectional occupational health survey. *International Statistical Review,* 45: 51–62.

Higgs N.T. (1991). Practical and innovative uses of correspondence analysis. *The Statistician,* 40: 183–194.

Hochberg Y., Weiss G. & Hart S. (1982). On graphical procedures for multiple comparisons. *Journal of the American Statistical Association,* 77(380): 767–772.

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

Howell D.C. (1992). *Statistical Methods for Psychology (3rd Ed.).* Duxbury Press, Belmont CA.

Huber P.J. (1987). Experiences with three-dimensional scatterplots. *Journal of the American Statistical Association,* 82(398): 448–453.

Huff D. (1973). *How to Lie with Statistics.* Penguin Books, Harmondsworth.

Jeffers J.N.R. (1967). Two case studies in the application of principal components analysis. *Applied Statistics,* 16: 225–236.

Jensen H.F. (1971). An upper bound for the rectilinear crossing number of the complete graph. *Journal of Combinatorial Theory*, 11: 212–216.

Jobson J.D. (1991). *Applied Multivariate Data Analysis Volume 1: Regression and Experimental Design.* Springer-Verlag, New York.

Jobson J.D. (1992). *Applied Multivariate Data Analysis Volume II: Categorical and Multivariate Methods.* Springer-Verlag, New York.

Joint Committee on Standards for Graphic Presentation (1915). Preliminary report published for the purpose of inviting suggestions for the benefit of the committee. *Journal of the American Statistical Association,* 14: 790–797.

Jolliffe I.T. (1986). *Principal Component Analysis.* Springer-Verlag, New York.

Jones B. & Kenward M.G. (1989). *Design and Analysis of Cross-Over Trials.* Chapman & Hall, London.

Jones S. (1988). Graphical interfaces for knowledge engineering: An overview of relevant literature. *The Knowledge Engineering Review,* 3(3): 221–247.

Kafadar K. & Spiegelman C.H. (1986). An alternative to ordinary q-q plots: Conditional q-q plots. *Computational Statistics & Data Analysis,* 4: 167–184.

Kainen P.C. (1974). Some recent results in topological graph theory. In: R.A. Bari & F. Harary (Eds.), *Graphs and Combinatorics: Proceedings of the Capital Conference on Graph Theory and Combinatorics*, pp.76–108. Springer-Verlag: Lecture Notes in Mathematics, 406.

Karger J.C. (1982). An application of logit analysis to the effective graphical representations of two-way tables of ratios. *The Statistician*, 31(2): 143–148

Kastenbaum M.A. (1974). Analysis of categorical data: Some well-known analogues and some new concepts. *Communications in Statistics,* 3: 401–417.

Kelly J.D. (1993). The effects of display format and data density on time spent reading statistics in text, tables and graphs. *Journalism Quarterly,* 70(1): 140–149.

Kernighan B.W. & Ritchie D.M. (1988). *The "C" Programming Language (2nd Ed.).* Prentice Hall, Englewood Cliffs NJ.

Knuiman M. (1978). Covariance Selection. *Supplement Advances in Applied Probability*, 10: 123–130.

Kokoska S. & Nevison C. (1992). *Statistical Tables and Formulae.* Springer-Verlag, New York.

Kolata G. (1984). The proper display of data. *Science,* 226: 156–157.

Kosslyn S.M. (1985). Graphics and human information processing: A review of five books. *Journal of the American Statistical Association,* 80(391): 499–512.

Kosslyn S.M. (1994). *Elements of Graph Design.* W.H. Freeman & Co., New York.

Kruskal W.H. (1982). Criteria for judging statistical graphics. *Utilitas Mathematica,* 21(B): 283–310.

Krzanowski W.J. (1988). *Principles of Multivariate Analysis: A User's Perspective.* Clarendon Press, Oxford.

Landwehr J.M., Pregibon D. & Shoemaker A.C. (1984). Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association,* 79(385): 61–83 (incl. comments)

Landwehr J.M. & Watkins A.E. (1985). Stem-and-leaf plots. *Mathematics Teacher,* 78: 528–531.

Lauritzen S.L. (1979). Lectures on contingency tables. *Technical Report,* Institute of Mathematical Statistics, University of Copenhagen.

Lauritzen S.L. (1989). Mixed graphical association models. *Scandinavian Journal of Statistics,* 16(4): 273–306.

Lauritzen S.L., Speed T.P. & Vijayan K. (1984). Decomposable graphs and hypergraphs. *Journal of the Australian Mathematics Society, Series A,* 36: 12–29.

Lauritzen S.L. & Spiegelhalter D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, B,* 50(2): 157–224 (incl. discussion).

Lauritzen S.L. & Wermuth N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics,* 17(1): 31–57.

Lauro N.C. & Decarli A. (1982). Correspondence analysis and log-linear models in multiway contingency tables study. Some remarks on experimental data. *Metron,* 40: 213–234.

Lee W. (1966). Experimental design symbolisation and model derivation. *Psychometrika,* 31(3): 397–412.

Lewandowsky S. & Spence I. (1989a). Discriminating strata in scatterplots. *Journal of the American Statistical Association,* 84(407): 682–688.

Lewandowsky S. & Spence I. (1989b). The perception of statistical graphs. *Sociological Methods and Research,* 18(2&3): 200–242.

Long J.S. (1987). A graphical method for the interpretation of multinomial logit analysis. *Sociological Methods and Research,* 15(4): 420–446.

MacDonald A.D. (1982). A stem-leaf plot: An approach to statistics. *Mathematics Teacher,* 75(1): 25,27–28.

Mage D.T. (1982). An objective graphical method for testing normal distributional assumptions using probability plots. *American Statistician,* 36(2): 116–120.

Mahon B.H. (1977). Statistics and decisions: The importance of communication and the power of graphical presentation. *Journal of the Royal Statistical Society, A,* 140(3): 298–323.

Mallows C.L. (1973). Some comments on Cp. *Technometrics,* 15: 661–675.

Manly B.F.J. (1994). *Multivariate Statistical Methods: A Primer (2nd Ed.).* Chapman & Hall, London.

Mardia K.V., Kent J.T. & Bibby J.M. (1979). *Multivariate Analysis.* Academic Press, London.

McGill R., Tukey J.W. & Larsen W.A. (1978). Variations of box plots. *American Statistician,* 32(1): 12–16.

Mezrich J.J., Frysinger S. & Slivjanovski R. (1984). Dynamic representation of multivariate time series data. *Journal of the American Statistical Association,* 79(385): 34–40.

Monlezun C.J. (1979). Two-dimensional plots for interpreting interactions in the three-factor analysis of variance model. *American Statistician,* 33(2): 63–69.

Morrison D.F. (1976). *Multivariate Statistical Methods (2nd Ed.).* McGraw-Hill Book Co., New York.

Myers J.L. (1979). *Fundamentals of Experimental Design (3rd Ed.).* Allyn & Bacon Inc., Boston.

Nicholson T.A.J. (1968). Permutation procedure for minimising the number of crossings in a network. *Proceedings of the IEEE,* 115(1): 21–26.

Paik M. (1985). A graphic representation of a three-way contingency table: Simpson's paradox and correlation. *American Statistician,* 39(1): 53–54.

Pittenger D.J. (1995). Teaching students about graphs. *Teaching of Psychology,* 22(2): 125–128.

Read R.C. (1970). Graph theory algorithms. In: B. Harris (Ed.), *Graph Theory and its Applications,* pp.51–78. Academic Press, New York.

Read R.C. (1979). Algorithms in Graph Theory. In: R.J. Wilson & L.W. Beineke (Eds.), *Applications of Graph Theory*, pp.381–418. Academic Press, London.

Reichmann W.J. (1964). *Use and Abuse of Statistics.* Penguin Books, Harmondsworth.

Saaty T.L. (1964). The minimum number of intersections in complete graphs. *Proceedings of the National Academy of Sciences*, 52: 688–690.

Schildt H. (1987). *"C": The Complete Reference.* Osborne McGraw-Hill, Berkeley CA.

Schmid C.F. (1983). *Statistical Graphics: Design Principles and Practices.* Wiley-Interscience, New York.

Schweder T. & Spjøtvoll E. (1982). Plots of P-values to evaluate many tasks simultaneously. *Biometrika,* 69(3): 493–502.

Scott A.J. & Knott M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics,* 30: 507–512.

Scott D.W. (1979). On optimal and data-based histograms. *Biometrika,* 66(3): 605–610.

Scott D.W. (1985). Frequency polygons: Theory and application. *Journal of the American Statistical Association,* 80(390): 348–354.

Seber G.A.F. (1984). *Multivariate Observations.* John Wiley & Sons, New York.

Simkin D. & Hastie R. (1987). An information-processing analysis of graph perception. *Journal of the American Statistical Association,* 82(398): 454–465.

Simpson E.H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, B,* 13(2): 238–241.

Snee R.D. (1974). Graphical display of two-way contingency tables. *American Statistician,* 28(1): 9–12.

Sparks D.N. (1970). Half normal plotting. Algorithm AS 30. *Applied Statistics,* 19: 192–196.

Speed T.P. (1978). Relations between models for spatial data, contingency tables, and Markov fields on graphs. *Supplement Advances in Applied Probability,* 10: 111–122.

Spjøtvoll E. (1977). Alternatives to plotting Cp in multiple regression. *Biometrika,* 64(1): 1–8.

Stuetzle W. (1987). Plot windows. *Journal of the American Statistical Association,* 82(398): 466–475.

Tasaki T., Yoden A. & Goto M. (1987). Graphical data analysis in comparative experimental studies. *Computational Statistics and Data Analysis,* 5: 113–125.

Taylor Jr W.H. & Hilton H.G. (1981). A structure diagram symbolization for analysis of variance. *American Statistician,* 35(2): 85–93.

Tenenhaus M. & Young F.W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika,* 50(1): 91–119.

Tufte E.R. (1983). *The Visual Display of Quantitative Information.* Graphics Press, Cheshire CT.

Tufte E.R. (1990). *Envisioning Information.* Graphics Press, Cheshire CT.

Tukey J.W. (1972). Some graphic and semigraphic displays. In: T.A. Bancroft (Ed.), *Statistical Papers in Honor of George W. Snedecor,* pp.292–316. Iowa State University Press, Ames, Iowa.

Tukey J.W. (1977). *Exploratory Data Analysis.* Addison-Wesley Publishing Company, Reading MA.

Tutte W.T. (1963). How to draw a graph. *Proceedings of the London Mathematical Society,* 3(13): 743–768.

Upton G.J.G. (1978). *The Analysis of Cross-Tabulated Data.* John Wiley & Sons, Chichester.

Upton G.J.G. (1986). Cross-classified data. In: A.D. Lovie (Ed.), *New Developments in Statistics for Psychology & the Social Sciences.* The British Psychological Society & Methuen, London.

Upton G.J.G. (1991). The exploratory analysis of survey data using log-linear models. *The Statistician,* 40: 169–182.

Van der Heijden P.G.M., de Falguerolles A. & de Leeuw J. (1989). A combined approach to contingency table analysis using correspondence analysis and log-linear analysis. *Applied Statistics,* 38(2): 249–292. (incl. Discussion).

Van der Heijden P.G.M. & de Leeuw J. (1985). Correspondence analysis used complementary to loglinear analysis. *Psychometrika,* 50(4): 429–447.

Wainer H. (1984). How to display data badly. *American Statistician,* 38(2): 137–147.

Wainer H. (1990). Graphical visions from William Playfair to John Tukey. *Statistical Science,* 5(3): 340–346.

Wainer H. & Francolini C.M. (1980). An empirical inquiry concerning human understanding of two-variable color maps. *American Statistician,* 34(2): 81–93.

Weihs C. & Schmidli H. (1990). OMEGA (Online Multivariate Exploratory Graphical Analysis): Routine searching for structure. *Statistical Science,* 5(2): 175–226.

Weinberg L. (1972). Planar graphs and matroids. In: Y. Alavi, D.R. Lick & A.T. White (Eds.), *Graph Theory and Applications: Proceedings of the Conference at Western Michigan University, 1972*, pp.313–329. Springer-Verlag: Lecture Notes in Mathematics, Volume 303.

Wermuth N. (1976a). Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics*, 32: 95–108.

Wermuth N. (1976b). Model search among multiplicative models. *Biometrics*, 32: 253–263.

Wermuth N. (1985). Data analysis and conditional independence structures. *Bulletin of the International Statistical Institute*, 51(24/2): 1–13.

Wermuth N. (1988). Introduction to the use of graphical chain models. In: *COMPSTAT 88: Tutorial Notes.*

Wermuth N. & Lauritzen S.L. (1983). Graphical and recursive models for contingency tables. *Biometrika*, 70(3): 537–552.

Wermuth N. & Lauritzen S.L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models. *Journal of the Royal Statistical Society, B*, 52(1): 21–50 (Discussion: 51–72).

Wermuth N., Wehner T. & Gönner H. (1976). Finding condensed descriptions for multi-dimensional data. *Computer Programs in Biomedicine*, 6: 23–38.

Whittaker J. (1982). GLIM syntax and simultaneous tests for graphical log linear models. In: R. Gilchrist (Ed.), *GLIM 82: Proceedings of the International Conference on Generalised Linear Models. Lecture Notes in Statistics*, 14: 98–108. Springer-Verlag, Berlin.

Whittaker J. (1984). Fitting all possible decomposable and graphical models to multiway contingency tables. In: T. Havranek, Z. Sidak & M. Novak (Eds.), *Compstat 1984: Proceedings in Computational Statistics, 6th Symposium, Prague, 1984.* pp. 401–406. Physics-Verlag, Vienna.

Whittaker J. (1988). ESRC workshop: "Graphical modelling: Transparencies". *Technical Report*, Department of Statistics, University of Lancaster.

Whittaker J. (1990). *Graphical Models in Applied Multivariate Statistics.* John Wiley & Sons, Chichester.

Whittaker J. & Aitkin M. (1978). A flexible strategy for fitting complex log-linear models. *Biometrics*, 34: 487–495.

Whittaker J., Iliakopoulos A. & Smith P.W.F. (1988). Graphical modelling with large numbers of variables: An application of principal components. In: D. Edwards & N.E. Raun (Eds.), *COMPSTAT 1988: Proceedings in Computational Statistics*, pp.73–79. Physica-Verlag, Heidelberg.

Wilk M.B. & Gnanadesikan R. (1964). Graphical methods for internal comparisons in multiresponse experiments. *Annals of Mathematical Statistics,* 35: 613–631.

Wilk M.B. & Gnanadesikan R. (1968). Probability plotting methods for the analysis of data. *Biometrika,* 55(1): 1–17.

Wilk M.B., Gnanadesikan R., & Huyett M.J. (1962). Probability plots for the gamma distribution. *Technometrics,* 4(1): 1–20.

Wilkinson G.N. & Rogers C.E. (1973). Symbolic description of factorial models for analysis of variance. *Applied Statistics*, 22(3): 392–399.

Wilson R.J. (1985). *Introduction to Graph Theory (3rd Ed.).* Longman Scientific and Technical, Harlow, England.

Winer B.J., Brown D.R. & Michels K.M. (1991). *Statistical Principles in Experimental Design (3rd Ed.).* McGraw-Hill Inc., New York.

Worrall S. (1991). Anyone for virtual tennis? *The Sunday Times Magazine*, May 26, pp.20–26.

Young F.W. & Rheingans P. (1991). Visualizing structure in high-dimensional multivariate data. *IBM Journal of Research and Development,* 35(1/2): 97–107.

Zahn D.A. (1975a). Modifications of and revised critical values for the half-normal plot. *Technometrics,* 17(2): 189–200.

Zahn D.A. (1975b). An empirical study of the half-normal plot. *Technometrics,* 17(2): 201–211.