17

Graphical modelling of archaeological data

A. Scott
J. Whittaker
M. Green
(Centre for Applied Statistics, University of Lancaster)
S. Hillson
(Institute of Archaeology, 31–34 Gordon Square, London WC1H OPY)

17.1 Introduction

The analysis of archaeological datasets presents statistics with a unique and taxing combination of difficulties. They are inherently multivariate, with large numbers of measured variables of both continuous and discrete type and in addition often suffer from high proportions of missing values. From such complex datasets the analyst, whether statistician or archaeologist, has to determine which of the many variables interact and how strongly, decide if the data can be condensed without loss of information and select one (or perhaps a few) parsimonious models or forms of analysis that adequately represent the data. Standard analyses for continuous data summarised by a variance or correlation matrix, such as principle components, have become well established. As have, to a lesser degree, log-linear methods for discrete data summarised in a contingency table. Until recently, however, multivariate techniques had not been generalised to mixed continuous and discrete data, nor to dealing with the problem of large scale missing data.

The principle difficulty with such data, and perversely often the primary objective of the analysis, is the assessment of association. In osteometrical studies, for example, attempts to characterise the shape of complex objects such as skulls has resulted in the definition of a multitude of standard measurements (Brothwell 1981, von den Driesch 1976). By their very nature many of these are strongly correlated, so that their value in distinguishing between sexes, species and populations is not always clear. Disentangling their interrelationships, complicated as they are by the presence of both continuous and discrete variables, is difficult, if not impossible, with standard multivariate methods. In contrast a new statistical technique known as graphical modelling can be used to summarise the patterns of interaction in such studies in an easily interpretable way.

In the remainder of this paper a brief description is given of the most important aspects of graphical modelling methods, illustrated by their application to a small archaeological data set. A more detailed exposition and bibliography can be found in Whittaker (1990). There still remain a number of practical problems which need to be overcome before the new technique is applicable to the bulk of archaeological data; in particular the difficulties caused by large numbers of variables and of missing values. The extent to which these problems have been resolved is discussed.

17.2 Graphical modelling

Graphical modelling, has a theoretical basis in the concept of conditional independence, and is so called because it summarises the patterns of interaction between variables by means of a graph. The technique originated in the work of Darroch, Lauritzen and Speed (1980) who showed how a subset of log-linear models, the graphical models, can be easily interpreted, theoretically and practically, from their associated independence graph. The approach of Lauritzen and Wermuth (1984,1989) extends the technique, allowing graphical models to be applied in a unified way not only to the separate continuous and discrete cases but also to mixed models. The theory of graphical modelling has two essential ingredients. Firstly a parametric family of distributions which can be used to model the joint distribution of discrete and continuous variables and whose parameters can be interpreted as measures of association or interaction. Secondly the concept of a conditional independence graph which is used to represent the most important qualitative aspects of the fitted model.

17.2.1 The conditional gaussian distribution

Many of the commonest parametric statistical techniques, e.g. correlation, regression and log-linear modelling, are based on models derived using either the multinomial distribution for discrete variables or the normal distribution for continuous variables. The conditional Gaussian, or CG, distribution is obtained by combining these distributions and is therefore a natural generalisation when dealing with data containing both types of variable. The discrete variables are regarded as forming a multidimensional contingency table whose cell probabilities are described by a multinomial distribution. Within each cell of this table, i.e. conditional on the values of the discrete variables, the continuous variables are described by a multivariate normal distribution whose mean and covariance matrix are allowed to vary from cell to cell. Thus the general CG distribution for continuous variables x and discrete variables y can be written as

$$f(x,y) = p(y)(2\pi)^{d/2}|\Sigma(y)|^{-1/2}\exp\{-\frac{1}{2}(x-\mu(y))'\Sigma(y)^{-1}(x-\mu(y))\} \quad (17.1)$$

where $p(y)$ is the probability that the discrete variables have value y, $\mu(y)$ and $\Sigma(y)$ are the within-cell mean and variance of the continuous variables, d is the number of continuous variables, and $'$ is used to denote vector or

matrix transposition. This representation makes explicit the derivation of the CG distribution as the product of a multinomial and a normal distribution.

Because of its derivation from these two standard distributions, the most natural way to parameterise the CG distribution is in terms of the triplet of parameters $\{p, \mu, \Sigma\}$. The main advantage of this natural or moment parameterisation is that it is couched in terms of familiar concepts, proportions, means and variances, and thus results are easy to interpret. Restricted forms of the distribution, although not explicitly described as CG, have long been used in statistics for techniques such as discriminant analysis and analysis of variance.

The moment parameterisation is not the only way of parameterising the distribution. Lauritzen and Wermuth (1984), who appear to have introduced the name CG, write the distribution as

$$f(x, y) = \exp\{\alpha(y) + \beta(y)'x - 1/2x'\Omega(y)x\} \quad (17.2)$$

using the triplet of parameters $\{\alpha, \beta, \Omega\}$, known as the canonical parameters. The advantage of this canonical parameterisation is its direct interpretation in terms of conditional independence. If any two variables do not occur together in any term of expression (17.2), then they are conditionally independent given the remaining variables (Lauritzen & Wermuth 1989), a property which forms the basis of the new technique. Edwards (1990) replaces each of the parameters $\{\alpha, \beta, \Omega\}$ by a hierarchical linear expansion in terms of the discrete variables y. Graphical modelling then involves setting groups of these expanded parameters to zero and testing if this is reasonable. Each imposed constraint can be interpreted as the conditional independence of two variables given the remainder. Thus the qualitative interpretation of graphical modelling is straightforward and easily understandable by non-statisticians. The technique can be regarded as examining associations between variables to see whether they can be explained by the remaining variables, a common investigative objective in many fields. In quantitative terms, however, interpretation is not always so easy. With variables of mixed type it is far from clear what the estimated values of the canonical parameters mean (apart from those set to zero), or what implications they have in terms of the more standard moment parameterisation.

17.2.2 Conditional independence graphs

As described so far graphical modelling might be regarded as merely an extension of log-linear modelling to include continuous variables. In fact it is the introduction of conditional independence graphs as a means of model representation that gives the techniques both its name and its wide applicability. Mathematically a graph consists of two components, a set of nodes or vertices representing variables and a set of edges connecting nodes and representing association. More precisely the absence of an edge between two variables implies that the variables are conditionally independent given the remainder. Fig. 17.1a shows a possible conditional independence graph for a model with four variables x, y, z and w.

Of the six possible edges in this graph three, representing different conditional independencies, are missing. Depending on the variable types these may be interpretable in terms of standard statistics. Thus for continuous variables the lack of an edge connecting x and z implies that the partial correlation coefficient of x and z given y and w is zero. Different symbols are commonly used for nodes of different types, circles for continuous variables and filled in circles or dots for discrete variables.

The most important tool for interpreting such independence graphs is the separation property. Two variables or sets of variables in a graph are said to be separated by a third set if every path between the two sets passes through the third. The separation property states that sets of variables in a graph are conditionally independent given any separating set. Thus, in Fig. 17.1a, x and z are independent given either y and w together, or y alone. Use of a graph does more however than just represent the set of conditional independencies in a model. In the present case it

- compactly represents the complete pattern of association among the variables,
- highlights y as the one crucial variable in analysing the interrelationships of the data, since x, z and w are all independent given y alone, and
- gives the set of best predictors of each variable. So that regression with x, z or w as dependent variable would require only y as an independent variable, whereas y dependent necessitates all of the other variables in the regression equation.

The usefulness of the independence graph is most apparent for more complex data sets where it is invaluable in disentangling relationships, often suggesting meaningful clusters or chains of variables and highlighting variables which are directly associated.

A further extension of graphical modelling theory is to the concept of a chain graph, where edges connecting variables may be arrows (Wermuth & Lauritzen 1990). This allows the incorporation, in the same diagram, of independence statements with different conditioning sets. The variables in such chain graphs can be divided into an ordered sequence of blocks, edges joining variables in different blocks being arrows while edges joining variables within the same block are lines. The direction of the arrows is restricted to be the same as the ordering of the blocks. Missing edges in chain graphs are interpreted as conditional independences given the other variables in the blocks containing and preceding the two unconnected variables. Thus in Fig. 17.1, b the variables form two blocks, one containing x and z, the other y and w. The variables x and z are marginally independent whereas the pairs x and w or z and w are conditionally independent given y.

The use of graphical chain models greatly extends the power and versatility of the technique. It also brings further interpretational advantages. Recent work on graphical modelling at Lancaster (Scott 1990) has concentrated on two different aspects, the problem of missing data, and methods of fitting a particular form of two block model in which all the discrete variables are in the first block and all the continuous variables in the second. The advantage of this form of model is that the corresponding subset of CG

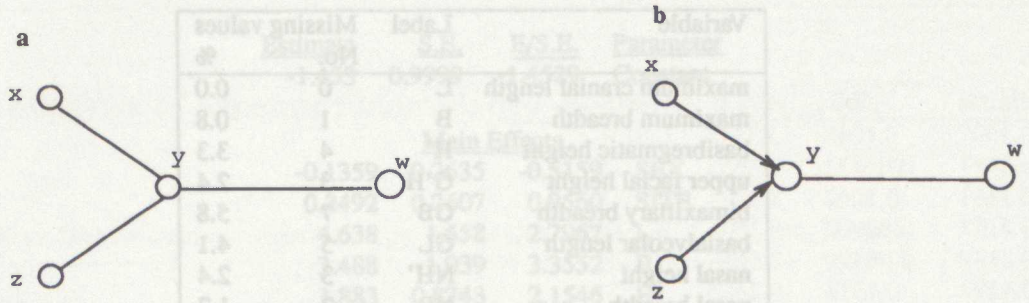

Figure 17.1: Graph examples

distributions can be represented and fitted in terms closer to the moment parameterisation. Thus parameter and model interpretation is particularly easy for this class of models.

17.2.3 Software

At present the theory of graphical modelling has far outstripped its practical implementation. The use of graphical chain models in their most general formulation promises to be a powerful data analysis tool, but efficient methods of fitting such models have still to be devised and current software is limited to restricted subsets of these more general models. The program MIM (Edwards 1987) was designed specifically to fit hierarchical interaction models (Edwards 1990), of which undirected graphical models are a subset. Unfortunately it is somewhat limited in scope, lacking, for example, any data transformation capability and not providing estimates of parameter standard errors. Packages such as Glim and Genstat can be used to fit limited forms of directed model through their regression facilities and the program EXA (Kreiner 1987) has been developed to fit chain models but for discrete data only. None of these allow for missing values. Software developed at Lancaster (the program ASP) can be used to fit the two stage models described in the previous section, although primarily intended for more traditional forms of analysis.

17.3 Example

The use of graphical modelling techniques is illustrated here by their application to a small archaeological data set. The data comprise thirteen measurements on a total of 121 human skulls from two Egyptian cemeteries: Badari, a predynastic site from Upper Egypt, and Sedment, a IXth Dynasty site from Middle Egypt. The data were originally published by Stoessiger (1927) and Woo (1930) and the variables used here are a selection from over seventy recorded for each collection. Only skulls from adult individuals have been included. Apart from sex differences the data show little evidence of grouping within collections. They therefore make useful homogeneous collections for experiments with statistical methods. All the skulls had been sexed for the original publications, although 27 males and 21 females were marked as questionable. Further examination by one of the authors for a previous study (Hillson 1978,1985) confirmed the majority of the original decisions but reclassified five of the questionable female skulls from Sedment as male. For the purposes of this paper, however, these 48 skulls are treated as being of unknown sex.

Table 17.1 lists the variables used, together with the number of values missing for each. For the continuous variables this ranges from 0% to almost 20% with an average of 5%. Even if sex were always known, over 30% of the cases would be incomplete. When sex is included this proportion rises to almost 60%. Apart from the missing information, however, this can be regarded statistically as a 'nice' dataset with no apparent peculiarities.

Table 17.2 shows the correlation matrix of the fifteen variables with the non-significant values underlined. Not unexpectedly with data of this type, the majority of the values are strongly significant. There is some indication that a number of the continuous variables do not vary from site to site but apart from this it is difficult to know exactly how to interpret the table.

As with other modelling procedures, the application of graphical modelling involves searching for parsimonious models; in this case by dropping or adding edges to the current graph. Edge type, whether directed or undirected, is not alterable by the search procedure but is fixed *a priori* by the initial specification of a block structure for the variables. Thus model comparison is always within a given structure. Sometimes this structure is determined by knowledge of the physical situation giving rise to the data. More often it is a matter of choice, and the wide variety of available models means that many models may fit equally well. For the present data the two block model described in section 17.2.2 was prefered for ease of parameter interpretation.

Tables 17.3 and 17.4 give the fitted values for the final model, separated into interaction and 'other' parameter sets. The entries in these tables can be interpreted in the same way as standard regression output. In particular the ratio of each parameter estimate to its standard error assesses significance and is used in the modelling process to decide which edges to drop from the current model. Fig. 17.2 shows the corresponding graph. Although at first sight this appears somewhat confusing, it actually conveys a considerable amount of information about the data structure. This is clarified by Fig. 17.3 which is obtained from Fig. 17.2 by omitting the discrete variables, sex and site, and all lines connected to them. Thus Fig. 17.3 shows the association graph of the continuous variables within each combination of sex and site. It is made up of four separate groups of variables, the variables comprising each of these being independent of variables in other groups. Furthermore each group can be associated with a particular aspect of skull morphology. The first is largely made up of variables measuring overall

Variable	Label	Missing values	
		No.	%
maximum cranial length	L	0	0.0
maximum breadth	B	1	0.8
basibregmatic height	H'	4	3.3
upper facial height	G'H	3	2.4
bimaxillary breadth	GB	7	5.8
basialveolar length	GL	5	4.1
nasal height	NH'	3	2.4
nasal breadth	NB	2	1.7
bidacryonic chord	DC	16	13.2
palatal length	G'1	13	10.7
palatal breadth	G2	22	18.2
orbital breadth	O'1	11	9.1
orbital height	O2	0	0.0
sex	SEX	48	39.7
site	SITE	0	0.0

Table 17.1: Variables used in the study

	Sex	Site	L	B	H'	G'H	GB	GL	NH'	NB	DC	G'1	G2	O'1	O2
Sex	1														
Site	.2	1													
L	.7	-.2	1												
B	.5	.5	.3	1											
H'	.7	.3	.5	.5	1										
G'H	.5	.4	.4	.4	.5	1									
GB	.6	-.1	.5	.2	.5	.4	1								
GL	.5	-.1	.5	.1	.3	.3	.6	1							
NH'	.5	.5	.4	.5	.5	.7	.4	.3	1						
NB	.4	-.1	.4	.0	.2	.2	.5	.4	.2	1					
DC	.4	-.3	.4	.1	.2	.2	.4	.4	.1	.4	1				
G'1	.4	-.1	.4	.1	.3	.4	.5	.8	.3	.3	.3	1			
G2	.4	.4	.2	.3	.4	.3	.4	.4	.4	.2	.2	.4	1		
O'1	.5	.1	.6	.4	.5	.5	.5	.4	.5	.4	.0	.3	.3	1	
O2	.4	.3	.2	.4	.4	.4	.3	.0	.6	.1	-.1	.0	.2	.5	1

Table 17.2: Variable correlation matrix

skull size, the second of variables measuring facial height, and the last of variables describing the skull base.

The main points of interest shown by the full graph can therefore be summarised as follows

1. Since sex and site are not directly connected, the relative proportions of the two sexes are the same for both sites.
2. The continuous measurements fall naturally into groups representing different aspects of skull size and shape, each group being independent of the others.
3. All of these groups show site differences and sexual dimorphism although this is mediated through different variables.

17.4 Concluding remarks

The importance of graphical models to both archaeology and statistics is that they provide a means of analysis in situations, such as those involving mixed data types, which were previously intractable. The simple graphical characterisation of such models is visually appealing and highly informative: the graph highlights those variables that are directly associated, may suggest meaningful clusters or chains of variables and identifies the set of best predictors of any particular variable. They also complement, rather than replace, more standard techniques, acting as an exploratory tool which can reduce complex data sets to manageable proportions and indicating further suitable forms of analysis.

The two main problems with applying the technique to archaeological data are the occurance of missing values and the large numbers of variables involved. The first of these difficulties has now been solved for some forms of model (Scott 1990), as shown by the analysis presented here. The second problem is mainly one of graph presentation and interpretation, which can be difficult for large numbers of variables. Some progress has been made in this area (Whittaker *et al.* 1988) and work is still continuing.

Acknowledgement

This work was performed under grant GR/E 94951 from the SERC

Estimate	S.E.	E/S.E.	Parameter
-1.453	0.9999	-1.4529	Constant
	Main Effects		
-0.1359	0.2635	-0.5159	SEX
0.2492	0.2607	0.9560	SITE
4.638	1.658	2.7967	L
3.488	1.039	3.3552	B
1.883	0.8743	2.1546	H'
1.069	0.7318	1.4610	G'H
1.331	0.8061	1.6521	GB
3.165	1.089	2.9058	GL
3.441	1.279	2.6906	NH'
6.224	2.086	2.9827	NB
5.906	1.146	5.1544	DC
0.5956	1.252	0.4755	G'1
6.591	2.173	3.0336	G2
1.350	1.928	0.7003	O'1
4.877	1.127	4.3238	O2
	Inverse Variance Parameters		
0.067261	0.011000	6.1145	L
0.051827	0.010402	4.9822	B
0.048252	0.007894	6.1124	H'
0.075222	0.004687	16.0484	G'H
0.073159	0.013764	5.3151	GB
0.120159	0.019700	6.0993	GL
0.257611	0.045574	5.6525	NH'
0.467055	0.083100	5.6203	NB
0.260354	0.049048	5.3080	DC
0.257775	0.042549	6.0582	G'1
0.249182	0.052339	4.7608	G2
0.279585	0.039095	7.1513	O'1
0.305647	0.041708	7.3281	O2

Table 17.3: Fitted parameters — main effects and inverse variances

Bibliography

BROTHWELL, D. R. 1981. *Digging up bones*. Oxford University Press, Oxford, third edition.

DARROCH, J., S. LAURITZEN, & T. SPEED 1980. "Markov fields and log-linear interaction models for contingency tables", *Ann. Stat.*, 8: 522–539.

EDWARDS, D. E. 1987. *A guide to MIM*. Research Report 87/1. Statistical Research Unit, University of Copenhagen.

EDWARDS, D. E. 1990. "Hierarchical interaction models (with discussion)", *Journal of the Royal Statistical Society*, 52: 3–20.

HILLSON, S. W. 1978. *Human biological variation in the Nile valley*. PhD thesis, London University.

HILLSON, S. W. 1985. "Missing information and collections of skeletal material", *in* Fieller, N. R. J. *et al*, (eds.), *Palaeoenvironmental Investigations*. British Archaeological Reports, Oxford.

KREINER, S. 1987. "Analysis of multidimensional contingency tables by exact conditional tests: techniques and strategies", *Scandinavian Journal of Statistics*, 14: 97–112.

LAURITZEN, S. L. & N. WERMUTH 1984. *Mixed interaction models*. Research report R-84-8. Inst. of Electr. Systems, Aalborg University.

LAURITZEN, S. L. & N. WERMUTH 1989. "Graphical models for associations between variables, some of which are qualitative and some quantitative", *Ann. Statist.*, 17: 31–57.

SCOTT, W. A. 1990. *Statistical modelling of incomplete data in archaeology*. PhD thesis.

STOESSIGER, B. N. 1927. "A study of the Badarian crania recently excavated by the British School of Archaeology in Egypt", *Biometrika*, 22: 65–83.

VON DEN DRIESCH, A. 1976. *A guide to the measurement of animal bones from archaeological sites*. Peabody Museum Bulletin, No 1. Harvard University Press, Cambridge, Mass.

WERMUTH, N. & S. L. LAURITZEN 1990. "On substantive research hypotheses, conditional independence graphs and graphical chain models", *Journal of the Royal Statististical Society*, 52: 21–72.

WHITTAKER, J. 1990. *Graphical models in applied multivariate statistics*. Wiley, Chichester.

WHITTAKER, J., A. ILIAKOPOULOUS, & P. W. F. SMITH 1988. "Graphical modelling with large numbers of variables: an application of principal components", *in* Edwards, D. & Raun, N. E., (eds.), *Proceedings in Computational Statistics*. Physica-Verlag, Heidelberg.

WOO, T. L. 1930. "A study of seventy-one 9th Dynasty skulls from Sedment", *Biometrika*, 22: 65–83.

Estimate	S.E.	E/S.E.	Parameter
-0.3507	0.1048	-3.3448	L.SEX
-0.3377	0.09471	-3.5653	L.SITE
0.3363	0.1042	3.2244	B.SITE
0.2197	0.06900	3.1844	H'.SITE
-0.2178	0.08490	-2.5658	GL.SEX
-0.4299	0.1316	-3.2666	NH'.SEX
0.4495	0.1181	3.8065	NHT.SITE(2)
-0.3855	0.1970	-1.9565	NB.SEX
-0.4971	0.2368	-2.0995	DC.SEX
-0.3563	0.1539	-2.3152	DC.SITE
-0.3270	0.1843	-1.7741	G2.SEX
0.3771	0.1369	2.7543	G2.SITE
-0.01814	0.005503	-3.2966	B.L
-0.02479	0.006384	-3.8827	H'.L
-0.03017	0.01010	-2.9862	GL.GB
-0.08319	0.01738	-4.7861	NH'.G'H
-0.05717	0.02006	-2.8495	NB.GB
-0.1189	0.02301	-5.1662	G'1.GL
-0.03136	0.01841	-1.7034	G2.GB
-0.05253	0.01392	-3.7730	O'1.L
-0.1034	0.02828	-3.6573	O2.NH'

Iterations	Likelihood	D.F.	Deviance	D.F.
38	-3813.177	1630	104.786	84

Table 17.4: Fitted parameters — interactions

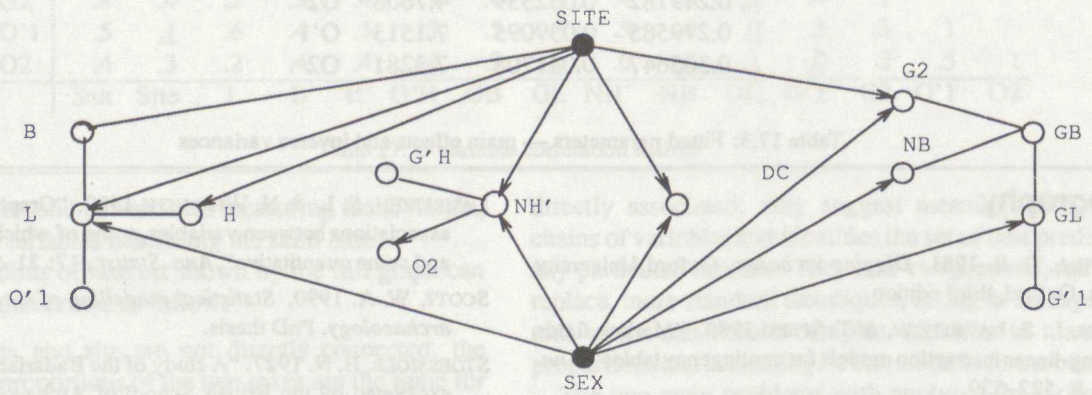

Figure 17.2: Graph of fitted model for skull data

Figure 17.3: Conditional graph of continuous variables