IMA Journal of Mathematics Applied in Business & Industry (1998) **9**, 241–266

Graphical models in credit scoring

PETE SEWART AND JOE WHITTAKER

Department of Mathematics and Statistics, Lancaster University,
Lancaster LA1 4YF, UK

Email: P.Sewart@lancaster.ac.uk, Joe.Whittaker@lancaster.ac.uk

Graphical models simplify the analysis of multivariate observations by summarizing conditional independences in the data. Variables are represented by nodes, and the absence of an edge between two nodes signifies their conditional independence. While graphical modelling has been used in several applications of statistics, credit scoring has only recently been suggested as a suitable candidate.

This paper suggests the following potential uses for graphical models: to display and interpret the associations between variables taken from a credit-card application form; to compare the credit scoring of subpopulations; to give a description of the credit-scoring selection process in terms of influence diagrams; and to assess the effect of selection bias and stratification on the interdependency of variables.

These methods are discussed in relation to the analysis of a subset of variables from a stratified sample of credit-card applicants. The large number of variables measured in an application form requires the statistical analysis of large sparse contingency tables. It is shown here that tractable graphical models can be extracted from fitting the relatively simple all-two-way interaction model.

1. Introduction

Graphical models simplify the analysis of multivariate data: associations between the variable comprising the data are summarized by a graph consisting of nodes and edges. Variables are represented by nodes, and the absence of edges between any two nodes signifies a conditional independence between the corresponding variables. While other techniques designed to analyse multivariate data, such as principle-component analysis (e.g. Krzanowski 1988), reduce dimension by combining the original variables into a (linear) score, with graphical models the original variables remain distinct, and it is the relationships between the variables that are examined.

The idea of representing log-linear models by independence graphs was originally introduced by Darroch *et al.* (1980). Whittaker (1990) explains in detail the properties of graphical models and includes numerous case studies. Edwards (1995) provides another recent introduction. The area of graphical models is rapidly expanding, with applications in many branches of statistics including regression analysis, log-linear models, time series, and probabilistic expert systems.

Statistical applications in credit scoring have mainly been concerned with the discrimination of 'good' and 'bad' credit risks. These include developing new methods (e.g. Boyle *et al.* 1992), finding ways of comparing these different methods (for a review, see Wilkie 1992), or improving or adapting current systems—for example, by reject inference (Hand & Henley 1993). Fitting logistic regression models is the standard way of building a good discriminant, or credit scorecard.

Log-linear models are a closely related technique used for the analysis of multiway contingency tables (e.g. Agresti 1990).

More recently the field of credit scoring has been suggested as an interesting application area for graphical models. The standard logistic regression model typically models the conditional distribution of Y given by $X = (X_1 ,..., X_k)$, where in the credit-scoring context, Y is behavioural performance such as the Good/Bad indicator, and X denotes the variables extracted from the application form. We contend, however, that the distribution of X is of interest (i) in its own right, and (ii) because it determines the statistical properties of the fitted regression for Y. A graphical model for X allows this distribution to be visually summarized, allows subpopulations to be compared, and relates independence and dependence to orthogonality and collinearity. An advantage of using a graphical model to describe the process is that it allows easy imputation of any missing responses in the X variables, which are a common occurrence in application forms and credit-bureau information.

Another interesting application of graphical models is to model the complete credit-scoring process, including the outcome variables of interest such as behavioural performance and revenue. Marshall & Oliver (1995) describe influence diagrams, which are directed independence graphs incorporating decision nodes and provide the necessary framework for a credit-scoring model. A similar avenue is pursued by Hand *et al.* (1997) who use a mixture of directed and undirected graphs to model the credit-scoring process. We do not consider such an exercise in this paper and instead utilize graphical models as a data-exploring tool to improve understanding of the credit-card applicant population.

The first purpose of this paper is, through illustration, to indicate the importance of conditional independence in credit scoring; to show how conditional independence relates to graphical models; and to distinguish between the use of directed graphs, undirected graphs, and influence diagrams.

The second purpose is to explain how undirected graphs may be used to model the joint distribution of the application-form variables. To illustrate the method, the analysis concentrates on modelling a subset of variables taken from the application form. There is much that can be gained from examining even a relatively small subset of variables. The undirected graphical model of these variables allows an improved understanding of the interrelationships between the application characteristics. This is useful for analysts keen to learn about the social, demographic, and economic behaviour of the population of interest. An insight into the differences between distinct populations is gained if they are modelled separately and their graphs compared. This highlights the extent of any differences, which in turn may suggest whether separate scorecards are likely to be necessary for each subpopulation.

Section 2 describes the properties of graphical models with definitions of conditional independence, and gives examples of Simpson's paradox, which emphasizes the importance of modelling the joint distribution. In Section 3 a discussion of the selection processes that occur in credit scoring is explained in terms of an influence diagram and conditional independence. The effects on measuring interaction are noted, and a similar discussion is given for the effect of using stratified samples. The use of the all-way log-linear model is described in Section 4, along with an explanation of the edge-exclusion deviance which is used to select acceptable models.

Section 5 shows that attempts to apply the log-linear all-way interaction model prove to be impractical for modelling a large number of variables. This problem is partially solved by constraining interactions of order higher than two to vanish, which still retains the conditional independence structure. Finally, in Section 6, our modelling approaches are illustrated on a stratified sample of credit-card applications to the credit-card division of a major bank. It is still necessary to reduce the dimension of the data, by decomposition, in order to construct the graphical model. To illustrate the methods, three graphical models are produced for the subset of application variables: the first constructed using all individuals in the sample, the second constructed on the subpopulation of 'young' applicants and the third on the separate subpopulation of 'old' applicants. Inferences are drawn from the first graph, and the other two graphs are briefly compared.

2. Conditional independence and graphical models

Graphical models are concerned with summarizing association and dependence between the variables in a multivariate data set. It is useful here to define independent and conditionally independent events.

Independence

Events A and B are independent, written $A \perp\!\!\!\perp B$, if and only if $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$, where $\mathrm{P}(\bullet)$ is the probability. An equivalent formulation is $\mathrm{P}(A \mid B) = \mathrm{P}(A)$, which states that the outcome of event B has no influence on the outcome of event A.

In applications, it is necessary to work in terms of the probability density or mass functions and use random variables or vectors rather than events. The random vectors X and Y are independent if and only if the joint density function $p_{XY}(\bullet,\bullet)$ satisfies $p_{XY}(x,y) = p_X(x)p_Y(y)$ for all x and y. It is straightforward to represent two random variables as graphs, as Fig. 1 illustrates.

Directed independent graphs (Bayes nets and influence diagrams) represent the dependence of Y on Y as opposed to the association between X and Y. Interest lies in the distribution p_X and the conditional distribution $p_{Y|X}$, rather than their joint distribution p_{XY}. This is shown in Fig. 2.

X
Y
X and Y dependent
X
Y
X and Y independent

FIG. 1. Graphical models representing dependent and independent events.

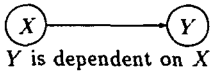

FIG. 2. An example of a directed graph.

Conditional independence

The definition of independent events can be rewritten, by replacing the unconditional probabilities with conditional probabilities, to give the definition of conditional independence. Events A and B are independent, conditional on C, written $A \perp\!\!\!\perp B \mid C$, if and only if $\mathrm{P}(A \cap B \mid C) = \mathrm{P}(A \mid C)\mathrm{P}(B \mid C)$. Expressed in terms of probability density functions, the random vectors X and Z are independent, conditional on the random vector Y, if and only if $p_{XZ|Y}(x, z \mid y) = p_{X|Y}(x \mid y)p_{Z|Y}(z \mid y)$ for all x, y, and z. With three random variables, the conditional independence is signified by the absence of a connecting edge, as in Fig. 3.

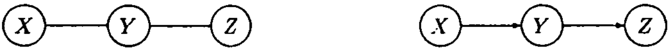

FIG. 3. Representing the conditional independence $X \perp\!\!\!\perp Z \mid Y$ in an undirected and in a directed graph.

More generally, in an arbitrary undirected graph, two subsets of variables are independent conditional upon any *separating* subset of variables. To predict the outcome of a variable, it is only necessary to use those variables that are directly joined to it with an edge. Variables that are separated from the outcome variable in the graph, in other words those that are not connected by an edge, are conditionally independent given the separating variables and hence provide no additional information for prediction.

Simpson's paradox

It is important to understand the difference between marginal and conditional independence. Examples 1 and 2 illustrate Simpson's paradox (Simpson 1951). It refers to seemingly contradictory interpretations of independence and association present simultaneously in the marginal and conditional distributions.

Example 1: marginal dependence does not imply conditional dependence. Consider a simple (though contrived) example related to sexual discrimination. A sample of 1200 bank-loan applicants are classified according to gender (male/female), loan status (rejected/accepted), and bank (A/B/C) to which they applied. Examining the marginal table between gender and loan status suggests that sexual discrimination is taking place; see Table 1. If the three bank categories refer to different branches of the same national bank, and a decision is made centrally on whether to grant loans,

TABLE 1
Loan applications by gender

	rejected	accepted	% accepted
male	250	350	58
female	350	250	42

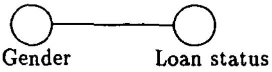

FIG. 4. The marginal dependence between gender and loan status.

TABLE 2
Loan applications by gender and bank

		rejected	accepted	% accepted
Bank A	male	75	25	25
	female	225	75	25
Bank B	male	100	100	50
	female	100	100	50
Bank C	male	75	225	75
	female	25	75	75

then this is the whole picture, and the conclusion of sexual discrimination is a valid one. The variables' gender and loan status are marginally correlated, and the graph of their distribution is shown in Fig. 4.

However, it may well be the case that the banks are competitors, or deal with different types of loan, and hence have different loan-granting policies. It is sensible then to break down the figures by bank, since the decision to grant loans is made locally, see Table 2. The truth is now that gender and loan status are independent conditional on the bank applied to. The suggestions of sexual discrimination is shown to be false, and instead it is apparent that males tend to apply to the banks with high acceptance rates and females to the banks with low acceptance rates. The reason the banks have different sex ratios applying for loans might perhaps be explained by the bank's marketing strategy, for instance, if directed towards a particular gender. Another possibility is that the banks specialize in loans designed for specific items which are traditionally gender-orientated. The acceptance rates for the three banks are likely to be predetermined by the bank's history, or the type of loan specialization.

It is evident that gender no longer provides any information about the outcome of the loan decision once it is known which bank the application was made to. The full graph, illustrating the conditional independence, is shown in Fig. 5. Collapsing the

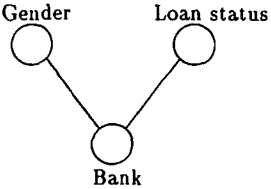

FIG. 5. Gender and loan status independent, given bank.

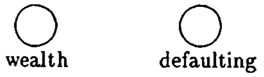

FIG. 6. Wealth independent of default status.

TABLE 3
Effect of wealth on defaulting

	defaulted		
wealth	no	yes	% default
poor	2450	60	2·4
rich	2200	54	2·4

full table over bank applied to gives the initial marginal table and induces the observed dependence between gender and loan status.

Example 2: marginal independence does not imply conditional independence. This fictional example illustrates how certain characteristics might influence the probability of defaulting on credit-card repayments. A sample of 4764 credit-card holders are classified according to wealth (poor/rich), default status (no/yes), and credit-card usage (light/heavy).

First, consider the effect of wealth on defaulting. Equal percentages of defaulters in both categories indicates that wealth provides no information about defaulting; see Table 3. The subgraph of Fig. 6, displays the marginal independence of the variables.

Breaking this table down by light and heavy credit-card users reverses the previous conclusion that wealth and defaulting are independent: they *are* dependent, conditional on credit-card usage; see Table 4. If it is known whether an individual is a light or a heavy user, then their wealth becomes an important factor in predicting whether they are likely to default or not. The graph of the three variables, shown in Fig. 7, has no edges missing.

TABLE 4
Defaulting against wealth and usage

		defaulted		
usage	wealth	no	yes	% default
light users	poor	2210	10	0·45
	rich	110	1	0·9
heavy users	poor	240	50	17·2
	rich	2090	53	2·5

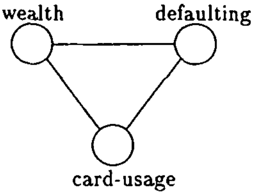

FIG. 7. Wealth and defaulting dependent in the joint distribution.

These two examples emphasize the importance of examining the whole picture. It is the joint distribution that is of interest if an understanding of how all the variables interact is desired. When only the marginal distributions are examined, interesting associations may be missed and spurious relationships may be unintentionally discovered.

3. Selection processes

The statistical analysis of a sample of data on credit-card applicants has to contend with the problem of selection. This may occur for several reasons, of which one, inbuilt into the credit-scoring process, is whereby applicants are accepted (selected) or rejected on the basis of their scorecard value. Another form of selection is the differential selection (or stratification) of cases on the basis of an observed variable, such as the Good/Bad indicator. We examine both these topics in this section.

Selection in the credit-scoring process

The graphical models discussed above relate only to random variables. An adequate description of the credit-scoring process requires the introduction of decision variables, as well as a variable to denote the resulting value of the outcome in the decision process. An influence diagram (Marshall & Oliver 1995) extends the directed graph described above to three types of node: square nodes to represent decisions, with an associated set of choices; circular nodes to represent random variables, each with an associated set of outcomes; and diamond nodes to represent results of the decision process. The direction of the arrows reflects the direction of the influence, and a missing edge portrays lack of direct influence or equivalently a conditional independence, as illustrated in Fig. 8.

There is a temporal sequencing from left to right. From the credit company's point of view, the 'Accept' node is a decision while the 'Apply' node is a random or chance variable. That there is no edge leading from Apply to Good signifies that Good $\perp\!\!\!\perp$ Apply | Score, and supposes that the Score contains all the information from the application form needed to predict the eventual status of Good. That there is no edge leading from Apply to Accept indicates that it is only the score that influences the decision. That there is no edge from Accept to Good states that the underlying probability of Good is not affected by accepting or rejecting the applicant. The only variables influencing profit are the decision Accept and whether the applicant turns out to be Good or Bad.

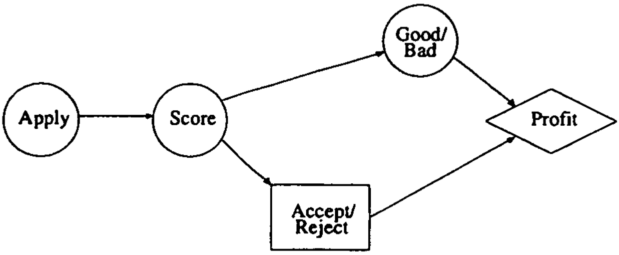

FIG. 8. An influence diagram representing the credit-scoring process.

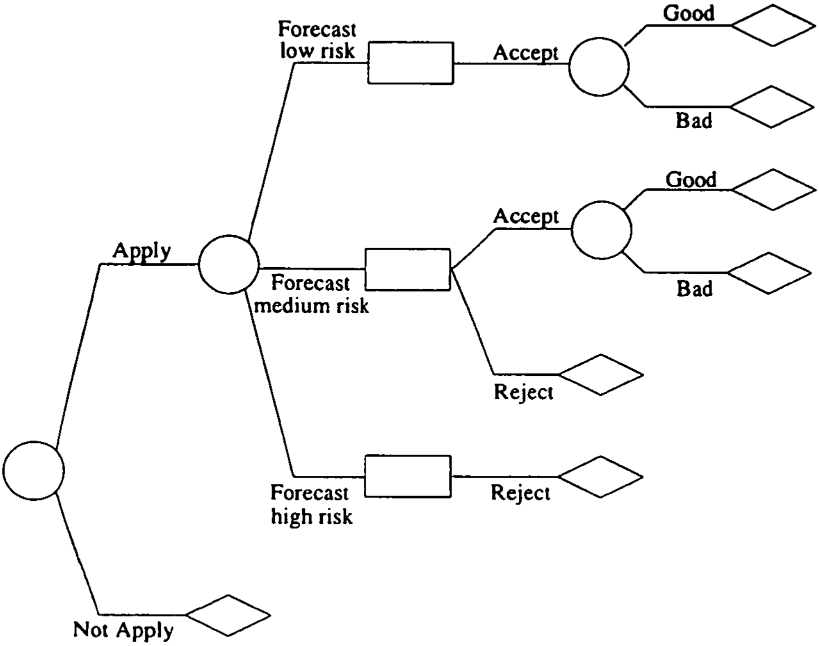

FIG. 9. The decision tree of the credit-scoring process. The profit resulting from each branch is recorded at the diamond nodes.

The influence diagram has an associated decision tree (Fig. 9) where every individual in the general population traverses along one of the branches with certain probabilities. These probabilities satisfy the independence constraints displayed in the influence diagram. It is apparent from the diagram that there are two selection processes occurring in the credit-scoring process. The accepted population is selected by the credit-card company from the through-the-door population via the credit-score forecasts of behavioural performance. This through-the-door population is itself self-selected from the general population, by the 'random' decisions of individuals to apply for a credit card. In constructing a graphical model of the application variables, as in the next section, it is important to be clear which is the population of interest. Graphical models of the through-the-door population or the population of accepted applicants are possibly of the most interest, but these populations have been conditioned on the decision variables.

The issue to consider here is the implication this inbuilt conditioning has on the dependence structure (or the independence graph) of the accepted or through-the-door populations. The structure of the graph may differ according to which sub-population is being examined, as the following example shows.

Example: how selection can influence the joint distribution. Consider a contrived example of reasons for applying for a credit card: X_1, the desire for credit (required/not required), and X_2, need for an alternative to cash (yes/no). The selection variable Y also has two levels (apply for card/do not apply). It is reasonable to suppose that the two reasons X_1 and X_2 are marginally independent in the overall population, and the graph is simple; see Fig. 10.

FIG. 10. Marginal independence of X_1 and X_2 in the overall population.

We represent the negative events (credit or alternative to cash not required, card not applied for) by 0, and the corresponding positive outcomes by 1. Let $P(X_1 = 1) = 0{\cdot}7$ and $P(X_2 = 1) = 0{\cdot}4$. By independence, $p(x_1, x_2) = p(x_1)p(x_2)$, and the marginal table of probabilities can be constructed (Table 5).

TABLE 5
$p(x_1, x_2)$ for the whole population

		x_2, cash alternative		
		not required (0)	required (1)	%
x_1	credit not required (0)	0·18	0.12	40
	credit required (1)	0.42	0.28	40

Another reasonable assumption is to suppose that the probability of the event that an individual applies or does not apply for a credit card, Y, jointly depends on X_1 and X_2. Let

$$P(Y = 1 \mid X_1 = 0, X_2 = 0) = 0.2, \qquad P(Y = 1 \mid X_1 = 0, X_2 = 1) = 0.5,$$
$$P(Y = 1 \mid X_1 = 1, X_2 = 0) = 0.7, \qquad P(Y = 1 \mid X_1 = 1, X_2 = 1) = 0.9.$$

The directed graph, shown in Fig. 11, displays this information.

However, when the population observed consists only of those individuals who applied for a credit card, we obtain an observed distribution, $p(x_1, x_2 \mid Y = 1)$, as shown in Table 6. These two observed variables are now dependent in this selected

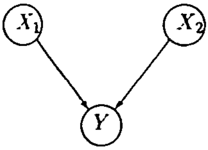

FIG. 11. The directed graph of the reasons for applying for a credit card.

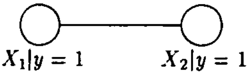

FIG. 12. Loss of independence of the two reasons in the subpopulation of credit-card applicants.

TABLE 6
$p(x_1, x_2 \mid Y = 1)$ *for the applicant population*

		x_2, cash alternative		
		not required (0)	required (1)	%
x_1	credit not required (0)	0.056	0.093	62.4
	credit required (1)	0.458	0.393	46.2

subpopulation, even though they are independent in the general population. For example, compared with 46.2% of those who require credit, of those who do not require credit, 62.4% require a cash alternative. In the general population, 40% required a cash alternative, regardless of whether they required credit or not.

The fact that X_1 and X_2 are marginally independent is no longer observable when their dependence is examined conditional on whether they applied for a credit card as shown by the graph in Fig. 12. One effect is that the independence graphs of the through-the-door or accepted subpopulations are likely to appear more complicated than graphs constructed on data from the general population.

Stratification

Although random sampling is the simplest method of sampling from a population, better estimates are obtained by stratification, where minority populations are oversampled in order to ensure a larger number of exceptional individuals in the sample. This the case in credit-scoring samples (Lewis 1992), where, although the Bads make up only a small proportion of the total population, a large proportion of the sample includes Bads in order to better estimate the regression coefficients in the score card.

The graphical model analysis of a sample of applicants requires several tests of independence and conditional independence. The test statistic—for example, Pearson's chi-squared or the likelihood-ratio test (here called the edge-exclusion deviance; see later)—not only provides a test of the hypothesis but also gives a scale to measure the relative importance of the dependence or association. Stratifica-

TABLE 7
Random sample of the population

	residential status	
creditworthiness	rent	own
Bad	195	65
Good	2964	1976

tion may well affect the values of these statistics and so influence the choice of model and the graph.

Example: how stratification can affect the observed relationships between the variables. Consider the perceived strength of association between creditworthiness and residential status under various stratification policies. Residential status has been classified into a binary variable with levels corresponding to renting and owning accommodation, and creditworthiness is classified as Good or Bad. Assume that a sample of 5200 is to be stratified into various proportions of Bad individuals. We examine the effect of stratifying a sample of 5200 individuals into various proportions of Goods and Bads.

If the sample was chosen randomly from the population, the proportion of the Bads in the sample might be as low as 5%; see Table 7.

The deviance against independence equals 24.7, and on the null hypothesis is to be compared against the chi-squared distribution on 1df. Choosing equal numbers of observations for each level is a common method of stratification. The deviance, in fact, may grow when this is the case as illustrated in Table 8. The deviance increases dramatically from the random sample case to 134.2.

Stratification can have major implications if the number of observations in each level become dissimilar. Consider the case of Bad customers making up only 0.7% of the sample, shown in Table 9. The deviance is now 3.6, which is in fact too low to be confident enough to reject the independence hypothesis. Although it appears that stratification actually increases or decreases the strength of the association, it should only affect the amount of information available to confidently reject the hypothesis of independence between the variables. When constructing the graph, the edge strengths and missing edges should be tested for sensitivity to the stratification. A

TABLE 8
Data for equally weighted stratification

	residential status	
creditworthiness	rent	own
Bad	1950	650
Good	1560	1040

TABLE 9
Data for unweighted sample with few Bads

	residential status	
creditworthiness	rent	own
Bad	27	9
Good	3098	2065

random sample is perhaps best if the relative edge strengths are to be compared sensibly. It is possible to infer back to the original sample from a stratified sample if the stratification proportions are known.

4. Modelling the joint distribution

There are various features of the credit scoring process that can be described by graphical models. We consider some of these here. Let X_a, X_b, and X_c denote three subvectors of application variables used to build the scorecard, with X_d denoting a variable subvector not used in the construction. For example, if $b = (4, 6, 8, 11)$, then $X_b = (X_4, X_6, X_8, X_{11})$. Let Y be the eventual credit performance indicator and S the Score or credit performance prediction.

The simplest situation is one which examines the joint distribution of the application variables X_a, X_b, X_c, X_d using undirected edges; see Fig. 13. Since the X variables are known for all applicants, inferences can be made about the through-the-door population. Inclusion of the credit performance indicator Y in the model only allows analysis of the accepted population, unless reject inference is successfully applied.

Including the performance indicator (see Fig. 14) requires a mixture of directed edges and undirected edges. Such a model allows interpretation of which variables directly influence credit behaviour, for example to check whether the variables X_d do not predict performance Y.

By further including the Score in this model (see Fig. 15), the main focus of interest is in checking whether credit performance is conditionally independent of

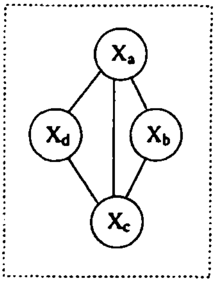

FIG. 13. Graph of the application variables.

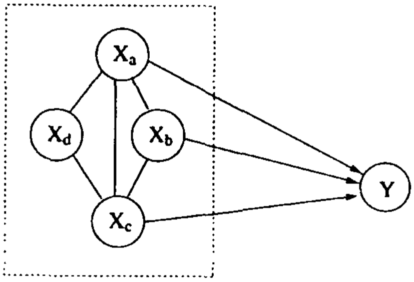

FIG. 14. Graph of the application variables and credit performance indicator.

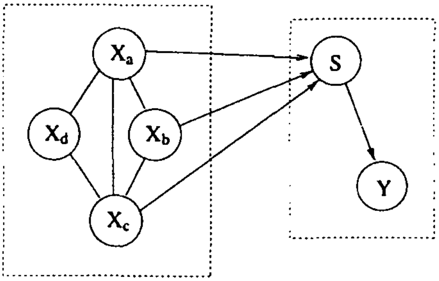

FIG. 15. Graph of the application variables, credit performance indicator, and the Score.

the application variables given the Score, or whether the X variables contain additional information which could have improved the predictions. The Score is a deterministic function of the observed application variables.

An alternative way of modelling the relationship between X and Y is by a latent-variable analysis. Essentially it is assumed that people possess an unobserved underlying characteristic that determines their credit behaviour. This variable, Z, is such that credit performance, Y, is conditionally independent of the X variables, given Z, as Fig. 16 illustrates.

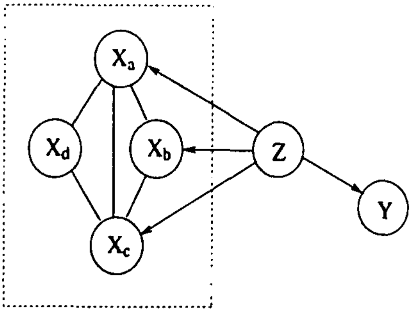

FIG. 16. Latent-variable model, showing the application variables' independence of credit performance, conditional on a latent (random) variable Z underlying credit behaviour.

Log-linear models

We concentrate on the simplest situation of building an undirected graphical model from the joint distribution of a subset of the application variables, the X's, as in Fig. 13. A systematic way of checking for conditional independences is to fit a log-linear model to the joint distribution (Agresti 1990); within the log-linear parametrization, the specification of independence graphs is relatively straightforward (Whittaker 1990).

To summarize this briefly, we note here that the log-linear all-way-interaction expansion for a three-way table is

$$\log p(x_1, x_2, x_3) = u_\varnothing + u_1(x_1) + u_2(x_2) + u_3(x_3) + u_{12}(x_1, x_2) + u_{13}(x_1, x_3) + u_{23}(x_2, x_3) + u_{123}(x_1, x_2, x_3), \qquad (4.1)$$

where, for example, $u_{12}(x_1, x_2)$ is the two-way interaction term between X_1 and X_2. To avoid overparametrization, constraints on the parameters are necessary.

A table in which the conditional independence, $X_2 \perp\!\!\!\perp X_3 \mid X_1$ is satisfied has all interaction terms containing x_2 and x_3 set to zero, i.e. $u_{23} = 0$ and $u_{123} = 0$. The density function corresponding to this conditional independence is then

$$\log p(x_1, x_2, x_3) = u_\varnothing + u_1(x_1) + u_2(x_2) + u_3(x_3) + u_{12}(x_1, x_2) + u_{13}(x_1, x_3). \qquad (4.2)$$

The parametrization is easily extended from three to k dimensions.

Undirected independence graphs are defined by the correspondence between a missing edge and the pairwise independence of two variables conditioned on the rest. Within the log-linear model, this corresponds to setting to zero all two-way and higher-order interactions containing that pair. Empirically these u terms are estimated by maximizing the appropriate likelihood function, and general-purpose software is widely available (e.g. SPSS, SAS, S+).

Model selection

For any contingency table, especially those in high dimensions, there are a large number of possible models to choose from. It is useful to have a technique to help select an acceptable model. A simple but rather naive model-selection procedure is to fit the saturated model, determine which u terms are negligible, and then use this to deduce the structure of the independence graph. An improvement is to drop the edge according to the site of the edge-exclusion deviance.

The deviance is just the maximized log-likelihood-ratio test statistic and measures the divergence between the observed values and the fitted values:

$$\text{deviance} = 2 \sum_{\text{cells}} \text{observed} \cdot \log\left(\frac{\text{observed}}{\text{fitted}}\right).$$

The fitted values are calculated from the specified log-linear model. Under the null hypothesis that the model specification is correct, the deviance has an asymptotic chi-squared distribution. The degrees of freedom are determined by the number of parameters which have been set to zero in the saturated model's log-linear expansion in order to derive the specified model.

Edge-exclusion deviances are those deviances corresponding to testing a pairwise conditional independence, and are calculated for each of the existing edges in a model, M, say. To calculate any one edge-exclusion deviance, that particular edge is dropped from the model M and a new model, M_1, is defined. The deviance calculated by fitting the model M_1 is then compared with the deviance from the fitted model M. The difference between these two deviances defines the edge-exclusion deviance, and is effectively a measure of how important that edge is in determining a good fit to the observed data. Small edge-exclusion devices correspond to near conditional independences while large deviances correspond to conditional dependences. The 'strength' of the edge reflects how confidently the null hypothesis that the conditional independence exists can be rejected. When a conditional independence is discovered, the corresponding edge is dropped from the graph and an updated model is defined. Edge-exclusion deviances are then calculated from this new model, and the process continues until a final model is chosen. These edge strengths are used to highlight the important interactions in the fitted graphs.

To add edges to the graph, edge-inclusion deviances can be calculated in a similar manner by comparing the deviances in the models with and without the edge of interest. The edge is added to the graph if a large drop in deviance is apparent in the model which includes this edge.

5. Sparsity in the contingency table

The log-linear all-way interaction model

First attempts to find a graphical model were carried out by fitting the all-way log-linear model to arbitrary subsets of application variables. This dataset, described in Section 6, includes 7702 credit-card applicants, and most of the variables contain between two and five levels. Subsets of between 4 and 8 dimensions were modelled, and an important problem soon became evident.

When attempts were made to drop edges from the saturated model containing all edges, by comparing the edge-exclusion deviance with a chi-squared distribution on the relevant number of degrees of freedom, it turned out that almost all the edges were significant, even at the 1% level. More worryingly it also appeared that the edge-exclusion deviances increased disproportionately when additional variables were introduced to the model, even after accounting for the extra degrees of freedom due to higher dimensions. Although this is theoretically possible, as Simpson's paradox illustrates, it is usually expected that the extra information an additional variable provides reduces the strength of the dependence between the two variables on that specific edge.

Bootstrapping can be used to estimate the variance of a function of the data by repeated replacement sampling from the empirical or the fitted distribution (Efron & Tibshirani 1993). By taking 100 bootstrap samples from the observed data and calculating the edge exclusion deviances for each sample, it is apparent that the sampling distribution of the deviance is misleading. The plots in Fig. 17 compare

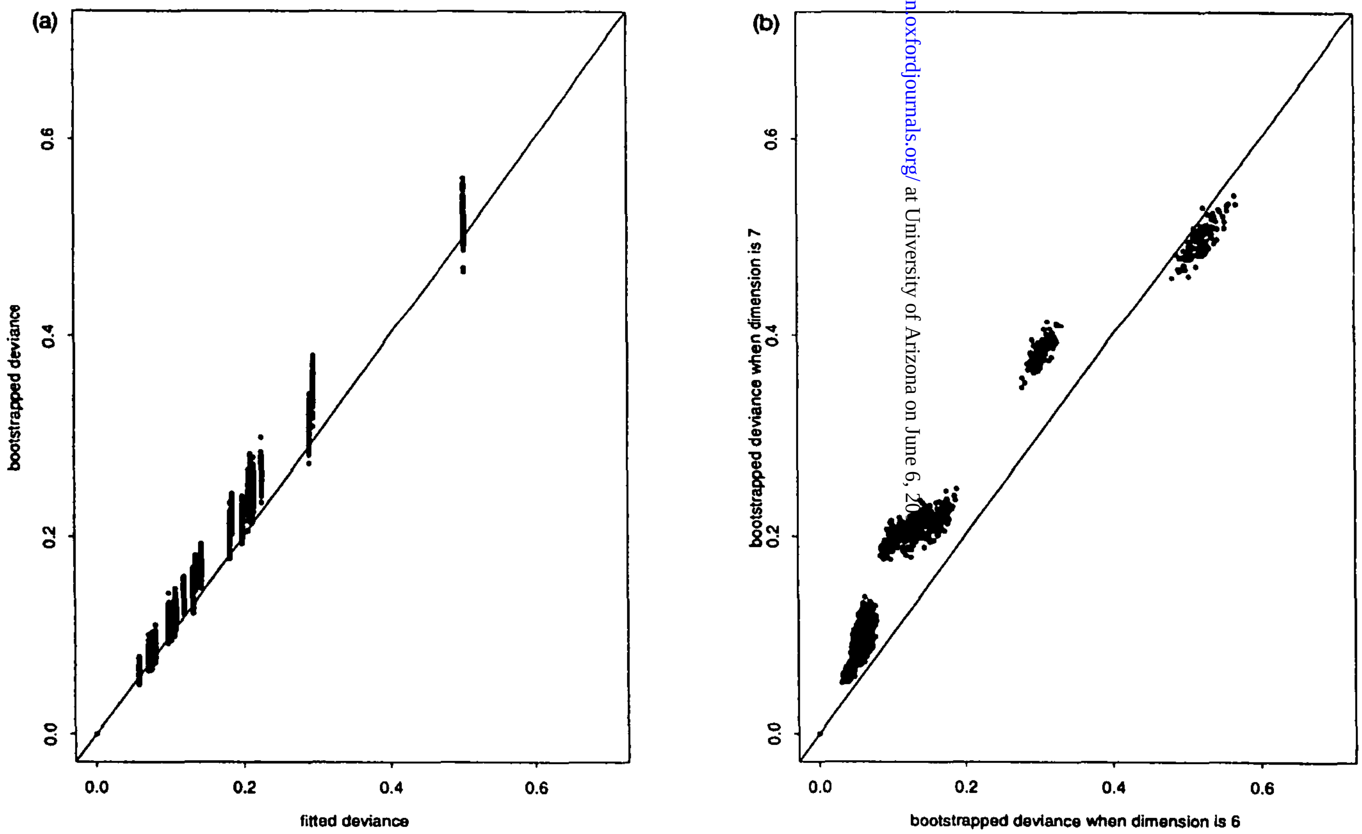

FIG. 17. Plot of the edge-exclusion deviances from the actual sample and from the bootstrapped samples: (a) for the original variables, and (b) showing the shift of the deviances when the dimension is increased from 6 to 7; the deviances have been divided by sample size.

the edge-exclusion deviances from the bootstrapped samples with the edge-exclusion deviances calculated from the actual data.

The first plot in Fig. 17 illustrates the bias existing in the distribution of edge-exclusion deviances. The bootstrap calculations of the deviances for the same edge should be symmetrically distributed about the straight line (the empirical edge-exclusion deviance); however, a positive bias for most edges is apparent. The second plot in Fig. 17 reveals that the edge-exclusion deviance for most of the edges is increasing as an additional variable is included in the model. It is expected that the deviances should either be unaffected or reduced upon further conditioning, so that the points lie on or below the line.

Since the number of cells in the contingency table increases with increasing dimension, it is sparsity that is causing the problems.

Sparsity

Sparse tables are caused by relatively small sample sizes, or by a high number of categories classifying the variables, or by a large number of variables. The failure to satisfy the large-sample assumptions causes the deviance to drift from the asymptotic chi-squared distribution and hence induce misleading model selection.

Example: how sparsity may cause the deviance to increase upon further conditioning. Consider the edge-exclusion deviance between variables X_1 and X_2. As the dimension of the table is increased, each partial table is divided over the levels of the additional conditioning variable. Each partial table contributes to the edge-exclusion deviance. Consider one such sparse observed partial table between the two variables, as in Table 10. This partial table's contribution to the edge-exclusion deviance is 0.138. Further conditioning on a variable X_3, independent of X_1 and X_2 with $P(X_3 = 1) = 0.5$, should not effect the edge-exclusion deviance between X_1 and X_2. However, the expected table includes non-integer cell counts, which cannot arise (Table 11). The sparsity of the table therefore forces the observations into one or the other of the levels of X_3, and dependences are induced. Two such possible tables are Tables 12(a, b).

The contribution to the edge-exclusion deviance is 1.73 and 1.05 respectively from these two tables. By considering all possible tables, the expected edge-exclusion deviance is calculated as 1.71: a large increase from the expected value of 0.138. A larger sample size would avoid this problem. However, with a fixed sample size, high dimensions induce many sparse partial tables which inevitably increase the edge-

TABLE 10
Observed incidences of X_1 and X_2

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	1	1
$X_2 = 1$	1	2

Table 11
Expected incidences conditional on X_3

	$X_3 = 0$		$X_3 = 1$	
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	0.5	0.5	0.5	0.5
$X_1 = 1$	0.5	1	0.5	1

Table 12(a)

	$X_3 = 0$		$X_3 = 1$	
X_2:	0	1	0	1
$X_1 = 0$	1	0	0	1
$X_1 = 1$	0	0	1	2

Table 12(b)

	$X_3 = 0$		$X_3 = 1$	
X_2:	0	1	0	1
$X_1 = 0$	0	1	1	0
$X_1 = 1$	0	1	1	1

exclusion deviances. We propose the examination of the all-two-way log-linear model as a remedy.

Two-way interactions

A method for dealing with sparse data is to restrict models to the class of all two-way interactions. All-way interaction models have the complete table as a sufficient statistic, but the two-way interaction class only requires the set of two-way marginal tables. In general these are not sparse. The number of parameters to estimate is correspondingly and drastically reduced. The conditional distributions derived from a two-way interaction model are necessarily additive in the log-odds-ratio scale, which is often assumed as a working hypothesis in logistic regression.

The two-way interaction models are constructed by constraining to zero all terms of interaction order higher than two in the log-linear expansion. So, in 3 dimensions, (4.1) becomes

$$\log p(x_1, x_2, x_3) \\ = u_\varnothing + u_1(x_1) + u_2(x_2) + u_3(x_3) + u_{12}(x_1, x_2) + u_{13}(x_1, x_3) + u_{23}(x_2, x_3),$$

and more generally, in k dimensions,

$$\log p(x_1, \ldots, x_k) = u_\varnothing + \sum_i u_i(x_i) + \sum_{i<j} u_{ij}(x_i, x_j). \qquad (5.1)$$

It is easily shown (e.g. Whittaker 1990) that the conditional independence $X_1 \perp\!\!\!\perp X_2 \mid (X_3, \ldots, X_k)$ is equivalent to $u_{12} = 0$, and so the independence graph is determined by the nonzero u_{ij} terms. Due to the reduced number of parameters to estimate, the tests for conditional independences in the two-way interaction model are determined by relatively small degrees of freedom yielding more powerful tests, and better asymptotic approximations.

A further justification for applying a two-way interaction model to the data is its close similarity with the linear logistic regression model commonly used in the credit-scoring discrimination process. The price to pay is that, if higher-order interactions are nonzero, this may not necessarily be recognized and the model may fit parts of the table poorly.

Applying two-way-interaction models to the application data subset

In practice, unlike the all-way interaction models, the two-way interaction model gives sensible results. The same subset of application-form variables is used here as was used with the all-way interaction models. Bootstrapping shows that the deviances are no longer biased and now reveals small edge-exclusion deviances indicating conditional independences. This is shown in the first plot of Fig. 18, where the edge-exclusion deviances from the bootstrapped samples are plotted against the edge-exclusion deviances calculated from the actual data. The second plot in Fig. 18 shows that the addition of further variables into the model now has the desired effect of either not affecting or of reducing the deviance.

Decomposition

Current computational constraints limit the dimension of long-linear models that can be fitted. One strategy for building the independence graph for all the variables is to search for a decomposition into subvectors (X_a, X_b, X_c). A decomposition requires that these subvectors be exclusive and exhaustive, that $X_a \perp\!\!\!\perp X_c \mid X_b$, and that the subgraph on X_b is complete as illustrated in Fig. 19.

To construct the whole graph, it is necessary to model only the subvectors (X_a, X_b) and (X_b, X_c). The joint distribution can then be obtained from the factorization $p_{abc} = p_{ab}p_{bc}/p_b$.

6. Modelling the application data

The application data

The data analysed in this section relate to a stratified sample of current account holders of a major bank who applied for a credit card to the bank's credit-card division between June and November 1992. The total sample size is 7702 which includes 'Goods', 'Bads', 'indeterminates', and 'rejects'. Note that the data are for a sample of the through-the-door population. Hence the credit performance variable is not known for some of the applicants, namely the rejects, and we do not consider it here.

The variables analysed in this paper, taken from the application form, are;

(1) bank account type, (2) own cheque guarantee card, (3) children, (4) employment status, (5) own telephone, (6) income band, (7) marital status, (8) residential status, (9) time at address, (10) time at employment, and (11) age.

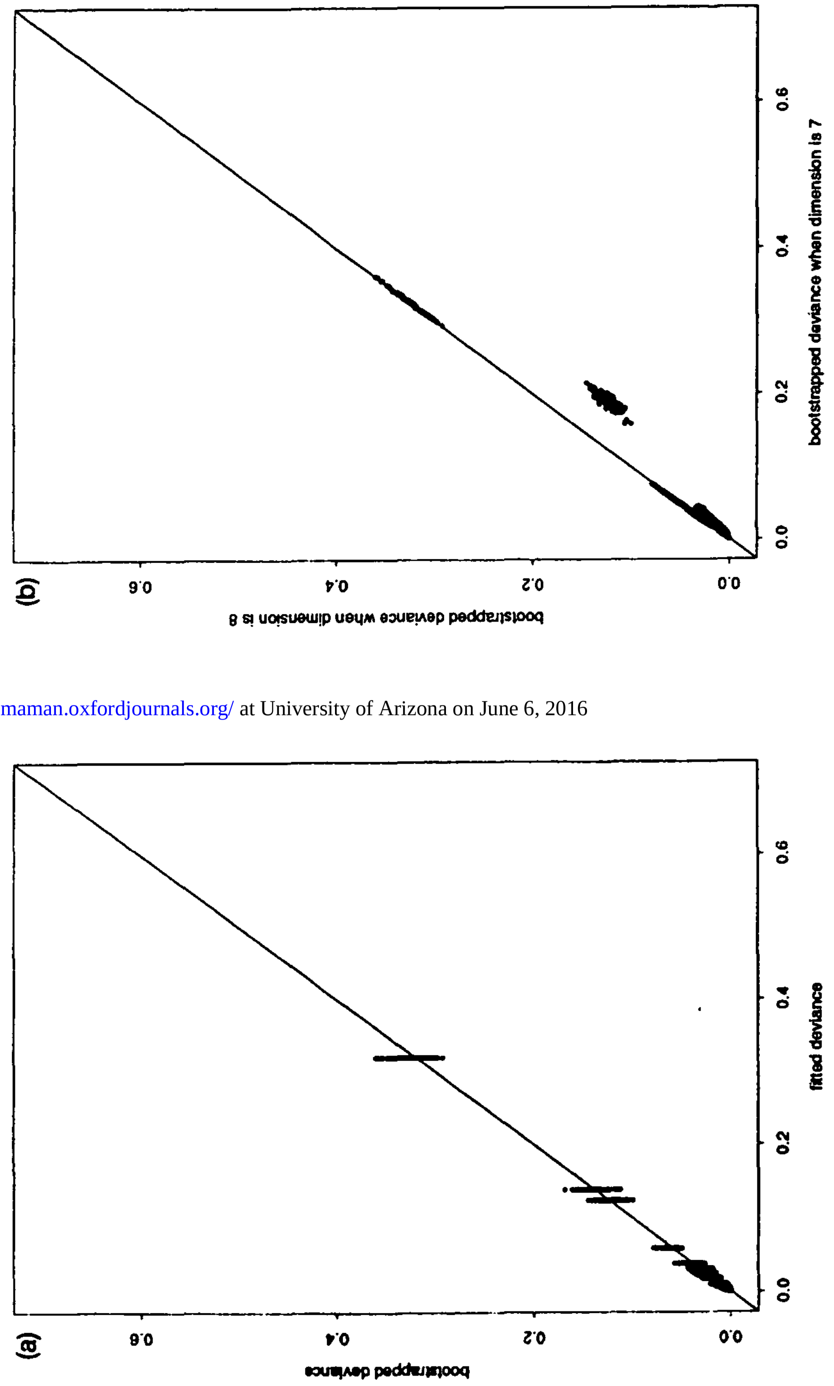

FIG. 18. Bootstrapped edge-exclusion deviances revealing (a) no bias and (b) no increase upon further conditioning.

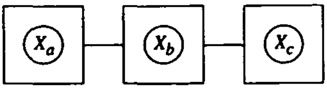

FIG. 19. Decomposition of X into subvectors X_a, X_b, and X_c.

The subset of variables selected here are not only potential scorecard covariates, but more importantly are chosen to describe the social and demographic characteristics of the population. The variables have been coarsely classified into the credit-card company's usual groupings in order to reduce the number of categories, and most variables contain between two and five levels.

The initial problem is to discover whether a decomposition as described in Fig. 19 exists. In practice we achieved this by reducing the number of categories within each variable so that all of the variables could be simultaneously modelled. The ordinal variables age, income, children, time at address, and time at employment were dichotomized into young/old, low income against high income, no children against children, less than three years against more than three years, and less than one year against more than one year respectively. Residential status was classified as rent/own and marital status was dichotomized as married against single or separated. Only the categories for employment and bank-account type were not combined. Although many of the complexities of the relationships may be disguised in this reduced classification, it should give an approximation to the structure of the graph.

Trial and error, together with guidance from credit-scoring experts, suggested the decomposition given below and illustrated in Fig. 20.

$$X_a = \text{children } (3),$$
$$X_b = \big(\text{employment status } (4), \text{ income band } (6), \text{ marital status } (7), \text{ age } (11)\big),$$
$$X_c = \big(\text{residential status } (8), \text{ time at address } (9), \text{ time at employment } (10)\big),$$
$$X_d = \big(\text{bank-account type } (1), \text{ own cheque guarantee card } (2), \text{ own phone } (5)\big).$$

The results of further modelling the subvectors (X_a, X_b), (X_b, X_c), and (X_c, X_d), using the original variables, reveals the graph in Fig. 21.

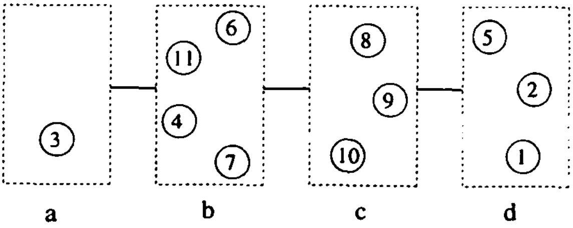

FIG. 20. A particular decomposition of the application variables into subsets. The decomposition was suggested by simultaneous modelling of the dichotomized variables.

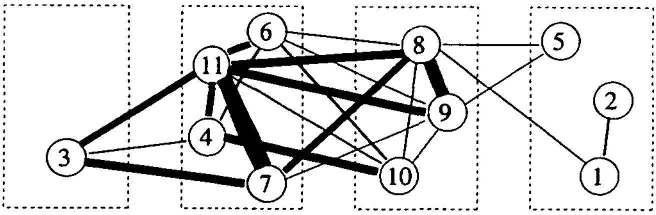

FIG. 21. Approximate independence graph of the categorical application variables constructed in two stages. First the decomposition was derived using the dichotomized variables, and then from further modelling using the original variables.

There are a number of observations worth noting.

- The banking variables, bank account type (1) and own cheque guarantee card (2), are conditionally independent from the rest of the variables given residential status. People who possess their own property are more likely to have both deposit and current accounts; tenants are more likely to have deposit accounts; and people who live with their parents are less likely to have both types of account. People with deposit accounts are less likely to have cheque guarantee cards. The other variables provide no additional information about the banking variables.
- Owning a telephone (5) depends only on residential status (8) and time at address (9). People who rent property are less likely to have their own telephone, as are people who have only recently moved into a new property.
- Once information is known about an individual's employment status (4), marital status (7), and age (11), the other variables have no additional power in predicting whether the person has children (3).
- Marital status (7) and age (11) are very strongly dependent. It is obvious that very young people have a lower probability of being married, let alone separated, divorced, or widowed.
- Residential status (8), time at address (9), and age (11) are strongly dependent. We can explain this association by considering that young people who live with their parents have generally lived at the same address for a number of years. Apart from this subpopulation, we would expect to see that, in the main, time at address increases linearly with age.
- Marital status (7) is conditionally independent of employment status (4), given income (6), residential status (8), and age (11).
- Note that age (11) and residential status (8) have important roles in the graph.

A comparison of distinct age groups

We now compare graphical models built for distinct subpopulation based on young and older age groups. It is apparent in Fig. 21 that age has an important role in the model. It is also likely that the two-way interaction model is not sufficient to explain some of the associations, and that some of the more complex relationships could be modelled by considering age groups separately. For example, the relationship between time at address and residential status is dependent on age: young people have not had the opportunity to live in rented or owned accommodation as long as older people. By modelling age groups separately, three-way interactions with age are included.

A separate scorecard is often used for applicants under a certain age. The graphs of the two distinct age groups can reveal the differences between the structure of the two populations, and hence suggest the importance of using separate scorecards. To carry out this comparison, the data have been split into two approximately equal-sized groups of 'young' and 'old' applicants.

The same decomposition is used as before in order to model the subsets. The graphs of the 'young' and 'old' applicants are shown in Figs 22 and 23 respectively. Note that edges present in the graph in Fig. 21 almost always exist in at least one of the subpopulation graphs. While the graphs are substantially similar, there are a number of evident differences, confirming that the interrelationships between these variables change with age. A discriminant function, built without including higher interactions with age, would not succeed in modelling these differences.

We comment on some of the differences between the two graphs here, although this list is by no means comprehensive:

- Bank-account type (1) and residential status (8) are independent of the other variables in the young subpopulation. This perhaps reflects a difference between generations and the more open attitudes of banks allowing all young people the same

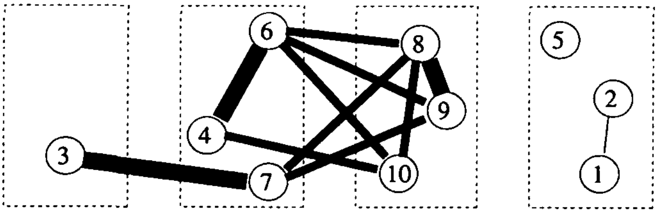

FIG. 22. Approximate independence graph of the young subpopulation. The model is built using the decomposition used previously, and edges with small edge-exclusion deviance are removed.

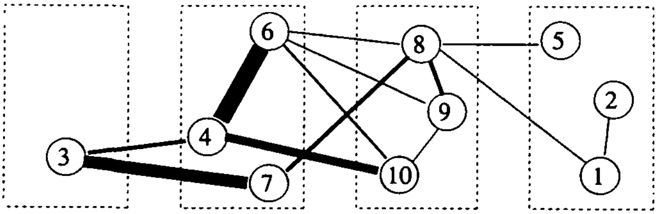

FIG. 23. Approximate independence graph of the older subpopulation. The model is built using the decomposition used previously, and edges with small edge-exclusion deviance are removed.

banking opportunities. The older subpopulation are perhaps more reluctant to change their ways and are less often targeted by the banks' marketing departments, and hence older tenants are less likely to have deposit accounts.

- The conditional dependence between children (3) and employment status (4) is only apparent in the older population, where the self-employed and nonworkers are more likely to have children. There are obviously less young people with children, so it is difficult to judge whether a relationship exists.
- Owing a telephone (5) and residential status (8) are independent in the young subpopulation. Young people are just as likely to have a telephone wherever they live.
- The relationship between income band (6) and residential status (8), time at address (9), and time at employment (10), is strong in the young population. For young people, a high income increases the opportunity for moving out of the parental home.
- Marital status (7) and time at address (9) are conditionally independent in the old population. People tend to change residence or ownership when they get married. In the young population, married people have only been married for a short time, and so marital status is associated with a short time at address.
- In the young population, time at employment (10) is associated with residential status (8), and is conditionally independent of time at address (9). The converse is true in the old population. We hypothesise that living with parents is associated with not having worked very long. Older people are more likely to live with their parents for reasons other than waiting to find a new job.

Extending the graphs to include the performance indicator is obviously of great importance to scorecard builders, though it raises the question of whether the graph is a model of the joint distribųtion of application variables X and credit performance Y, or of the conditional distribution of Y given X. In the latter case, the graph can be calculated from the logistic regression of Y on X. The credit performance variable cannot, however, be included in this investigation since the data analysed here were a sample of the through-the-door population, and, without making some kind of reject inference, we are without the performance realizations for the rejected applicants.

7. Summary

Graphical models and their associated conditional-independence statements have a role to play in credit scoring. We have argued that, to avoid interpreting spurious relationships and to understanding the more complex associations, it is necessary to examine the joint probability distribution of the variables of interest. Through didactic and practical examples set in the credit-scoring context, we have shown that graphical models provide the following advantages.

- They enable the display and interpretation of multivariate interactions between variables taken from a credit card application form.

- They allow the direct comparison of credit scoring subpopulations. In the example given, the comparison of young and old subpopulations highlights a number of differences. It is necessary to include higher interactions or to consider age groups separately in order to model these differences, and we suggest that it is worth investigating the use of separate scorecards based on age groups.
- They give a simplified description of the credit-scoring selection process in terms of influence diagrams, used in collaboration with decision trees.
- They illustrate possible effects of selection bias and stratification on variable interdependency.
- Furthermore we note that, while practical difficulties still remain in fitting graphs in high dimensions, we have shown that this is a feasible proposition.

There is further potential for the use of graphical models to model various aspects of the credit-scoring process, for example by including the score, credit performance indicator, or latent variables. In this paper we have only considered an application to the through-the-door population in order to improve understanding of the interrelationships between the variables that characterize the applicants. Hopefully our results should lead to further investigation of the role of graphical models in scoring technology.

We end on a technical note. Modelling in high dimensions inevitably leads to sparsity and misleading model selections from fitting the all-way log-linear interaction model. The asymptotic approximation to the chi-squared distribution becomes invalid and it is not possible to obtain reasonable parameter estimates. The all-two-way model overcomes the problems of sparsity while still retaining the conditional-independence structure. It has strong links with logistic regression, the interactions are easily interpretable, and it provides sensible estimates of the edge-exclusion deviances. On the downside, it may neglect higher-order interactions.

There are computational difficulties with large number of cells, since most algorithms for fitting the two-way interaction model requires storage of the complete contingency table. Even when the categories have been reduced into coarse categories, a typical contingency table of eight variables may consist of around 50000 cells, and a table with ten dimensions has around 750000 cells. Seven or eight dimensions with a variety of levels appears to be the limit for the current generation of workstations. However, we believe these constraints will rapidly prove less onerous as time goes by.

Acknowledgements

This work was carried out in collaboration with the credit-card division of a major bank and support from the ALCD (Analysis of Large and Complex Datasets) initiative of the ESRC, grant no. asa6116. We should like to acknowledge the comments of a referee which have led to very substantal improvements in this paper.

References

AGRESTI, A., 1990. *Categorical data analysis*. Wiley, New York.

BOYLE, M., CROOK, J. N., HAMILTON, R., & THOMAS, L. C., 1992. Method of credit scoring applied to slow payers. *Credit scoring and credit control*. Oxford University Press, Oxford.

DARROCH, J. N., LAURITZEN, S. L., & SPEED, T. P., 1980. Markov fields and log-linear models for contingency tables. *Ann. Statist*. **8**, 522–39.

EDWARDS, D., 1995. *An introduction to graphical modelling*. Springer Verlag.

EFRON, B., & TIBSHIRANI, J., 1993. *An introduction to the bootstrap*. Chapman & Hall.

HAND, D. J., & HENLEY, W. E., 1993. Can reject inference ever work? *IMA Journal of Mathematics Applied in Business & Industry* **5**, 45–55.

HAND, D. J., MCCONWAY, K. J., & STRANGHELLINI, E., 1997.Graphical models of applicants for credit. To appear in *IMA Journal of Mathematics Applied in Business & Industry* **8**, 143–55.

KRZANOWSKI, W. J., 1988. *Principles of multivariate analysis*. Oxford University Press, Oxford.

LEWIS, E. M., 1992. *An introduction to credit scoring*. Athena Press: San Rafael.

MARSHALL, K. T., & OLIVER, R. M., 1995, *Decision making and forecasting*. McGraw-Hill, USA.

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

WILKIE, A. D., 1992. Measures for comparing scoring systems. *Credit scoring and credit control*. Oxford University Press, Oxford.