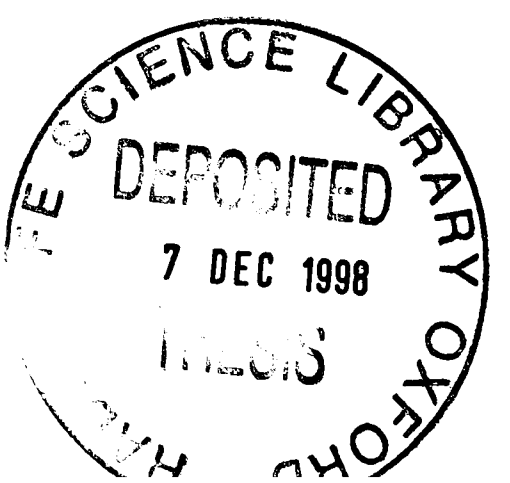
Graph-Theoretic Methods in Discrimination and Classification

R.A. Saldanha

Oriel College, Oxford

A thesis submitted to the Faculty of Mathematical Sciences for the degree of Doctor of Philosophy in the University of Oxford

Hilary Term 1998



To Jane and her cat Kizzy

Abstract

This thesis is concerned with the graphical modelling of multivariate data. The main aim of graphical modelling is to provide an easy to understand visual representation of, often complex, data relationships by fitting graphs to data. The graphs consist of nodes denoting random variables and connecting lines or edges are used to depict variable dependencies. Equivalently, the absence of particular edges in a graph describe conditional independencies between random variables. The resulting structure is called a conditional independence graph.

The use of conditional independence graphs as a guide to discrete (mainly binary), normal and mixed conditional Gaussian model building is described. The problem of parameter estimation in fitting conditional Gaussian models is considered. A FORTRAN 77 program called CGM is developed and used to fit conditional Gaussian models. Submodel specification, model selection criteria and goodness-of-fit are explored.

A procedure for discriminating between groups is constructed using fitted conditional Gaussian models. A Bayesian classification procedure is considered and is used to compute posterior classification probabilities. Standard bias-correcting error rates are used to test the performance of estimated classification rules.

The graph-theoretic methodology described in this thesis is applied to a Scandinavian study of intrauterine foetal growth retardation also known as a small-for-gestational age (SGA) birth. Possible pre-pregnancy risk factors associated with SGA births are investigated using conditional independence graphs and an attempt is made to classify SGA births using fitted conditional Gaussian models.

Acknowledgements

I would like to thank my supervisors Dr Francis Marriott and Professor Brian Ripley for their time, support and guidance throughout all stages of this research study.

I would also like to thank Ms Susan Hutchinson and Mr David Flitney for their very good advice about computing.

I am grateful to Dr Geir Jacobsen and to Dr Mette Langaas at the University of Trondheim for discussing the small-for-gestational age (SGA) births study with me and for helping me obtain the data. Permission to use data from the SGA births study was granted by the SGA Coordinating Centre at the University of Trondheim, Norway.

This research was funded by a studentship from the Science and Engineering Research Council (SERC). (SERC became the Engineering and Physical Sciences Research Council in 1993.)

Contents

Chapter 1

Introduction

One use of graph-theoretic methods in statistics is an attempt to describe the inter-relationships between several random variables by conditioning or controlling for other factors. These inter-relationships are displayed pictorially by a graph, whose vertices represent random variables and whose edges represent variable dependencies. This structure is called a conditional independence graph.

The use of graphs as a description of data structure was first introduced by Wright (1921, 1934) in the form of path analysis. Path analysis uses networks or path diagrams to represent possible models of cause and effect among variables. Path diagrams have occurred in causal econometric and social models, as 'influence' diagrams in decision analysis, as pedigrees in genetics, and more recently, as network diagrams in probabilistic expert systems.

This chapter introduces some of the theory underpinning the application of graph-theoretic methods in statistics. It includes a description of conditional independence, a review of some graph theory and the definition of a graphical model. The aim of this chapter is to show how graphs may be used to represent dependencies between random variables in a compact way. Before embarking on a description of these concepts we give an overview of the material contained in this thesis.

1.1 Overview

This thesis is concerned with the concept of *conditional independence* and the depiction of conditional independence relationships in terms of a *conditional independence graph*. Given a finite set of random variables defined on the vertices of a graph we construct the edges of the graph from statements about pairwise conditional independence among the set of random variables. By application of the *Markov properties* (see page 6) we may generalize pairwise conditional independence and infer conditional independence relationships among sets of random variables. In addition, random variables that are connected by edges are deemed to be non-independent.

Let us assume that $\mathcal{M}$ is a suitable model for some observed data. We also assume that $\mathcal{M}$ is made up of a set of interactions between variables and that these interactions are *hierarchical* in the sense that zero low-order interactions imply that all higher-order interactions involving the lower-order interactions are also zero. We are primarily concerned with zero pairwise interactions as, by the hierarchy rule just stated, this implies zero higher-order interactions. (A zero single-factor interaction simply means that the corresponding random variable is not included in the model.) If the set of zero interactions among the variables in $\mathcal{M}$ is exactly the set of conditional independence relationships read from the corresponding conditional independence graph, then $\mathcal{M}$ is a *graphical model* (see page 9). Note that the graph in itself does not uniquely de-

fine the model. Ideally, however, we would like to restrict ourselves to the class of graphical models as this makes interpretation of graphs and models easier. However, the non-graphical conditional independence graph still provides a useful description of data structure.

Chapter 1 reviews the foundations for interpreting conditional independence graphs. It links standard model definitions with conditional independence graphs by describing *graphical log-linear models* for discrete data and *graphical Gaussian models* for continuous data.

In Chapter 2 we look at the *conditional Gaussian interaction parametrization* as described by Lauritzen & Wermuth (1989). This embeds the class of graphical Gaussian models within the class of log-linear models allowing discrete and continuous data to be modelled explicitly in the joint framework. A maximum likelihood parameter estimation scheme using a *quasi-Newton* algorithm is used to maximize the conditional Gaussian (CG) likelihood. We look at CG sub-model specification and focus on models in which the covariance structure is constant. In particular, we are interested in the *location model* due to Tate (1954) and Olkin & Tate (1961). The location model describes the joint distribution of a set of mixed discrete and continuous data. A *location* (or *cell*) is specified by the values of the discrete variables. At each location the conditional distribution of the continuous variables is assumed to be multivariate normal with the same covariance matrix. The location model may be used as a method of *discrimination* and this aim was developed by Krzanowski (1975, 1980). We show how more general CG models may be used as a basis for discrimination. However, one of the main problems in fitting CG models is the presence of observed cell frequencies of zero when the number of cells formed by the arrangement of the discrete variables is large. Reduced models, such as models that just contain all two-factor interactions, are used to overcome this problem. We look at standard asymptotic chi-squared tests of deviance and AIC 'An Information Criterion' (Akaike, 1973) in model selection.

In Chapter 3 we look at estimating discriminant functions using fitted CG models. *Estimative* and *predictive* classification rules are defined using standard results for multivariate normal distributions and as part of our predictive procedure with CG models we adopt a logistic model as a basis for refitting the cell probabilities. Classification rules based on CG models are compared with rules based on logistic discrimination and k-nearest neighbour methods. We look at bias-correcting classification error rates using standard methods.

We attempt to justify the methods described in this thesis in Chapter 4. This describes the analysis of a Scandinavian study of intrauterine foetal growth deviations also known as small-for-gestational (SGA) age births. The data consist of 5,722 Swedish and Norwegian women expecting their second or third child between January 1986 and March 1988. Possible pre-pregnancy risk factors associated with SGA births are investigated using conditional independence graphs. A loss function approach is described that avoids the problem of classification results being dominated by the larger non-SGA group.

The emphasis of this thesis is on the practical application of graphical modelling. It is hoped that it will serve as a useful guide to data discrimination and classification. Conclusions and suggestions for further work are given in Chapter 5.

1.2 Independence and conditional independence

Let X_1 and X_2 be random variables. We shall assume throughout this thesis (where appropriate) that the following densities exist: the joint density of X_1 and X_2 denoted by $f_{X_1X_2}(x_1, x_2)$, the conditional density of X_1 given X_2 denoted by $f_{X_1|X_2}(x_1 \mid x_2)$, and the marginal densities of X_1 and X_2 denoted by $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$ respectively. Here X_1 and X_2 may be discrete or continuous random variables. Alternatively, $\{X_A\}$ and $\{X_B\}$ may be discrete or continuous sets of random variables and we assume that the above definitions still hold. We shall also assume throughout this thesis that all probability densities are strictly positive.

Two random variables are independent if their joint density factorizes into a product of their marginal densities, i.e.

$$f_{X_1X_2}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2).$$

In general, the joint density of $X_1, \ldots, X_k$ mutually independent random variables may be expressed as

$$f_{X_1 \cdots X_k}(x_1, \ldots, x_k) = f_{X_1}(x_1) \cdots f_{X_k}(x_k).$$

Throughout this thesis we denote independence between random variables by the symbol '$\perp$', e.g. if X_1 is independent of X_2 then this is represented by $X_1 \perp X_2$. In general, mutual independence for k random variables is expressed as

$$X_1 \perp X_{V\setminus\{1\}}, \ldots, X_k \perp X_{V\setminus\{k\}},$$

where $V = \{1, \ldots, k\}$ and $X_{V\setminus\{i\}}$ denotes the set containing all k random variables except X_i. Independence of X_1 and X_2 implies that the conditional density of X_1 given X_2 may be written in terms of X_1 alone, i.e.

$$f_{X_1|X_2}(x_1|x_2) = f_{X_1}(x_1).$$

Independence of X_1 and X_2 also implies that the conditional density of X_2 given X_1 may be written in terms of X_2 alone, i.e.

$$f_{X_2|X_1}(x_2|x_1) = f_{X_2}(x_2).$$

Pairwise conditional independence between random variables X_1 and X_2 is defined by introducing one or more conditioning variables, e.g. if X_1 and X_2 are conditionally independent given the set of k random variables not including X_1 and X_2, i.e. $X_{V\setminus\{1,2\}}$, then the joint density of X_1 and X_2 given $X_{V\setminus\{1,2\}}$ may be expressed as

$$f_{X_1X_2|X_{V\setminus\{1,2\}}}(x_1, x_2|x_{V\setminus\{1,2\}}) = f_{X_1|X_{V\setminus\{1,2\}}}(x_1|x_{V\setminus\{1,2\}})f_{X_2|X_{V\setminus\{1,2\}}}(x_2|x_{V\setminus\{1,2\}}).$$

The joint density is factorized into a product of two conditional densities, one not involving X_2 and the other not involving X_1. For arbitrary X_i and X_j conditional independence may be expressed as

$$X_i \perp X_j \mid X_{V\setminus\{i,j\}} \qquad (i \neq j). \tag{1.1}$$

Conditional independence between a random variable X_i and a set of random variables $X_A = \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_j\}$, say, conditional on the set of remaining

variables $X_{V\setminus\{i,A\}} = \{X_{j+1}, \ldots, X_k\}$ implies that the joint density of X_i and $\{X_A\}$ may be factorized as

$$f_{X_1 \cdots X_i X_{i+1} \cdots X_j | X_{j+1} \cdots X_k}(x_1, \ldots, x_i, x_{i+1}, \ldots, x_j | x_{j+1}, \ldots, x_k) =$$
$$f_{X_i | X_{j+1} \cdots X_k}(x_i | x_{j+1}, \ldots, x_k)$$
$$\times f_{X_1 \cdots X_{i-1} X_{i+1} \cdots X_j | X_{j+1} \cdots X_k}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_j | x_{j+1}, \ldots, x_k).$$

For arbitrary X_i and $\{X_A\}$ conditional independence may be expressed as

$$X_i \perp X_A \mid X_{V\setminus\{i,A\}} \qquad (i \notin A). \tag{1.2}$$

Conditional independence between sets of random variables $X_A = \{X_1, \ldots, X_i\}$ and $X_B = \{X_{i+1}, \ldots, X_j\}$, say, conditional on the set of remaining variables $X_{V\setminus\{A,B\}} = \{X_{j+1}, \ldots, X_k\}$ implies that the joint density of $\{X_A\}$ and $\{X_B\}$ may be factorized as

$$f_{X_1 \cdots X_i X_{i+1} \cdots X_j | X_{j+1} \cdots X_k}(x_1, \ldots, x_i, x_{i+1}, \ldots, x_j | x_{j+1}, \ldots, x_k) =$$
$$f_{X_1 \cdots X_i | X_{j+1} \cdots X_k}(x_1, \ldots, x_i | x_{j+1}, \ldots, x_k)$$
$$\times f_{X_{i+1} \cdots X_j | X_{j+1} \cdots X_k}(x_{i+1}, \ldots, x_j | x_{j+1}, \ldots, x_k).$$

For arbitrary $\{X_A\}$ and $\{X_B\}$ conditional independence may be expressed as

$$X_A \perp X_B \mid X_{V\setminus\{A,B\}} \qquad (A \cap B = \varnothing). \tag{1.3}$$

See Dawid (1979, 1980) for further details about conditional independence.

1.3 Graph theory

Here we briefly describe some graph-theoretic terms used in this thesis. A good introduction to graph theory may be found in Wilson (1985) and a more extensive treatment of the subject in Berge (1973).

A *graph* $G(V, E)$, or simply G, consists of a finite nonempty set of *vertices*, *points* or *nodes*, V, and a (possibly empty) set of unordered pairs of distinct vertices, E, called *edges*. This definition of a graph does not permit multiple edges or loops. A graph is represented by a diagram in an obvious way: each vertex v belonging to V is represented by a dot (or circle) and each edge e belonging to E is represented by a line that connects its endpoints v_i and v_j. As an example, consider Figure 1.1 showing a graph with four vertices, $V = \{A, B, C, D\}$, and five edges, $E = \{e_1 = [A, B], e_2 = [B, C], e_3 = [C, D], e_4 = [A, C], e_5 = [D, B]\}$.

If v is an endpoint of an edge e, then e is said to be *incident* on v. The *degree* of a vertex v, is equal to the number of edges incident on v. The graph shown in Figure 1.1 has vertices A and D that are each of degree 2 and vertices B and C that are each of degree 3. Two vertices are said to be *adjacent* or *neighbours* if there is a connecting edge. A graph is termed *complete* if each vertex is connected to every other vertex. A *path* consists of a sequence of vertices, $v_0, v_1, \ldots, v_n$ for which $[v_i, v_{i+1}]$ is in the edge set E, for each $i = 1, \ldots, n-1$. The path forms a *cycle* if the end points are the same, i.e. $v_0 = v_n$. A cycle of length k is called a k-cycle. Any cycle must have length three or more. A cycle is said to be *chordless* if no pairs other than successive pairs of

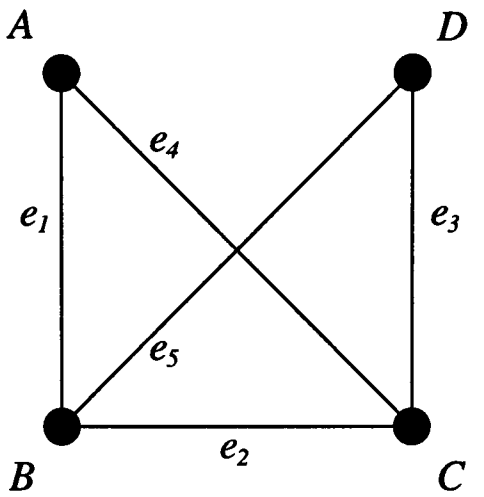


Figure 1.1 A graph.

vertices are adjacent. A graph is *triangulated* if it has no chordless cycles of length four or more. A graph is *connected* if there is a path between every pair of vertices.

The induced *subgraph* of a subset of vertices V_1, i.e. $V_1 \subset V$, is the graph obtained by deleting all the vertices not in V_1 from the graph $G(V, E)$, together with all the edges that do not join two elements of V_1. (Throughout this thesis we use the symbol '$\subset$' to denote a subset of elements that may also be equal to the set itself but will usually denote some reduced collection.) A *clique* is defined to be a maximal complete subset of vertices. A subset of vertices V_1 *separates* two vertices v_i and v_j in V, if every path joining the two vertices contains at least one vertex from the separating subset. A subset of vertices separates two subsets V_1 and V_2 of vertices in V if it separates every pair of vertices v_i belonging to V_1 and v_j belonging to V_2. The *boundary* of a subset of vertices $V_1 \subset V$, denoted by $\mathrm{bd}(V_1)$, are those vertices in V not contained in V_1, that are adjacent to a vertex in V_1.

1.4 Conditional independence graphs

DEFINITION 1.1 *The conditional independence graph of a set of k random variables $\{X_V\}$, where $V = \{1, \ldots, k\}$, is the graph $G(V, E)$ with vertex set V and edge set E where (i, j) is not in the edge set if and only if $X_i \perp X_j \mid X_{V \setminus \{i,j\}}$.*

The definition of the conditional independence graph combines the statement of pairwise conditional independence for k random variables from Equation (1.1) with the definition of a graph on k vertices. Throughout this thesis we shall interpret unconnected vertices in an independence graph as evidence of conditional independence between the corresponding random variables. Conversely, we will interpret connected vertices in an independence graph as no evidence of conditional independence between the corresponding random variables. Therefore, unconnected vertices are guaranteed to represent conditionally independent random variables and connected vertices will actually represent non-independent random variables.

Given a fixed number of vertices or dimensions $k \geqslant 2$ there will be total of $2^{\binom{k}{2}}$ different independence graphs. If the enumeration is extended to include subgraphs, obtained by considering all distinct graphs with $k, k-1, \ldots, 1$ vertices, then the number of possible independence graphs increases to

$$k + \sum_{l=2}^{k} \binom{k}{l} 2^{\binom{l}{2}}, \qquad k \geqslant 2. \tag{1.4}$$

When $k = 1$, the trivial graph consisting of a single vertex and no edges is the only possible graph.

1.4.1 Markov properties

A conditional independence graph may be interpreted through its *Markov properties*. Three Markov properties exist with respect to a graph on a finite number of vertices:

1. The *pairwise* Markov property states that non-adjacent pairs of variables are independent conditional on the remaining variables,
2. the *local* Markov property states that conditional on the adjacent variables, any variable is independent of all the remaining variables,
3. and the *global* Markov property states that any two subsets of variables separated by a third subset are independent conditionally only on the variables in the third subset.

(See Whittaker, 1990, pp. 56–57, 70–71.) The pairwise, local and global Markov properties are all equivalent if the joint distribution of the random variables is strictly positive. The truth of this statement is contained in the Hammersley–Clifford theorem (Hammersley & Clifford, 1971).

EXAMPLE 1.1 *Jamaican lizards.*

This example is based on data from Schoener (1970), concerning the structural habitat for two species of Jamaican lizard. The original data are in the form of counts in a $2 \times 2 \times 2 \times 3 \times 2$ contingency table and are analysed by Bishop *et al.* (1975, pp. 164–165). Figure 1.2 shows a conditional independence graph based on the final conclusions of Bishop *et al.* in their analysis.

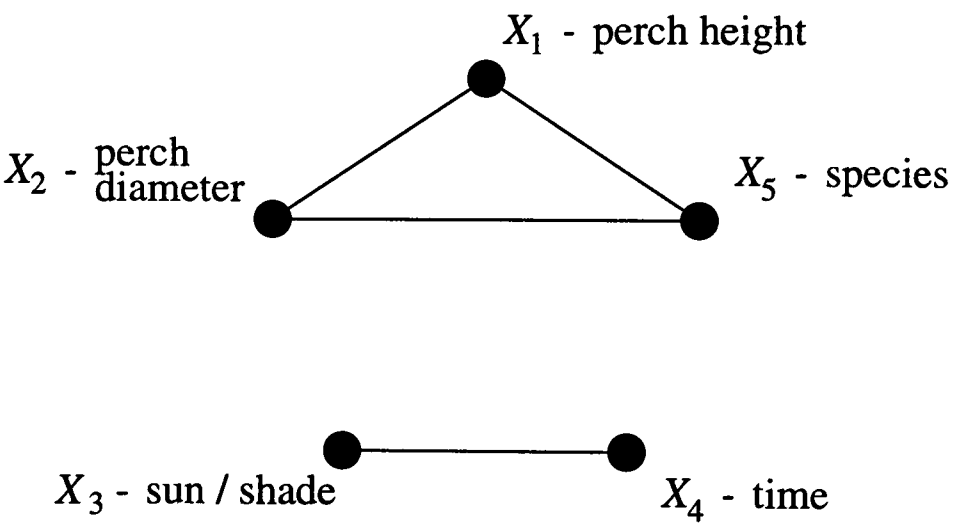


Figure 1.2 A conditional independence graph for Jamaican lizards: variable X_1 is perch height ($1 = <$ 5ft, $2 = \geq$ 5ft), variable X_2 is perch diameter ($1 = \leq$ 2in, $2 = >$ 2in), variable X_3 is sun versus shade exposure (1=sun, 2=shade), variable X_4 is time (1=early, 2=midday, 3=late) and variable X_5 is species (1=*grahami*, 2=*opalinus*).

The main conclusion based on the conditional independence graph is that the two species of lizard (variable X_5) differ in their choice of perch height and perch diameter, but the choice is independent of sun and shade conditions and time of day, i.e. $X_5 \perp X_3 \mid \{X_1, X_2, X_4\}$ and $X_5 \perp X_4 \mid \{X_1, X_2, X_3\}$. Also, $X_1 \perp X_3 \mid \{X_2, X_4, X_5\}$,

$X_1 \perp X_4 \mid \{X_2, X_3, X_5\}$, $X_2 \perp X_3 \mid \{X_1, X_4, X_5\}$ and $X_2 \perp X_4 \mid \{X_1, X_3, X_5\}$. These conditional independence relations illustrate the *pairwise* Markov property, that non-adjacent pairs of variables X_i, X_j are independent conditional on the remaining variables $X_{V\setminus\{i,j\}}$ $(V = \{1, \ldots, 5\})$.

In Example 1.1, the empty set, $\varnothing$, may be used as the conditioning set since the graph for the data is disconnected, i.e. $G = G_1 \cup G_2$, where G_1 and G_2 have vertex sets $V_1 = \{1, 2, 5\}$ and $V_2 = \{3, 4\}$ respectively. Then $X_{V_1} \perp X_{V_2} \mid \varnothing$, or simply $X_{V_1} \perp X_{V_2}$. These conditional independence relations illustrate the global Markov property, that for disjoint subsets V_1, V_2, V_3, whenever V_1 and V_2 are separated by V_3 in the graph, then X_{V_1} and X_{V_2} are independent given X_{V_3}, i.e.

$$X_{V_1} \perp X_{V_2} \mid X_{V_3}.$$

We also have $X_1 \perp \{X_3, X_4\} \mid \{X_2, X_5\}$, $X_2 \perp \{X_3, X_4\} \mid \{X_1, X_5\}$, $X_3 \perp \{X_1, X_2, X_5\} \mid \{X_4\}$, $X_4 \perp \{X_1, X_2, X_5\} \mid \{X_3\}$ and $X_5 \perp \{X_3, X_4\} \mid \{X_1, X_2\}$. These conditional independence relations illustrate the local Markov property, that for every vertex i, if $V_2 = \text{bd}(i)$ is its boundary set, and V_1 is the set of remaining vertices then

$$X_i \perp X_{V_1} \mid X_{V_2}.$$

1.5 Log-linear models for contingency tables

Let $\{X_V\}$ $(V = \{1, \ldots, p\})$ be a set containing p discrete random variables, which we shall call *factors* and let $v \in V$ index X_V. We define $\mathcal{I}_v$ to be the set of *levels* of v, thus $\mathcal{I} = \prod_{v \in V} \mathcal{I}_v$ is the full set of all levels or equivalently the full set of *cells*. An individual cell is defined by the $p \times 1$ dimensional integer vector $i = (i_1, \ldots, i_p)'$ where $i \in \mathcal{I}$. A set of n objects is classified according to the set of p factors and the number of objects in each cell is given by the set of *counts*, $n(i)$, such that $n = \sum_{i \in \mathcal{I}} n(i)$. We call the structure formed by classifying the n objects according to the set of p factors a *contingency table*. For a subset $U \subset V$ we have a *marginal cell* $i_U \in \mathcal{I}_U = \prod_{v \in U} \mathcal{I}_v$ and $n(i_U)$ is the corresponding *marginal count*. The probability that an object belongs to cell i is given by $p(i)$. Similarly, the marginal probability for $U \subset V$ that an object belongs to the marginal cell i_U is given by $p(i_U)$.

If we consider the classifications of the n objects to be n independent observations of the distribution P then the distribution of the counts is multinomial.

DEFINITION 1.2 *The multinomial density P is defined as*

$$P\{N(i) = n(i), i \in \mathcal{I}\} = \frac{n!}{\prod_{i \in \mathcal{I}} n(i)!} \prod_{i \in \mathcal{I}} p(i)^{n(i)}, \tag{1.5}$$

where the $i = (i_1, \ldots, i_p)' \in \mathcal{I}$ are $p \times 1$ integer vectors defining cells with associated scalar cell counts $n(i)$ and scalar parameters giving the cell probabilities $p(i)$, subject to $p(i) \geqslant 0$ and $\sum_{i \in \mathcal{I}} p(i) = 1$.

Maximum likelihood estimation of $p(i)$, if the cell probabilities are unrestricted, yields $\hat{p}(i) = n(i)/n$ (see Rice, 1988, pp. 238–239). Based on this sampling scheme, for fixed n, the general log-linear model involves specification of the unknown distribution P as follows

$$\log p(i) = \sum_{U \subset V} \lambda_U(i_U), \tag{1.6}$$

where $V = \{1, \ldots, p\}$, $\sum_{U \subset V}$ denotes summation over all possible subsets of variables U of V and the λ_U are functions of i_U that only depend on i_U through the integer values of those random variables $U \subset V$. The λ_U are called interactions among the factors belonging to V, and for $U = \varnothing$, $\lambda_U = \lambda$ is a constant. Let $|U|$ denote the number of elements in the set U. If $|U| = 1$ we call λ_U a main effect, if $|U| = 2$ we call λ_U a first-order interaction (also called a two-factor interaction) and, in general, if $|U| = m$ we call λ_U an interaction of order $m - 1$ (or m-factor interaction).

In order to identify interactions uniquely it is necessary to employ constraints on the estimated interactions. The choice of constraints alters the specific value of the estimates but does not alter the overall contribution to the fit of the model. There are any number of choices for the type of constraints used. Typical examples of constraints often used include symmetric constraints where the estimated interactions are constrained to sum to zero and treatment constraints where the estimated interactions measure the difference between some base level (or control) and subsequent levels of each of the factors, e.g. if p factor levels are coded so as to take integer values from 1 to r_j for $j = 1, \ldots, p$, then $\lambda_U(i_U)$ is set to zero whenever all values in i_U take value 1 ($U \subset V$ and $U \neq \varnothing$). This last type of constraint has been used in the computer program GLIM (Aitkin *et al.*, 1989). Further details about parameter constraints may be found in McCullagh & Nelder (1989, pp. 63–65).

Goodman (1970, 1971) describes how by setting certain terms in a model equal to zero, various independence hypotheses connected with a contingency table may be tested. Goodman accomplishes this by restricting the class of models to the class of *hierarchical models*. Hierarchical models are defined by ensuring that whenever a higher-order effect is included in a model, the lower-order effects composed from variables in the higher-order effect are also included. Conversely, higher-order effects are set to zero if they are composed from one or more variables whose lower-order effects are zero. Consider Figure 1.2, if $X_1, \ldots, X_5$ are indexed by $i_1, \ldots, i_5$ respectively then the log-linear model implied by the conditional independence graph is given by

$$\begin{aligned} \log p(i) \;=\; & \lambda + \lambda_1(i_1) + \lambda_2(i_2) + \lambda_3(i_3) + \lambda_4(i_4) + \lambda_5(i_5) \\ & + \lambda_{12}(i_1, i_2) + \lambda_{15}(i_1, i_5) + \lambda_{25}(i_2, i_5) + \lambda_{34}(i_3, i_4) \\ & + \lambda_{125}(i_1, i_2, i_5). \end{aligned} \tag{1.7}$$

(Here we denote an interaction by its subscripts only, e.g. $\{1, 3\}$ denotes the first-order interaction λ_{13}.) For hierarchical models, zero first-order interactions $\{13\}$, $\{14\}$, $\{23\}$, $\{24\}$, $\{35\}$ and $\{45\}$ mean that the second-order interactions $\{123\}$, $\{124\}$, $\{134\}$, $\{135\}$, $\{145\}$, $\{234\}$, $\{235\}$, $\{245\}$, $\{345\}$; third-order interactions $\{1234\}$, $\{1235\}$, $\{1245\}$, $\{1345\}$, $\{2345\}$ and fourth-order interaction $\{12345\}$ must also be set equal to zero.

1.5.1 The generating class of a hierarchical log-linear model

DEFINITION 1.3 *Consider a hierarchical log-linear model, $\mathcal{M}$, defined on a set of p discrete random variables $\{X_V\}$, where $V = \{1, \ldots, p\}$. The generating class of $\mathcal{M}$ is defined to be the set K consisting of all distinct maximal subsets of interactions between variables in $\{X_V\}$.*

EXAMPLE 1.2 *The generating class for the hierarchical log-linear model* (1.7) *is given by the set*

$$K = \{\{125\}\{34\}\}.$$

The definition of the generating class, K, for a hierarchical log-linear model also implies that for all subsets $U \subset V$ discrete interactions $\lambda_U = 0$ if and only if there is no set $W \in K$ such that $U \subset W$. From now on we shall restrict our attention to hierarchical log-linear models.

1.6 Graphical log-linear models

DEFINITION 1.4 *Consider a set of p discrete random variables,* $\{X_V\}$ *(where* $V = \{1, \ldots, p\}$*), defined on the vertices of a graph G. The set of cliques of G is given by C. A log-linear model,* $\mathcal{M}$*, with generating class K defined on* $\{X_V\}$ *is a called a graphical log-linear model if and only if* $C \equiv K$*.*

If $\mathcal{M}$ is a graphical log-linear model then all conditional independence relationships among the random variables may be read directly from the associated conditional independence graph constructed using the generating class K for $\mathcal{M}$ (Darroch *et al.*, 1980).

The simplest example of a log-linear model that is non-graphical is the three factor model with missing second order-interaction, i.e. the model with generating class $K = \{\{12\}\{13\}\{23\}\}$. The independence graph for this model is the complete 3-graph with clique set $C = \{\{123\}\}$.

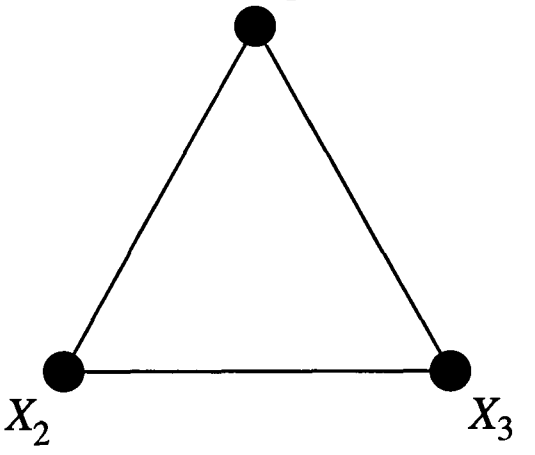


Figure 1.3 A complete 3-graph.

1.7 Graphical Gaussian models

DEFINITION 1.5 *The multivariate normal density is defined as*

$$f(y) = (2\pi)^{-q/2}|\Sigma|^{-1/2}\exp\{-\tfrac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\}, \tag{1.8}$$

for a vector $y \in \mathbb{R}^q$*, where* μ *is a* $(q \times 1)$ *real valued vector of means of* y *and* Σ *is a positive definite (symmetric)* $(q \times q)$ *covariance matrix of* y*.*

Let $\{Y_\Gamma\}$ $(\Gamma = \{1, \ldots, q\})$ be a set of q continuous random variables and let γ, ζ index the random variables belonging to Γ. Let $\Omega = \Sigma^{-1}$, i.e. Ω is the inverse covariance matrix or *concentration matrix*. The diagonal elements of Σ are called the *variances* $(\sigma_{\gamma\gamma})$ and the diagonal elements of Ω are called the *precisions* $(\omega_{\gamma\gamma})$ $(\gamma \in \Gamma)$. The off-diagonal elements of Σ are called the *covariances* $(\sigma_{\gamma\zeta})$ and the off-diagonal elements

of Ω are called the *concentrations* $(\omega_{\gamma\zeta})$ $(\gamma, \zeta \in \Gamma)$. A *marginal correlation*, $\rho_{\gamma\zeta}$, is expressible in terms of elements of the covariance matrix in the following way

$$\rho_{\gamma\zeta} = \frac{\sigma_{\gamma\zeta}}{\sqrt{\sigma_{\gamma\gamma}\sigma_{\zeta\zeta}}}.$$

Similarly, a *partial correlation* between Y_γ and Y_ζ, given all the remaining variables Y_Λ $(\Lambda = \Gamma \backslash \{\gamma\zeta\})$, $\rho_{\gamma\zeta.\Lambda}$, is expressible in terms of elements of the concentration matrix in the following way

$$\rho_{\gamma\zeta.\Lambda} = \frac{-\omega_{\gamma\zeta}}{\sqrt{\omega_{\gamma\gamma}\omega_{\zeta\zeta}}}$$

(see Cox & Wermuth, 1996, §3.4). For a set of q continuous random variables, $\{Y_\Gamma\}$, whose joint distribution is multivariate normal we can state that

1. $Y_\gamma \perp Y_\zeta$ if and only if $\sigma_{\gamma\zeta} = 0$,
2. $Y_\gamma \perp Y_\zeta \mid Y_{\Gamma\backslash\{\gamma\zeta\}}$ if and only if $\omega_{\gamma\zeta} = 0$.

DEFINITION 1.6 *Consider a set of q continuous random variables,* $\{Y_\Gamma\}$, *(where* $\Gamma = \{1, \ldots, q\}$*), whose joint distribution is multivariate normal defined on the vertices of a graph* $G(\Gamma, E)$. *A graphical Gaussian model exists for* $\{Y_\Gamma\}$ *when* (γ, ζ) *is not in the edge set E if and only if* $Y_\gamma \perp Y_\zeta \mid Y_{\Gamma\backslash\{\gamma\zeta\}}$ $(\gamma, \zeta \in \Gamma)$.

Graphical Gaussian models are based on the covariance selection models introduced by Dempster (1972). Covariance selection models aim to simplify the structure of multivariate data by setting the off-diagonal elements of an inverse covariance matrix equal to zero. Graphical Gaussian models depict the continuous random variables in the covariance matrix as vertices in a conditional independence graph. The non-zero off-diagonal inverse covariances determine the graph's edges. Speed & Kiiveri (1986) show that for graphical Gaussian models the rules for reading a conditional independence graph are a direct analogue of those used to interpret conditional independence graphs associated with graphical log-linear models.

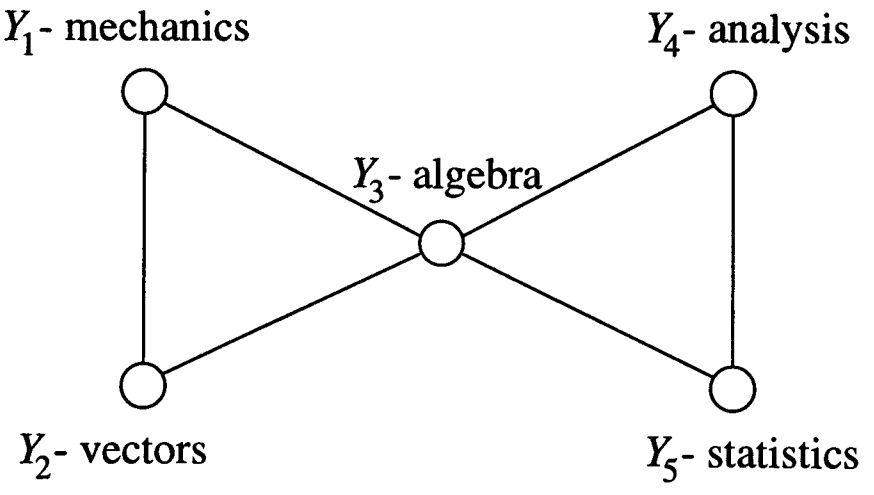


Figure 1.4 A conditional independence graph for the mathematics marks (Whittaker, 1990).

A simple example of a graphical Gaussian model taken from Whittaker (1990, pp. 1–6) concerns the marks obtained in five mathematics examinations for 88 students. (The dataset itself is taken from Mardia *et al.*, 1979.) Whittaker derives the following independence graph based on the sample inverse correlation matrix, which is scaled to produce unit entries on the leading diagonal. The off-diagonal elements of the

scaled inverse correlation matrix are then the negatives of the partial correlations. The variables are of course assumed to have a joint normal distribution. The graph shown in Figure 1.4 is constructed by noting the zero, or near zero, partial correlations and leaving out the corresponding edges. (Note that we use circles rather than dots for the vertices of the conditional independence graph shown in Figure 1.4. Circles indicate that the vertices represent continuous random variables.)

The conditional independence graph for a graphical Gaussian model may be interpreted using the equivalent Markov properties (see Section 1.4.1). Figure 1.4 indicates that the scores in {mechanics, vectors} are independent of the scores in {analysis, statistics} given the score on algebra, i.e. $\{Y_1, Y_2\} \perp \{Y_4, Y_5\} \mid Y_3$.

Chapter 2

Parameter Estimation in Conditional Gaussian Models

This chapter describes the implementation of a general-purpose optimization procedure for estimating the parameters in conditional Gaussian (CG) models for mixed discrete and continuous data. The main advantages of the parameter estimation method used are that models may be easily defined and the precision of the parameter estimates may be obtained directly once the procedure has converged. Practical examples are used to illustrate how CG models may be employed.

The definition of a CG model based on Lauritzen & Wermuth (1989) for mixed discrete and continuous data is presented first. We examine the likelihood for the general CG model and CG submodels. We then show how maximum likelihood parameter estimates may be obtained using a general-purpose optimization procedure. We look briefly at model search, model search criteria and present some worked data analysis examples.

2.1 Conditional Gaussian models for mixed data

Log-linear models were described in the preceding chapter for discrete data assumed to have arisen from a multinomial distribution. We also described the class of graphical Gaussian models for continuous data. CG models provide a joint framework that may be used to model explicitly both discrete and continuous data. If a particular CG model is graphical then all conditional independencies inherent in the model may be read directly from its associated conditional independence graph.

2.1.1 The conditional Gaussian distribution

Let $\{X_V\}$ be a set of k random variables partitioned into a discrete set $\{X_\Delta\}$ and a continuous set $\{X_\Gamma\}$ ($V = \Delta \cup \Gamma$). Let p be the number of random variables belonging to the discrete set and let q be the number of random variables belonging to the continuous set ($k = p + q$). A particular observation vector is denoted by $x = (i', y')'$, where $i = (i_1, \ldots, i_p)'$ is a $p \times 1$ vector of integer values for discrete random variables belonging to $\{X_\Delta\}$ and $y = (y_1, \ldots, y_q)'$ is a $q \times 1$ vector of real values for continuous random variables belonging to $\{X_\Gamma\}$. A particular combination of values i is called a cell and the set of all possible values of i is given by $\mathcal{I}$. We assume that the joint distribution of the full set of k random variables $\{X_V\}$ is defined by the CG density.

DEFINITION 2.1 *The conditional Gaussian (CG) density f is defined as*

$$f(x) = f(i,y) = p(i)(2\pi)^{-q/2}|\Sigma(i)|^{-1/2}\exp\Big\{-\tfrac{1}{2}[y-\mu(i)]'\Sigma(i)^{-1}[y-\mu(i)]\Big\}, \tag{2.1}$$

for $x = (i',y')' \in \mathcal{I}\times\mathbb{R}^q$, where $i = (i_1,\ldots,i_p)'$ is a $p\times 1$ integer vector and $y = (y_1,\ldots,y_q)'$ is a $q\times 1$ vector of real values, the $p(i)$ are positive scalar parameters giving the probability of y for cells $i\in\mathcal{I}$ and $\sum_{i\in\mathcal{I}} p(i) = 1$, $\mu(i)$ are $(q\times 1)$ real valued vectors of means of y for cells $i\in\mathcal{I}$ and $\Sigma(i)$ are positive-definite (symmetric) $(q\times q)$ covariance matrices of y for cells $i\in\mathcal{I}$.

Therefore conditional on cell i, the distribution of the continuous random variables is multivariate normal, i.e.

$$X_\Gamma \mid (X_\Delta = i) \sim N_q\{\mu(i),\Sigma(i)\}.$$

When the covariance matrix does not depend on the values of the discrete random variables, i.e. when $\Sigma(i)\equiv\Sigma$, the probability density is called *homogeneous conditional Gaussian* (HCG). The models studied by Olkin & Tate (1961) and Krzanowski (1975), discussed towards the end of this chapter, are based on particular HCG densities. (We adopt the convention of Lauritzen (1996) of not subscripting cell parameters with i but rather placing i in parentheses in normal-sized type to emphasize, where appropriate, the dependence of parameter values on the value of i.)

The *exponential family* representation of the CG density is given by

$$f(x) = f(i,y) = \exp\Big\{\alpha(i) + \beta(i)'y - \tfrac{1}{2}y'\Omega(i)y\Big\}. \tag{2.2}$$

The relationship between (2.2) and (2.1) is given by

$$\begin{aligned}
\alpha(i) &= \log p(i) - \tfrac{1}{2}\log|\Sigma(i)| - \tfrac{1}{2}\mu(i)'\Sigma(i)^{-1}\mu(i) - \tfrac{1}{2}q\log(2\pi), && (2.3)\\
\beta(i) &= \Sigma(i)^{-1}\mu(i), && (2.4)\\
\Omega(i) &= \Sigma(i)^{-1} && (2.5)
\end{aligned}$$

and conversely

$$\begin{aligned}
p(i) &= (2\pi)^{q/2}|\Omega(i)|^{-1/2}\exp\Big\{\alpha(i) + \tfrac{1}{2}\beta(i)'\Omega(i)^{-1}\beta(i)\Big\}, && (2.6)\\
\mu(i) &= \Omega(i)^{-1}\beta(i), && (2.7)\\
\Sigma(i) &= \Omega(i)^{-1}. && (2.8)
\end{aligned}$$

Here the $\{\alpha(i)\}_{i\in\mathcal{I}}$ are scalar parameters, the $\{\beta(i)\}_{i\in\mathcal{I}}$ are $q\times 1$ vectors of real values, and the $\{\Omega(i)\}_{i\in\mathcal{I}}$ are positive-definite (symmetric) $q\times q$ matrices. The $\alpha(i)$, $\beta(i)$ and $\Omega(i)$ are called, respectively, the discrete, linear and quadratic *canonical* parameters.

2.1.2 The conditional Gaussian interaction parametrization

We shall refer to a subset of discrete random variables as $\{X_A\}$ $(A\subset\Delta)$ with particular integer-valued observation vector $x_A = i_A$. Similarly, we refer to a set of continuous random variables as $\{X_\Lambda\}$ $(\Lambda\subset\Gamma)$ with particular real-valued observation vector $x_\Lambda = y_\Lambda$. Let $U = A\cup\Lambda$ then a particular observation vector for $\{X_U\}$ $(U\subset V)$ is given

by $x_U = (i'_A, y'_\Lambda)'$. (In what follows we often omit brackets $\{\}$, emphasizing a set of objects, if it is clear that a set is being referred to.)

An important feature of the CG distribution, illustrated by Lauritzen & Wermuth (1989), is that it may be directly parametrized in terms of interactions between variables. Variables that are conditionally independent have zero interaction parameters. Interaction expansions are obtained by firstly expanding the terms in (2.2) as follows

$$y = (y_1, \ldots, y_\gamma, \ldots, y_q)',$$
$$\beta(i) = (\beta_1(i), \ldots, \beta_\gamma(i), \ldots, \beta_q(i))'$$

and

$$\Omega(i) = \begin{pmatrix} \omega_{11}(i) & \cdots & \omega_{1\gamma}(i) & \cdots & \omega_{1q}(i) \\ \vdots & \ddots & \vdots & & \vdots \\ \omega_{\gamma 1}(i) & \cdots & \omega_{\gamma\gamma}(i) & \cdots & \omega_{\gamma q}(i) \\ \vdots & & \vdots & \ddots & \vdots \\ \omega_{q1}(i) & \cdots & \omega_{q\gamma}(i) & \cdots & \omega_{qq}(i) \end{pmatrix}.$$

Substituting the individual elements in (2.2) now gives

$$f(i,y) = \exp\left\{\alpha(i) + \sum_{\gamma\in\Gamma} \beta_\gamma(i)\, y_\gamma - \tfrac{1}{2}\sum_{\gamma\in\Gamma}\sum_{\zeta\in\Gamma} \omega_{\gamma\zeta}(i)\, y_\gamma y_\zeta\right\}. \tag{2.9}$$

Expanding each term in (2.9) over all subsets of discrete random variables X_A $(A \subset \Delta)$, yields

$$\alpha(i) = \sum_{A\subset\Delta} \lambda_A(i_A), \tag{2.10}$$

$$\beta_\gamma(i) = \sum_{A\subset\Delta} \eta_{\gamma;A}(i_A) \qquad (\gamma \in \Gamma) \tag{2.11}$$

and

$$\omega_{\gamma\zeta}(i) = \sum_{A\subset\Delta} \psi_{\gamma\zeta;A}(i_A) \qquad (\gamma, \zeta \in \Gamma). \tag{2.12}$$

The $\lambda_A(i_A)$, $\eta_{\gamma;A}(i_A)$ and $\psi_{\gamma\zeta;A}(i_A)$ are called interaction expansions, which are functions of i_A that only depend on i_A through the integer values of the discrete random variables X_A $(A \subset \Delta)$. In the above expansions we define $\lambda = \lambda_\varnothing$ to be a *normalizing constant*, $\eta_\gamma = \eta_{\gamma;\varnothing}$ to be a *constant main effect* of a continuous variable and $\psi_{\gamma\zeta} = \psi_{\gamma\zeta;\varnothing}$ to be a *constant pure quadratic interaction*. The interaction expansions for $\alpha(i)$, $\beta(i)$ and $\Omega(i)$ are formed in an analogous way to the discrete interactions in the log-linear model. As in the log-linear model a constraint is needed to ensure identifiability of the parameters and this role is taken by $\lambda = \lambda_\varnothing$. The CG interaction parametrization is finally obtained by substituting these interaction expansions in (2.9) to give

$$f(i,y) = \exp\left\{\sum_{A\subset\Delta} \lambda_A(i_A) + \sum_{\gamma\in\Gamma}\sum_{A\subset\Delta} \eta_{\gamma;A}(i_A)\, y_\gamma - \tfrac{1}{2}\sum_{\gamma\in\Gamma}\sum_{\zeta\in\Gamma}\sum_{A\subset\Delta} \psi_{\gamma\zeta;A}(i_A)\, y_\gamma y_\zeta\right\}. \tag{2.13}$$

Density (2.13) consists of

1. a finite number of discrete interactions involving discrete variables only,
2. a finite number of linear interactions involving discrete variables and one continuous variable only,
3. a finite number of quadratic interactions involving discrete variables and two continuous variables only.

(Note that an HCG interaction parametrization is obtained when the quadratic interactions do not depend on the values of the discrete variables.) Pairwise conditional independence relations are specified by setting certain interactions equal to zero. More formally, for a CG distribution a variable pair is conditionally independent given the remaining variables if and only if all interaction terms containing this particular variable pair are zero (Lauritzen & Wermuth, 1989, Proposition 3.1). Note that when $\Gamma = \varnothing$ (i.e. $q = 0$) CG models reduce to the class of log-linear models and when $\Delta = \varnothing$ (i.e. $p = 0$) CG models reduce to the class of graphical Gaussian models. For p discrete random variables each with r_j levels ($j = 1, \ldots, p$) the maximum number of discrete interactions that may be estimated is $r = \prod_{j=1}^{p} r_j - 1$. One discrete interaction is fixed since $\sum_{i \in \mathcal{I}} p(i) = 1$. If in addition to p discrete random variables there are q continuous random variables then the maximum number of linear interactions that may be estimated is given by $r \times q$ and the maximum number of estimated quadratic interactions is given by $r \times q(q+1)/2$. Clearly, for moderately sized p the situation is likely to be over-parametrized.

Here we consider hierarchical models (see page 7), i.e. we assume the following rules for the interaction expansions in a CG model: let $A \subset \Delta$, then, for all $B \supset A$ and $\gamma, \zeta \in \Gamma$

1. zero discrete interactions $\{\lambda_A(i_A)\}$ imply that $\{\lambda_B(i_B)\}$, $\{\eta_{\gamma;B}(i_B)\}$ and $\{\psi_{\gamma\zeta;B}(i_B)\}$ are also zero;
2. zero linear interactions $\{\eta_{\gamma;A}(i_A)\}$ imply that $\{\eta_{\gamma;B}(i_B)\}$ and $\{\psi_{\gamma\zeta;B}(i_B)\}$ are also zero;
3. zero quadratic interactions $\{\psi_{\gamma\gamma;A}(i_A)\}$ imply that $\{\psi_{\gamma\gamma;B}(i_B)\}$ are also zero and zero $\{\psi_{\gamma\zeta;A}(i_A)\}$ imply that $\{\psi_{\gamma\zeta;B}(i_B)\}$ are also zero.

2.1.3 The generating class of a conditional Gaussian model

DEFINITION 2.2 *Consider a CG model, $\mathcal{M}$, defined on a set of k random variables, X_V, where X_V is partitioned into a set of p discrete random variables X_Δ and a set of q continuous random variables X_Γ ($V = \Delta \cup \Gamma$, $k = p+q$). The generating class of $\mathcal{M}$ is defined to be the set of interactions K consisting of the union of three sets of distinct types of maximal subsets of interactions between the full set of random variables given by*

1. *a discrete set D consisting of all distinct maximal subsets of interactions between the discrete random variables belonging to X_Δ,*
2. *a linear set L consisting of all distinct maximal subsets of interactions between discrete random variables belonging to X_Δ and individual continuous random variables belonging to X_Γ,*

3. *a quadratic set Q consisting of all distinct maximal subsets of interactions between discrete random variables belonging to X_Δ and two continuous random variables belonging to X_Γ.*

We shall assume the interactions contained in K are hierarchical. In order to simplify model definition we adopt Edwards' (1995) notation for 'hierarchical interaction models'. (Edwards' definition of a hierarchical interaction model is simply another name for the CG interaction parametrization with hierarchical constraints on the interactions.) Suppose the maximal set of discrete interactions is represented by terms $d_1, \ldots, d_r$, the maximal set of linear interactions is represented by terms $l_1, \ldots, l_s$ and the maximal set of quadratic interactions is represented by $q_1, \ldots, q_t$. We call the maximal sets of discrete, linear and quadratic interactions the *generators* of the model. The CG model may then be represented by a list of maximal discrete, linear and quadratic model generators. Generators within the discrete, linear or quadratic part of the model are separated by commas, and the discrete linear and quadratic parts of the model are separated by a forward slash, '/', i.e. the CG model $\mathcal{M}$ may be represented by

$$d_1, \ldots, d_r \;/\; l_1, \ldots, l_s \;/\; q_1, \ldots, q_t. \tag{2.14}$$

The discrete part specifies the expansion of $\alpha(i)$ in terms of maximal interactions among the discrete variables. The linear part specifies the expansion of $\beta(i)$ in terms of maximal interactions among the discrete variables and one continuous variable. The quadratic part specifies the expansion of maximal interactions among the discrete variables and (at most) two continuous variables. Rather than writing down the actual interactions in (2.14) we simply use the names of the random variables. (Single uppercase characters are used to denote random variables when defining CG models throughout this thesis.) The constraint that the CG models are hierarchical defines the syntax for the shorthand in an obvious way. For further details regarding model syntax see Edwards (1995). The following examples show the use of Edwards' notation. (Note that in the examples λ is a constant term.)

EXAMPLE 2.1 *Three CG models involving two discrete and one continuous variable.*

Let $\Delta = \{A, B\}$, $\Gamma = \{Y\}$ and let A and B be indexed by i_1 and i_2 respectively. From (2.13) the CG density is given by

$$\begin{aligned} f(i_1, i_2, y) \;=\; \exp\Big\{ & \lambda + \lambda_{A}(i_1) + \lambda_{B}(i_2) + \lambda_{AB}(i_1, i_2) \\ & + \Big[\eta_{Y} + \eta_{Y;A}(i_1) + \eta_{Y;B}(i_2) + \eta_{Y;AB}(i_1, i_2)\Big] y \\ & - \Big[\psi_{YY} + \psi_{YY;A}(i_1) + \psi_{YY;B}(i_2) + \psi_{YY;AB}(i_1, i_2)\Big] y^2 \;/\; 2 \Big\} \end{aligned}$$

then it may be represented in Edwards' shorthand notation as $AB \;/\; ABY \;/\; ABY$. From the above density, if we now specify that Y only varies with the value of A, i.e. $B \perp Y \mid A$, then the density is defined by

$$\begin{aligned} f(i_1, i_2, y) \;=\; \exp\Big\{ & \lambda + \lambda_{A}(i_1) + \lambda_{B}(i_2) + \lambda_{AB}(i_1, i_2) \\ & + \Big[\eta_{Y} + \eta_{Y;A}(i_1)\Big] y - \Big[\psi_{YY} + \psi_{YY;A}(i_1)\Big] y^2 \;/\; 2 \Big\} \end{aligned}$$

and may be represented as $AB \;/\; AY \;/\; AY$. Alternatively, from the first density in this example, the conditional independence constraint $A \perp B \mid Y$ assumes a density

$$\begin{aligned} f(i_1, i_2, x) &= \exp\Big\{\lambda + \lambda_A(i_1) + \lambda_B(i_2) + \Big[\eta_Y + \eta_{Y;A}(i_1) + \eta_{Y;B}(i_2)\Big]y \\ &\quad - \Big[\psi_{YY} + \psi_{YY;A}(i_1) + \psi_{YY;B}(i_2)\Big]y^2 \;/\; 2\Big\}, \end{aligned}$$

which may be represented by $A, B \;/\; AY, BY \;/\; AY, BY$ using Edwards' notation.

Edwards abbreviates the definition of a CG model further by allowing more than two continuous variables to appear together in the list of quadratic interactions. However, only pairwise interactions are implied between any number of continuous variables appearing together. This is illustrated by the following two examples.

EXAMPLE 2.2 *Two CG models involving three continuous random variables.*

Let $\Delta = \{\varnothing\}$ and let $\Gamma = \{X, Y, Z\}$ then the density

$$\begin{aligned} f(x, y, z) &= \exp\Big\{\lambda + \eta_X x + \eta_Y y + \eta_Z z - (\psi_{XX}x^2 + \psi_{YY}y^2 + \psi_{ZZ}z^2) \;/\; 2 \\ &\quad - (\psi_{XY}xy + \psi_{XZ}xz + \psi_{YZ}yz)\Big\} \end{aligned}$$

may be represented by $// \; XYZ$. The conditional independence constraint $X \perp Y \mid Z$ assumes the following density

$$\begin{aligned} f(x, y, z) &= \exp\Big\{\lambda + \eta_X x + \eta_Y y + \eta_Z z - (\psi_{XX}x^2 + \psi_{YY}y^2 + \psi_{ZZ}z^2) \;/\; 2 \\ &\quad - (\psi_{XZ}xz + \psi_{YZ}yz)\Big\} \end{aligned}$$

and may be represented by $// \; XZ, YZ$.

EXAMPLE 2.3 *A CG model involving four continuous random variables.*

Let $\Delta = \{\varnothing\}$ and let $\Gamma = \{W, X, Y, Z\}$ then the density given the conditional independence constraints $W \perp X \mid Z$ and $X \perp Y \mid Z$ is given by

$$\begin{aligned} f(w, x, y, z) &= \exp\Big\{\lambda + \eta_W w + \eta_X x + \eta_Y y + \eta_Z z \\ &\quad - (\psi_{WW}w^2 + \psi_{XX}x^2 + \psi_{YY}y^2 + \psi_{ZZ}z^2) \;/\; 2 \\ &\quad - (\psi_{WY}wy + \psi_{WZ}wz + \psi_{YZ}yz + \psi_{XZ}xz)\Big\} \end{aligned}$$

and may be represented by $// \; WYZ, XZ$.

EXAMPLE 2.4 *Three CG models involving two discrete and two continuous random variables.*

Let $\Delta = \{A, B\}$ and let $\Gamma = \{Y, Z\}$ and let A and B be indexed by i_1 and i_2, respectively. The density for the saturated model is given by

$$\begin{aligned} f(i_1, i_2, y, z) \;=\; \exp\Big\{ & \lambda + \lambda_{A}(i_1) + \lambda_{B}(i_2) + \lambda_{AB}(i_1, i_2) \\ & + \Big[\eta_{Y} + \eta_{Y;A}(i_1) + \eta_{Y;B}(i_2) + \eta_{Y;AB}(i_1, i_2)\Big] y \\ & + \Big[\eta_{Z} + \eta_{Z;A}(i_1) + \eta_{Z;B}(i_2) + \eta_{Z;AB}(i_1, i_2)\Big] z \\ & - \Big[\psi_{YY} + \psi_{YY;A}(i_1) + \psi_{YY;B}(i_2) + \psi_{YY;AB}(i_1, i_2)\Big] y^2 \,/\, 2 \\ & - \Big[\psi_{YZ} + \psi_{YZ;A}(i_1) + \psi_{YZ;B}(i_2) + \psi_{YZ;AB}(i_1, i_2)\Big] yz \\ & - \Big[\psi_{ZZ} + \psi_{ZZ;A}(i_1) + \psi_{ZZ;B}(i_2) + \psi_{ZZ;AB}(i_1, i_2)\Big] z^2 \,/\, 2 \Big\}, \end{aligned}$$

may be written as $AB \;/\; ABY, ABZ \;/\; ABYZ$. If we now assume that $A \perp B \mid \{Y, Z\}$ then the density reduces to

$$\begin{aligned} f(i_1, i_2, y, z) \;=\; \exp\Big\{ & \lambda + \lambda_{A}(i_1) + \lambda_{B}(i_2) \\ & + \Big[\eta_{Y} + \eta_{Y;A}(i_1) + \eta_{Y;B}(i_2)\Big] y \\ & + \Big[\eta_{Z} + \eta_{Z;A}(i_1) + \eta_{Z;B}(i_2)\Big] z \\ & - \Big[\psi_{YY} + \psi_{YY;A}(i_1) + \psi_{YY;B}(i_2)\Big] y^2 \,/\, 2 \\ & - \Big[\psi_{YZ} + \psi_{YZ;A}(i_1) + \psi_{YZ;B}(i_2)\Big] yz \\ & - \Big[\psi_{ZZ} + \psi_{ZZ;A}(i_1) + \psi_{ZZ;B}(i_2)\Big] z^2 \,/\, 2 \Big\} \end{aligned}$$

and may be written $A, B \;/\; AY, BY, AZ, BZ \;/\; AYZ, BYZ$. If both A, B and Y, Z independence is assumed then we lose the yz terms in the above density giving

$$\begin{aligned} f(i_1, i_2, y, z) \;=\; \exp\Big\{ & \lambda + \lambda_{A}(i_1) + \lambda_{B}(i_2) \\ & + \Big[\eta_{Y} + \eta_{Y;A}(i_1) + \eta_{Y;B}(i_2)\Big] y + \Big[\eta_{Z} + \eta_{Z;A}(i_1) + \eta_{Z;B}(i_2)\Big] z \\ & - \Big[\psi_{YY} + \psi_{YY;A}(i_1) + \psi_{YY;B}(i_2)\Big] y^2 \,/\, 2 \\ & - \Big[\psi_{ZZ} + \psi_{ZZ;A}(i_1) + \psi_{ZZ;B}(i_2)\Big] z^2 \,/\, 2, \Big\} \end{aligned}$$

which is succinctly written $A, B \;/\; AY, BY, AZ, BZ \;/\; AY, BY, AZ, BZ$.

We adopt Edwards' notation for CG models throughout the remainder of this thesis.

2.2 Graphical and non-graphical CG models

DEFINITION 2.3 *The conditional independence graph of a set of k random variables, X_V, where X_V is partitioned into a set of p discrete random variables, X_Δ, and a set of q continuous random variables, X_Γ ($V = \Delta \cup \Gamma$, $k = p + q$), is the graph $G = G(V, E)$ with vertex set V and (i, j) is not in the edge set, E, if and only if $X_i \perp X_j \mid X_{V\setminus\{i,j\}}$. In addition, the vertices in G corresponding to the discrete random variables are represented by dots and the vertices in G corresponding to the continuous random variables are represented by circles.*

DEFINITION 2.4 *Consider a set of k random variables, X_V, defined on the vertices of a graph G, where X_V is partitioned into a set of p discrete random variables, X_Δ, and a set of q continuous random variables, X_Γ ($V = \Delta \cup \Gamma$, $k = p + q$). The set of cliques in G is given by $C = C_1 \cup C_2 \cup C_3$, where C_1 is the set of cliques in G for all $A \subset \Delta$, C_2 is the set of cliques in G for all $A \cup \gamma$ ($A \subset \Delta$ and $\gamma \in \Gamma$) and C_3 is the set of cliques in G for all $A \cup \{\gamma, \zeta\}$ ($A \subset \Delta$ and $\gamma, \zeta \in \Gamma$). A CG model, $\mathcal{M}$, defined on X_V with generating class K is called a graphical CG model if and only if $C \equiv K$.*

Thus, a non-graphical model is one in which the cliques of a graph G are not exactly the same as the generating class K, that is, for the set of k random variables, X_V, defined on the vertices of a graph G and a CG model $\mathcal{M}$ defined on X_V.

EXAMPLE 2.5 *A graphical CG model involving two binary variables and one continuous variable.*

Let $\Delta = \{A, B\}$, $\Gamma = \{Y\}$ and let A and B be indexed by i_1 and i_2 respectively. If A and B are both binary there will be four cells. The saturated graphical model consists of 11 fitted parameters, i.e. 3 discrete since $\sum p(i) = 1$, 4 linear and 4 quadratic. The density for this model is given by

$$\begin{aligned} f(i_1, i_2, y) \;=\; & \exp[\lambda + \lambda_A(i_1) + \lambda_B(i_2) + \lambda_{AB}(i_1, i_2) \\ & + \{\eta_Y + \eta_{Y;A}(i_1) + \eta_{Y;B}(i_2) + \eta_{Y;AB}(i_1, i_2)\}y \\ & - \tfrac{1}{2}\{\psi_{YY} + \psi_{YY;A}(i_1) + \psi_{YY;B}(i_2) + \psi_{YY;AB}(i_1, i_2)\}y^2], \end{aligned}$$

which is formed by expanding equation (2.13) in elements of Δ and Γ. Using (2.14) the model may be represented by AB / ABY / ABY. The conditional independence graph of the model is shown in Figure 2.1. If we impose the conditional independence constraint $A \perp B \mid Y$ then we remove three parameters (assuming A and B are binary) from the model, i.e. the density for the model is now given by

$$\begin{aligned} f(i_1, i_2, y) \;=\; & \exp[\lambda + \lambda_A(i_1) + \lambda_B(i_2) \\ & + \{\eta_Y + \eta_{Y;A}(i_1) + \eta_{Y;B}(i_2)\}y \\ & - \tfrac{1}{2}\{\psi_{YY} + \psi_{YY;A}(i_1) + \psi_{YY;B}(i_2)\}y^2]. \end{aligned}$$

Using (2.14) we may abbreviate this expansion to A, B / AY, BY / AY, BY. The corresponding conditional independence graph is shown in Figure 2.2.

The independence graph is useful because it summarizes the conditional independence structure of a graphical CG model. By application of the pairwise Markov property (see Section 1.4.1) the following statements made about any nonadjacent nodes or pair of variables from the set of variables $\Delta \cup \Gamma$ in the conditional independence graph G of a graphical CG model are equivalent:

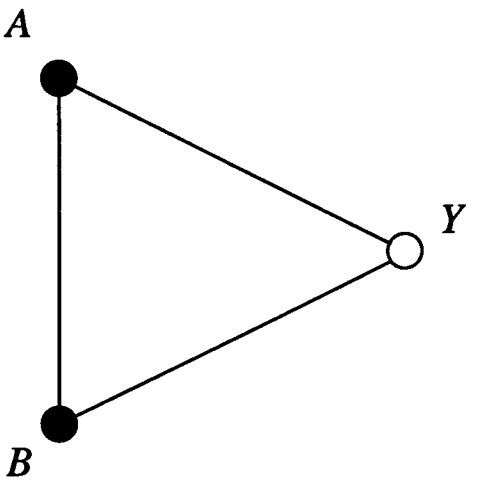


Figure 2.1 The conditional independence graph for the saturated graphical model for discrete random variables $\{A, B\}$ and continuous random variable $\{Y\}$.

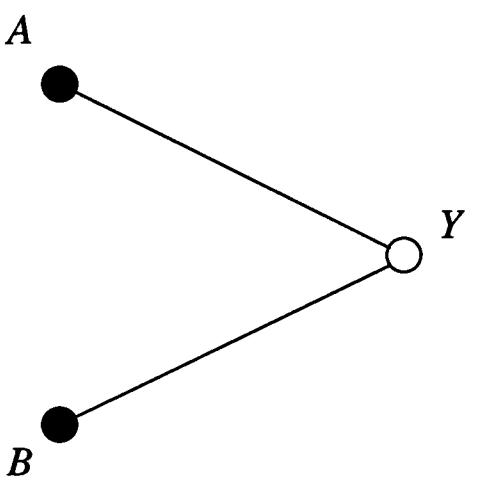


Figure 2.2 The conditional independence graph for discrete random variables $\{A, B\}$ and continuous random variable $\{Y\}$, where $A \perp B \mid Y$.

1. the variable pair is conditionally independent given all the remaining variables,
2. all interactions involving the variable pair are equal to zero in the interaction parametrization of the CG distribution,
3. the edge of the variable pair is missing in the conditional independence graph

(see Wermuth & Lauritzen, 1990).

2.2.1 Graphical homogeneous CG models

DEFINITION 2.5 *A homogeneous CG model, $\mathcal{M}_{\mathcal{H}}$, is a CG model, $\mathcal{M}$, with generating class $K = D \cup L \cup Q$ in which there are no discrete random variables in Q.*

The conditional independence graph giving rise to a graphical HCG or graphical CG model still allows us to read off the exact set of conditional independencies among the variables. Thus, the term graphical is not out of place here. However, the conditional independence graph no longer identifies the set of model interactions unambiguously due to the absence of discrete and continuous interactions in Q when discrete and continuous variables are non-independent. We can get around this problem by simply stating whether the model implied by the conditional independence graph is graphical CG or graphical HCG.

2.2.2 *Non-graphical CG models*

A graphical model allows us to read the complete set of model interactions from its associated independence graph. This is because the cliques of the graph correspond directly with the model generators. However, it is often necessary to fit non-graphical models as these provide the best fit to the data. In this case the associated independence graph still provides a useful description of data structure even though it does not give us as complete a model description as we might like. We will use non-graphical models extensively in Chapter 4.

2.3 Likelihood, gradient and Hessian

In this section we look at the specification of the likelihood for saturated CG models, sub-model specification and Hessian.

2.3.1 *The likelihood for the general saturated CG model*

Recall the expression for the CG density (2.2) given on page 13. To simplify the algebra and subsequent estimation procedure we replace $\alpha(i)$ by $\varphi(i)$ in density (2.2) and include κ as a normalizing constant, i.e. we now express the CG density as

$$f(x) = f(i,y) = \kappa \exp\left\{\varphi(i) + \beta(i)'y - \tfrac{1}{2}y'\Omega(i)y\right\}. \tag{2.15}$$

Here we define $\varphi(i)$ as the expansion over all subsets $A \subset \Delta$ excluding the normalizing constant λ, i.e.

$$\varphi(i) = \sum_{A\subset\Delta:\, A\neq\varnothing} \lambda_A(i_A). \tag{2.16}$$

In the above expression the summation is taken as being over all subsets $A \subset \Delta$ excluding A equal to the empty set. For example, if $\Delta = \{A, B, C\}$ and A, B and C are indexed by i_1, i_2 and i_3, respectively, then $\varphi(i)$ is expanded as follows:

$$\begin{aligned}\varphi(i_1,i_2,i_3) &= \lambda_A(i_1) + \lambda_B(i_2) + \lambda_C(i_3) \\ &\quad + \lambda_{AB}(i_1,i_2) + \lambda_{AC}(i_1,i_3) + \lambda_{BC}(i_2,i_3) \\ &\quad + \lambda_{ABC}(i_1,i_2,i_3).\end{aligned}$$

Note that in (2.16) $\varphi(i)$ equals $\alpha(i) - \lambda$ and this change does not alter the estimated values of $\beta_\gamma(i)$ and $\omega_{\gamma\zeta}(i)$ because of the presence of κ. Also, the interaction expansions for $\beta(i)$ and $\Omega(i)$ remain unchanged and are as defined in Equations (2.11) and (2.12) on page 14. (We can still use Edwards' notation to define the discrete, linear and quadratic parts of the model without any loss of generality. It is simply necessary to remember that the discrete part of the model does not include the $\lambda = \lambda_\varnothing$ normalizing constant.)

The value of κ in Equation (2.15) is obtained by integrating $f(i,y)$ over y to obtain the marginal density $f(i)$, i.e.

$$f(i) = \int_{\mathbb{R}^q} f(i,y)dy = \kappa(2\pi)^{q/2}|\Omega(i)|^{-1/2} \exp\left\{\varphi(i) + \tfrac{1}{2}\beta(i)'\Omega(i)^{-1}\beta(i)\right\}. \tag{2.17}$$

The above result is obtained by re-writing $f(i, y)$ as

$$f(i,y) = \exp\left[\varphi(i)^* - \tfrac{1}{2}\{y-\mu(i)\}'\Sigma(i)^{-1}\{y-\mu(i)\}\right],$$

where

$$\varphi(i)^* = \varphi(i) + \tfrac{1}{2}\beta(i)'\Omega(i)^{-1}\beta(i).$$

Setting $z = y - \mu(i)$ and integrating over $z \in \mathbb{R}^q$ where $\Omega(i)$ is a positive-definite (symmetric) matrix, then

$$\int_{\mathbb{R}^q} \exp\{-z'\Omega(i)z/2\}dz = (2\pi)^{q/2}|\Omega(i)|^{-1/2}$$

(see also Lauritzen, 1996, p. 159). From (2.17), noting that $\sum_{i\in\mathcal{I}} f(i) = 1$, we obtain

$$\kappa = \frac{1}{(2\pi)^{q/2}\sum_{i\in\mathcal{I}}\left[|\Omega(i)|^{-1/2}\exp\left\{\varphi(i) + \tfrac{1}{2}\beta(i)'\Omega(i)^{-1}\beta(i)\right\}\right]}. \tag{2.18}$$

Let θ denote the combined vector of unknown parameters composed of scalar parameters $\{\varphi(i)\}_{i\in\mathcal{I}}$, $q\times 1$ vectors $\{\beta(i)\}_{i\in\mathcal{I}}$ and $q(q+1)/2\times 1$ vectors $\{\text{svec}[\Omega(i)]\}_{i\in\mathcal{I}}$. (Here 'svec[$\Omega(i)$]' denotes the $q(q+1)/2\times 1$ vector obtained by stacking the lower-triangular elements of $\Omega(i)$ one underneath the other in columnwise fashion.) We assume that the vectors $\{\beta(i)\}_{i\in\mathcal{I}}$ are real valued and that the $\{\Omega(i)\}_{i\in\mathcal{I}}$ are positive-definite (symmetric) $q\times q$ matrices. Also, let ν index an observation $x^{(\nu)} = (i', y^{(\nu)'})'$ so that $y^{(\nu)} = (y_1^{(\nu)}, \ldots, y_q^{(\nu)})'$ denotes a $q\times 1$ vector of continuous observations belonging to cell i. From (2.15) we obtain the following expression for the likelihood

$$\text{lik}(\theta; x) = \prod_{i\in\mathcal{I}}\prod_{\nu=1}^{n(i)} \kappa\exp\left\{\varphi(i) + \beta(i)'y^{(\nu)} - \tfrac{1}{2}y^{(\nu)'}\Omega(i)y^{(\nu)}\right\}. \tag{2.19}$$

Taking the logarithm of the above expression gives

$$\begin{aligned} L(\theta;x) &= \sum_{i\in\mathcal{I}}\sum_{\nu=1}^{n(i)}\left\{\log\kappa + \varphi(i) + \beta(i)'y^{(\nu)} - \tfrac{1}{2}y^{(\nu)'}\Omega(i)y^{(\nu)}\right\} \\ &= n\log\kappa + \sum_{i\in\mathcal{I}}\left[n(i)\varphi(i) + \beta(i)'\sum_{\nu=1}^{n(i)}y^{(\nu)} - \tfrac{1}{2}\text{tr}\left\{\Omega(i)\sum_{\nu=1}^{n(i)}y^{(\nu)}y^{(\nu)'}\right\}\right] \\ &= n\log\kappa + \sum_{i\in\mathcal{I}}\left[n(i)\varphi(i) + \beta(i)'t(i) - \tfrac{1}{2}\text{tr}\left\{\Omega(i)S(i)\right\}\right], \end{aligned} \tag{2.20}$$

where $n = \sum_{i\in\mathcal{I}} n(i)$ is the total number of observations, $t(i) = \sum_{\nu=1}^{n(i)} y^{(\nu)}$ is a vector of totals for cell i, and $S(i) = \sum_{\nu=1}^{n(i)} y^{(\nu)}y^{(\nu)'}$ is an uncorrected sums of squares and products (SSP) matrix for cell i. In the above expression we use 'tr' to denote the trace of a matrix.

2.3.2 *The general saturated homogeneous CG model*

The HCG density is given by

$$f(x) = f(i,y) = \kappa \exp\left\{\varphi(i) + \beta(i)'y - \tfrac{1}{2}y'\Omega y\right\}, \tag{2.21}$$

where Ω is a common concentration matrix. We obtain κ by integration over the continuous variables and summation over i as in (2.18) but with common concentration matrix, i.e.

$$\kappa = \frac{1}{(2\pi)^{q/2}|\Omega|^{-1/2}\sum_{i\in\mathcal{I}}\left[\exp\left\{\varphi(i) + \tfrac{1}{2}\beta(i)'\Omega^{-1}\beta(i)\right\}\right]}. \tag{2.22}$$

The log-likelihood is given by

$$L(\theta;x) = n\log\kappa - |\mathcal{I}|\mathrm{tr}\left\{\Omega S\right\}/2 + \sum_{i\in\mathcal{I}}\left\{\varphi(i)n(i) + \beta(i)'t(i)\right\}, \tag{2.23}$$

where $|\mathcal{I}|$ equals the number of cells $i \in \mathcal{I}$ and $S = \sum_{\nu=1}^{n} y^{(\nu)}y^{(\nu)'}$ is an uncorrected SSP matrix based on all continuous observations.

2.3.3 *Sub-model specification*

DEFINITION 2.6 *A CG sub-model is a non-saturated hierarchical CG or HCG model.*

As the definition states, the class of sub-models is restricted by ensuring that the models defined by setting parameters in the saturated model equal to zero are hierarchical, i.e. we ensure that all higher-order interactions containing zero lower-order interactions are also set equal to zero. This makes interpretation easier, however, it is not always the case that we want to examine only hierarchical models. For example, two drugs may exhibit little or no effect when taken separately but (although rarely found in practice) their combined effect may be highly significant. Thus, in such situations it may be desirable to consider non-hierarchical models. We restrict our attention to hierarchical models as non-hierarchical models are beyond the scope of this thesis.

EXAMPLE 2.6 *CG models involving two discrete and one continuous random variable.*

Let $\Delta = \{A,B\}$, $\Gamma = \{Y\}$ and let A and B be indexed by $i_1 = 1,\ldots,I_1$ and $i_2 = 1,\ldots,I_2$, respectively (I_1 need not be equal to I_2). From (2.15) the log-density for the saturated model is given by

$$\begin{aligned}\log f(i_1,i_2,y) &= \log\kappa + \lambda_A(i_1) + \lambda_B(i_2) + \lambda_{AB}(i_1,i_2)\\ &\quad + \{\eta_Y + \eta_{Y;A}(i_1) + \eta_{Y;B}(i_2) + \eta_{Y;AB}(i_1,i_2)\}y\\ &\quad - \tfrac{1}{2}\{\psi_{YY} + \psi_{YY;A}(i_1) + \psi_{YY;B}(i_2) + \psi_{YY;AB}(i_1,i_2)\}y^2,\end{aligned}$$

where,

$$\log\kappa = -\tfrac{1}{2}\log(2\pi) - \log\sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2}\left[\left\{\frac{1}{\sqrt{\omega_{YY}(i_1,i_2)}}\right\}\exp\left\{\varphi(i_1,i_2) + \frac{\beta_Y^2(i_1,i_2)}{2\omega_{YY}(i_1,i_2)}\right\}\right]$$

from (2.18). We assume that the first level of each of the factors does not define an interaction so that $\lambda_A(i_1 = 1)$, $\lambda_B(i_2 = 1)$, $\lambda_{AB}(i_1 = 1, i_2 = 1)$, $\eta_{Y;A}(i_1 = 1)$, $\eta_{Y;B}(i_2 = 1)$, $\eta_{Y;AB}(i_1 = 1, i_2 = 1)$, $\psi_{YY;A}(i_1 = 1)$, $\psi_{YY;B}(i_2 = 1)$ and $\psi_{YY;AB}(i_1 = 1, i_2 = 1)$ do not exist. Thus, the discrete, linear and quadratic parameters are expanded as follows: first, for the discrete parameters we have

$$\begin{aligned}
\varphi(2,1) &= \lambda_A(2) \\
\varphi(1,2) &= \lambda_B(2) \\
\varphi(2,2) &= \lambda_A(2) + \lambda_B(2) + \lambda_{AB}(2,2) \\
&\vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \\
\varphi(i_1,i_2) &= \lambda_A(i_1) + \lambda_B(i_2) + \lambda_{AB}(i_1,i_2) \\
&\vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \\
\varphi(I_1,I_2) &= \lambda_A(I_1) + \lambda_B(I_2) + \lambda_{AB}(I_1,I_2)
\end{aligned}$$

second, for the linear parameters we have

$$\begin{aligned}
\beta_Y(1,1) &= \eta_Y \\
\beta_Y(2,1) &= \eta_Y + \eta_{Y;A}(2) \\
\beta_Y(1,2) &= \eta_Y + \eta_{Y;B}(2) \\
\beta_Y(2,2) &= \eta_Y + \eta_{Y;A}(2) + \eta_{Y;B}(2) + \eta_{Y;AB}(2,2) \\
&\vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \\
\beta_Y(i_1,i_2) &= \eta_Y + \eta_{Y;A}(i_1) + \eta_{Y;B}(i_2) + \eta_{Y;AB}(i_1,i_2) \\
&\vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \\
\beta_Y(I_1,I_2) &= \eta_Y + \eta_{Y;A}(I_1) + \eta_{Y;B}(I_2) + \eta_{Y;AB}(I_1,I_2)
\end{aligned}$$

and third, for the quadratic parameters we have

$$\begin{aligned}
\omega_{YY}(1,1) &= \psi_{YY} \\
\omega_{YY}(2,1) &= \psi_{YY} + \psi_{YY;A}(2) \\
\omega_{YY}(1,2) &= \psi_{YY} + \psi_{YY;B}(2) \\
\omega_{YY}(2,2) &= \psi_{YY} + \psi_{YY;A}(2) + \psi_{YY;B}(2) + \psi_{YY;AB}(2,2) \\
&\vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \\
\omega_{YY}(i_1,i_2) &= \psi_{YY} + \psi_{YY;A}(i_1) + \psi_{YY;B}(i_2) + \psi_{YY;AB}(i_1,i_2) \\
&\vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \\
\omega_{YY}(I_1,I_2) &= \psi_{YY} + \psi_{YY;A}(I_1) + \psi_{YY;B}(I_2) + \psi_{YY;AB}(I_1,I_2)
\end{aligned}$$

In the above expansions we write φ, β and ω together with the cell values given by the 2×1 vector $i = (i_1, i_2)'$, where i_1 (the first element in i) gives the value of A and i_2 (the second element in i) gives the value of B. If the individual interaction depends on just A then only the value of i_1 appears in parentheses. Similarly, if the individual interaction depends on just B then only the value of i_2 is given. Of course, if the interaction depends on both A and B values then the value of both levels are given in the order i_1 then i_2.

If we now specify that A is independent of B given Y then the log-density becomes

$$\begin{aligned}
\log f(i_1,i_2,y) \;=\; & \log\kappa + \lambda_A(i_1) + \lambda_B(i_2) \\
& + \{\eta_Y + \eta_{Y;A}(i_1) + \eta_{Y;B}(i_2)\}y \\
& - \tfrac{1}{2}\{\psi_{YY} + \psi_{YY;A}(i_1) + \psi_{YY;B}(i_2)\}y^2,
\end{aligned}$$

which removes all terms involving AB so that the interaction expansions are now given by

$$
\begin{array}{lllll}
\varphi(2,1) &= \lambda_A(2), & & \beta_Y(1,1) &= \eta_Y \\
\varphi(1,2) &= & \lambda_B(2), & \beta_Y(2,1) &= \eta_Y + \eta_{Y;A}(2) \\
\varphi(2,2) &= \lambda_A(2) + & \lambda_B(2), & \beta_Y(1,2) &= \eta_Y + \eta_{Y;B}(2) \\
\vdots & \vdots & \vdots & \beta_Y(2,2) &= \eta_Y + \eta_{Y;A}(2) + \eta_{Y;B}(2) \\
\varphi(i_1,i_2) &= \lambda_A(i_1) + & \lambda_B(i_2), & \vdots & \vdots \quad \vdots \quad \vdots \\
\vdots & \vdots & \vdots & \beta_Y(i_1,i_2) &= \eta_Y + \eta_{Y;A}(i_1) + \eta_{Y;B}(i_2) \\
\varphi(I_1,I_2) &= \lambda_A(I_1) + & \lambda_B(I_2), & \vdots & \vdots \quad \vdots \quad \vdots \\
 & & & \beta_Y(I_1,I_2) &= \eta_Y + \eta_{Y;A}(I_1) + \eta_{Y;B}(I_2)
\end{array}
$$

and

$$
\begin{array}{ll}
\omega_{YY}(1,1) &= \psi_{YY} \\
\omega_{YY}(2,1) &= \psi_{YY} + \psi_{YY;A}(2) \\
\omega_{YY}(1,2) &= \psi_{YY} + \psi_{YY;B}(2) \\
\omega_{YY}(2,2) &= \psi_{YY} + \psi_{YY;A}(2) + \psi_{YY;B}(2) \\
\vdots & \vdots \quad \vdots \quad \vdots \\
\omega_{YY}(i_1,i_2) &= \psi_{YY} + \psi_{YY;A}(i_1) + \psi_{YY;B}(i_2) \\
\vdots & \vdots \quad \vdots \quad \vdots \\
\omega_{YY}(I_1,I_2) &= \psi_{YY} + \psi_{YY;A}(I_1) + \psi_{YY;B}(I_2)
\end{array}
$$

i.e. $A \perp B \mid Y$ specifies $\lambda_{AB}(i_1,i_2) = \eta_{Y;AB}(i_1,i_2) = \omega_{Y;AB}(i_1,i_2) = 0$. We can of course specify other conditional independence constraints, e.g. $B \perp Y \mid A$, which would have the effect of removing all interactions containing BY.

From (2.20) we obtain the log-likelihood as

$$
L(\theta;x) = n\log\kappa + \sum_{i\in\mathcal{I}}\left[\varphi(i)n(i) + \{\beta_Y(i)\}'t(i) - \tfrac{1}{2}\omega_{YY}(i)s(i)\right], \tag{2.24}
$$

where $n(i)$, $t(i)$ and $s(i)$ are the relevant counts, totals and uncorrected sum of squares for each cell $i \in \mathcal{I}$. For the saturated model we simply substitute the first set of interactions into Equation (2.24) and for the sub-model, specified by the conditional independence constraint $A \perp B \mid Y$, the second set of interaction expansions. We obtain κ by summation over i_1 and i_2 as in Equation (2.18) again by substituting in the the relevant set of interaction expansions.

Higher-order interactions

So far we have only looked at a simple models containing a small number of variables. We now need to describe models that contain more terms. Perhaps the hardest situation to deal with, from a computational perspective, is when we have a large number of factors. The continuous interactions are handled fairly easily. The situation is best illustrated using an example. Rather than writing out the interaction expansions in full we shall simplify the interaction expansions by only using the variable label and drop the explicit use of i. Let us assume that we have three factors A, B, C and one continuous variable Y. An interaction between two or more variables is assumed if they appear together, e.g. AB indicates an interaction between factors A and B. (We shall only denote random variables using a single letter from the alphabet.) For example, let

$$
\begin{aligned}
\beta_{Y} &= Y \\
\beta_{Y;A} &= Y + AY \\
\beta_{Y;B} &= Y + BY \\
\beta_{Y;C} &= Y + CY \\
\beta_{Y;AB} &= Y + AY + BY + ABY \\
\beta_{Y;AC} &= Y + AY + CY + ACY \\
\beta_{Y;BC} &= Y + BY + CY + BCY \\
\beta_{Y;ABC} &= Y + AY + BY + ABY + CY + ACY + BCY + ABCY
\end{aligned}
$$

be the full set of interaction expansions for the linear parameters $\{\beta(i)\}_{i\in\mathcal{I}}$ $(\gamma \in \Gamma)$ in a saturated model. Note that the Y appearing on its own is a constant. If we now specify that $A \perp B \mid Y$ then this removes all interactions between A and B from the above expansions, i.e. we are left with

$$
\begin{aligned}
\beta_{Y} &= Y \\
\beta_{Y;A} &= Y + AY \\
\beta_{Y;B} &= Y + BY \\
\beta_{Y;C} &= Y + CY \\
\beta_{Y;AB} &= Y + AY + BY \\
\beta_{Y;AC} &= Y + AY + CY + ACY \\
\beta_{Y;BC} &= Y + BY + CY + BCY \\
\beta_{Y;ABC} &= Y + AY + BY + CY + ACY + BCY
\end{aligned}
$$

so that $\beta_{Y;AB} = \beta_{Y;A} + \beta_{Y;B} - \beta_{Y}$ and $\beta_{Y;ABC} = \beta_{Y;AC} + \beta_{Y;BC} - \beta_{Y;C}$. (In fact, it is these linear relationships among the parameters, when we define conditional independence relationships, that is exploited in the fitting procedure. Thus, enabling the relevant parameters, in this case the linear parameters, to be estimated directly. In particular, the model for no three-factor interaction is quite useful and is fitted via $\beta_{Y;ABC} = \beta_{Y;AB} + \beta_{Y;AC} + \beta_{Y;BC} - \beta_{Y;A} - \beta_{Y;B} - \beta_{Y;C} + \beta_{Y}$.) If, in addition to $A \perp B \mid C, Y$, we specify $A \perp C \mid B, Y$ then the interaction expansions reduce to

$$
\begin{aligned}
\beta_{Y} &= Y \\
\beta_{Y;A} &= Y + AY \\
\beta_{Y;B} &= Y + BY \\
\beta_{Y;C} &= Y + CY \\
\beta_{Y;AB} &= Y + AY + BY \\
\beta_{Y;AC} &= Y + AY + CY \\
\beta_{Y;BC} &= Y + BY + CY + BCY \\
\beta_{Y;ABC} &= Y + AY + BY + CY + BCY
\end{aligned}
$$

and $\beta_{Y;AC} = \beta_{Y;A} + \beta_{Y;C} - \beta_{Y}$ and $\beta_{Y;ABC} = \beta_{Y;BC} + \beta_{Y;A} - \beta_{Y}$. If $A \perp B \mid C, Y$; $A \perp C \mid B, Y$ and $B \perp C \mid A, Y$ then the interaction expansions are given by

$$
\begin{aligned}
\beta_{Y} &= Y \\
\beta_{Y;A} &= Y + AY \\
\beta_{Y;B} &= Y + BY \\
\beta_{Y;C} &= Y + CY \\
\beta_{Y;AB} &= Y + AY + BY \\
\beta_{Y;AC} &= Y + BY + CY \\
\beta_{Y;BC} &= Y + BY + CY \\
\beta_{Y;ABC} &= Y + AY + BY + CY
\end{aligned}
$$

so that $\beta_{Y;BC} = \beta_{Y;B} + \beta_{Y;C} - \beta_{Y}$ and $\beta_{Y;ABC} = \beta_{Y;A} + \beta_{Y;B} + \beta_{Y;C} - 2\beta_{Y}$. We can take this example further by specifying that $A \perp Y \mid B, C$, i.e. Y is independent of the levels of A so that (given all the previous conditional independence relationships) we obtain

$$
\begin{aligned}
\beta_{Y} &= Y \\
\beta_{Y;A} &= Y \\
\beta_{Y;B} &= Y + BY \\
\beta_{Y;C} &= Y + CY \\
\beta_{Y;AB} &= Y + BY \\
\beta_{Y;AC} &= Y + CY \\
\beta_{Y;BC} &= Y + BY + CY \\
\beta_{Y;ABC} &= Y + BY + CY
\end{aligned}
$$

so that $\beta_{Y;ABC} = \beta_{Y;BC} = \beta_{Y;B} + \beta_{Y;C} - \beta_{Y}$, $\beta_{Y;AC} = \beta_{Y;C}$ and $\beta_{Y;AB} = \beta_{Y;B}$.

If we add another continuous variable then this will give rise to an additional set of $\beta_{\gamma;D}$ $(\gamma \in \Gamma, D \subset \Delta)$ interaction expansions in a completely analogous way to those expansions defined above. Alternatively, if φ_{D} is used to reflect the discrete interaction expansions then there is no initial constant interaction. If $\omega_{\gamma\zeta;D}$ is used to reflect quadratic interaction expansions then there are (at most) pairwise interactions between the continuous variables together with any number of factors. For example, take the case of two factors A, B and two continuous variables Y, Z. The saturated model specifies the following set of generic interaction expansions for the quadratic part of the model:

$$
\begin{aligned}
\omega_{YY} &= Y^2 \\
\omega_{YY;A} &= Y^2 + AY^2 \\
\omega_{YY;B} &= Y^2 + BY^2 \\
\omega_{YY;AB} &= Y^2 + AY^2 + BY^2 + ABY^2 \\
\omega_{YZ} &= YZ \\
\omega_{YZ;A} &= YZ + AYZ \\
\omega_{YZ;B} &= YZ + BYZ \\
\omega_{YZ;AB} &= YZ + AYZ + BYZ + ABYZ \\
\omega_{ZZ} &= Z^2 \\
\omega_{ZZ;A} &= Z^2 + AZ^2 \\
\omega_{ZZ;B} &= Z^2 + BZ^2 \\
\omega_{ZZ;AB} &= Z^2 + AZ^2 + BZ^2 + ABZ^2
\end{aligned}
$$

where the leading Y^2, YZ and Z^2 terms are constant. If we specify that $Y \perp Z \mid A, B$ then this removes all $\omega_{YZ;.}$ parameters. Alternatively, specifying $A \perp Y \mid B, Z$ has the following effect on the interaction expansions:

$$\begin{aligned}
\omega_{YY} &= Y^2 \\
\omega_{YY;A} &= Y^2 \\
\omega_{YY;B} &= Y^2 + BY^2 \\
\omega_{YY;AB} &= Y^2 + BY^2 \\
\omega_{YZ} &= YZ \\
\omega_{YZ;A} &= YZ \\
\omega_{YZ;B} &= YZ + BYZ \\
\omega_{YZ;AB} &= YZ + BYZ \\
\omega_{ZZ} &= Z^2 \\
\omega_{ZZ;A} &= Z^2 + AZ^2 \\
\omega_{ZZ;B} &= Z^2 + BZ^2 \\
\omega_{ZZ;AB} &= Z^2 + AZ^2 + BZ^2 + ABZ^2
\end{aligned}$$

so that $\omega_{YY;AB} = \omega_{YY;B}$, $\omega_{YY;A} = \omega_{YY}$, $\omega_{YZ;AB} = \omega_{YZ;B}$ and $\omega_{YZ;A} = \omega_{YZ}$. In addition, specifying $A \perp B \mid Y, Z$ yields the following interaction expansions:

$$\begin{aligned}
\omega_{YY} &= Y^2 \\
\omega_{YY;A} &= Y^2 \\
\omega_{YY;B} &= Y^2 + BY^2 \\
\omega_{YY;AB} &= Y^2 + BY^2 \\
\omega_{YZ} &= YZ \\
\omega_{YZ;A} &= YZ \\
\omega_{YZ;B} &= YZ + BYZ \\
\omega_{YZ;AB} &= YZ + BYZ \\
\omega_{ZZ} &= Z^2 \\
\omega_{ZZ;A} &= Z^2 + AZ^2 \\
\omega_{ZZ;B} &= Z^2 + BZ^2 \\
\omega_{ZZ;AB} &= Z^2 + AZ^2 + BZ^2
\end{aligned}$$

so that $\omega_{ZZ;AB} = \omega_{ZZ;A} + \omega_{ZZ;B} - \omega_{ZZ}$.

In summary, the interaction expansions are determined by the pairwise conditional independence constraints. These expansions may be used to evaluate κ via Equation (2.18) and then κ and the same interaction expansions substituted into the general log likelihood (2.20) to calculate the log likelihood for the sub-model. Alternatively, Equations (2.22) and (2.23) may be used to calculate the log-likelihood of a sub-model in the HCG case; see pages 22–23.

2.3.4 The gradient and Hessian

Differentiating the log-likelihood, $L(\theta; x)$, given by (2.20) yields the gradient vector

$$L^{(1)} = \left(L'_{\varphi}, L'_{\beta}, L'_{\Omega}\right)'. \tag{2.25}$$

For the saturated (heterogeneous) model, given r cells ($s = r - 1$) and q continuous variables, $L^{(1)}$ is an $\{s + rq + rq(q+1)/2\} \times 1$ real-valued vector, where L_φ is of dimension $s \times 1$, L_β is of dimension $rq \times 1$ and L_Ω is of dimension $\{rq(q+1)/2\} \times 1$. The components of L_φ, L_β, and L_Ω are given by

$$\begin{aligned}
\{L_{\varphi(i)}\}_{i\in\mathcal{I}} &= n(i) - a(i), \\
\{L_{\beta(i)}\}_{i\in\mathcal{I}} &= t(i) - \Omega(i)^{-1}\beta(i)a(i), \\
\{L_{\Omega(i)}\}_{i\in\mathcal{I}} &= -A(i) + \{C(i) + D(i)\}a(i),
\end{aligned}$$

respectively; in which $\{L_{\varphi(i)}\}_{i\in\mathcal{I}}$ are scalar parameters, $\{L_{\beta(i)}\}_{i\in\mathcal{I}}$ are $q \times 1$ vectors and $\{L_{\Omega(i)}\}_{i\in\mathcal{I}}$ are $q(q+1)/2$ vectors. In the above expressions

$$a(i) = n\kappa(2\pi)^{q/2}|\Omega(i)|^{-1/2}\exp\left[\varphi(i) + \tfrac{1}{2}\mathrm{tr}\left\{\Omega(i)^{-1}B(i)\right\}\right]$$

is a scalar quantity, where $B(i) = \beta(i)\beta(i)'$ is a $q \times q$ symmetric matrix. Also,

$$A(i) = \mathrm{svec}\left[S(i) - \tfrac{1}{2}\mathrm{diag}\left\{S(i)\right\}\right],$$

$$C(i) = \mathrm{svec}\left[\Omega(i)^{-1} - \tfrac{1}{2}\mathrm{diag}\left\{\Omega(i)^{-1}\right\}\right],$$

and

$$D(i) = \mathrm{svec}\left[\Omega(i)^{-1}B(i)\Omega(i)^{-1} - \tfrac{1}{2}\mathrm{diag}\left\{\Omega(i)^{-1}B(i)\Omega(i)^{-1}\right\}\right]$$

are $q(q+1)/2 \times 1$ vectors. Note that $\Omega(i)$ and $S(i)$ are as defined on page 22, 'diag' denotes a diagonal matrix and 'svec' denotes the $q(q+1)/2$ lower-triangular elements of a $q \times q$ matrix stored as a vector.

The gradient vector for sub-models is obtained in the following way. First, we create matrices $\mathcal{X}^d$, $\mathcal{X}^l_\gamma$ and $\mathcal{X}^q_{\gamma\zeta}$ ($\gamma, \zeta \in \Gamma$), which we shall refer to as *parameter matrices* for the discrete, linear and quadratic parameters, respectively. Note that each $\gamma \in \Gamma$ for the linear parameters generates a different $\mathcal{X}^l_\gamma$ and each pair $\gamma, \zeta \in \Gamma$ generates a different $\mathcal{X}^q_{\gamma\zeta}$. We could of course combine all these matrices together but it is easier to think of them separately at this stage. We index each of the $\mathcal{X}$'s using i and j where the rows indexed by i correspond to the full combination of factor levels $i \in \mathcal{I}$ and the columns indexed by j correspond to some subset of factor levels $j \in \mathcal{J}$. In a CG sub-model the $\mathcal{X}$'s contain the linear combinations of non-zero parameters that we use as a purely organizational device to 'fill-in' values for the full set of discrete, linear and quadratic parameters over all $i \in \mathcal{I}$ and to simplify computing the gradient vector. It is important to note that we only estimate the non-zero parameters directly hence preserving the reduced parameter dimensionality determined by a CG sub-model. The zero parameters implied by the conditional independence relationships are either zero off-diagonal elements in the case of conditionally independent continuous random variables only or linear combinations of estimated parameters for conditional independence between discrete or discrete and continuous random variables as described in Section 2.3.3. If the full set of parameters is estimated then $|\mathcal{J}|$ will equal $|\mathcal{I}|$, otherwise $|\mathcal{J}|$ will be less than $|\mathcal{I}|$ for one or more $\mathcal{X}$ matrices. Second, we let

$$\begin{aligned}
\{L^*_{\varphi(i)}\}_{i\in\mathcal{I}} &= a(i), \\
\{L^*_{\beta(i)}\}_{i\in\mathcal{I}} &= \Omega(i)^{-1}\beta(i)a(i), \\
\{L^*_{\Omega(i)}\}_{i\in\mathcal{I}} &= \{C(i) + D(i)\}a(i),
\end{aligned}$$

where the individual elements for $\varphi(.)$, $\beta(.)$ and $\Omega(.)$ over all $i \in \mathcal{I}$, $\gamma \in \Gamma$ and $\gamma, \zeta \in \Gamma$ are obtained via

$$\begin{aligned} \varphi(i) &= \sum_{j \in \mathcal{J}} \mathcal{X}^d(i,j)\varphi(j), \\ \beta_\gamma(i) &= \sum_{j \in \mathcal{J}} \mathcal{X}^l_\gamma(i,j)\beta_\gamma(j), \\ \omega_{\gamma\zeta}(i) &= \sum_{j \in \mathcal{J}} \mathcal{X}^q_{\gamma\zeta}(i,j)\omega_{\gamma\zeta}(j), \end{aligned}$$

i.e. by summation over the (potentially) reduced set of factor levels j for fixed i. Let the individual elements of $\{L^*_{\varphi(i)}\}_{i \in \mathcal{I}}$, $\{L^*_{\beta(i)}\}_{i \in \mathcal{I}}$ and $\{L^*_{\Omega(i)}\}_{i \in \mathcal{I}}$ be given by $l^*_{\varphi(i)}$, $l^*_{\beta_\gamma(i)}$ and $l^*_{\omega_{\gamma\zeta}(i)}$, respectively, and let

$$\begin{aligned} l_{\varphi(j)} &= \sum_{i \in \mathcal{I}} \mathcal{X}^d(i,j)n(i) - \sum_{i \in \mathcal{I}} \mathcal{X}^d(i,j)l^*_{\varphi(i)}, \\ l_{\beta_\gamma(j)} &= \sum_{i \in \mathcal{I}} \mathcal{X}^l_\gamma(i,j)t_\gamma(i) - \sum_{i \in \mathcal{I}} \mathcal{X}^l_\gamma(i,j)l^*_{\beta_\gamma(i)} \end{aligned}$$

and

$$l_{\omega_{\gamma\zeta}(j)} = -\sum_{i \in \mathcal{I}} \mathcal{X}^q(i,j)s_{\gamma\zeta}(i)/2 + \sum_{i \in \mathcal{I}} \mathcal{X}^q_{\gamma\zeta}(i,j)l^*_{\omega_{\gamma\zeta}(i)} \quad (\text{for } \gamma = \zeta),$$

or

$$l_{\omega_{\gamma\zeta}(j)} = -\sum_{i \in \mathcal{I}} \mathcal{X}^q(i,j)s_{\gamma\zeta}(i) + \sum_{i \in \mathcal{I}} \mathcal{X}^q_{\gamma\zeta}(i,j)l^*_{\omega_{\gamma\zeta}(i)} \quad (\text{for } \gamma \neq \zeta);$$

where $n(i)$ is a cell count, $t_\gamma(i)$ is a cell total corresponding to $\gamma \in \Gamma$ and $\omega_{\gamma\zeta}$ is a cell sum of squares for $\gamma = \zeta$ or cell sum of cross-products for $\gamma \neq \zeta$ ($\gamma, \zeta \in \Gamma$). Finally, by substituting the above elements into $\{L_{\varphi(j)}\}_{j \in \mathcal{J}}$, $\{L_{\beta(j)}\}_{j \in \mathcal{J}}$ and $\{L_{\Omega(j)}\}_{j \in \mathcal{J}}$ and using these terms as the components of L_φ, L_β and L_Ω, respectively, in (2.25) gives the reduced dimension gradient vector $L^{(1)}$.

The gradient vector calculations were checked for a range of examples using a finite difference approximation to the gradient (see Appendix C, Section C.2) and by comparison with the output of MIM (see Section 2.4.1 on page 88). (The matrix results employed here are given in Appendix A. Most of the details concerning matrix and vector differentiation, including proofs, may be found in Magnus & Neudecker (1988).)

The general form of the Hessian matrix (or matrix of second partial derivatives with respect to the parameters) of $L(\theta; x)$ for the saturated (heterogeneous) model is given by

$$L^{(2)} = \begin{bmatrix} L_{\varphi\varphi} & L'_{\beta\varphi} & L'_{\Omega\varphi} \\ (s \times s) & (s \times rq) & (s \times rt) \\ L_{\beta\varphi} & L_{\beta\beta} & L'_{\Omega\beta} \\ (rq \times s) & (rq \times rq) & (rq \times rt) \\ L_{\Omega\varphi} & L_{\Omega\beta} & L_{\Omega\Omega} \\ (rt \times s) & (rt \times rq) & (rt \times rt) \end{bmatrix}, \tag{2.26}$$

where the dimensions of each of the sub-matrices are given in parentheses, r is the number of cells, q is the number of continuous variables, $s = r - 1$ and $t = q(q +$

1)/2. The above matrix is useful in measuring the curvature of the log-likelihood at the maximum, which in turn indicates the precision of $\hat{\theta}$. High curvature indicates a strongly concentrated log-likelihood and therefore high precision; see Cox & Wermuth (1996, pp. 402–5). To measure curvature we can use the second derivative of $L(\theta; x)$ evaluated at $\theta = \hat{\theta}$. This is called the *observed information*, $J(\hat{\theta})$, where

$$J(\hat{\theta}) = \left[-\frac{\partial^2 L(\theta; x)}{\partial \theta^2} \right]_{\theta=\hat{\theta}} \tag{2.27}$$

using the definition given by Cox & Wermuth. Allied to this we have the *Fisher information* defined by

$$I(\theta) = \text{Var} \left[\frac{\partial f(x_i; \theta)}{\partial \theta} \right], \tag{2.28}$$

where $f(x_i; \theta)$ is the density function corresponding to the parameter value θ on the space of a single observation. If the x_i's constitute independent random observations from the same distribution then the information is n times the information provided by each observation. Thus,

$$I(\theta) = \frac{1}{n} \text{Var} \left[\frac{\partial L(\theta; x)}{\partial \theta} \right] \tag{2.29}$$

(see Cox & Hinkley, 1974, p.108). Note that (2.28) or (2.29) may be more appropriate to work with from a theoretical perspective. When the amount of information about θ in the data is large, the log-likelihood will usually be quite tightly concentrated around $\hat{\theta}$; and the distribution of θ around θ will be approximately multivariate normal with covariance matrix given by either $J^{-1}(\hat{\theta})$ or $nI^{-1}(\hat{\theta})$.

Here we rely on the observed information but rather than evaluating the observed information (2.27) directly via $L^{(2)}$, we build-up a numerical approximation to the inverse Hessian via the parameter estimation procedure. In the saturated case, the maximum number of parameters are estimated and we obtain a full approximate inverse Hessian. Sub-model specification means that the Hessian will be of reduced dimension, i.e. we lose one row and one column corresponding to each parameter lost from the saturated model.

2.4 Parameter estimation

We now look at maximizing the likelihood for a sample of n independent identically distributed (IID) observations assumed to have come from a CG distribution. Instead of maximizing the likelihood directly we choose to (equivalently) minimize the negative log-likelihood. We do this using a general quasi-Newton minimization algorithm. The use of quasi-Newton procedures are generally well known in the literature on numerical analysis and so the details about the particular algorithm we employ are relegated to Appendix B. The most important feature of the algorithm is its Hessian approximation. Our choice is the BFGS updating rule, which is known to have superior performance in terms of roundoff error to other quasi-Newton procedures. The use of quasi-Newton procedures in statistics is generally advocated by a number of researchers, e.g. Bunday & Kiri (1987), Bishop (1995) and Ripley (1996).

Lauritzen & Wermuth (1989) and Lauritzen (1996, pp. 171–172, 205) show that densities of the form (2.1) expanded in terms of full and reduced sets of interaction parameters constitute a regular exponential family (see Barndorff-Nielsen, 1978). For sufficiently regular problems, i.e. where the likelihood is differentiable and the maximum occurs within the interior of the parameter space (for $\theta \in \Theta$), the maximum likelihood estimate (MLE) is the unique solution of the likelihood equation $L^{(1)} = 0$. For the saturated CG case this is true when $\widehat{\Sigma}(i)$ is positive-definite for all $i \in \mathcal{I}$ and for the saturated HCG case when $n(i) > 0$ for all i and $\widehat{\Sigma}$ is positive-definite; see Lauritzen, (1996, pp. 169–70).

We employ a quasi-Newton minimization procedure to estimate the MLEs for saturated CG, HCG and their sub-models obtained by setting certain interaction expansion terms equal to zero. The objective function is given by the negative of the log-likelihood, $-L(\theta; x)$, and its gradient vector given by $-L^{(1)}$, which will be model specific. At each evaluation of the log-likelihood a value for κ is calculated based on the current values of the parameters. Since we are minimizing the negative of the log-likelihood we obtain an approximation to the observed information, i.e. $-\hat{L}^{(2)^{-1}}(\theta; x)$, once the procedure has converged. This approximation is used to give approximate standard errors for the maximum likelihood estimates of a CG model. (See also Roverato & Whittaker (1996) who show how to obtain standard errors for parameters in graphical Gaussian models.)

2.4.1 Practical considerations

It would be wrong to suggest that CGM worked perfectly well in all situations. Although theory states that the quasi-Newton optimization procedure is guaranteed to converge to a local minimum there is no absolute certainty that this corresponds with the maximum likelihood estimate. We do benefit from having the program MIM, which provides a check on the solution obtained using a different algorithm; see Edwards (1995). MIM uses an iterative proportional scaling algorithm that is analogous to the procedure of Deming & Stephan (1940) for fitting log-linear models. The main problem encountered with our parameter estimation scheme is that when the likelihood is particularly flat we get a slow rate of convergence. If the pre-specified maximum number of steps is exceeded before the algorithm converges then the procedure is stopped and is flagged as not having converged. The practical solution to this problem in a model search situation is to retain the edge being tested in the independence graph. The only time this problem was encountered was in analysing the high-dimensional SGA dataset of Chapter 4. We encountered about four or five sub-models that could not be fitted successfully, usually at the start of the model search procedure. However, as more parameters were removed from the saturated model we found that this improved the speed and convergence of the algorithm. The same problem was also apparent in MIM but with a different set of sub-models. Of course, apart from the obvious saturated cases where we can calculate analytic maximum likelihood estimates for the parameters as a check of the algorithm, we are assuming that the procedure we employ gives the unique solution or overall maximum.

The BFGS algorithm we employ ensures that the estimate of the inverse Hessian remains positive-definite throughout the iteration process. For the algorithm to always generate downhill search directions positive definiteness of the Hessian is essential. In fact, it is often better to use an approximation to the inverse Hessian than to use

the actual Hessian itself since when we are a long way from the minimum there is no guarantee that the Hessian is positive-definite. For more details about quasi-Newton algorithms see Gill *et al.* (1981, pp 63–65).

2.5 Program CGM

The quasi-Newton procedure for general CG model parameter estimation described in the preceding section and Appendix B has been implemented in a FORTRAN 77 computer program called CGM. CGM simply stands for Conditional Gaussian Modelling. Some of the programming details are given in Appendix C.

2.5.1 An example CGM session

As an example of a typical CGM session we look at the analysis of data taken from Morrison (1967, p. 167) concerning the effect of a drug on the level of three biochemical compounds found in the brains of mice. Twenty-two mice were randomly divided into two groups. The second group were given the drug periodically and the first group were used as a control. Both case and control groups were reported to have received the same care and diet, although two of the control group mice died of natural causes during the experiment. The data consist of one binary variable A indicating either a case ($A = 2$) or a control ($A = 1$) and three continuous variables X, Y and Z measuring the amounts of three compounds in micrograms per gram of brain tissue. The data are as follows:

A	X	Y	Z	A	X	Y	Z
1	1.21	0.61	0.70	2	1.40	0.50	0.71
1	0.92	0.43	0.71	2	1.17	0.39	0.69
1	0.80	0.35	0.71	2	1.23	0.44	0.70
1	0.85	0.48	0.68	2	1.38	0.42	0.71
1	0.98	0.42	0.71	2	1.17	0.45	0.70
1	1.15	0.52	0.72	2	1.31	0.41	0.70
1	1.10	0.50	0.75	2	1.30	0.47	0.67
1	1.02	0.53	0.70	2	1.22	0.29	0.68
1	1.18	0.45	0.70	2	1.19	0.37	0.72
1	1.09	0.40	0.69	2	1.12	0.27	0.72
				2	1.09	0.35	0.73
				2	1.00	0.30	0.70

A parameter file (`params.dat`) used to control program execution for this example is shown below. The first line of this file specifies a title for the session. A datafile is specified by `file mice.dat`. One discrete variable (A) is declared using `fact A`, which has two levels specified by `lev 2`. Three continuous variables X, Y and Z are declared using `cont XYZ` and the read order for the declared variables in `mice.dat` is given by `read AXYZ`. Five models are specified using the `model` command and comments appear after the character # naming the type of model to be fitted. Full details about all commands available in CGM are given in Appendix D.

```
title Mice data from Morrison (1967)
file  mice.dat
fact  A
lev   2
cont  XYZ
read  AXYZ
```

```
# Saturated model
model A/AX,AY,AZ/AXYZ
# Independence given A
model A/AX,AY,AZ/AX,AY,AZ
# Saturated HCG model
model A/AX,AY,AZ/XYZ
# Independence from A
model A/X,Y,Z/XYZ
# Mutual independence
model A/X,Y,Z/X,Y,Z
```

Table 2.1 Initial CGM model fits for the mice data. For each model $-2L$ gives the value of minus twice the maximized log-likelihood, d gives the total number of independently adjusted parameters, deviance is the difference between $-2L$ for the current model and $-2L$ for the saturated model, Df gives the corresponding model degrees of freedom and p-value is the probability of observing a chi-squared value for deviance on Df degrees of freedom.

Model	$-2L$	d	deviance	Df	p-value
Saturated model	−184.61	19	0.00	0	—
Independence given A	−170.10	13	14.51	6	0.0245
Saturated HCG model	−182.78	13	1.83	6	0.9350
Independence from A	−154.49	10	30.11	9	0.0004
Mutual independence	−153.78	7	30.82	12	0.0021

An initial attempt at modelling these data compares the saturated CG model with the models for mutual independence, independence given A, independence from A and the saturated HCG model. (The saturated HCG model specifies the maximum number of discrete and linear interactions, together with the maximum number of pairwise quadratic interactions between the continuous variables only.) A summary of the fit for each model is given in Table 2.1 and the corresponding conditional independence graphs for the models are shown in Figure 2.3.

The deviance for a submodel $\mathcal{M}_k \subset \mathcal{M}$ with $d_k < d$ parameters is defined as

$$D_{\mathcal{M}_k} = -2(L_{\mathcal{M}_k} - L_{\mathcal{M}}), \tag{2.30}$$

where $L_{\mathcal{M}}$ is the value of the maximized log-likelihood for some assumed maximal or saturated model $\mathcal{M}$ and $L_{\mathcal{M}_k}$ is the value of the maximized log-likelihood for a reduced model $\mathcal{M}_k$ (e.g. Aitkin *et al.*, 1989, p. 110). We assume that the deviance defined by (2.30) is asymptotically distributed as a chi-squared random variable on $(d - d_k)$ degrees of freedom (see Cox & Hinkley, 1974, pp. 327–328). (Since the log-likelihood is only defined up to a constant $-2L$ may be negative.)

The next stage in the analysis might be to consider the removal of edges from the saturated HCG model. CGM is able to perform backwards model selection using chi-squared tests of deviance. Deleting all `model` definitions apart from the saturated HCG model in `params.dat` and including the commands `step chi` and `crit 0.05` specifies backwards model selection using chi-squared tests of deviance with a critical value set at 0.05.

Backwards model selection proceeds as a series of stages. Based on the conditional independence graph for some assumed maximal CG model each edge (in turn) is considered for removal. The models generated by the removal of a single edge are

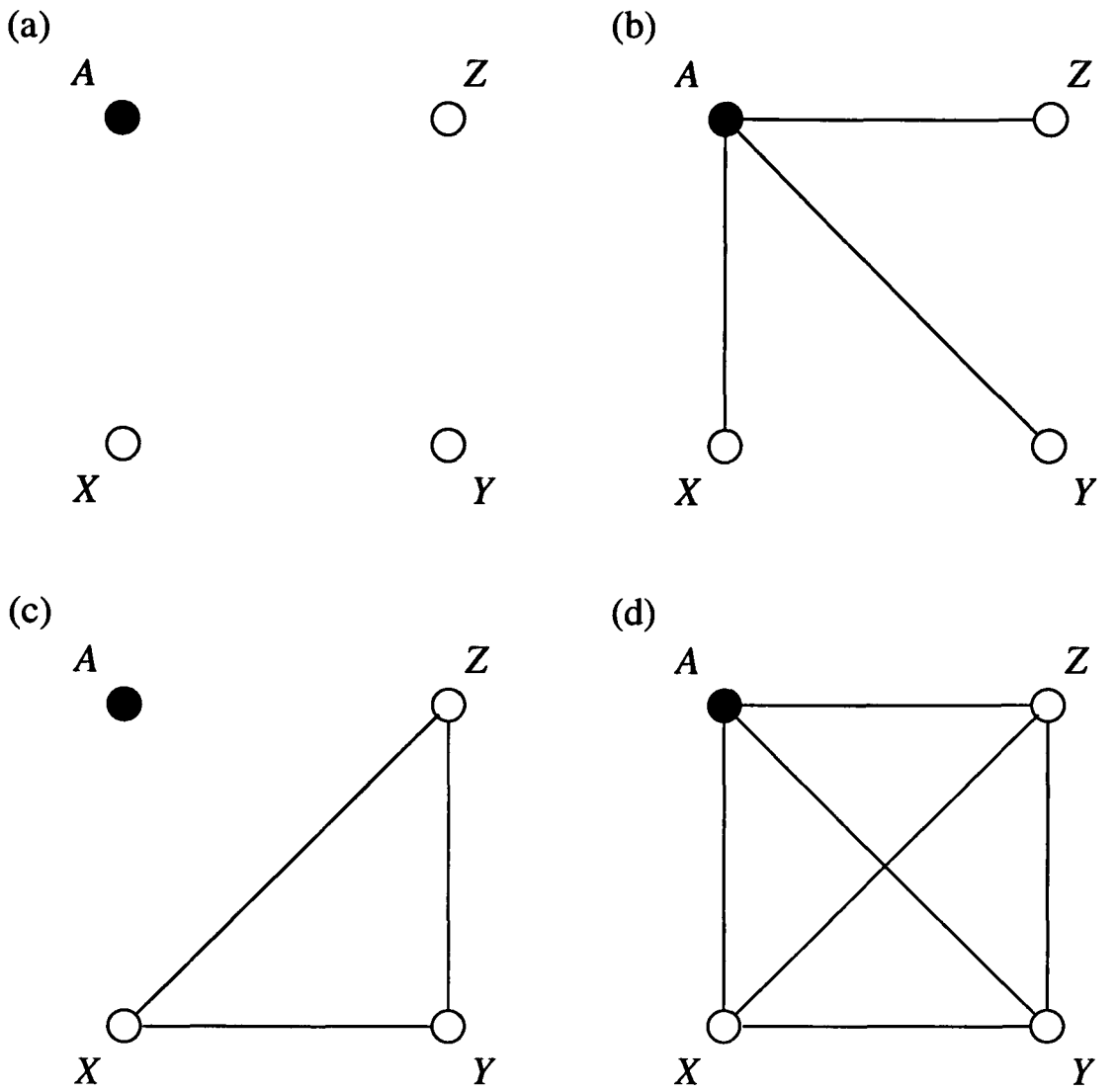


Figure 2.3 Conditional independence graphs for model (a) $A \mid X, Y, Z \mid X, Y, Z$ — mutual independence, (b) $A \mid AX, AY, AZ \mid AX, AY, AZ$ — independence given A, (c) $A \mid X, Y, Z \mid XYZ$ — independence from A and (d) $A \mid AX, AY, AZ \mid XYZ$ — saturated HCG. The vertex corresponding to the discrete random variable is denoted by a filled circle or dot and the vertices corresponding to the continuous random variables are denoted by circles.

compared using a chi-squared test of deviance. The model chosen is the one whose chi-squared value for deviance produces the largest p-value greater than the critical value. The model chosen then becomes the current model and the procedure continues by considering the removal of each edge in the conditional independence graph of the current model. The procedure is stopped when no further edges may be removed from the current model. The current model is then chosen as the final model.

Backwards model selection for the mice data starting with the saturated HCG model removes edges AZ, XY and YZ. The final model is given by $A \mid AX, AY, Z \mid XY, Z$. Comparison of the final model with the saturated HCG model gives a deviance of 2.43 on 9 degrees of freedom. Table 2.2 gives the results from the final stage in the model selection process.

Clearly, no further edges should be removed from the model. The conditional independence graph for the final model is shown in Figure 2.4. The relative strengths of the pairwise interactions between the variables are indicated by the thickness of each of the edges. The main conclusion is that the level of compound Z is independent of treatment and compounds X and Y but there is a significant interaction between brain compounds X and Y and the treatment. Parameter estimates and corresponding approximate standard errors are given in Table 2.3 on page 36. These data are also analysed by Edwards (1995, pp. 77–79) using MIM and give the same results.

Table 2.2 Results from CGM backwards model selection at step 3 for the mice data. Comparison of model A / AX, AY, Z / XY, Z with the models generated by the removal of edges AX, AY and XY. For each model, based on the removal of an edge, $-2L$ gives minus twice the maximized log-likelihood, d gives the number of independently adjusted parameters and p-value is the probability of observing a chi-squared value for deviance on Df degrees of freedom.

Edge	$-2L$	d	deviance	Df	p-value
—	−182.18	10	0.00	0	—
$[AX]$	−160.48	9	21.70	1	0.000003
$[AY]$	−164.20	9	17.98	1	0.000022
$[XY]$	−169.72	9	12.45	1	0.000417

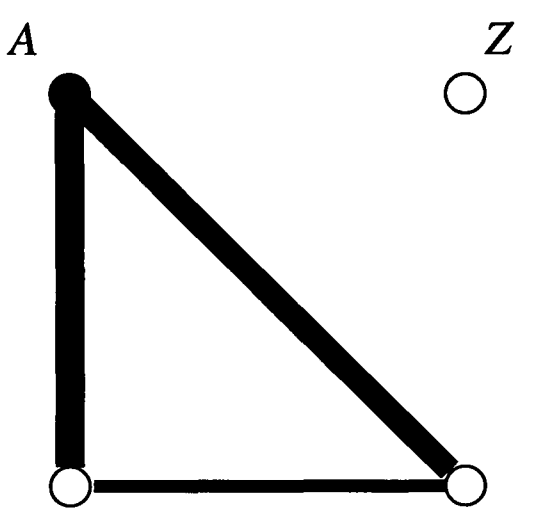


Figure 2.4 The conditional independence graph corresponding to the final graphical CG model selected $(A/AX, AY, Z/XY, Z)$ for the mice data. Edge thickness corresponds to the significance of the edge deletion deviance in the model, the more significant the deviance the thicker the edge.

Table 2.3 Estimated canonical parameters (inverse covariances, linear parameters and discrete parameters) for the final mice data model with corresponding approximate standard errors in parentheses.

$A=1$	discrete	–		
	linear	58.12 (23.82)	27.01 (34.93)	2368.09 (675.34)
	quadratic	X	Y	Z
	X	116.97 (35.63)		
	Y	−132.94 (52.10)	349.56 (107.11)	
	Z	0.00	0.00	3361.17 (904.98)
$A=2$	discrete	−13.51 (8.07)		
	linear	90.49 (30.68)	−25.78 (40.54)	2368.09 (675.34)
	quadratic	X	Y	Z
	X	116.97 (35.63)		
	Y	−132.94 (52.10)	349.56 (107.11)	
	Z	0.00	0.00	3361.17 (959.04)

2.6 Model selection

The usual approach to model selection involves the selection of variables. We might do this in order to avoid measuring unimportant variables, which may in turn improve the precision of the estimated parameters; model simplification yielding easier interpretation or to provide a more finely tuned classification procedure. Variable selection is typically performed using some stepwise scheme together with a suitably defined selection criterion. If we have the appropriate asymptotic theory then significance tests may be appropriate here. Alternatively, if our goal is classification, we may opt to select variables that seek to minimize the error rate. A detailed practical account of variable-selection in discriminant analysis is given by McKay & Campbell (1982a, b). An alternative approach considers all subsets of variables. The aim is the selection of the 'best' subset of variables with respect to some pre-determined criterion. By nature this procedure is computationally intensive but is generally practical if there are not too many variables at the start; see, e.g. Draper & Smith (1981).

The approach used to identify suitable CG models does not seek the removal of variables. Instead a model is sought that adequately describes the multivariate data structure. Model selection is based on considering the set of pairwise interactions between the full set of variables. The approach that we adopt is stepwise backwards selection, which is usually preferred to stepwise forwards inclusion. This is because backwards selection starts with a model that is usually consistent with the data. Simpler models are chosen that are also data consistent. In contrast, forwards inclusion starts with an inconsistent data model that is successively enlarged until an acceptable model is obtained. For a discussion of this last point see Edwards (1995, Ch. 6). Edwards also describes a procedure in which subsets of models are considered. The procedure seeks the simplest set of models consistent with the data. The algorithm is due to Edwards & Havránek (1985, 1987) and may be thought of as being loosely based on the all-subsets approach to model selection.

2.6.1 An alternative model selection criteria

An alternative to using asymptotic chi-squared tests of deviance for testing the removal of edges from a CG model may be based on the information criterion proposed by Akaike (1973, 1974). This is given by

$$\text{AIC} = -2\times \text{ maximized log-likelihood} + 2\times \text{no. of independently adjusted parameters.}$$

AIC is derived by considering the expected difference between the true and estimated log-likelihoods for θ. A Taylor-series expansion about θ leads to an AIC correction term for the maximized log-likelihood of twice the number of fitted parameters. Details may be found in Akaike (1973, 1974); see also Stone (1977). The practical implication of using AIC is that models with a large numbers of parameters are penalized in favour of models with fewer parameters.

Stepwise model selection using AIC may be based on model deviance, $D_{\mathcal{M}_k}$, plus twice the number of fitted parameters (see e.g. Hastie & Pregibon, 1992, pp. 233–234), i.e. use

$$\text{AIC} = D_{\mathcal{M}_k} + 2d_k. \tag{2.31}$$

Model selection is based on minimizing AIC. A related model selection criterion known as NIC (or Network Information Criterion) was implemented in CGM. However, the calculation of NIC proved to be unstable for certain sub-models in high-dimensional problems. With smaller datasets NIC did reasonably well but gave much the same results as AIC. This instability in high-dimensional problems was probably due to the approximations used to estimate the components of NIC. Details about NIC are given in Section C.5 of Appendix C.

2.7 The location model

The location model was introduced by Olkin & Tate (1961) and later developed by Krzanowski (1975, 1980). In the location model it is assumed that p discrete random variables expressed as a single variable i is multinomial. The pattern of discrete variable values uniquely determines a multinomial cell (or 'location'). The continuous variables are assumed to follow a multivariate normal distribution but the mean parameters of the distribution depend on the cells defined by the values of the discrete variables. A common covariance matrix is assumed for all cells. The location model defined by Olkin & Tate (1961) assumes that the cell probabilities are unrestricted so that the maximum likelihood estimates are given by $\hat{p}(i) = n(i)/n$. If the discrete data are sparse then Krzanowski restricts $\hat{p}(i)$ by imposing a second-order log-linear model on the multinomial cell probabilities. The location model may be fitted as an HCG model with a complete graph on the discrete variables leading to the recovery of the observed cell frequencies. The location model does in fact give rise to a complete graph on the variables. Such structures are decomposable yielding analytic maximum likelihood estimates (see Frydenberg & Lauritzen, 1989).

As indicated on page 15, given p discrete random variables each with r_j levels ($j = 1, \ldots, p$) there will be $r = \prod_{j=1}^{p} r_j$ cells in the contingency table. Unless the dataset is large, for moderately large p there are likely to be observed cell frequencies of zero. In this situation, we can use a restricted model for the discrete variables. Krzanowski (1975) suggests an approximate parameter estimation scheme that uses multivariate regression to estimate the cell means, using the residual matrix as an estimate of Σ and a log-linear model incorporating main effects and first-order interactions to estimate the probabilities in the contingency table. Krzanowski (1975) advocates restricting $\hat{p}(i)$ further by dropping first-order interactions in those samples which give rise to very sparse tables.

An attempt was made to model sparse binary data together with continuous data in a graph-theoretic context by smoothing observed cell probabilities using a kernel. The details of this approach are described in Appendix E. There are, however, problems with the method from a model selection point of view but it may have some validity if we wanted to fit a specific model, say, and were only prevented from doing so because of one or a small number of empty cells.

2.8 Example datasets

In this section we look at applying the CG modelling procedure to two datasets: the Anderson–Fisher *Iris* data and the low birth weights dataset from a study that took place in 1986 at the US Baystate Medical Center in Massachusetts.

2.8.1 *Iris data*

The following well known dataset consists of four measurements on three species of *Iris*. The measurements taken are: sepal length, sepal width, petal length and petal width (the raw measurements are in centimetres). There are 50 observations in each group. The data were collected by Anderson (1935) and are analysed by Fisher (1936). In most of the literature the data are treated as a problem in classification (e.g. Mardia *et al.*, 1979, Ch. 11). However, the original paper by Fisher investigated and found evidence in favour of a genetic hypothesis that placed *Iris versicolor* as a hybrid two-thirds of the way between *Iris setosa* and *Iris virginica*. The aim of our initial analysis will be to describe the conditional independence structure of the data using a CG model. These data are also analysed in this way by Whittaker (1990, pp. 359–363). We then look at a classification rule for *Iris* species using the fitted canonical parameters of the model in the next Chapter.

As a first step in the analysis it is probably appropriate to take the logarithm of each of the biological variables. A comparison of the transformed measurements is made in Table 2.4. There are large relative differences between petal length and petal width

Table 2.4 Empirical correlations (lower diagonal), partial correlations given all remaining variables (upper diagonal), means and standard deviations for the *Iris* data. There are 50 of each species of *Iris*. (All measurements are on a $\log_e$ scale.)

I. setosa	W = sepal length	1	0.710	0.195	0.154
	X = sepal width	0.730	1	−0.052	0.003
	Y = petal length	0.295	0.181	1	0.243
	Z = petal width	0.293	0.208	0.309	1
		W	X	Y	Z
	Mean	1.608	1.226	0.373	−1.485
	Standard deviation	0.070	0.113	0.121	0.409
I. versicolor	W = sepal length	1	0.247	0.624	−0.180
	X = sepal width	0.528	1	−0.089	0.460
	Y = petal length	0.765	0.560	1	0.624
	Z = petal width	0.569	0.663	0.784	1
		W	X	Y	Z
	Mean	1.777	1.012	1.443	0.271
	Standard deviation	0.087	0.117	0.116	0.153
I. virginica	W = sepal length	1	0.260	0.827	−0.103
	X = sepal width	0.461	1	−0.061	0.486
	Y = petal length	0.857	0.405	1	0.164
	Z = petal width	0.295	0.546	0.329	1
		W	X	Y	Z
	Mean	1.881	1.084	1.709	0.697
	Standard deviation	0.097	0.108	0.098	0.139

means for all the species. Petal length and sepal length appear to be highly correlated for species *I. versicolor* and *I. virginica*. Sepal length and width are highly correlated for *I. setosa*. There are small partial correlations between sepal width and petal length given the other variables for all three species. The partial correlation between sepal length and width given the other variables is large for *I. setosa* in comparison with *I. versicolor* and *I. virginica*. Species *I. versicolor* and *I. virginica* appear to be the most similar overall, tending to support Fisher's genetic hypothesis.

In modelling the *Iris* data we start with the saturated graphical CG model and use the backwards model selection procedure (described on page 33) employing asymptotic chi-squared tests of deviance as the model selection criterion. (In the following analysis I is a ternary variable indicating class, i.e. $I = 1$ labels the *setosa* group, $I = 2$ labels the *versicolor* group and $I = 3$ labels the *virginica* group.) Table 2.5 shows the final stage in selecting a model for the *Iris* data. At this point in the model selection procedure, two edges, $[XY]$ and $[WZ]$ have been removed from the saturated model. No further reduction in the model is possible as all the edges in the corresponding conditional independence graph are significant.

Table 2.5 Results from CGM backwards model selection using asymptotic chi-squared tests of deviance as the model fitting criterion. The table gives a comparison of model I / IW, IX, IY, IZ / IWX, IXZ, IWY, IYZ with the models generated by the removal of edges listed under 'Edge'. For each model $-2L$ gives minus twice the maximized log-likelihood, d gives the number of independently adjusted parameters and p-value is the probability of observing a chi-squared value for deviance on Df degrees of freedom.

Edge	$-2L$	d	deviance	Df	p-value
—	−779.78	38	0.00	0	—
$[IW]$	−716.82	30	62.96	8	0.0000
$[IX]$	−675.89	30	103.88	8	0.0000
$[IY]$	−563.12	30	216.66	8	0.0000
$[IZ]$	−624.87	30	154.90	8	0.0000
$[WX]$	−733.45	35	46.32	3	0.0000
$[WY]$	−685.13	35	94.65	3	0.0000
$[XZ]$	−750.88	35	28.89	3	0.0000
$[YZ]$	−741.34	35	38.43	3	0.0000

The conditional independence graph corresponding to the selected model is shown in Figure 2.5. The graph is easily interpreted, sepal length (W) is adjacent to petal length (Y), sepal width (X) is adjacent to petal width (Z), petal length and width and sepal length and width are adjacent to each other. Whittaker (1990) obtains the same results with the untransformed *Iris* measurements using Edwards' MIM program.

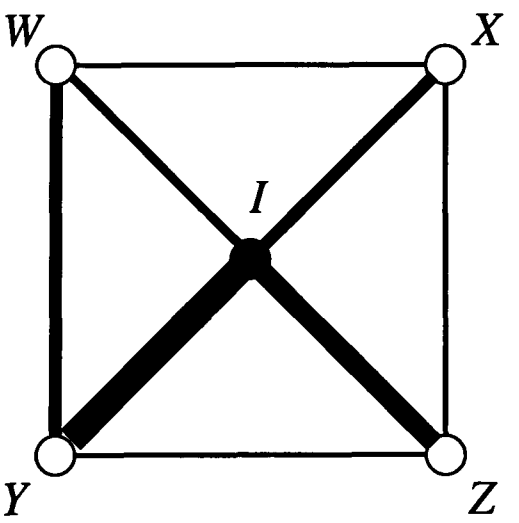


Figure 2.5 The conditional independence graph corresponding to the graphical CG model for the *Iris* data selected using asymptotic chi-squared tests. Edge thickness corresponds to the significance of the edge deletion deviance in the model, the more significant the deviance the thicker the edge.

Parameter estimates, obtained from CGM, together with their standard errors are given in Table 2.6 below. A comparison of parameter estimates divided by their stan-

Table 2.6 Discrete, linear and quadratic parameter estimates and corresponding approximate standard errors (in parentheses) for the fitted CG model of the *Iris* data, I / IW, IX, IY, IZ / IWX, IXZ, IWY, IYZ.

discrete	$i=1$	$i=2$	$i=3$
$\hat{\varphi}(i)$	—	−10.3 (76.6)	134.1 (68.4)
linear	$i=1$	$i=2$	$i=3$
$\hat{\beta}_W(i)$	475.1 (94.4)	308.8 (63.4)	167.7 (60.0)
$\hat{\beta}_X(i)$	−101.6 (47.0)	63.4 (38.3)	26.9 (29.2)
$\hat{\beta}_Y(i)$	−15.8 (30.0)	81.9 (58.6)	38.1 (55.0)
$\hat{\beta}_Z(i)$	−17.0 (5.0)	−179.9 (35.0)	−22.7 (20.4)
quadratic	$i=1$	$i=2$	$i=3$
$\hat{\omega}_{WW}(i)$	455.3 (84.0)	333.1 (60.1)	428.9 (76.6)
$\hat{\omega}_{XW}(i)$	−199.2 (44.3)	−34.0 (22.1)	−49.5 (18.0)
$\hat{\omega}_{YW}(i)$	− 34.9 (32.2)	−172.4 (24.7)	−342.3 (26.1)
$\hat{\omega}_{XX}(i)$	174.0 (18.8)	137.8 (40.0)	142.0 (69.3)
$\hat{\omega}_{ZX}(i)$	−3.6 (15.0)	−57.9 (47.6)	−48.7 (74.3)
$\hat{\omega}_{YY}(i)$	82.6 (3.1)	289.7 (16.3)	404.4 (15.1)
$\hat{\omega}_{ZY}(i)$	−6.4 (3.2)	−110.1 (25.3)	−13.5 (13.0)
$\hat{\omega}_{ZZ}(i)$	6.9 (1.3)	138.4 (23.2)	76.4 (14.4)

dard error show that all but $\hat{\omega}_{YW}(1)$, $\hat{\omega}_{ZX}(1)$, $\hat{\omega}_{XW}(2)$, $\hat{\omega}_{ZX}(2)$, $\hat{\omega}_{ZX}(3)$ and $\hat{\omega}_{ZY}(3)$ have ratios greater 2 for the quadratic interactions. The largest ratios are observed for $\{\hat{\omega}_{YY}(.)\}$, $\{\hat{\omega}_{ZZ}(.)\}$ and $\{\hat{\omega}_{XX}(.)\}$. The precision of the parameter estimates is consistent with the strong bonds between I and Y, I and Z, I and X and weaker bonds between X and Z, Y and Z, X and W illustrated by the independence graph shown in Figure 2.5. Linear interactions $\{\beta_Y(.)\}$, $\beta_Z(1)$, $\beta_Z(2)$ and $\beta_X(1)$ yield the largest values relative to their standard error, which is again consistent with the independence graph.

2.8.2 *Low birth weights*

Hosmer & Lemeshow (1989) analyse data on 189 births at the US Baystate Medical Center, Massachusetts during 1986. The aim of the study was to determine possible risk factors associated with low infant birth weight. These data are also analysed by Ripley (1994a) and Venables & Ripley (1997). The variables and their descriptions are given in Table 2.7.

Table 2.7 Low birth weights dataset variable names and descriptions.

Name	Description
I - LBW	1 = normal birth weight / 2 = birth weight < 2.5kg
A - smoke	smoking status during pregnancy (1 = no / 2 = yes)
B - ht	has history of hypertension (1 = no / 2 = yes)
C - ui	has uterine irritability (1 = no / 2 = yes)
D - race	1 = white / 2 = black / 3 = other
E - ptl	number of previous premature labours (range 1–4)
F - ftv	number of physician visits in the first trimester (range 1–7)
X - age	age of mother in years
Y - lwt	weight of mother (lbs) at last menstrual period
Z - bwt	actual birth weight (grams)

Ripley (1994a) notes that five pairs of rows contain identical values and so we delete

one of each of these pairs. Following Venables & Ripley (1997), we recode *ptl* as a binary variable (1 = no *ptl* / 2 = 1+ *ptl*) and *ftv* as a ternary variable (1 = no *ftv* / 2 = 1 *ftv* / 3 = 2+ *ftv*) since there were only a few high values for both these variables. Natural log transformations of *age* and *lwt* were taken in order to improve marginal normality (see Figure 2.6). The normal probability plots for X-*age* clearly indicate data normality with the log scale giving a slightly better overall fit. There is an outlier at a mother's age of 45 years which is pulled in by the transformation. Note that the discreteness in age measurements is highlighted by the plots. The normal probability plots of Y-*lwt* (mother's weight at last menstrual period) indicate a longer right-hand tail than the normal. Clearly, the log-transformed data are to be preferred.

Typical approaches to analysing these data would be to model actual birth weight using multiple regression, or alternatively to fit a logistic regression with the binary response *LBW*. Since there is not enough data to fit a saturated graphical CG model the aim is to fit a reduced model that describes the interaction between birth weight grouping (*LBW*) and the other variables. Univariate significance tests comparing *LBW* with each of the other variables were performed to determine if any of the variables might be removed. More specifically, a chi-squared test was used to compare *LBW* with each of the discrete variables and a two-sample t-test used to compare the low-weight and normal-weight *age* and *lwt* means on a natural log scale. The p-values for these tests were p=0.019 (*smoke*), p=0.044 (*ht*), p=0.017 (*ui*), p=0.084 (*race*), $p<0.001$ (*ptl*), p=0.230 (*ftv*), p=0.070 (*age*) and p =0.008 (*lwt*). Based on these results it looks as if we could drop *race*, *ftv* and *age* from the analysis. However, simply dropping variables on the basis of a univariate significance test does ignore possibly significant interactions. It is hoped that the removal of these variables will still allow reasonably accurate classification of birth weights. (This part of the analysis is discussed in Section 3.8.2 on page 46.)

CG model selection

We start by examining the observed counts in the 2^5=32 cells. The data are very sparse with twelve empty cells and ten cells with fewer than 5 observations. Thus, it is not possible to fit a saturated homogeneous nor heterogeneous model to these data. One approach to modelling such sparse data is to fit a reduced model. Whittaker (1990, p. 278) discusses the fitting of an all two-way interaction log-linear model to a sparse dataset consisting of eight binary variables. We adopt a similar approach to modelling our main dataset in Chapter 4. However, with this dataset, due to sparsity, it is still not possible to fit the all two-way interaction model. Instead we resort to fitting a model of independence given *I-LBW* with a common (inverse) covariance structure in the quadratic part. Thus, the starting model is simply given by $AI, BI, CI, EI \ / \ IX, IY \ / \ X, Y$ — i.e. a model of independence given I. The final stage in the analysis in shown in Table 2.8. Figures 2.7(a)–(c) show the two stages in the model selection process. The main features of the final graph (c) are that *LBW* is conditionally independent of *ht* and *age* and that *LBW* and *ptl* appear to be strongly related.

The last menstrual period mean weight for mothers with non-*LBW* infants is 134lbs compared with 122lbs for mothers whose infants have *LBW*.

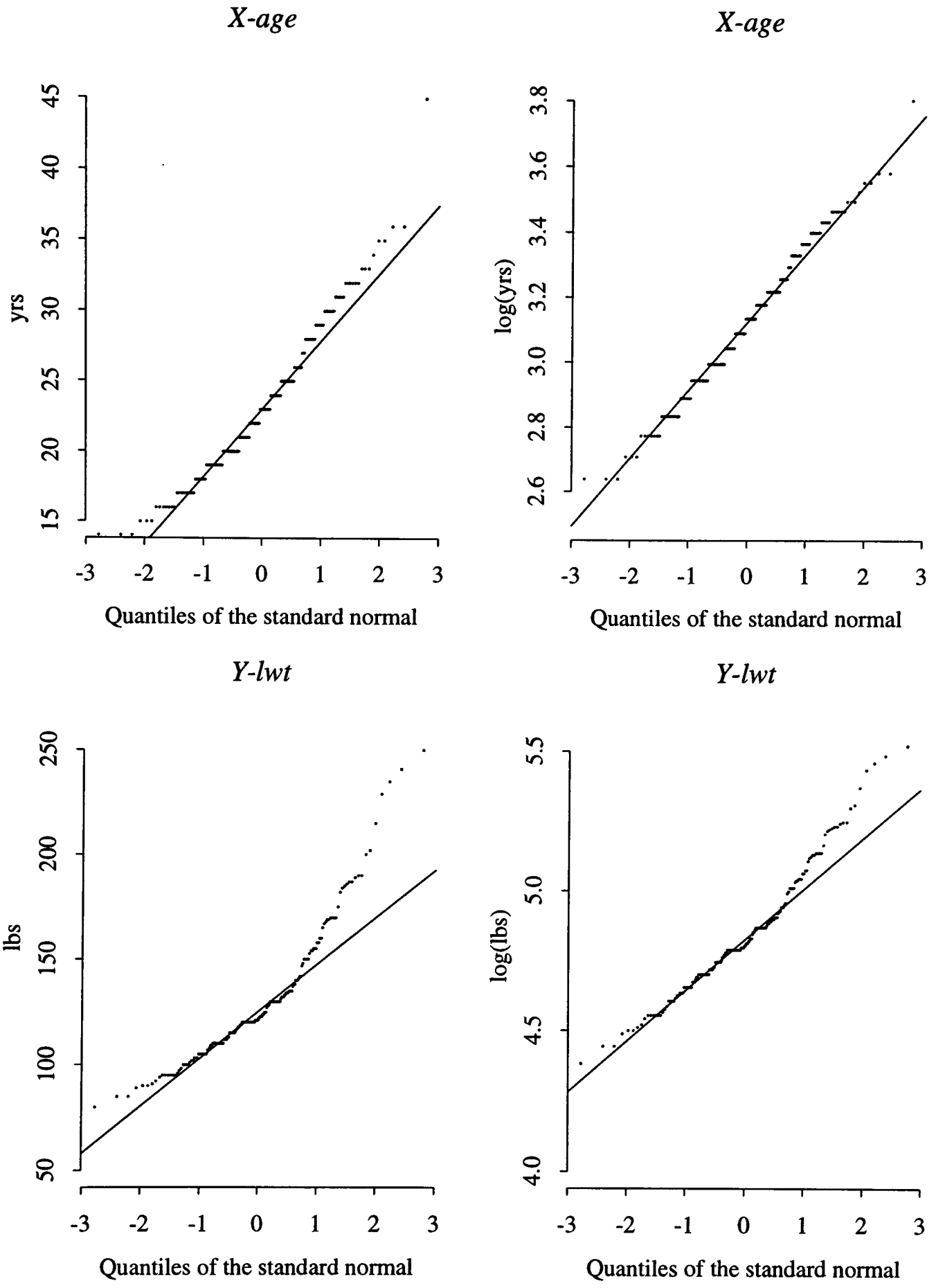


Figure 2.6 Normal probability plots for untransformed and $\log_e$ transformed variables *age* (age of mother in years) and *lwt* (weight of mother at last menstrual period). In each graph the straight line is drawn through the upper and lower quartiles.

Table 2.8 Comparison of model $AI, CI, EI \mid X, IY \mid X, Y$ with the models generated by the removal of edges listed under 'Edge'. For each model, based on the removal of an edge, $-2L$ gives minus twice the maximized log-likelihood, d gives the number of independently adjusted parameters and Df the number of degrees of freedom.

Edge	$-2L$	d	deviance	Df	p-value
—	776.00	13	.00	0	—
$[AI]$	781.44	12	5.44	1	.019642
$[CI]$	781.36	12	5.36	1	.020630
$[EI]$	788.01	12	12.01	1	.000529
$[IY]$	783.16	12	7.16	1	.007462

Relative risk

It is useful to analyse the subtables formed by the two-way interactions AI and IE. A measure of relative risk may be used in this situation. Given a cross classification of *LBW* with some potential aetiological factor, e.g.

		Factor	
		yes	no
LBW	yes	a	b
	no	c	d

relative risk is calculated by forming the ratio $r = ad/bc$. This is also known as the *odds ratio* (since it is the ratio of a/c to b/d, and these two quantities may be thought of as the odds in favour of *LBW*). An approximate confidence interval may be obtained by noting that asymptotically

$$\log \hat{r} = \log\left(\frac{ad}{bc}\right) \sim N(\log r, \sigma^2),$$

where an estimate of the variance of $\log \hat{r}$ is given by

$$\hat{\sigma}^2 = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

(see Armitage & Berry, 1987, §16.2).

The relative risk estimates for *I-LBW* given *A-smoke*, *I* given *C-ui* and *I* given *E-ptl* together with their corresponding 95% confidence intervals are: 2.1 with 95% CI (1.1, 4.0); 2.1 with 95% CI (1.2, 6.1) and 4.1 with 95% CI (1.8, 9.3) All three estimates of relative risk are significant at the 5% level since their 95% CIs do not enclose 1. (We shall look again at this dataset in the next chapter with the aim of classifying non-*LBW* and *LBW* infants using those variables identified here as being associated with *LBW*.)

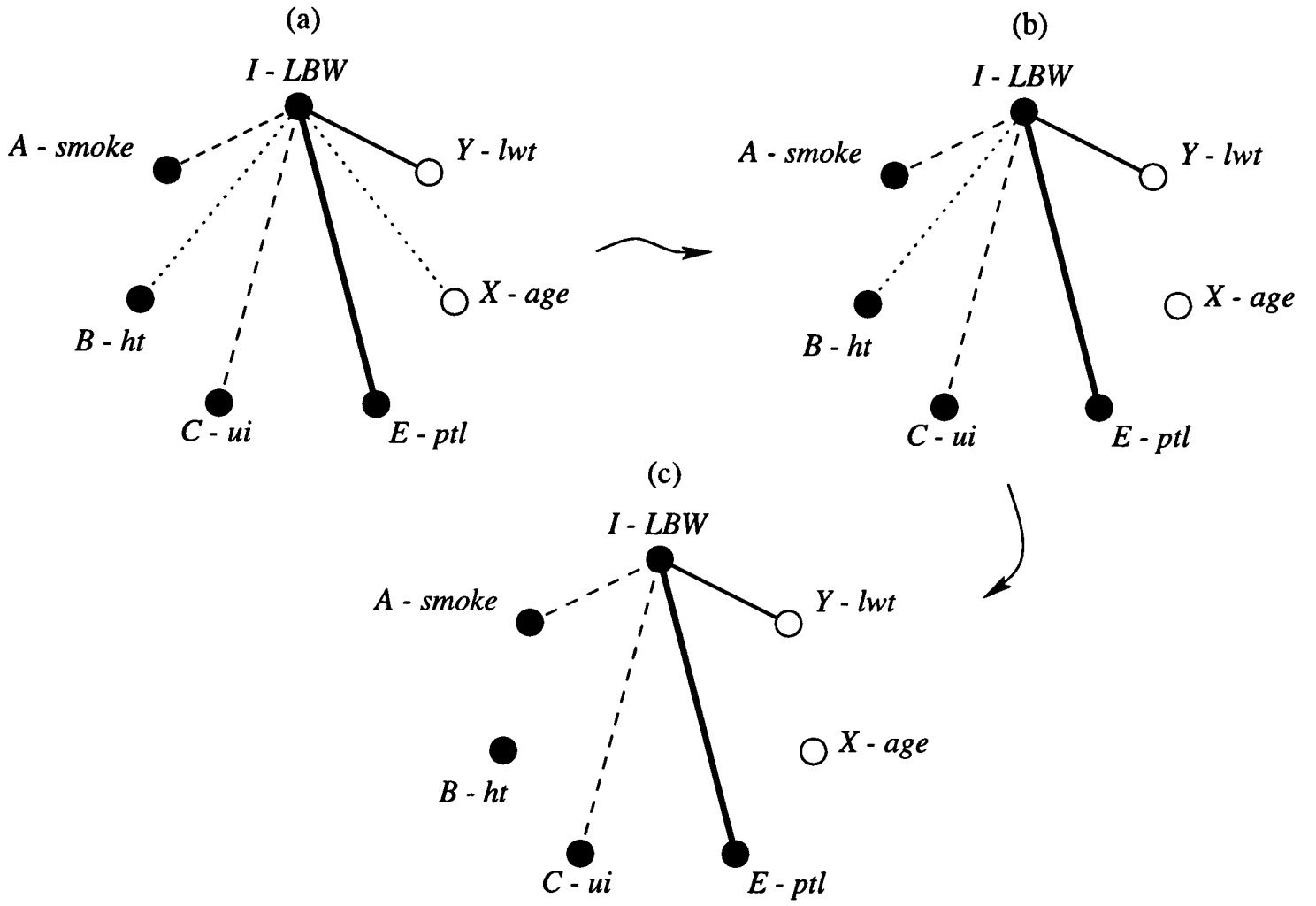


Figure 2.7 Conditional independence graphs showing the stages in the model selection process for the low birth weights dataset. Edge deletion is based on chi-squared tests of deviance with a critical value set at 0.05. Edge thickness corresponds to the significance of the edge deletion deviance for the model, the more significant the deviance the thicker the edge. The dotted edges indicate an edge that is non-significant at the 5% level ($p > 0.05$) and the dashed edges indicate a non-significant edge at the 1% level ($0.01 < \mathrm{p} \leqslant 0.05$) but one that is significant at the 5% level.

Chapter 3

Discrimination and Classification

Discriminant analysis is concerned with how best to summarize the basic differences between two or more populations, groups or classes of objects based on a set of explanatory variables. The classical method of linear discrimination was described by Fisher (1936) for two populations and later extended to more than two populations by Rao (1948). Classification deals specifically with the problem of assigning future observations to populations. The validity of the classification procedure is measured in terms of its ability to assign observations to the correct populations. Modern methods of discrimination and classification aim to distinguish between an arbitrary number of populations and often employ non-linear classification rules in order to allocate future observations. For details of some of the more advanced discrimination and classification techniques available see McLachlan (1992) and also Ripley (1996).

In Chapter 2 we saw how CG models may be used to discriminate between species of Iris, and between normal and low birth weight infants. In this chapter we show how models estimated in the joint CG framework may be used to define classification rules. The predictive approach to classification is considered and comparisons are made with logistic discrimination and k-nearest neighbour methods of classification.

3.1 Bayes rule

Consider the problem of classifying an observation to one of g classes on the basis of q measurements $y = (y_1, \ldots, y_q)'$. The basic assumption is that y has a different probability distribution in each of the g classes. Let the probability density of y in class c be $f_c(y)$ and let π_c be the prior probability that an observation comes from class c. The posterior probability that a new observation, with observed $q \times 1$ measurement vector y_0, comes class c is obtained using Bayes formula in the following way:

$$p(C = c \mid Y = y_0) = \frac{\pi_c f_c(y_0)}{\sum_{l=1}^{g} \pi_l f_l(y_0)}, \quad \text{for } c \in \{1, \ldots, g\}. \tag{3.1}$$

A Bayes classification rule is formed by assigning y_0 to the class with the largest posterior probability, i.e.

$$\text{allocate } y_0 \text{ to class } c \text{ if } p(c \mid y_0) = \max_{l \leqslant g} p(l \mid y_0). \tag{3.2}$$

This Bayes classification rule selects the class with the largest posterior probability of population membership, which is equivalent to minimizing the total probability of misallocation of an observation.

Consider the problem of classifying an observation to one of two populations. The probability densities for these populations are given by $f_1(y)$ and $f_2(y)$ respectively.

From (3.1) and (3.2) (for $g = 2$) a Bayes classification rule may be expressed as follows:

$$\begin{aligned} &\text{allocate } y_0 \text{ to class 1 if } f_1(y_0)/f_2(y_0) > \pi_2/\pi_1, \\ &\text{and to class 2 if } f_1(y_0)/f_2(y_0) \leqslant \pi_2/\pi_1. \end{aligned} \tag{3.3}$$

The standard parametric approach assumes that $f_1(y)$ and $f_2(y)$ are known probability densities involving unknown parameters θ. The ratio $f_1(y)/f_2(y)$ is then also a function of the unknown parameters θ. Let $R^*(y, \theta)$ denote this ratio and replace θ by its estimate, $\hat{\theta}$, based on observed data. The classification rule for a new observation may now be expressed as

$$\begin{aligned} &\text{allocate } y_0 \text{ to class 1 if } R^*(y_0, \hat{\theta}) > \pi_2/\pi_1, \\ &\text{and to class 2 if } R^*(y_0, \hat{\theta}) \leqslant \pi_2/\pi_1. \end{aligned} \tag{3.4}$$

This is known as an *estimative* or *plug-in* classification rule. The main problem with an estimative rule is that it ignores the sampling variation of $\hat{\theta}$. In Section 3.6 we look at overcoming this problem by using predictive densities.

3.1.1 Error rates

In this chapter, we will want to assess classification procedures that have been estimated from sample data. It is well known that an optimistic bias in the error rate exists in attempting to use the same data to both train and test a classification rule. This suggests splitting datasets into two parts, one for training the classification procedure and the other for testing. This is fine if there are enough data to satisfy the estimation procedure but if the data are sparse then this method of performance assessment may prove problematic. There are two bias-correcting performance assessment methods that have been implemented in CGM. The first is v-fold error rate estimation, which divides a dataset of size n into $v = \min(10, \sqrt{n})$ pieces. In turn, each piece is left out, the classification rule estimated and then used to classify those observations excluded from estimation procedure. Performance assessment is based on averaging the error rates over all v pieces left out of the estimation step of the analysis. The second method is a bootstrap estimate due to Efron (1983). We found that with the datasets analysed in this thesis v-fold error rates gave very nearly the same results as bootstrap error rates. For this reason and because there can be a considerable computational overhead using bootstrap error rates we generally report v-fold error rates. Methods for estimating error-rates are described more fully in Appendix F.

3.2 CG discrimination

From (2.15) on page 21, we write the CG density for some class c by setting the first element of i, i.e. i_1, equal to the class indicator c and specifying

$$f_c(x) = f\{(i|i_1{=}c), y\} = \kappa_c \exp\left\{\varphi(i|i_1{=}c) + [\beta(i|i_1{=}c)]'y - \tfrac{1}{2}y'[\Omega(i|i_1{=}c)]y\right\}, \tag{3.5}$$

where $x = \{(i|i_1{=}c), y\}$ is a $(p+q) \times 1$ observation vector (described in Section 2.1.2 on page 13), the $\{\varphi(i|i_1 = c)\}_{i\in\mathcal{I}}$ are scalar parameters, the $\{\beta(i|i_1 = c)\}_{i\in\mathcal{I}}$ are $q \times 1$ real-valued vectors, the $\{\Omega(i|i_1{=}c)\}_{i\in\mathcal{I}}$ are positive-definite (symmetric) $q \times q$ matrices

and κ_c is a normalizing constant. Let the logarithm of the ratio of two CG densities be given by

$$R_{c1}(x) = \log\left\{\frac{f_c(x)}{f_1(x)}\right\} \tag{3.6}$$

so that from (3.5) where $c = 2, \ldots, g$;

$$\begin{aligned} R_{c1}(x) &= \log(\kappa_c/\kappa_1) + \Big[\varphi(i|i_1{=}c) - \varphi(i|i_1{=}1)\Big] + \Big[\beta(i|i_1{=}c) - \beta(i|i_1{=}1)\Big]' y \\ &\quad - y'\Big[\Omega(i|i_1{=}c) - \Omega(i|i_1{=}1)\Big] y/2 \\ &= \varphi(k) + \beta(k)'y - \tfrac{1}{2} y'\Omega(k) y/2, \end{aligned} \tag{3.7}$$

where $k = (k_1, \ldots, k_{p-1})'$ is a $(p-1) \times 1$ integer-valued vector giving the specific value of factor levels $i_2, \ldots, i_p$; the $\{\varphi(k)\}_{k \in \mathcal{K}}$ are scalar quantities giving the difference between the discrete interactions for the two classes for each value of k and also incorporating the logarithm of the ratio of the normalizing constants κ_c and κ_1; the $\{\beta(k)\}_{k \in \mathcal{K}}$ are $q \times 1$ real-valued vectors giving the difference between linear interactions for the two classes for each value of k; and the $\{\Omega(k)\}_{k \in \mathcal{K}}$ are $q \times q$ symmetric matrices giving the difference between quadratic interactions for the two classes for each value of k. Finally, $\mathcal{K}$ denotes the full set of levels $i_2 \otimes i_3 \otimes \cdots \otimes i_p$. Note that (3.7) is computed by expanding the individual elements of $\varphi(i)$, $\beta(i)$ and $\Omega(i)$ as described in on page 13. We shall call (3.7) a *CG discriminant function*. If a common concentration matrix is assumed then (3.7) simply reduces to

$$\begin{aligned} R_{c1}(x) &= \log(\kappa_c/\kappa_1) + \Big[\varphi(i|i_1{=}c) - \varphi(i|i_1{=}1)\Big] + \Big[\beta(i|i_1{=}c) - \beta(i|i_1{=}1)\Big]' y \\ &= \varphi(k) + \beta(k)'y \end{aligned} \tag{3.8}$$

which we shall call an *HCG discriminant function*.

In general, If there are g classes and q continuous random variables we obtain $g(g-1)/2$ CG discriminant functions of the form (3.7), of which the $\min(g-1, q)$ are linearly-independent for each cell $i \in \mathcal{I}$. Usually (and especially when $g > 2$) it will be more convenient to employ classification rule (3.2) replacing y_0 with x_0, rather than use rules of the form (3.4). (The two allocation schemes are of course equivalent.) Parameter estimates for the interaction parameters may be obtained using a suitable estimation procedure, such as maximum likelihood.

If the number of cells is large or if sample size is small then $\Omega(i|i_1{=}c)$ is likely to be poorly estimated. We can overcome this problem by constraining the model at the outset, e.g. we might assume that the concentration matrix is constant over cells within each population separately by taking

$$\Omega(i|i_1{=}c) = \Omega(i_1{=}c), \qquad \text{for all } c = 1, \ldots, g$$

leading to constrained CG discriminant functions at cells $i \in \mathcal{I}$; or assume that the concentration matrix is constant over populations within each cell by taking

$$\Omega(i|i_1{=}c) = \Omega(k), \qquad \text{for all } i \in \mathcal{I} \text{ and } k \in \mathcal{K}$$

leading to HCG discriminant functions at cells $i \in \mathcal{I}$; or assume a constant concentration throughout by taking

$$\Omega(i|i_1{=}c) = \Omega, \qquad \text{for all } c = 1, \ldots, g \text{ and } i \in \mathcal{I}$$

leading to constrained HCG discriminant functions at cells $i \in \mathcal{I}$.

The standard approaches to discriminant analysis are clearly special cases of CG discrimination, e.g the ratio of two q-variate multivariate normal densities for class $1 \sim N_q(\mu_1, \Sigma_1)$ and class $2 \sim N_q(\mu_2, \Sigma_2)$ is given by

$$\begin{aligned} f_1(y)/f_2(y) &= |\Sigma_2|^{1/2}|\Sigma_1|^{-1/2} \exp\left[-\tfrac{1}{2}\{y'(\Sigma_1^{-1} - \Sigma_2^{-1})y \right. \\ &\quad \left. - 2y'(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2) + \mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2\}\right], \end{aligned} \tag{3.9}$$

where in the usual notation μ_l is a $q \times 1$ mean vector and Σ_l is a $q \times q$ positive-definite (symmetric) covariance matrix for class l. Taking the logarithm of the above expression gives

$$\begin{aligned} \log\{f_1(y)/f_2(y)\} &= \tfrac{1}{2}\log|\Sigma_2| - \tfrac{1}{2}\log|\Sigma_1| - \tfrac{1}{2}\{y'(\Sigma_1^{-1} - \Sigma_2^{-1})y \\ &\quad -2y'(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2) + \mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2\}, \end{aligned} \tag{3.10}$$

If the simplifying assumption of common covariance matrices is used, i.e. $\Sigma_1 = \Sigma_2 = \Sigma$, then

$$\log\{f_1(y)/f_2(y)\} = (\mu_1 - \mu_2)'\Sigma^{-1}y - \tfrac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2). \tag{3.11}$$

If (3.9) is used in (3.3) the classification rule for a new observation, $y = y_0$, will be based on a *quadratic discriminant function*. Alternatively, if the exponentiated version of (3.11) is used in (3.3) the classification rule for a new observation will be based on a *linear discriminant function*. Thus, the quadratic discriminant function may be viewed as a CG discriminant containing continuous variables only, with unrestricted heterogeneous covariance matrix; and the linear discriminant function may be viewed as an HCG discriminant containing continuous variables only with unrestricted homogeneous covariance matrix. Also related is the location model as defined by Krzanowski for discrete and continuous variables as described in Section 2.7 on page 38; also see Krusińska (1991).

An alternative to choosing either the linear or quadratic discriminant function is obtained by shrinking individual class covariance matrices towards a common covariance. A single parameter is used to control the degree of shrinkage. Shrinkage methods of this kind, which reduce the variance of the parameter estimates at the expense of introducing bias, are investigated in the context of discriminant analysis by Campbell (1980) and Kimura *et al.* (1987). (Shrinking covariance matrices in this way is completely analogous to the use of a ridge parameter in regression problems where the standard maximum likelihood parameter estimates are unstable, see Draper & Smith (1981, §6.7).) An additional shrinkage parameter may be used to adjust possible eigenvalue distortion in the sample class covariance matrices. This increases the values of the smaller eigenvalues whilst reducing the values of the larger ones (see e.g. James & Stein (1961), also McLachlan (1992, §5.2) and the references therein). Friedman (1989) adopts both types of shrinkage in a study of discriminant analysis, choosing to determine the values of the shrinkage coefficients using cross-validation.

3.3 Model selection

In Section 2.6 we described the difference between the standard linear model building approach based on variable selection and the graph-theoretic approach that is concerned

with modelling the structure of the dataset as a whole. In the latter case, although (potentially) redundant variables were removed from an analysis prior to model fitting (e.g. see Section 2.8.2) no attempt was made to remove variables during the subsequent iterative model selection process. The model selection process itself was based upon the removal of edges. Particular emphasis was placed on the variable that provided an assigned classification (if any) and this was justified on the basis of identifying those variables that were useful in discriminating between classes. The more usual approach to discrimination centres on choosing those variables that more usefully discriminate between classes. For classification purposes it may be sensible to measure only those variables that usefully contribute to correct classification. From this point of view, model selection is usually based on minimizing classification error rates.

In contrast, the Bayesian approach specifies a prior distribution for an unknown parameter vector θ. A prior distribution is introduced as part of the model and is supposed to express a state of knowledge or ignorance about θ before the data are obtained (see Box & Tiao, 1973, p. 6). Given a prior distribution, $p(\theta)$, a probability model $p(D \mid \theta)$ and observed data, D, we can calculate posterior model probabilities for class c via (3.1) with

$$p_c(y) = \int p_c(y;\theta)p(\theta \mid D)d\theta,$$

where $f_c(y) = p_c(y)$, $p_c(y;\theta)$ is a probability density parametrized in terms of θ and

$$p(\theta \mid D) \propto p(D \mid \theta)p(\theta) = L(\theta;D)p(\theta),$$

is a posterior distribution for θ given the data, which is proportional to a prior for θ multiplied by the likelihood, $L(\theta;D)$. Thus a saturated or maximal model is chosen at the outset and we avoid the problem of model selection by integration over the unknown parameters. For further discussion see Ripley (1996, §2.4).

3.4 Logistic discriminant analysis

Here we are concerned with the use of the logistic regression model in discrimination and classification. This type of analysis is typically called *logistic discrimination*. A more complete history of the use of the logistic function may be found in Cox & Snell (1989, pp. 24–25, 102–104). The earliest ideas in logistic discrimination may be found in Cox (1958) as a means of describing the distribution over two classes; Minsky (1961) in an epidemiological study with independent explanatory binary variables and binary response; and Cornfield (1962) and Truett *et al.* (1967) who adopt the logistic discrimination model in their studies of prospective heart disease. In addition, Cox (1966) and Day & Kerridge (1967) both suggested the logistic form for posterior probabilities as a basis for classification. A review of logistic discrimination may be found in Anderson (1982).

Logistic discrimination specifies the logistic response curve as a suitable model for the posterior probabilities, i.e.

$$p(C = 2 \mid X = x) = \frac{\exp(\alpha + \beta' x)}{1 + \exp(\alpha + \beta' x)}, \tag{3.12}$$

$$p(C = 1 \mid X = x) = 1 - p(C = 2 \mid x) = \frac{1}{1 + \exp(\alpha + \beta' x)},$$

where β is a $p \times 1$ parameter vector and $C = c$ is a random variable denoting population membership, so that the log odds ratio is a linear function of the observed variables, i.e.

$$\log\left\{\frac{p(C=2 \mid x)}{p(C=1 \mid x)}\right\} = \alpha + \beta' x.$$

Extension to more than two classes is given by

$$p(C=c \mid X=x) = \exp(\alpha_c + \beta_c' x) p(1 \mid x), \tag{3.13}$$

for $c = 2, \ldots, g$ and

$$p(C=1 \mid X=x) = 1 \Big/ \left[1 + \sum_{k=2}^{g} \exp\left\{\alpha_k + \beta_k' x\right\}\right], \tag{3.14}$$

where the β_k are $g-1$ vectors each having p parameters. It is easy to see that

$$\log\left\{\frac{p(C=c \mid x)}{p(C=1 \mid x)}\right\} = \alpha_c + \beta_c' x. \tag{3.15}$$

Thus, logistic discrimination assumes that the log odds for population c against a baseline population (here population 1) are linear in elements of x for $c = 2, \ldots, g$.

The logistic approach is widely applicable in a number of situations. Anderson (1982) notes that the logistic approach satisfies any of a rich set of assumptions, in particular:

1. multivariate normal distributions with equal covariance matrices;
2. multivariate discrete distributions following the log-linear model with equal (not necessarily zero) interaction terms;
3. joint distributions of 1 and 2, not necessarily independent;
4. selective and truncated versions of 1–3;
5. versions of 1-4 with quadratic, log or specified functions of x.

3.5 The relation between normal-based discrimination and logistic regression

Let π_c be the prior probability for population c then from Bayes theorem we have

$$p(c \mid y_0) = \frac{\pi_c f_c(y_0)}{\sum_{l=1}^{g} \pi_l f_l(y_0)}$$

so that

$$\log\left\{\frac{p(c \mid y_0)}{p(1 \mid y_0)}\right\} = \log\left(\frac{\pi_c}{\pi_1}\right) + \log\left\{\frac{f_c(y_0)}{f_1(y_0)}\right\} \tag{3.16}$$

If we assume that the $\{f_\cdot(y_0)\}$ are multivariate normal densities then

$$\begin{aligned} \log\left\{\frac{p(c \mid y_0)}{p(1 \mid y_0)}\right\} &= \alpha_{(c)} - \alpha_{(1)} + \{\beta_c - \beta_1\}' y_0 - y_0' \{\Omega_c - \Omega_1\} y_0/2 \\ &= \alpha_c + \beta_c' y_0 - y_0' \Omega_c y_0/2. \end{aligned} \tag{3.17}$$

This is a quadratic logistic discriminant, where Ω_c is a $q \times q$ symmetric matrix. Equation (3.17) is linear in the elements of β_c and the distinct elements of Ω_c so that it may be written in the the form (3.15) with $1 + q + q(q+1)/2$ parameters. The additional $q(q+1)/2$ elements of y_0, i.e. the squared and cross-product terms, enter into (3.15) in the same way as we might add such terms into a multiple linear regression. Anderson (1975) suggests some approximations to the quadratic form, $y_0'\Omega_c y_0$, which are useful when q is greater than 4 or 5.

Efron (1975) explored the asymptotic efficiency of logistic discriminant analysis relative to normal-based linear discriminant analysis for two groups. If the distributions of the observed variables are truly normal then the estimated discriminant coefficients will be more efficient (i.e. less variable) than the corresponding logistic estimates. This is because in the logistic case we lose information by conditioning on the observation vector x (or y), which is not the case in normal-based discrimination (see page 48). Thus, maximization of the conditional likelihood in the logistic case does not give the same answer as plugging in the normal maximum likelihood estimators for the logistic regression parameters $\hat{\beta}_c = (\hat{\mu}_c - \hat{\mu}_1)'\widehat{\Sigma}^{-1}$ and $\hat{\alpha}_c = \log(\pi_c/\pi_1) - (\hat{\mu}_c + \hat{\mu}_1)'\widehat{\Sigma}^{-1}(\hat{\mu}_c - \hat{\mu}_1)/2$. For further details and a wider discussion see Cox & Snell (1989, pp.135–136) and Ripley (1996, pp. 44–45).

Under the assumption of multivariate normal densities for two classes with common covariance matrix, equal priors and group sizes the parameter estimates determined by a regression of a 0/1 group indicator on the data are directly proportional to the substitution of maximum likelihood estimates into the linear discriminant function (see Cox & Snell, 1989, p. 136 and Ripley, 1996, pp. 101–105). An extension to more than two groups is given by Breiman & Ihaka (1984). (See also Hastie *et al.* (1994) and Ripley (1996) for their reinterpretations of Breiman & Ihaka's approach.)

Table 3.1 briefly summarizes the properties of the various discrimination models described in this thesis so far.

Table 3.1 Comparison of logistic, CG and HCG models. The table describes the most general model followed by a common submodel (if appropriate); the type of variates: mixed discrete (mainly binary) and continuous or continuous only; whether normality is assumed for the continuous measurements; and whether or not the group covariances are assumed to be equal. (In the table QDF stands for quadratic discriminant function, LDF for linear discriminant function and 'Location' for Krzanowski's location model.)

model	submodel	variates	normality	covariance
Quadratic Logistic	–	mixed	yes	non-homogeneous
Linear Logistic	–	mixed	yes	homogeneous
CG	–	mixed	yes	non-homogeneous
CG	QDA	continuous	yes	non-homogeneous
HCG	–	mixed	yes	homogeneous
HCG	Location	mixed	yes	homogeneous
HCG	LDA	continuous	yes	homogeneous

3.6 Predictive classification in the CG framework

We now look at obtaining predictive densities as a basis for CG classification. According to Ripley (1996, §2.4) this the correct thing to do if we intend to apply Bayes rule. We start by following the approach of Geisser & Cornfield (1963) (see also Geisser,

1993) for a normal population in which the predictive density for class c, $f_c(y)$, is obtained by assuming independence of the population parameters μ_c and Σ_c and then seeking a non-informative prior for Σ_c alone (for discussion see Box & Tiao, 1973, pp. 425–427). This is also the approach adopted by Aitchison & Dunsmore (1975). A non-informative prior on Σ_c is then taken as being proportional to $|\Sigma_c|^{-a/2}$, where $a \leqslant n_c$ is an adjustable parameter. Geisser & Cornfield's calculations make it more convenient to work with Σ_c^{-1} yielding a non-informative prior for Σ_c^{-1} proportional to $|\Sigma_c|^{a/2}$. In the absence of prior information Jeffreys' (1961) rule, which takes the prior distribution as being proportional to the square root of the Fisher information matrix, leads to $a = q+1$ (with q continuous random variables) — this is Geisser & Cornfield's choice. Choosing $a = 2$ makes the final predictive density equivalent to a fiducial predictive density (Fisher, 1959, Ch. V).

From Geisser & Cornfield (1963) we obtain the predictive density

$$\begin{aligned} f_c(y) &= \left\{\frac{1}{\pi(n_c+1)}\right\}^{q/2} \frac{\Gamma\{\frac{1}{2}(n_c+1-a+q)\}}{\Gamma\{\frac{1}{2}(n_c+1-a)\}} |\widehat{\Sigma}_c|^{-1/2} \\ &\quad \times \left[1 + \frac{1}{n_c+1}(y-\hat{\mu}_c)'\widehat{\Sigma}_c^{-1}(y-\hat{\mu}_c)\right]^{-(n_c+q-a+1)/2}, \end{aligned} \tag{3.18}$$

where $\hat{\mu}_c$ and $\widehat{\Sigma}_c$ are the usual maximum likelihood estimators for μ_c and Σ_c, respectively. In the univariate case (i.e. for $q = 1$), Jeffreys' rule and Fisher's fiducial argument yield the same prior specification and hence the same predictive density, namely a (scaled) t distribution centred on $\hat{\mu}_c$. The general form of (3.18) is a multivariate t distribution on $\nu = n_c + 1 - a$ degrees of freedom with location parameter $\hat{\mu}_c$ and scale matrix $\nu/(\nu-1)\,\Sigma_c$, provided that $\nu > 2$; see Geisser & Cornfield (1963) and Ripley (1996, pp. 49–50). We shall assume that unique estimates of the relevant parameters are obtained so that from Section 2.1.1 (on page 12) $\widehat{\Sigma} = \widehat{\Omega}^{-1}$ and $\hat{\beta} = \widehat{\Sigma}^{-1}\hat{\mu}$ and by writing $\widehat{\Sigma}_c = \widehat{\Omega}^{(c)-1}$ and $\hat{\mu}_c = \widehat{\Omega}^{(c)-1}\hat{\beta}^{(c)}$ density (3.18) may also be expressed as

$$\begin{aligned} f_c(y) &= \left\{\frac{1}{\pi(n_c+1)}\right\}^{q/2} \frac{\Gamma\{\frac{1}{2}(n_c+1-a+q)\}}{\Gamma\{\frac{1}{2}(n_c+1-a)\}} |\widehat{\Omega}^{(c)}|^{1/2} \\ &\quad \times \left(1 + \frac{1}{n_c+1}\left[y'\widehat{\Omega}^{(c)}y - 2y'\hat{\beta}^{(c)} + \hat{\beta}^{(c)\prime}\widehat{\Omega}^{(c)-1}\hat{\beta}^{(c)}\right]\right)^{-(n_c+q-a+1)/2}. \end{aligned} \tag{3.19}$$

The effect of using (3.18), or equivalently (3.19), rather than a normal density with covariance $\widehat{\Sigma}_c$ is to inflate the estimated posterior probabilities in the tails of the distribution. Estimated densities for four samples from a univariate normal distribution with mean 0 and variance 1 are shown in Figure 3.1. The predictive density, in comparison with the plug-in estimates, is less sharp with greater mass in the tails of the density, at least for small to moderate sample size. The effect of increased spread in the predictive density may also be seen in two dimensions (see Figure 3.2). The contours for the predictive density show greater spread and lower peak than for the two plug-in estimates. The predictive methods described here often give nearly the same class classification as their non-predictive analogues but the estimated probabilities of class membership may be quite different.

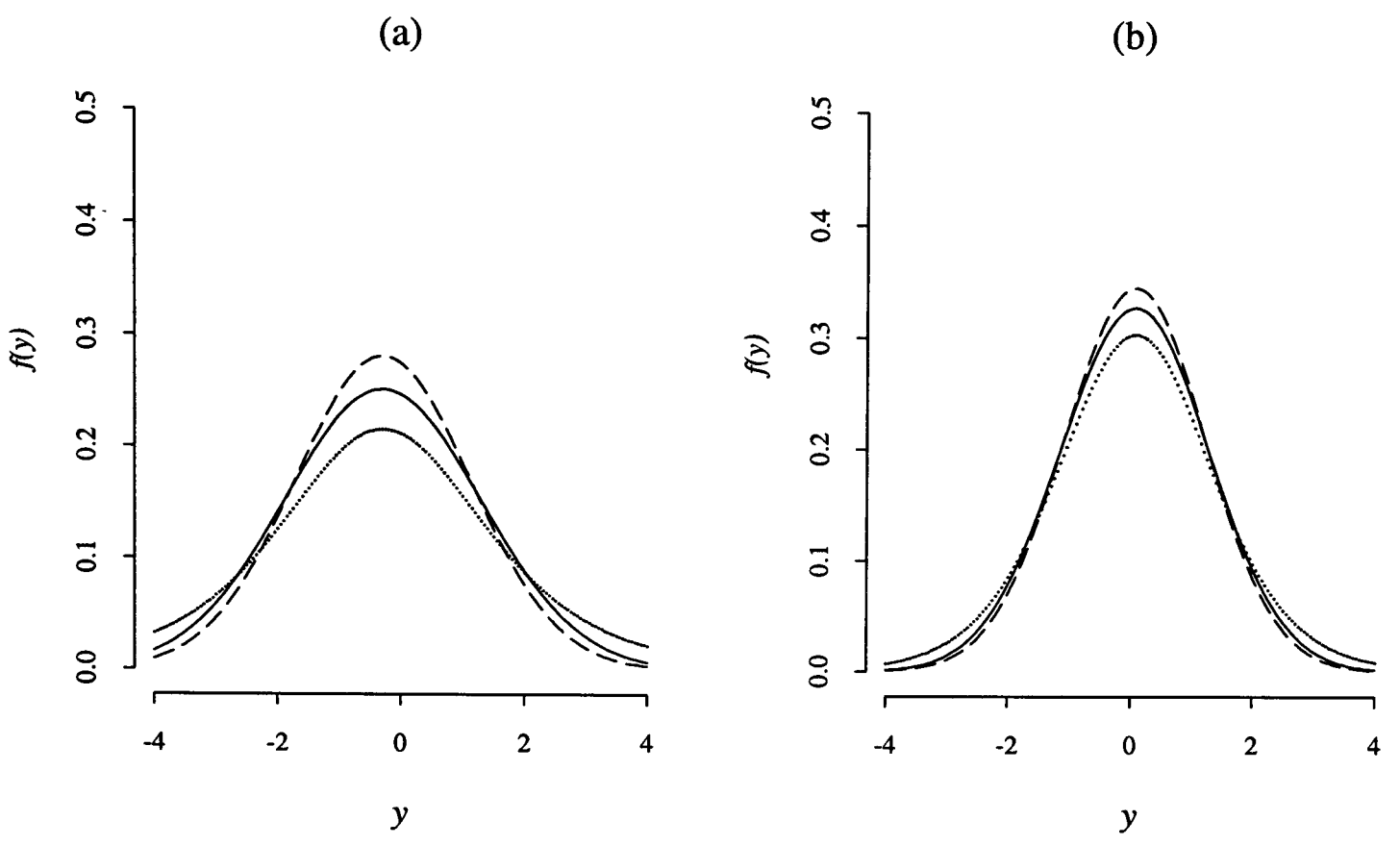


Figure 3.1 Estimates of the density based on a random sample of size n from a $N(0,1)$ distribution for (a) $n = 5$, (b) $n = 10$, (c) $n = 20$ and (d) $n = 40$. In each plot, the predictive estimate is shown with a dotted line (3.18), the 'plug-in' estimate using the usual unbiased estimate for the variance by a solid line and the 'plug-in' estimate using the maximum likelihood estimate with a dashed line.

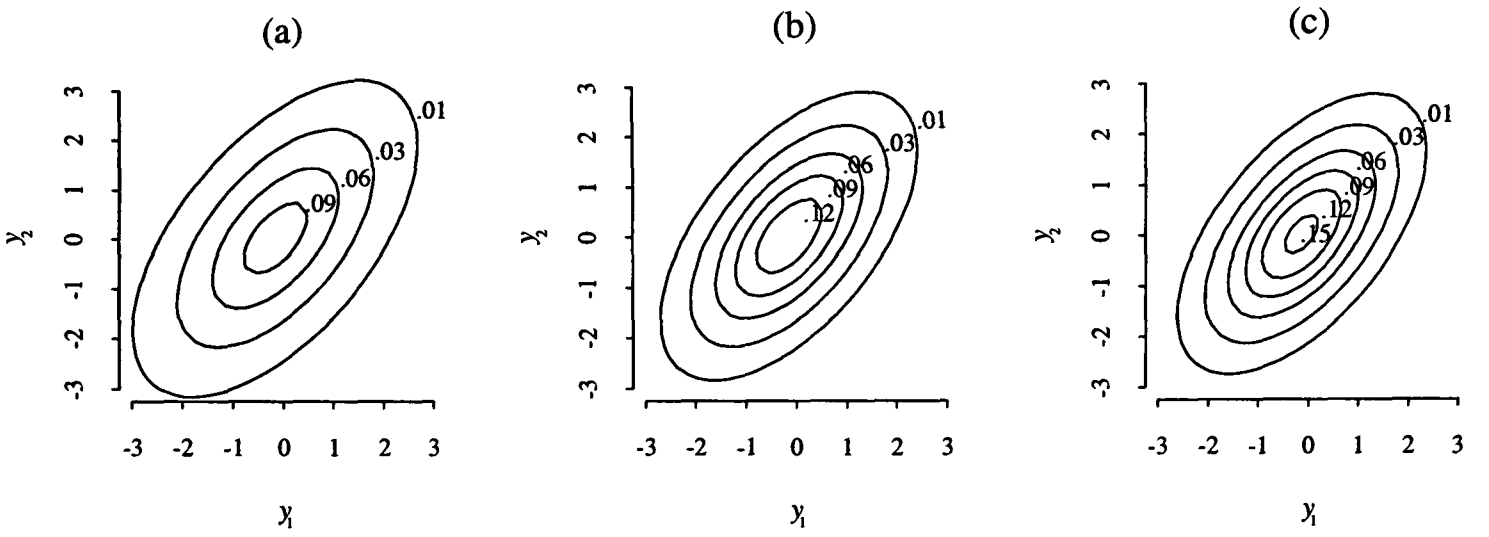


Figure 3.2 Contours of the estimated density based on a random sample of size 10 from a bivariate normal distribution with mean vector $\mu = (0,0)'$ and covariance matrix $\text{vec}\{\Sigma\} = (1,.5,.5,1)'$. Plot (a) shows the predictive density (3.18), (b) the 'plug-in' estimate using the unbiased estimate for Σ and (c) the 'plug-in' estimate using the maximum likelihood estimate for Σ.

If we assume that $\Sigma_c = \Sigma$ for all $c = 1, \ldots, g$, then a similar argument to that used previously yields the predictive density

$$f_c(y) = \left\{\frac{n_c}{n\pi(n_c+1)}\right\}^{q/2} \frac{\Gamma\{\frac{1}{2}(n-g+q-a+2)\}}{\Gamma\{\frac{1}{2}(n-g-a+2)\}} |\widehat{\Sigma}|^{-1/2} \times \left[1 + \frac{n_c}{n(n_c+1)}(y-\hat{\mu}_c)'\widehat{\Sigma}^{-1}(y-\hat{\mu}_c)\right]^{-(n-g+q-a+2)/2}, \tag{3.20}$$

where $n = \sum_c n_c$. This is a multivariate t distribution on $n-g-a+2$ degrees of freedom with location vector $\hat{\mu}_c$ and scale matrix

$$\frac{(1+1/n_c)n}{n-g-a+2}\widehat{\Sigma},$$

Ripley (1996, p. 51). By substituting Ω^{-1} for Σ and $\Omega^{-1}\beta^{(c)}$ for μ_c, density (3.20) may be written as

$$f_c(y) = \left\{\frac{n_c}{n\pi(n_c+1)}\right\}^{q/2} \frac{\Gamma\{\frac{1}{2}(n-g+q-a+2)\}}{\Gamma\{\frac{1}{2}(n-g-a+2)\}} |\widehat{\Omega}|^{1/2} \times \left(1 + \frac{n_c}{n(n_c+1)}\left[y'\widehat{\Omega}y - 2y'\hat{\beta}^{(c)} + \hat{\beta}^{(c)'}\widehat{\Omega}^{-1}\hat{\beta}^{(c)}\right]\right)^{-(n-g+q-a+2)/2}. \tag{3.21}$$

Unknown class priors

A suitable prior density for the unknown $\{\pi_c\}$ is given by the Dirichlet form

$$p(\pi_1, \ldots, \pi_{g-1}) \propto \prod_{l=1}^{g} \pi_l^{\xi_l - 1}, \tag{3.22}$$

where the ξ_l are positive constants that may be chosen so as to reflect previous frequencies or intuitive impressions. (Note that $\sum \pi_l = 1$ so that π_g is replaced by $1 - \sum_{l=1}^{g-1} \pi_l$.) In the absence of prior information we can simply set $\xi_c = \xi$ for all c. Using Jeffreys' rule, we might choose ξ to be $1/2$ (see Box & Tiao, 1973, p. 55).

If we assume the multinomial density for the class counts, n_c (see Definition 1.2 on page 7), then the posterior density of the π_c's is given by

$$p(\pi_1, \ldots, \pi_{g-1} \mid n_1, \ldots, n_{g-1}) \propto \prod_{l=1}^{g} \pi_l^{n_l + \xi_l - 1}, \tag{3.23}$$

which is formed by multiplying the Dirichlet prior by the multinomial likelihood. (Note that $n_g = n - \sum_{l=1}^{g-1} n_l$.) From Geisser (1964, p. 74) using (3.22), conditioning (3.23) on y and integrating out over the unknown parameters we obtain

$$p(c \mid y) = \frac{(n_c + \xi_c) f_c(y)}{\sum_{l=1}^{g} (n_l + \xi_l) f_l(y)}, \tag{3.24}$$

where $f_\cdot(y)$ is obtained from either (3.18) or (3.20). The predictive classification rule is then formed by using (3.24) in (3.2) for a new observation $y = y_0$ and replacing $\pi_\cdot$ with $(n_l + \xi_\cdot)$. In practice, the simplest choice for the Dirichlet is to take $\xi_\cdot \equiv 0$, giving the plug-in rule $\hat{\pi}_c = n_c/n$, see Geisser & Cornfield (1963) and also Ripley (1996, pp. 53–54). The approach we adopt for CG predictive classification is different.

3.6.1 CG predictive classification

In the CG case, we can extend the results of the previous section. From (3.19) we obtain

$$\begin{aligned} f_c(i, y) &= \left\{ \frac{1}{\pi(n_c(i)+1)} \right\}^{q/2} \frac{\Gamma\{\frac{1}{2}(n_c(i) + 1 + q - a)\}}{\Gamma\{\frac{1}{2}(n_c(i) + 1 - a)\}} |\widehat{\Omega}(i)^{(c)}|^{1/2} \\ &\times \left(1 + \frac{1}{n_c(i)+1} \left[y' \Omega(i)^{(c)} y - 2 y' \hat{\beta}(i)^{(c)'} + \hat{\beta}(i)^{(c)'} \Omega(i)^{(c)-1} \hat{\beta}(i)^{(c)} \right] \right)^{-(n_c+q-a+1)/2}, \end{aligned} \tag{3.25}$$

where $n_c = \sum_i n_c(i)$. In the linear case (i.e. with common group covariance matrices) we have from (3.21)

$$\begin{aligned} f_c(i, y) &= \left\{ \frac{n_c(i)}{n(i)\pi(n_c(i)+1)} \right\}^{q/2} \frac{\Gamma\{\frac{1}{2}(n(i) - g + q - a + 2)\}}{\Gamma\{\frac{1}{2}(n(i) - g - a + 2)\}} |\widehat{\Omega}(i)|^{1/2} \\ &\times \left(1 + \frac{n_c(i)}{n(i)(n_c(i)+1)} \left[y' \widehat{\Omega}(i) y - 2 y' \hat{\beta}(i)^{(c)} \right.\right. \\ &\qquad \left.\left. + \hat{\beta}(i)^{(c)'} \widehat{\Omega}(i)^{-1} \hat{\beta}(i)^{(c)} \right] \right)^{-(n(i)-g+q-a+2)/2}. \end{aligned} \tag{3.26}$$

With a common covariance throughout

$$f_c(i,y) = \left\{\frac{n_c(i)}{n\pi(n_c(i)+1)}\right\}^{q/2} \frac{\Gamma\{\frac{1}{2}(n-g+q-a+2)\}}{\Gamma\{\frac{1}{2}(n-g-a+2)\}} |\widehat{\Omega}|^{1/2} \times \left(1+\frac{n_c(i)}{n(n_c(i)+1)}\left[y'\widehat{\Omega}y - 2y'\hat{\beta}(i)^{(c)} + \hat{\beta}(i)^{(c)\prime}\widehat{\Omega}^{-1}\hat{\beta}(i)^{(c)}\right]\right)^{-(n-g+q-a+2)/2}. \tag{3.27}$$

One way of obtaining posterior cell probabilities for CG distributions is to extend the approach described on page 56 for the unknown $\{\pi_c\}$. This assumed a Dirichlet form for the unknown class probabilities. Vlachonikolis (1990) used this approach with Krzanowski's location model by fitting a second-order log-linear model for the cell counts after taking a common value of $\xi_\cdot = \xi_c + 1$ for all cells within group c. In this section we adopt an alternative approach that yields posterior cell probabilities. The procedure we employ uses a predictive logistic model applied to the discrete data. The method is one of five methods compared by Aitken (1978) for discrimination based on multivariate binary data. We concentrate on the multivariate binary case with two groups.

Recall the logistic response curve (3.12) given on page 50 in which β was a $p \times 1$ parameter vector and α was a scalar parameter. Let $\theta = (\alpha, \beta')'$ and let x be replaced by a $p \times 1$ binary observation vector i. Set $j = (1, i')'$ and let $p_2(j) = p(2 \mid j)$ so that

$$\begin{aligned} p_2(j) &= \exp(\theta' j)/\{1+\exp(\theta' j)\} \\ &\approx \Phi\left\{(\theta' j)/\sqrt{a}\right\}, \end{aligned} \tag{3.28}$$

where according to an empirical study of Aitchison & Begg (1976) a sensible choice for a is the value 2.942. The approximation is based on the fact that the logistic function $e^x/(1+e^x)$ is well approximated by the cumulative distribution function of a zero-mean normal random variable with a suitably selected quantile (see Cox & Snell, §1.5). The value 2.942 gives agreement of the two curves at the 90% point. In the predictive framework we have from Aitken (1978)

$$\begin{aligned} p_2(j,D) &= \int p_2(j,\theta)p(\theta \mid D)d\theta \\ &= \int \exp(\theta' j)/\{1+\exp(\theta' j)\}\phi(\theta \mid \hat{\theta}, \hat{V})d\theta \\ &\approx \Phi\left\{\hat{\theta}' j \Big/ (a + j'\hat{V}j)^{1/2}\right\}, \end{aligned} \tag{3.29}$$

where we assume $\theta \sim N(\hat{\theta}, \hat{V})$, $\hat{\theta}$ is the maximum likelihood estimate of the coefficients in the logistic regression, $\hat{V}$ is an estimate of the covariance matrix of the maximum likelihood estimate and D is the data-matrix containing all observations. The fitting procedure is relatively straightforward and may be accomplished by maximizing the likelihood for the logistic model using the quasi-Newton procedure outlined in Appendix B. Thus, we seek to maximize the conditional log likelihood

$$\log L = \sum_j \left[n_1(j)\log p_1(j) + n_2(j)\log p_2(j)\right], \tag{3.30}$$

where $p_2(j) = \exp(\theta' j)/\{1 + \exp(\theta' j)\}$ and $p_1(j) = 1 - p_2(j)$. Use of the procedure outlined in Appendix B yields $\hat{V}$ as a by-product of the estimation scheme. Note that the derivatives of $\log L$ may be easily obtained via

$$\frac{\partial \log L}{\partial \theta_k} = \sum_j \Big[n_1(j) - n(j)p_1(j)\Big] j_k, \quad k = 0, ..., p;$$

see Anderson (1982).

There are two pitfalls that may be encountered during maximization: the first, is the problem of 'complete separation', which occurs when all points in each population are separated from points of all other populations by a hyperplane. Such configurations lead to non-unique maxima at infinity; the second problem occurs when there are zero marginal proportions with discrete data, which also leads to multiple maxima at infinity. Both these problems, their avoidance and features are discussed by Albert & Anderson (1984). From a practical viewpoint we did not encounter either of these problems with the datasets analysed in this thesis. However, it was not always possible to obtain a reasonable matrix, $\hat{V}$, in which case we resorted to using Equation (3.28) rather than (3.29). Finally, we estimate the posterior probabilities over the full set of variables via

$$\hat{p}(2 \mid x) = \hat{p}(2 \mid i, y) = \frac{\hat{p}_2(j)\hat{f}_1(i, y)}{\hat{p}_1(j)\hat{f}_1(i, y) + \hat{p}_2(j)\hat{f}_2(i, y)}, \tag{3.31}$$

where $j = (1, i')'$.

3.7 Nearest neighbour methods of classification

Nearest neighbour (NN) methods of classification are based on the assumption that observations which lie close to each other are likely to belong to the same class. A suitable metric, e.g. Euclidean distance, is used to determine how close one observation is to another on the basis of its observed measurements. Suppose that we have a training set, $\mathcal{T}$ on g classes each with n_c observations in each class so that $n = \sum n_c$. Here we assume that the prior probability is proportional to the number of observations in each class or that there are equal numbers in each of the classes. We now want to classify a new observation $x_{(0)}$, say, independent of $\mathcal{T}$ we first determine the nearest neighbour $x_{(*)}$ in $\mathcal{T}$ to $x_{(0)}$. We then assign $x_{(0)}$ to the class of $x_{(*)}$, which we shall assume is c. Thus,

$$\hat{c} = c \text{ if } \delta(x_{(0)}, x_{(*)}) = \min_{\nu=1,...,n} \delta(x_{(0)}, x_{(\nu)}),$$

where δ is some metric of the feature space. This defines the nearest-neighbour rule. The k-NN rule involves searching the k nearest observations in $\mathcal{T}$ to $x_{(0)}$ and assigning $x_{(0)}$ to one of the g classes by majority vote amongst the k-NN, or, equivalently by estimating the posterior probabilities via

$$\hat{p}(c \mid x_0) = \frac{k_c}{k},$$

where k_c is the number of points that occur from class c in k-nearest neighbours. In order to render the analysis feasible, it is necessary to impose some theoretical restrictions

that ensure the convergence to $x_{(0)}$ of the nearest neighbour $x_{(*)}$ as the cardinality of the set $\mathcal{T}$ grows arbitrarily large. These restrictions are discussed in detail by Devijver & Kittler (1982). The main advantages of k-NN methods is that they can be used with any sort of distance or similarity metric, with both discrete and continuous variables, and also with missing values (given an appropriate metric).

3.7.1 A measure of similarity for mixed data types

Here we use a general measure of similarity for binary and continuous observations suggested by Gower (1971). This measure of similarity is suitable for use in k-NN procedures when the measured variables are of mixed type.

Let $x^{\nu_1} = (i^{\nu_1}, y^{\nu_1})'$ and $x^{\nu_2} = (i^{\nu_2}, y^{\nu_2})'$ be the measurement vectors for two individuals ($\nu_1 \neq \nu_2$), where i^{ν_1}, i^{ν_2} are $p \times 1$ binary-valued vectors and y^{ν_1} and y^{ν_2} are $q \times 1$ real-valued vectors. Measurement vectors x^{ν_1} and x^{ν_2} are compared on the basis of a variable $u \in V$ (where V is the full set of random variables) and assigned a score $s(x_u^{\nu_1}, x_u^{\nu_2})$. We set $s(x_u^{\nu_1}, x_u^{\nu_2}) = 0$ when the measurements are considered completely different, otherwise $s(x_u^{\nu_1}, x_u^{\nu_2})$ is some positive fraction. Let $\delta(x_u^{\nu_1}, x_u^{\nu_2}) = 1$ when $x_u^{\nu_1}$ and $x_u^{\nu_2}$ may be compared on variable u and 0 otherwise (possibly because of missing data, or 0–0 binary matches). Similarity between ν_1 and ν_2 on the basis of $u \in V$ is then defined as

$$S(\nu_1, \nu_2) = \sum_{u \in V} s(x_u^{\nu_1}, x_u^{\nu_2}) \; / \sum_{u \in V} \delta(x_u^{\nu_1}, x_u^{\nu_2}). \tag{3.32}$$

$S(\nu_1, \nu_2)$ is undefined for all $\delta(x_u^{\nu_1}, x_u^{\nu_2}) = 0$. When all comparisons are possible the denominator in (3.32) is equal to the number of elements in V, otherwise it is the number of elements over which the comparison is made. The scores, $s(x_u^{\nu_1}, x_u^{\nu_2})$, are assigned as follows:

(a) for a binary random variable $u \in V$ assign scores

$i_u^{\nu_1}$	+	+	−	−
$i_u^{\nu_2}$	+	−	+	−
$s(i_u^{\nu_1}, i_u^{\nu_2})$	1	0	0	0
$\delta(i_u^{\nu_1}, i_u^{\nu_2})$	1	1	1	0

(b) for a qualitative random variable $u \in V$ assign scores

$$s(y_u^{\nu_1}, y_u^{\nu_2}) = 1 - |y_u^{\nu_1} - y_u^{\nu_2}|/r_u,$$

where r_u is the range of the continuous variable $u \in V$ in the sample. If population values are known then r_u could be chosen using these values. When $y_u^{\nu_1} = y_u^{\nu_2}$, $s(y_u^{\nu_1}, y_u^{\nu_2}) = 1$ and when $y_u^{\nu_1}$ and $y_u^{\nu_2}$ are at the opposite ends of their range $s(y_u^{\nu_1}, y_u^{\nu_2})$ is a minimum (equal to 0 when r_u is determined from the sample). With intermediate values $s(y_u^{\nu_1}, y_u^{\nu_2})$ is a positive function.

3.8 Examples

We can fit CG discrimination models using CGM by including a discrete variable that gives the (assumed true) class of each observation. The data may then be modelled in the joint framework and then by collapsing over the levels of the classification variable

we can obtain CG discriminants. For example, the Iris dataset contains four continuous measurement variables W, X, Y and Z. We also have the initial classifications of the *Irises* into three species. We label the classification variable as I. We then model $f(I, W, X, Y, Z)$ via the CG density and form

$$\begin{aligned} R_{12} &= \log\{f(W,X,Y,Z|I=1) \,/\, f(W,X,Y,Z|I=2)\}, \\ R_{23} &= \log\{f(W,X,Y,Z|I=2) \,/\, f(W,X,Y,Z|I=3)\}, \\ R_{13} &= \log\{f(W,X,Y,Z|I=1) \,/\, f(W,X,Y,Z|I=3)\}, \end{aligned}$$

which are the CG discriminants.

3.8.1 Classification of Iris species

We start by forming a training set by taking a random sample of 25 observations from each of the three groups (*I.setosa*, *I.versicolor* and *I.virginica*). The remaining 75 observations will be used for testing the classification procedure. These data were analysed in Chapter 2. Here we perform a similar analysis but assume a common inverse covariance matrix. The HCG model we start with is $I \;/\; IW, IX, IY, IZ \;/\; WXYZ$. Model selection via chi-squared tests yields the HCG model $I \;/\; IW, IX, IY, IZ \;/\; WX, XZ, WY, YZ$. Clearly, the structure of the model is similar to that found previously but here the grouping variable does not appear in the quadratic part of the model. Of course, the parameter estimates are different. The CG discriminants $\widehat{R}_{12}$, $\widehat{R}_{23}$ and $\widehat{R}_{13}$ estimated using the training set of 75 observations are given below in Table 3.2. There are clearly only two linearly independent HCG discriminants as $\widehat{R}_{12} + \widehat{R}_{23} = \widehat{R}_{13}$. Note that approximate standard errors are obtained from the computed inverse Hessian by assuming

$$\mathrm{Var}(U - V) = \mathrm{Var}(U) + \mathrm{Var}(V) - 2\mathrm{Cov}(U, V),$$

for parameters U and V. Comparison of the coefficients in relation to their (approximate) standard errors shows all the coefficients of the first HCG discriminant to be important. For the second function, the coefficients for variables Y and Z are more important than the coefficients for variables W and X. This feature is also revealed in Figure 2.5 (on page 40) in which the edges IY and IZ are thicker than the edges between IW and IX.

Table 3.2 Estimated HCG discriminant function coefficients with approximate standard errors given in parentheses. The underlying CG model is given by $I \;/\; IW, IX, IY, IZ \;/\; WX, XZ, WY, YZ$.

Variable	$\widehat{R}_{12}$	$\widehat{R}_{23}$	$\widehat{R}_{13}$
constant	−48.3 (26.2)	23.0 (8.3)	−25.3 (31.3)
sepal length (W)	45.8 (20.5)	5.2 (5.7)	50.9 (24.4)
sepal width (X)	53.8 (10.8)	4.6 (3.5)	58.4 (12.3)
petal length (Y)	−110.5 (19.5)	−22.5 (5.2)	−133.0 (23.7)
petal width (Z)	−18.8 (6.0)	−3.9 (1.7)	−22.6 (7.2)

Figure 3.3 overleaf gives frequency histograms for function scores for the 75 test observations via the estimated HCG discriminants $\widehat{R}_{12}$ and $\widehat{R}_{23}$. We assume equal prior

probabilities which means the classification boundaries occur at zero. The first discriminant distinguishes the *I. virginica* and *I. versicolor* from the *I. setosa* species, and the second discriminant separates the *I. virginica* species from *I. versicolor* and *I. setosa*. Clearly, there is some overlap between *I. virginica* and *I. versicolor*. However, the *I. setosa* species is well separated from the other two. The superimposed density estimates indicate that it is not completely clear whether there are two or three species present. The less smoothed or lower bandwidth density estimates suggest three groups, whereas the more smoothed density estimates suggest two distinct groups only. Note that Fisher's hypothesis of *I. versicolor* being two-thirds of the way between *I. setosa* and *I. virginica* is not true because we have transformed the raw measurements by taking natural logarithms. Fisher used a hypothesis test in finally deciding that the plants were in agreement with the two-thirds hypothesis after adjusting for the different species covariance matrices. Applying the first two fitted discriminant functions to the test data dataset gives 1 *I. versicolor* and 1 *I. virginica* misclassified. Using the predictive density given by (3.21) on page 55 but this time with the re-fitted saturated HCG model we get the same 2 plants misclassified.

3.8.2 Classification of low birth weights

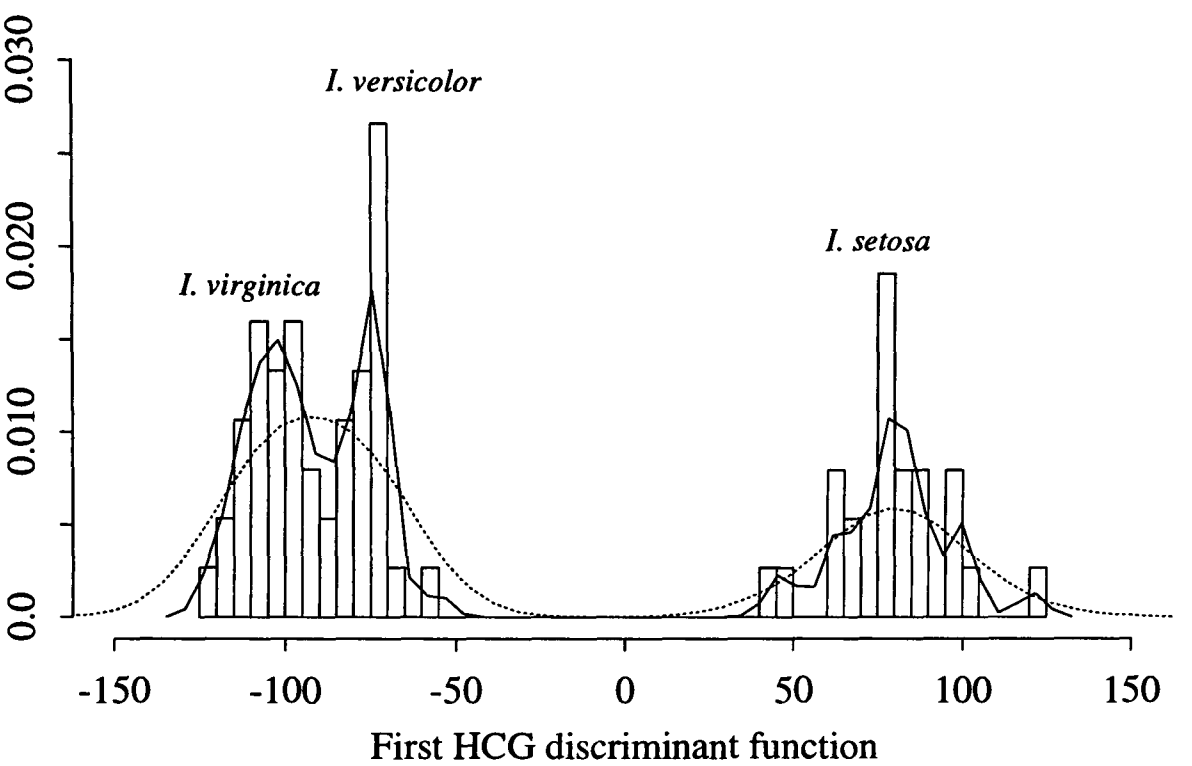
We now turn our attention again to the low birth weights example considered in Section 2.8.2. In order to classify observations in this example we employ posterior densities (3.27), (3.29) together with (3.31); see pages 57–58. We use the initial CG model defined by $IA, IB, IC, IE \mid IX, IY \mid X, Y$ in the classification procedure. The resulting 10-fold error rates are given in Table 3.3. (Error rate estimation is described in Appendix F.) Clearly, the larger non-*LBW* class tends to dominate the classification procedure with around 53% of cases being misclassified in the smaller *LBW* class. There is very good agreement between the raw and smoothed error-rates. Note that the smoothed error-rates have a much smaller standard error. If the error rate is found by counting R errors in a test set of size N, then R has a binomial(N,P) distribution. The standard error of $\hat{p} = R/N$ is then $\sqrt{\hat{p}(1-\hat{p})/N}$; see, e.g. Ripley (1996). The standard errors for the smoothed rates are found by calculating the standard deviation of the error probabilities. The class conditional error rates are obtained by dividing by the class size the sum of those probabilities less than 0.5. Fewer misclassifications will give a lower overall error-rate and so to will smaller misclassification error probabilities. (Note that we do not rely on the a priori assigned classifications for the smoothed rates.)

Table 3.3 Percentage error rates for the HCG model $IA, IB, IC, IE \mid IX, IY \mid X, Y$ used to classify low birth weights. Corresponding standard errors are given in parentheses. The column labelled d gives the total number of fitted parameters. The raw error rates are calculated using the proportion of misclassified observations made in the assumed true classes. The smoothed error rates are calculated by averaging over the posterior probabilities alone. The class sizes are 125 (non-LBW) and 59 (LBW).

	v-fold error rates % (v =10)					
	raw			smoothed		
d	non-*LBW*	*LBW*	overall	non-*LBW*	*LBW*	overall
15	21.0 (3.6)	53.0 (6.5)	31.0 (3.4)	21.0 (0.2)	52.9 (0.2)	31.2 (0.1)

Logistic regression models may be fitted using the generalized linear model routine (`glm`) in the computer package S-Plus; see Hastie & Pregibon (1992) and also Venables

(a)



(b)

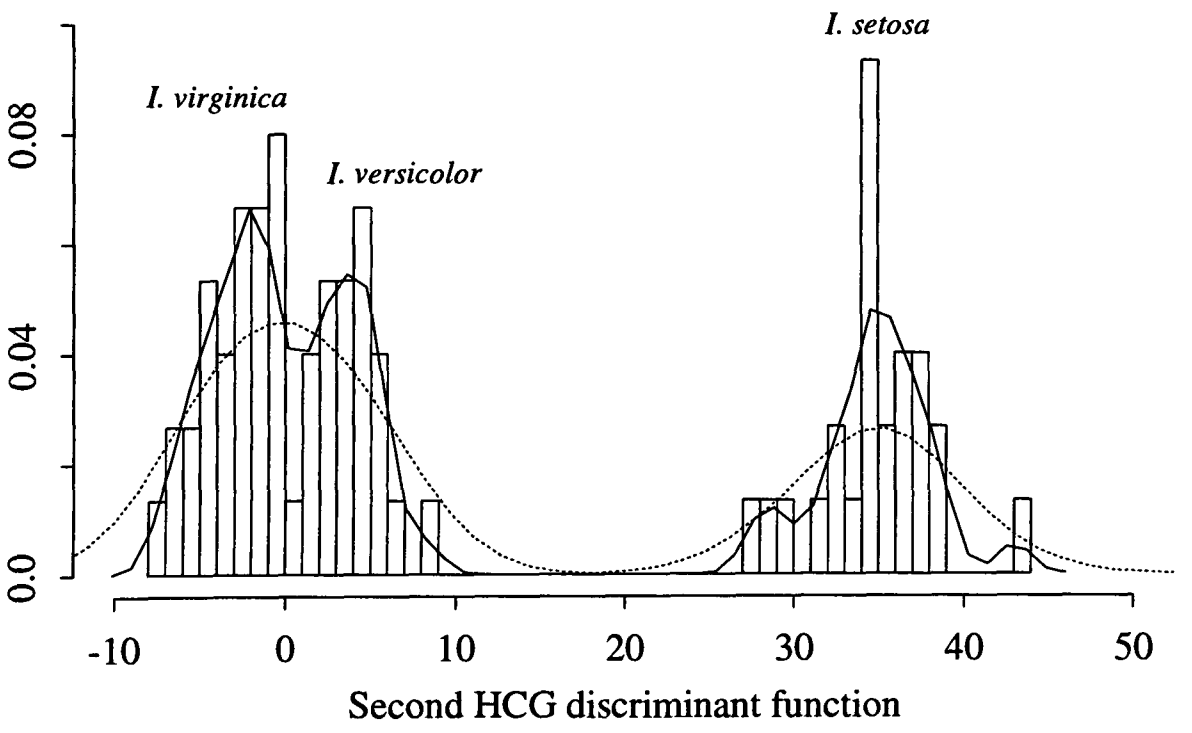


Figure 3.3 Frequency histograms and superimposed density estimates of the first ($\widehat{R}_{12}$) and second ($\widehat{R}_{23}$) HCG discriminant function scores for the test set of irises. The densities drawn in (a) use a bandwidth of 16 (solid line) and 64 (dotted line), and in (b) a bandwidth of 4 (solid line) and 16 (dotted line).

& Ripley (1997). Table 3.4 gives the error rates for two logistic regression models fitted to the low birth weights dataset. The first model was found using a stepwise model selection routine via AIC, initially including variables *race* and *ftv*, which were not included in the CG analysis. The final model selected is given by

$$LBW = \text{const} + age + lwt + smoke + ptl + ht + ui + ftv + age \times ftv + smoke \times ui.$$

Note that *race* was not included in the final model. In addition, no three-way interactions were selected. The second model was again found using stepwise model selection but excludes variables *race* and *ftv*. Here only main effects were chosen, the final model selected was

$$LBW = \text{const} + age + lwt + smoke + ptl + ht + ui.$$

Both *age* and *lwt* are on a natural log scale for both models.

Table 3.4 Percentage error rates for two logistic regression models used to classify low birth weights. Model 1 includes main effects *age*, *lwt*, *smoke*, *ptl*, *ht*, *ui*, *ftv* and two-way interactions *age*×*ftv*, *smoke*×*ui*. Model 2 includes main effects *age*, *lwt*, *smoke*, *ptl*, *ht* and *ui* only. Corresponding standard errors are given in parentheses. The column labelled d gives the total number of fitted parameters for each model. The class sizes are: 125 (non-*LBW*) and 59 (*LBW*).

		v-fold error rates % (v=10)					
		raw			smoothed		
model	d	non-*LBW*	*LBW*	overall	non-*LBW*	*LBW*	overall
1	12	14.4 (3.1)	57.6 (6.4)	28.3 (3.3)	11.0 (1.3)	49.7 (6.6)	23.6 (1.0)
2	8	12.8 (3.0)	67.8 (6.1)	30.4 (3.4)	9.0 (1.2)	59.8 (7.9)	23.6 (0.9)

Referring to Table 3.4 we see that logistic discrimination appears to perform better than the HCG model with fewer errors overall. Comparing the results of the CG analysis with the logistic regression analysis, shows fewer errors in the LBW class for the CG analysis. The standard errors on the raw rates are much the same as those estimated for the CG model. However, the logistic smoothed rates show more variability.

Results for k-NN classification of the low birth weights dataset are given in Table 3.5. Overall error rates are worse than for the graphical HCG models and logistic discrimination. As with the other methods the classification error rates for the *LBW* class are rather poor in comparison with the larger non-*LBW* class. The poor performance of k-NN is probably due to the fact that we have a small sample where the classes appear to be heavily inter-mingled.

Table 3.5 Percentage error rates using k-nearest neighbour classification for low birth weights. Corresponding standard errors are given in parentheses. The variables are A - *smoke*, B - *ht*, C - *ui*, E - *ptl*, X - *age* and Y - *lwt*. The class sizes are: 125 (non-*LBW*) and 59 (*LBW*). (Both *age* and *lwt* are on a $\log_e$ scale.)

Raw v-fold error rates % (v =10)

	1-NN			3-NN		
Variables	non-*LBW*	*LBW*	overall	non-*LBW*	*LBW*	overall
A, B, C, E, X, Y	32.0 (4.2)	76.3 (5.5)	46.2 (3.7)	25.6 (3.9)	78.0 (5.4)	42.4 (3.6)
A, E, Y	36.0 (4.3)	72.9 (5.8)	47.8 (3.7)	20.8 (3.6)	79.7 (5.2)	39.7 (3.6)
E, Y	36.0 (4.3)	72.9 (5.8)	47.8 (3.7)	22.4 (3.7)	78.0 (5.4)	40.2 (3.6)

	5-NN			9-NN		
Variables	non-*LBW*	*LBW*	overall	non-*LBW*	*LBW*	overall
A, B, C, E, X, Y	17.6 (3.4)	84.7 (4.7)	39.1 (3.6)	6.4 (2.2)	88.1 (4.2)	32.6 (3.5)
A, E, Y	14.4 (3.1)	84.7 (4.7)	37.0 (3.6)	13.6 (3.1)	78.0 (5.4)	34.2 (3.5)
E, Y	15.2 (3.2)	78.0 (5.4)	35.3 (3.5)	12.8 (3.0)	79.7 (5.2)	34.2 (3.5)

Chapter 4

Scandinavian Small-for-Gestational Age Births Study

To study the aetiology and consequences of intrauterine growth retardation, a prospective study was organized by the US National Institute of Child Health and Human Development. This study was conducted by the Universities of Trondheim and Bergen in Norway, Uppsala in Sweden and Alabama in the US. Preliminary results from this collaborative study are given by Bergsjø *et al.* (1989). Here we concentrate on the Scandinavian portion of the study. These data are also analysed by Jacobsen (1992) and Bakketeig *et al.* (1993). Bakketeig *et al.* suggest that populations of pregnant women from Scandinavian countries are most suitable for the study of foetal growth. This is because women are generally in good health and health care is available to all women on equal terms.

A diagnosis of foetal growth retardation also known as small-for-gestational age (SGA) birth is typically assigned to newborn infants with a birth weight below the 10th percentile for gestational age. The weight versus gestational age reference standards used in this study are for women who have given birth to one or more children based on last menstrual period (LMP) dating and infant sex, constructed using data from the Norwegian Medical Birth Registry (see Bjerkedal & Skjærven, 1980). Bakketeig *et al.* also take newborn infants with a birth weight below the 15th percentile for gestational age as an indicator of SGA birth. This is because ultrasound, rather than LMP, dating of gestational age was used in the study, which according to Bakketeig *et al.* gives a lower estimate of gestational age by between three to seven days when compared with LMP dating. Thus, Bakketeig *et al.* argue that the nominal weight versus gestational age 15th percentile, based on the Norwegian Medical Birth Registry LMP estimates, better represents a current population based 10th percentile (see the entries for SGA births below the 10th and 15th percentiles in Table 4.1).

The aim of this chapter is to show how conditional independence graphs may be used to discriminate between SGA and non-SGA infants. Due to lack of data we choose to fit reduced (non-graphical) hierarchical CG models. We show that these models give a reasonable description of data structure and are useful in helping to identify pre-pregnancy risk factors for SGA birth. The Bayesian classification procedure described in the preceding chapter is applied to the SGA data in an attempt to derive a rule for classifying SGA infants based on pre-pregnancy risk factors. Classification error rates for the Bayesian procedure are compared with error rates for logistic regression and k-nearest neighbour allocation rules.

4.1 Background

Women who were expecting their second or third child between January 1986 and March 1988, had a singleton pregnancy, were of White Scandinavian origin, spoke either Norwegian or Swedish and were registered by a study centre prior to the 20th gestational week were considered eligible for the study. Study centres were the university hospitals at Trondheim and Bergen in Norway and Uppsala in Sweden. A total of 6,354 women were recruited to the study at their first prenatal visit.

At the time of the first study visit, 432 women were excluded because they did not fulfil the study criteria. Of these, 34 had a multiple pregnancy, 229 aborted (215 spontaneously and 14 by inducement) and 169 were considered to be ineligible because of ethnic or language reasons, were not expecting their second or third child, or had a pregnancy that had gone beyond 20 weeks. In addition to the 432 women considered ineligible, 200 women failed to come to the first study visit. This left a total of 5,722 eligible women in the study.

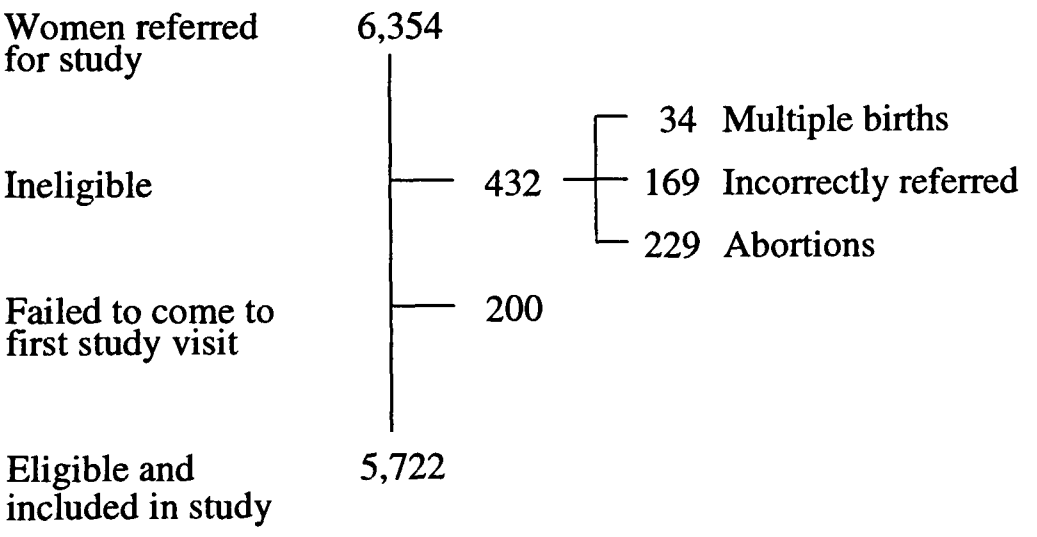


Figure 4.1 A flowchart illustrating the allocation of study subjects to groups.

The full database comprises 64 measurements on 5,722 women. (A large proportion of the measurements are longitudinal.) Of the full sample of 5,722 women, 1,945 women were selected for detailed follow up at four prenatal visits, delivery and during the first year of life (see Bakketeig *et al.*, 1993, for further details). Our analysis of the SGA data concentrates on all the 5,722 women eligible and included in the study based on a subset of the 64 database measurements. Descriptions of the measurement variables considered are given in Table 4.1. (Note that the missing values for the SGA entries in Table 4.1 will include stillbirths.) In the study, a diagnosis of *previous low birth weight* was assigned to a prior first birth of a baby girl below 2,700 grams or a baby boy below 2,800 grams, or a prior second birth of a baby girl below 2,800 grams or a baby boy below 2,900 grams. Previous low birth weight is determined without reference to gestational age. In contrast, a *previous preterm birth* is determined by gestational age alone, i.e. a birth prior to 37 weeks.

The aim of our analysis will be to model the independence structure between the variables in an attempt to identify pre-pregnancy risk factors associated with SGA birth. In addition, we apply the Bayesian method of classification (developed in the previous chapter) in order to classify SGA births. We concentrate on analysing SGA births below both the 15th and 10th birthweight percentiles for gestational age.

Table 4.1 SGA dataset variable names and descriptions. Cell counts are given for the two levels of each binary variable (percentages in parentheses). The mean and standard deviation 'stdev' are given for each continuous variable. Total sample size is 5,722. (Variable *D-cmd* chronic maternal disease includes chronic renal disease, essential hypertension, heart disease or any other chronic condition that might complicate the pregnancy.)

Binary variables				
name	description	no	yes	missing
A - pbstatus	previous infant death	5608 (98.0)	95 (1.7)	19 (0.3)
B - pblbwt	previous low birth weight infant	5191 (90.7)	531 (9.3)	0 (0.0)
C - pbpreterm	previous preterm birth	5284 (92.3)	386 (6.7)	52 (0.9)
D - cmd	chronic maternal disease	5512 (96.3)	197 (3.4)	13 (0.2)
E - disoth	med. disorders other than *D*	4656 (81.4)	1038 (18.1)	28 (0.5)
F - smoke	smoking at time of conception	3701 (64.7)	1938 (33.9)	83 (1.5)
H - SGA10	SGA below 10th percentile	5302 (92.7)	361 (6.3)	59 (1.0)
I - SGA15	SGA below 15th percentile	5083 (88.8)	581 (10.2)	58 (1.0)
		male	female	
G - sex	sex of infant	2889 (50.4)	2803 (49.0)	30 (0.5)
Continuous variables				
name	description	mean	stdev	missing
W - weight	mother's pre-pregnancy weight (kg)	61.1	9.6	97 (1.7)
X - age	mother's age (yrs)	29.8	4.4	27 (0.5)
Y - height	mother's height (cm)	166.7	5.7	70 (1.2)
Z - bwt	actual birth weight of infant (g)	3604.5	583.9	32 (0.6)

We start by looking at normal probability plots of the continuous measurements pre-pregnancy weight, and age and height of the mother. Figures 4.2 and 4.3 (see pages 68 and 69) certainly suggest taking a logarithmic transformation of weight and it is probably sensible to transform age and height using logarithms as well. Note that the normal probability plots highlight the discreteness in weight (measured to the nearest kilogram), height (measured to the nearest centimetre) and age (measured to the nearest year). Height appears to be very nearly normal, so too does age but with a small departure from normality in the right-hand tail after transformation. The *weight* variable, even after transformation, suggests a longer right-hand tail than the normal. The correlations between the transformed variables were measured as 0.037 (*age*, *weight*), 0.455 (*weight*, *height*) and 0.036 (*age*, *height*). It was thought that height might be used as a proxy for weight on the basis that remembered pre-pregnancy weight might prove unreliable. (Note that only pregnant women were selected for the study.) However, in view of the low correlation between the two measurements both are retained in the following analyses. From now on we shall refer to variables *weight*, *age* and *height* using the same labels as those given in Table 4.1 on their natural logarithmic scales unless stated otherwise.

4.2 Modelling SGA births below the 15th percentile for gestational age

Initially, binary variables A–F and G, and continuous variables W, X and Y were considered together with the group indicator, I. As a guide, a saturated heterogeneous graphical CG model describing the joint distribution of all 11 variables would require the estimation of $2^8-1=255$ discrete parameters, 256×3 linear parameters, 256×3 precisions (the diagonal elements of $\Omega(i)$) and 256×3 concentrations (the off-diagonal elements of $\Omega(i)$). Model search in such high dimensions is computationally rather time

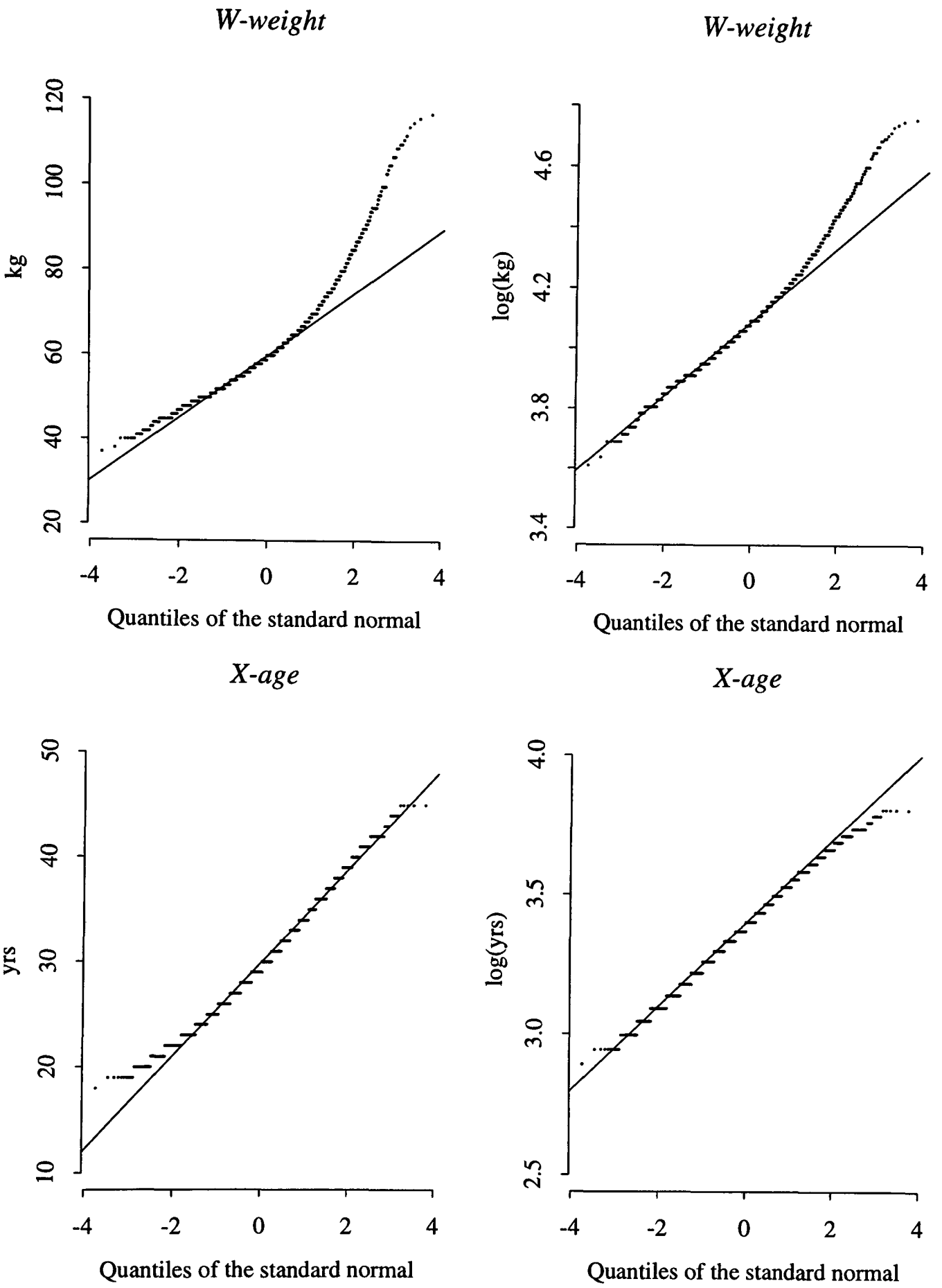


Figure 4.2 Normal probability plots for untransformed and $\log_e$ transformed variables *weight* (mother's pre-pregnancy weight) and *age* (mother's age). In each graph the straight line is drawn through the upper and lower quartiles.

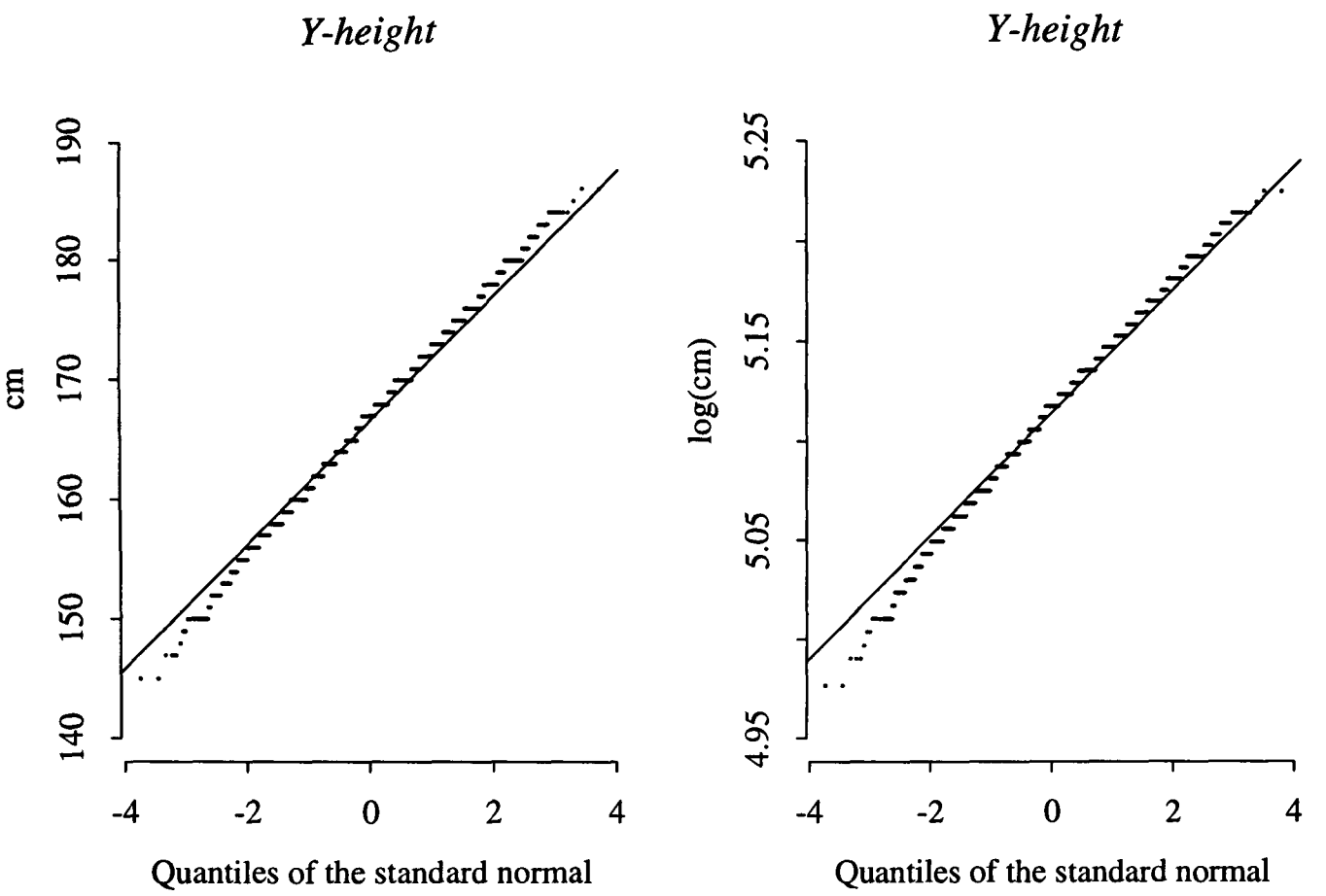


Figure 4.3 Normal probability plots for untransformed and $\log_e$ transformed variable *height* (mother's height). In each graph the straight line is drawn through the upper and lower quartiles.

consuming so some initial reduction in the number of variables was sought. Univariate significance tests were computed so as to give an approximate idea about which variables should be included in a model of the data. For comparison of I with each of the 7 explanatory binary variables a 2×2 contingency table chi-squared test was employed. For the 3 explanatory continuous variables a standard two-sample t-test was used to compare the *SGA15* and non-*SGA15* means. (Missing values were ignored when computing both chi-squared and t-tests.) The p-values for these tests were: p=0.902 (*A-pbstatus*), p<0.001 (*B-pblbwt*), p<0.015 (*C-pbpreterm*), p<0.031 (*D-cmd*), p=0.362 (*E-disoth*), p<0.001 (*F-smoke*), p=0.613 (*G-sex*), p<0.001 (*W-weight*), p<0.593 (*X-age*) and p<0.001 (*Y-height*). Variables A, E, G and X with large p-values were removed from the subsequent analysis.

4.2.1 CG modelling

Casewise deletion of missing values was performed (on the basis of those variables retained in the analysis) leaving a total of 4,896 non-*SGA15* and 552 *SGA15* observations in the sample. Examination of the observed cell counts showed ten cells with a small number of observations (i.e. less than 5) of which five cells had fewer than 2 observations and two cells were empty. It was therefore not possible to fit a saturated heterogeneous nor homogeneous graphical model to the data. In view of this, we defined an initial model containing all two-factor interactions between the binary variables together with these same two-factor interactions and their interaction with mother's weight, W, and height, Y. A check of the cell counts, totals and SSP matrices in each of the 2×2 tables indicated that there should be no problem fitting the specified sub-model. The smallest marginal cell count, i.e. smallest count over all 2×2 tables,

was found to be 19.

The starting model we chose was a heterogenous CG model. The heterogeneity in the concentration matrix was confirmed by testing a homogeneous model in W and Y in each of the two-way tables defined by the binary interactions against the corresponding heterogeneous model. Significant heterogeneity was found between XY and all two-factor binary interactions except BI, CI and DI. The p-values for the tests were found to be: $p<0.001$ for IF, $p=0.013$ for BC, $p<0.023$ and BD, $p<0.001$ for BF, $p<0.017$ for CD, $p<0.001$ for CF, $p<0.001$ for DF. Interactions BI, CI and DI were found to be non-significant with $p=0.568$, $p=0.656$ and $p=0.056$, respectively. Box's test (Box, 1949) implemented in MIM was used to test for variance homogeneity. The test is a multivariate generalization of a test of variance homogeneity due to Bartlett (1937).

Although no evidence of heterogeneity for BI, CI and DI was found, we still retained these interactions in the quadratic part of the model as we did not want to over-prune the model at this stage. Thus, the initial model was given by:

$$
\begin{aligned}
&BI, CI, DI, FI, BC, BD, BF, CD, CF, DF/\\
&BIW, CIW, DIW, FIW, BCW, BDW, BFW, CDW, CFW, DFW,\\
&BIY, CIY, DIY, FIY, BCY, BDY, BFY, CDY, CFY, DFY/\\
&BIWY, CIWY, DIWY, FIWY, BCWY, BDWY, BFWY, CDWY, CFWY, DFWY;
\end{aligned}
\tag{4.1}
$$

which is a non-graphical CG model. Model selection was performed using chi-squared tests of deviance with critical values set at 5% and 10%, and AIC. Model selection using chi-squared tests with 5% and 10% critical levels produced the same results. The resulting independence graphs for the chi-squared and AIC selected models are shown in Figure 4.4.

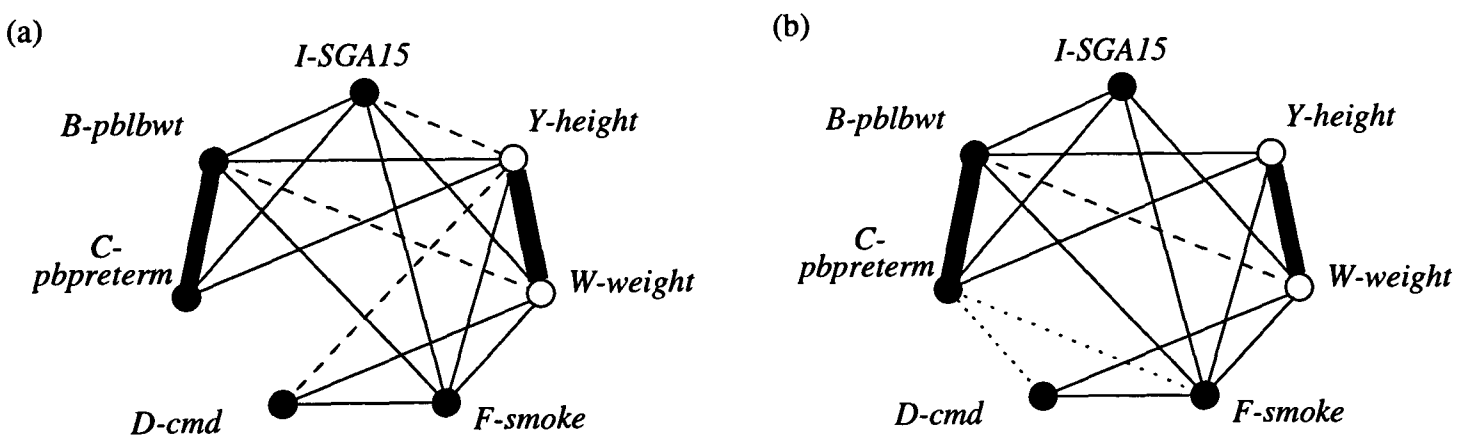


Figure 4.4 Conditional independence graphs corresponding to two (non-graphical) CG models for SGA births below the 15th birth weight percentile for gestational age. Graph (a) was obtained using chi-squared tests of deviance as the model selection criterion (both 10% and 5% critical values giving the same results), graph (b) was obtained using AIC as the model selection criterion. Edge thickness corresponds to the significance of the edge deletion deviance, the more significant the deviance the thicker the edge. The dotted edges indicate an edge that is non-significant at the 5% level ($p>0.05$) and the dashed edges indicate a non-significant edge at the 1% level but one which is significant at the 5% level (i.e. $0.01<p\leqslant 0.05$).

The conditional independence graphs for the two models are formed by removing five edges from the conditional independence graph associated with the starting model. Figure 4.4 (a) chosen using chi-squared tests of deviance. The edges removed were (in order): $[DI]$, $[CD]$, $[CF]$, $[CW]$ and $[BD]$. In comparision, AIC based model selection producing graph (b) removed (in order) edges $[CW]$, $[DY]$, $[BD]$, $[IY]$ and

$[DI]$. Both graphs are dominated by the $[BC]$ and $[WY]$ edges and for this reason we include dotted and dashed edges indicating non-significance at the 5% and 1% levels, respectively. Doing this gives an idea of the relative weight that should be placed on the less dominant edges. (Note that although the AIC selected graph (b) includes edge thicknesses drawn with respect to edge deletion deviance this is only for comparison with graph (a). Model selection was based solely on minimizing AIC.) There is a large positive correlation between B (previous low birth weight) and C (previous preterm birth), which is highlighted by the relative risk estimate given in Table 4.2 on page 72. From a practical point of view, a low birth weight infant is more likely to occur if birth takes place before term than if birth took place at 37 or more weeks. Although, as mentioned previously, a diagnosis of low birth weight excludes any reference to gestational age. It is worth noting that edge D had the least significant p-value of those variables retained in the analysis when tested against I prior to model fitting and it is the only variable unconnected with I in Figure 4.4. Note that the differences between the chi-squared and AIC selected models differ in the dashed and dotted edges, i.e. edges $[CD]$ and $[CF]$ removed in (a) are non-significant at the 5% level in (b) and edges $[DY]$ and $[DI]$ removed in (b) are non-significant at the 1% level in (a). The final chi-squared selected model is given by:

$$
\begin{aligned}
&BI, CI, FI, BC, BF, DF/\\
&BIW, FIW, BFW, DFW, BIY, CIY, FIY, BCY, BFY, DFY/\\
&BIWY, CIY, FIWY, BCY, BFWY, DFWY;
\end{aligned}
$$

which has a deviance of 41.32 on 33 degrees of freedom. The final AIC selected model is given by

$$
\begin{aligned}
&BI, CI, FI, BC, BF, CD, CF, DF/\\
&BIW, FIW, BFW, DFW, BCY, BFY, CFY/\\
&BIW, FIW, BCY, BFWY, CFY;
\end{aligned}
$$

which has a deviance of 63.56 on 46 degrees of freedom. Tables G.1 and G.2 in Appendix G give model fitting details for those edges retained in the conditional independence graph using the two model selection schemes used to produce Figure 4.4.

It is instructive to estimate relative risks among the binary variables (see Table 4.2). There is a clear threefold increase in risk of an SGA baby given a previous low birth weight baby. In addition, there is a twofold increase in SGA-risk for mothers who smoked cigarettes around the time of conception. There is also a significant increase in SGA-risk of one-and-a-half times for mothers who had previously experienced a preterm birth of one and a half times. The presence of chronic maternal diseases (essential hypertension, heart disease or renal disease) did not appear to increase SGA-risk significantly, although its relative risk estimate is on the borderline of significance at the 5% level. Looking at the relative risk estimate for previous low birth weight given smoking indicates a significant in increase in risk of one-and-a-half times. The highly significant previous low birth weight with previous preterm birth can be accounted for by the strong positive correlation in the associated 2×2 table with approximately only 6% presence/absence, absence/presence mismatches in the two factors. Although, previous preterm birth and previous low birth weight are not exactly proxies for one another there is some suggestion that one or other could be dropped from the analysis. Finally, the mean pre-pregnancy weights (with weight on its original scale) for *SGA15*

and non-*SGA15* are 58.12kg and 61.5kg with standard deviations of 9.1kg and 9.6kg, respectively; the mean heights for *SGA15* and non-*SGA15* are 165cm and 167cm (on the original height scale), respectively, both with standard deviations of 6cm.

Table 4.2 Relative risk estimates (RR) together with 95% confidence interval (95% CI). A confidence interval enclosing 1 indicates a non-significant relative risk.

Interaction	RR	95% CI
I-SGA15 with *B-pblbwt*	3.4	(2.7, 4.3)
I-SGA15 with *F-smoke*	2.2	(1.9, 2.7)
I-SGA15 with *C-pbpreterm*	1.5	(1.1, 2.1)
I-SGA15 with *D-cmd*	1.6	(1.0, 2.4)
B-pblbwt with *C-pbpreterm*	40.4	(31.4, 51.9)
B-pblbwt with *F-smoke*	1.6	(1.3, 1.9)
B-pblbwt with *D-cmd*	1.5	(1.0, 2.3)
C-pbpreterm with *D-cmd*	1.6	(1.0, 2.6)
C-pbpreterm with *F-smoke*	1.1	(0.9, 1.3)
D-cmd with *F-smoke*	1.3	(0.9, 1.7)

4.3 Modelling SGA births below the 10th percentile for gestational age

A similar analysis to that carried out with *I-SGA15* as the primary variable of interest was performed using *H-SGA10*. Chi-squared tests of H versus the other binary variables yielded p-values of p=0.667 (*A-pbstatus*), p<0.001 (*B-pblbwt*), p=0.094 (*C-pbpreterm*), p=0.002 (*D-cmd*), p=0.051 (*E-disoth*), p<0.001 (*F-smoke*) and p=0.632 (*G-sex*). For the continuous measurements standard two-sample t-tests were employed yielding p-values of p<0.001 (*W-weight*), p=0.902 (*X-age*) and p<0.001 (*Y-height*). Variables A, C, G and X were removed from the subsequent analysis. (Note that variable E has a borderline p-value, at the 5% level, that we decided to retain in the analysis.) The set of variables selected at this point is much the same as that selected for *SGA15* except that variable *D-cmd* is included instead of *C-pbpreterm*. (Again, we ignored missing values when computing the univariate tests.)

4.3.1 CG modelling

Casewise deletion of missing values was performed (on the basis of those variables retained in the analysis) leaving a total of 5125 non-*SGA10* and 350 *SGA10* observations in the sample. There were found to be a total of ten cells with fewer than 5 observations and, of these, two cells were empty. (Note that the configuration of the sparse cells was different to that for *SGA15*.) We choose to define an initial model consisting of all two-way binary interactions in the discrete part of the model, all two-way binary interactions and one continuous variable in the linear part of the model and all two-way binary interactions and the two continuous variables in the quadratic part of the model. The smallest marginal cell count for this model was found to be 21. Variance heterogeneity was checked using Box's test described previously. Significant heterogeneity was found for BD (p=0.016), BF (p<0.001), DE (p=0.009), DF (p<0.001), EF (p<0.001) and FH (p<0.001). There was no evidence of heterogeneity for BE (p=0.878) and BH (p=0.481). We selected a suitable starting model as:

$$HB, HD, HE, HF, BD, BE, BF, DE, DF, EF/$$
$$HBW, HDW, HEW, HFW, BDW, BEW, BFW, DEW, DFW, EFW/$$
$$HBY, HDY, HEY, HFY, BDY, BEY, BFY, DEY, DFY, EFY/$$
$$HBWY, HDWY, HEWY, HFWY, BDWY, BEWY, BFWY, DEWY, DFWY, EFWY. \quad (4.2)$$

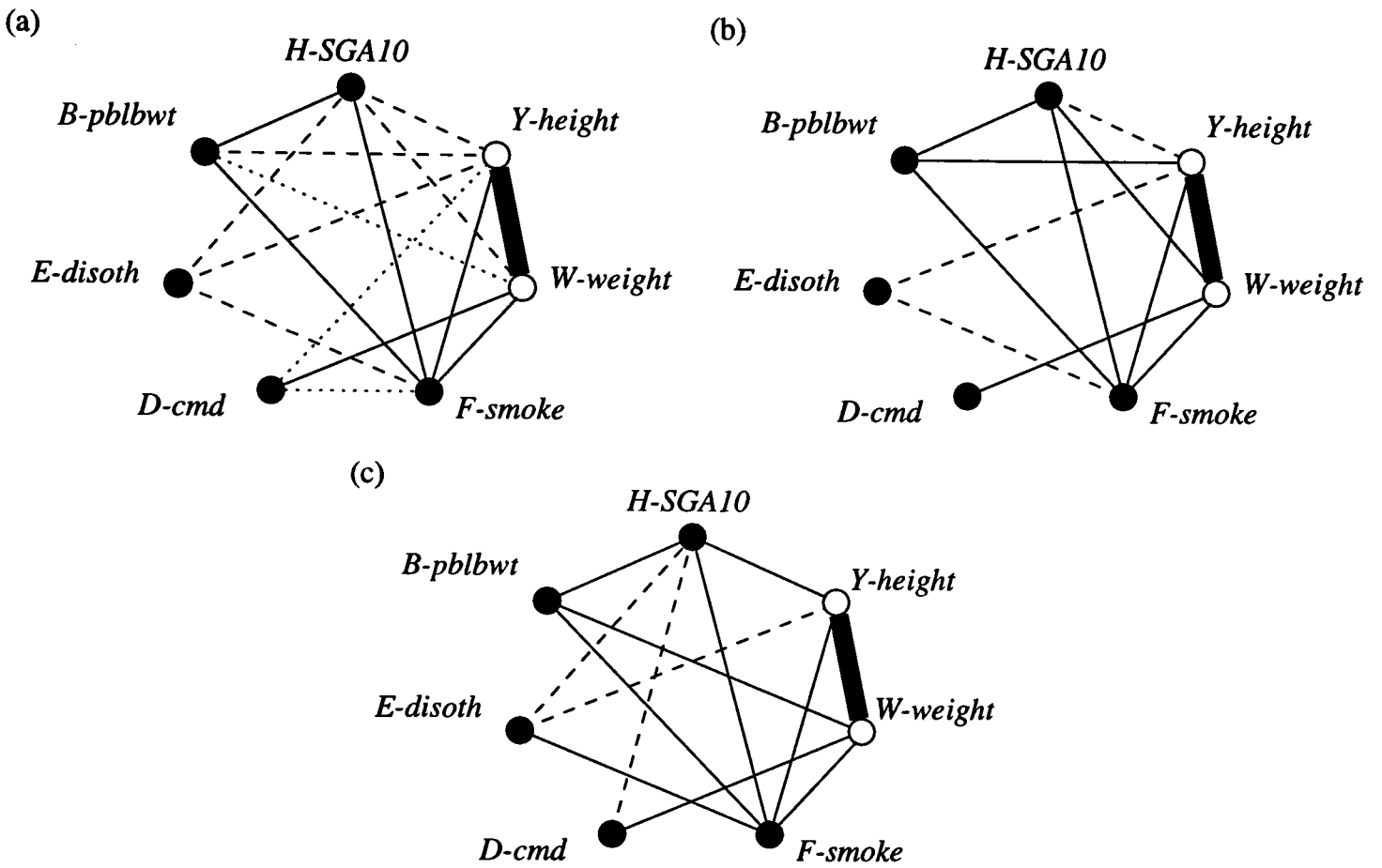


Figure 4.5 Conditional independence graphs corresponding to three (non-graphical) hierarchical CG models for SGA births below the 10th birth weight percentile for gestational age. Graph (a) was obtained using chi-squared tests of deviance as the model selection criterion with a 10% critical value, graph (b) was obtained using chi-squared tests of deviance as the model selection criterion with a 5% critical value and graph (c) was obtained using AIC as the model selection criterion. Edge thickness corresponds to the significance of the edge deletion deviance in the model, the more significant the deviance the thicker the edge. The dotted edges indicate a edge that is non-significant at the 5% level and the dashed edges indicate a non-significant edge at the 1% level but one which is significant at the 5% level.

Independence graphs corresponding to three (non-graphical) hierarchical CG models are shown in Figure 4.5(a)–(c). The three graphs were obtained using chi-squared tests of deviance with 10% and 5% significance levels, and AIC. Using a significance level of 10% we get graph (a), obtained by the removal of five edges from the independence graph associated with the initial model (in order): $[DE]$, $[BE]$, $[EW]$, $[DH]$ and $[BD]$. Graph (b) was obtained using a significance level of 5%, which removed nine edges from the initial graph, these were the five edges removed that gave graph (a) and edges $[BW]$, $[DY]$, $[DF]$ and $[EH]$. The AIC selected model removed eight edges from the initial graph (in order): $[EW]$, $[DY]$, $[BY]$, $[HW]$, $[DF]$, $[BE]$, $[BD]$ and $[DE]$. Notice that all the non-significant edges at the 5% level (indicated by the dotted lines) plus $[EH]$ in graph (a) are removed in graph (b). Apart from edge $[DH]$, the edges retained in graph (c) are a subset of those contained in (a). The first model selected using a 10% significance level is given by

$BF, BH, DF, EF, EH, FH/$
$BFW, BHW, DFW, FHW, BFY, BHY, DFY, EFY, EHY, FHY/$
$BFWY, BHWY, DFWY, EFY, EHY, FHWY/$

and has a deviance of 31.70 on 33 degrees of freedom. The second model selected using a 5% significance level is given by

$BF, BH, D, EF, FH/$
$DW, FHW, BFY, BHY, EFY, FHY/$
$BFY, BHY, DW, EFY, FHWY/$

and has a deviance of 71.84 on 54 degrees of freedom. The AIC selected model is given by

$BF, BH, DH, EF, EH, FH/$
$DW, BFW, BH, DH, EFY, EHY, FHY/$
$DW, BFW, EFY, EHY, FHY, FWY/$

and has a deviance of 70.36 on 58 degrees of freedom. The final stage for each of the three model fits is detailed in Appendix G.

We can again look at relative risks on the binary variables to test whether or not the CG model selection process has highlighted significant SGA risks (see Table 4.3). As with the *SGA15* analysis there are large relative risk estimates of SGA with previous preterm birth and also with cigarette smoking. The *SGA10* sample exhibits nearly a four-fold increase in risk given a previous preterm birth and a two-and-a-half times increase in risk given cigarette smoking. Both risks are higher than the relative risks estimated using *SGA15* with the same factors. There is also a significant SGA risk with chronic maternal diseases. Finally, there is a significant relative risk of previous low birth weight with cigarette smoking.

Table 4.3 Relative risk estimates (RR) together with 95% confidence interval (95% CI). A confidence interval enclosing 1 indicates a non-significant relative risk.

Interaction	RR	95% CI
H-SGA10 with *B-pblbwt*	3.9	(3.0, 5.1)
H-SGA10 with *F-smoke*	2.5	(2.0, 3.1)
H-SGA10 with *D-cmd*	2.0	(1.3, 3.2)
H-SGA10 with *E-disoth*	1.3	(1.0, 1.7)
B-pblbwt with *F-smoke*	1.6	(1.4, 2.0)
B-pblbwt with *D-cmd*	1.5	(1.0, 2.3)
B-pblbwt with *E-disoth*	1.0	(0.8, 1.3)
D-cmd with *F-smoke*	1.5	(1.0, 2.3)
D-cmd with *E-disoth*	1.1	(0.7, 1.6)
E-disoth with *F-smoke*	1.1	(0.9, 1.2)

4.4 Modelling actual birthweight

Although our primary interest lies in examining SGA risk factors it is perhaps worth taking a brief look at actual infant birthweight. Figure 4.6 shows normal probability

plots for actual birthweight on its original and logarithmic scales. The data are clearly skewed to the left as a number of infants have particularly low birth weights. Taking the natural logarithm of the data helps to alleviate this skew but it is still clearly visible. Note that apart from the particularly low birth weights (roughly below 2100 grams) the data are reasonably normal.

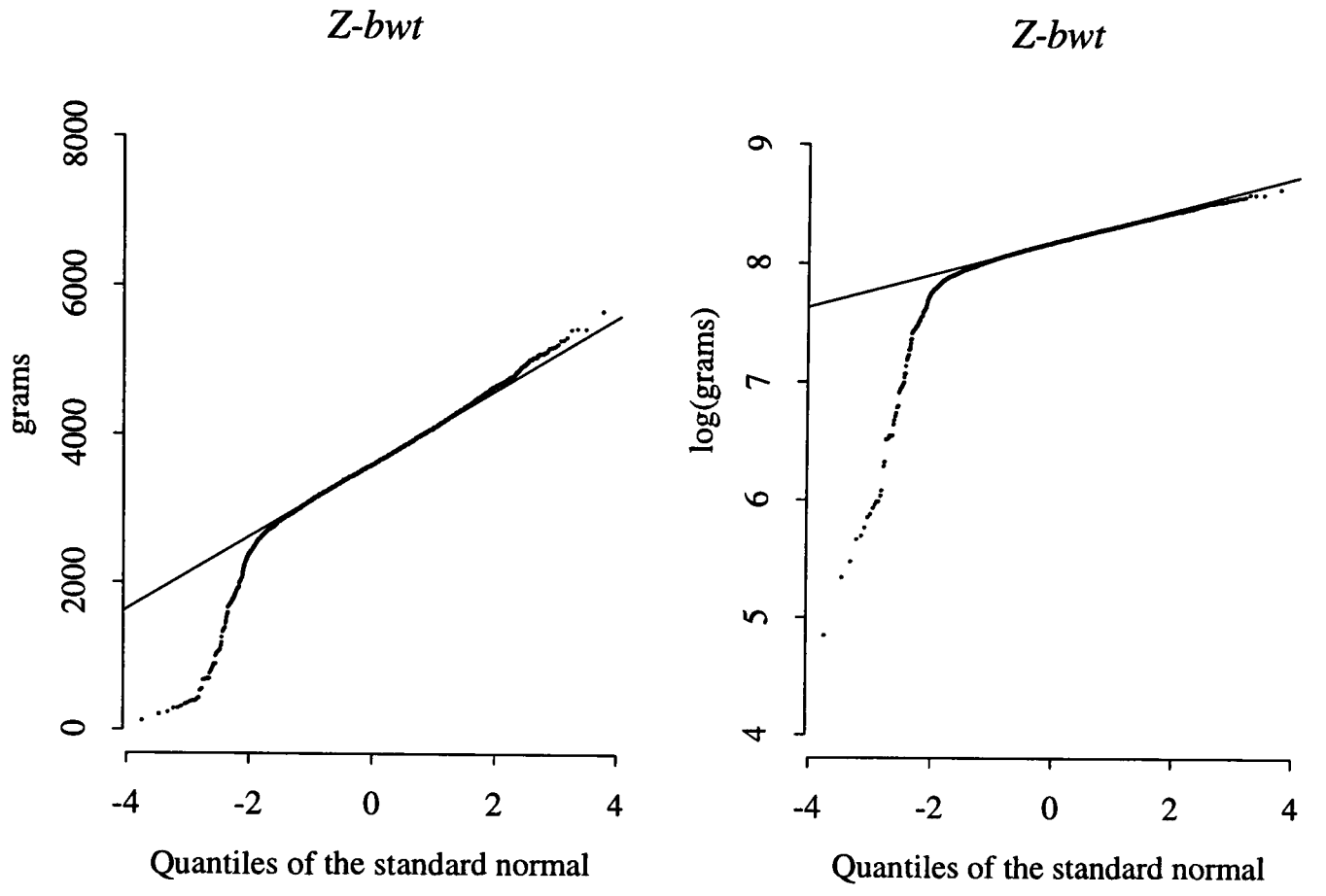


Figure 4.6 Normal probability plots for untransformed and $\log_e$ transformed variable *bwt* (infant birthweight). In each graph the straight line is drawn through the upper and lower quartiles.

Although, marginal normality for *Z-bwt* is somewhat suspect we proceed by analysing the data in a similar fashion to *SGA15* and *SGA10*. In an attempt to obtain some reduction in the number of variables standard two-sample t-tests were computed giving the following p-values for the variables within parentheses: p=0.003 (*A-pbstatus*), p<0.001 (*B-pblbwt*), p<0.001 (*C-pbpreterm*), p<0.001 (*F-smoke*), p=0.067 (*D-cmd*), p= 0.726 (*E-disoth*). Based on these results we retain binary variables A, B, C and F together with continuous variables W, X, Y and Z (after transformation using natural logarithms) in the analysis.

Examination of the empirical cell counts based on the cross-classification of the four binary variables indicates four cells with 3 or fewer observations but there are no empty cells. We can actually proceed with the analysis by fitting a saturated homogeneous graphical model to the data.

Graph (a) shown in Figure 4.7 is obtained by the removal of the following edges (in order): $[AF]$, $[FW]$, $[AX]$, $[AY]$, $[AW]$, $[CW]$, $[XY]$ and $[XZ]$ using AIC model selection. The same edges are removed by using chi-squared tests of deviance using a 10% significance level but the order in which the edges are removed is different, i.e. the edges are removed in the following order: $[FW]$, $[XY]$, $[AX]$, $[AF]$, $[AY]$, $[XZ]$, $[AW]$ and $[CW]$. The second graph (b) is selected using chi-squared tests with a 5% significance level, which simply removes the 8 edges as with a 10% significance level

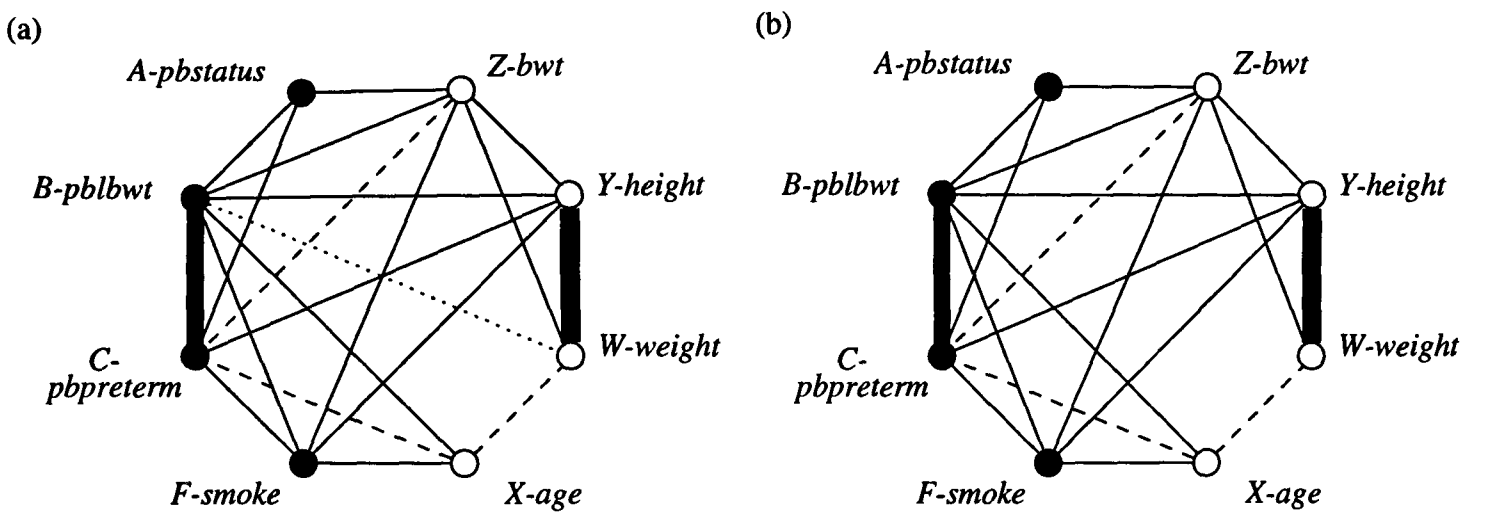


Figure 4.7 Homogeneous graphical conditional independence graphs. Graph (a) was obtained using chi-squared tests of deviance with a 10% critical value. The same graph was also independently selected using AIC. Graph (b) was obtained using chi-squared tests of deviance with a 5% critical value. Edge thickness indicates the relative strength of the pairwise interaction between variables based on rescaling p-values, the smaller the p-value the thicker the edge. The dotted lines indicate an edge that is non-significant at the 5% level and the dashed edges indicate a non-significant edge at the 1% level but one that is significant at the 5% level.

plus $[BW]$ (the only edge that is not significant at the 5% level in (a)). The final models selected upon which the above graphs are drawn are for (a):

$$ABC, BCF/BW, BCFX, BCFY, ABCZ, BCFZ/WX, WYZ;$$

which has a deviance of 46.40 on 40 degrees of freedom and for (b)

$$ABC, BCF/W, BCFX, BCFY, ABCZ, BCFZ/WX, WYZ;$$

which has a deviance of 49.15 on 41 degrees of freedom.

There are marked similarities between the graphs selected here and the previous analyses for *SGA15* and *SGA10*, i.e. there is a strong bond between the height and weight variables, the *bwt* variable is connected to previous low birth weight, the smoking variable and weight. In common with the *SGA15* analysis *Z-bwt* is connected to previous preterm birth. The analysis shows that *X-age* is conditionally independent of birth-weight given the other variables and so it does not seem too unusual that this variable was dropped in the previous analyses. Note that there appears to be significant interaction three-way interactions between X, and the three-way interaction B, C and F.

4.5 Classification of SGA births

It is often true in medical studies that when data are randomly sampled from a population the majority of the cases turn out to be 'normal'. Clearly, the more uncommon the medical condition the harder it is to obtain a sufficient number of individuals exhibiting the condition. The effect of having a much larger normal group is that we are more likely to misclassify a diseased case as normal, i.e. declaring a *false negative result*. If the medical condition being studied is possibly life-threatening then a false negative result is potentially disastrous. To overcome this problem, we can attach a loss of declaring a false negative result at ℓ times higher, say, than declaring a *false positive result* (i.e. declaring a non-diseased case as being diseased).

As we have seen, the majority of the data are non-SGA and therefore this group is likely to dominate any objective statistical procedure used to classify those women whose babies are SGA. The method we adopt in using CG models to classify SGA births correctly is to shift the classification boundary using a loss, ℓ. This simply multiplies the estimated posterior probabilities in the SGA class by some factor greater than 1. This method is easy to implement and directly interpretable by a clinician. However, Ripley (1996, pp. 58–59) points out that there is potential for estimation bias in the posterior probabilities themselves, i.e. the larger normal group could have the effect of underestimating the posterior probabilities of the smaller SGA group. This potential problem might be alleviated by using a weighted estimation procedure. This uses a factor $w < 1$ to down-weight each observation occurring in the normal sample. Another approach that could be implemented is to randomly sub-sample the larger non-SGA group.

4.5.1 CG classification

The approach we adopt here is to use the initial models for *SGA15* and *SGA10* for classification. (The models are given by 4.1, 4.2 on pages 70 and 73, respectively). These are heterogeneous CG models and so we employ densities (3.19) and (3.29) in (3.31) (pages 53–57). Raw and smoothed v-fold error rates are given below in Table 4.4. The raw overall error rates are low but we get nearly 70% of SGA15 cases and 64%

Table 4.4 Percentage error rates for CG models used to classify *SGA15* and *SGA10* births. Corresponding standard errors are given in parentheses. The raw error rates are calculated using the proportion of misclassified observations made in the assumed true classes. The smoothed error rates are calculated by averaging over the posterior probabilities alone.

	v-fold error rates % (v =10)					
	non-*SGA15*	*SGA15*	overall	non-*SGA10*	*SGA10*	overall
raw	12.2 (0.5)	69.9 (2.0)	18.0 (0.5)	13.2 (0.5)	63.7 (2.6)	16.5 (0.5)
smoothed	1.8 (0.1)	34.0 (1.9)	7.9 (0.2)	1.4 (0.1)	24.4 (1.4)	5.5 (0.1)

of SGA10 cases wrong. The predicted probabilities are seriously out of step with the raw rates. This suggests that the model for $p(c \mid x)$ is actually incorrect. In the absence of a weighted estimation procedure it is probably best to ignore the smoothed rates and continue by using the raw rates alone.

Clearly, we do a lot better at classifying the normal group. However, our primary aim is the correct identification of SGA cases. It would seem appropriate to specify a loss due to incorrect classification of SGA, e.g.

$$\text{Classify } x_0 \text{ as SGA if and only if } \ell p(\text{SGA}|x_0) > p(\text{non-SGA}|x_0)$$

for some suitably chosen loss, ℓ, and assuming a constant loss multiplier of 1 on the non-SGA group. Table 4.5 gives the raw error rates for *SGA15* and *SGA10* for values of ℓ ranging from 1 to 50. We get roughly equal numbers misclassified when using a loss of between 10 and 15 for the *SGA15* group and 30 for the *SGA10* group. Increasing the loss from this point onwards has the effect of correctly identifying more individuals in the SGA group but increases the incorrect classifications of non-*SGA* infants.

Table 4.5 Raw percentage error rates for CG models used to classify *SGA15* and *SGA10* births incorporating a loss, ℓ of incorrect non-SGA classification. Corresponding standard errors are given in parentheses. (The original v-fold error rates are used throughout.)

ℓ	non-*SGA15*	*SGA15*	overall	non-*SGA10*	*SGA10*	overall
1	12.2 (0.5)	69.9 (2.0)	18.0 (0.5)	13.2 (0.5)	63.7 (2.6)	16.5 (0.5)
5	21.6 (0.6)	54.9 (2.1)	25.0 (0.6)	21.5 (0.6)	52.3 (2.7)	23.5 (0.6)
10	28.5 (0.6)	48.2 (2.1)	30.5 (0.6)	26.6 (0.6)	47.1 (2.7)	27.9 (0.6)
15	45.4 (0.7)	33.2 (2.0)	44.1 (0.7)	30.0 (0.6)	43.1 (2.6)	30.8 (0.6)
30	66.4 (0.7)	19.9 (1.7)	61.7 (0.7)	37.2 (0.7)	37.1 (2.6)	37.2 (0.6)
50	75.3 (0.6)	13.6 (1.5)	69.1 (0.6)	51.5 (0.7)	26.0 (2.3)	49.9 (0.7)

4.5.2 Classification using logistic regression

It is possible to fit weighted logistic regression models using S-Plus by specifying a vector of weights when calling the `glm` function. This is described by Hastie & Pregibon (1992, Ch. 6). We first weighted the non-*SGA15* sample by $\frac{1}{10}$ retaining a weight of 1 on the *SGA15* sample. The non-*SGA10* sample was weighted by $\frac{1}{15}$ with a weight of 1 on *SGA10*.

We started with the all two-factor interaction model using the same set of variables that was used to define the starting model for the CG analysis of *SGA15*, i.e. *B-pblbwt*, *C-pbpreterm*, *D-cmd*, *F-smoke*, *W-weight* and *Y-height*. Stepwise model selection was based on an approximation to the AIC fitting criterion, which is automatically computed by S-Plus. For *SGA10* we again fitted an all two-factor interaction model based on variables *B-pblbwt*, *D-cmd*, *E-disoth*, *F-smoke*, *W-weight* and *Y-height* and used AIC based model selection. The models chosen using this procedure were given by

$$\begin{aligned} SGA15 \;=\; & pblbwt + pbpreterm + smoke + weight + height + pblbwt \times pbpreterm \\ & + pbpreterm \times height \end{aligned}$$

and

$$SGA10 \;=\; pblbwt + cmd + smoke + weight + height$$

Classification error rates for the starting models and for the AIC selected models are given in Table 4.6. In terms of error rates there is little difference in the two-factor interaction models and their AIC selected alternatives for *SGA15* and *SGA10*, respectively. The AIC selected models are perhaps to be preferred on the basis of simplicity. The weighted logistic regression approach does much to alleviate the asymmetry encountered with the CG procedure but gives larger standard errors for both classes, suggesting more extreme estimates of posterior probability.

It is worth noting that the weights used at the parameter estimation stage in fitting the logistic regression models must be applied to the estimated posterior probabilities in calculating smoothed error rates. There is no such problem in calculating the raw error rates since S-Plus can be forced to return a confusion matrix for actual class versus predicted class with the correct fixed sample size.

4.5.3 Classification using k-nearest neighbour methods

Comparison of the preceding methods with k-nearest neighbour methods of classification was made using binary variables B, C, D and F and continuous variables W and

Table 4.6 Percentage error rates for four logistic regression models used to classify *SGA15* and *SGA10* births. Corresponding standard errors are given in parentheses. The raw error rates are calculated using the proportion of misclassified observations made in the assumed true classes. The smoothed error rates are calculated by averaging over the posterior probabilities alone. non-SGA15 4896, SGA15 552

	v-fold error rates % (v =10)					
	Starting models					
	non-*SGA15*	*SGA15*	overall	non-*SGA10*	*SGA10*	overall
raw	39.8 (0.6)	35.3 (2.6)	39.3 (0.6)	31.6 (0.9)	41.4 (2.6)	32.2 (0.9)
smoothed	39.3 (1.0)	34.7 (3.1)	36.8 (2.2)	30.7 (0.9)	40.1 (6.4)	35.4 (3.4)
	AIC selected models					
	non-*SGA15*	*SGA15*	overall	non-*SGA10*	*SGA10*	overall
raw	38.5 (0.7)	37.0 (2.1)	38.3 (0.7)	32.9 (0.7)	41.7 (2.6)	33.4 (0.6)
smoothed	37.5 (1.0)	36.0 (3.2)	36.7 (2.2)	31.2 (0.9)	39.9 (6.3)	35.6 (3.4)

Y for *SGA15*. The procedure uses the general coefficient of similarity for binary and continuous variables described by Gower (1971) (for details of its use here see Chapter 3, Section 3.7).

Due to the large number of observations making up the SGA dataset, it was not possible, for computational reasons, to apply the k-nearest neighbour procedure to the entire dataset. Instead, we sub-sampled the larger non-*SGA15* group and combined this with the full *SGA15* sample. A total of 552 observations were randomly sampled from the full non-*SGA15* set. The random sampling was stratified using the observed proportions for each of the categories formed by B, C and F in the full non-*SGA15* set. (Variable D was dropped as previous analyses implied that it had a weak connection with *SGA15*. A total of three stratified random samples were taken from the full set of controls and each combined with the full set of cases. The classification results for the three samples using k-nearest neighbour methods are given in Table 4.7.

Table 4.7 Percentage error rates for k-nearest neighbour methods used to classify SGA births. Corresponding standard errors are given in parentheses. The variables used are *B-pblbwt*, *C-pbpreterm*, *F-smoke*, *W-weight*, *Y-height*.

	Raw v-fold error rates % (v =10)					
	1-NN		3-NN		5-NN	
Sample	non-*SGA15*	*SGA15*	non-*SGA15*	*SGA15*	non-*SGA15*	*SGA15*
1	46.6 (2.1)	50.7 (2.1)	37.7 (2.1)	52.5 (2.1)	37.5 (2.1)	48.9 (2.1)
2	48.4 (2.1)	48.2 (2.1)	40.8 (2.1)	48.2 (2.1)	37.0 (2.1)	49.3 (2.1)
3	48.6 (2.1)	52.0 (2.1)	42.2 (2.1)	55.4 (2.1)	39.5 (2.1)	54.0 (2.1)
	9-NN		13-NN		19-NN	
Sample	non-*SGA15*	*SGA15*	non-*SGA15*	*SGA15*	non-*SGA15*	*SGA15*
1	33.3 (2.0)	52.0 (2.1)	35.3 (2.0)	55.3 (2.1)	33.9 (2.0)	53.3 (2.1)
2	34.4 (2.0)	54.3 (2.1)	30.1 (2.0)	53.4 (2.1)	30.3 (2.0)	54.5 (2.1)
3	35.3 (2.0)	58.2 (2.1)	32.8 (2.0)	57.4 (2.1)	31.0 (2.0)	57.6 (2.1)

Nearest neighbour methods do badly, particularly in the *SGA15* class with around 50% misclassified. The *SGA15* misclassification error rate rises as k increases, whereas non-*SGA15* errors rates fall to around 30%. A nearest-neighbour analysis of *SGA10* was also performed but this gave similarly poor results.

4.5.4 Comments

The use of CG models was useful in identifying the main risk factors associated with an SGA birth. The main factors appeared to be a previous low birth weight infant, previous preterm birth, smoking at or during the time of conception and some evidence of lower average pre-pregnancy weight than in the non-SGA class. As reported by Bakketeig *et al.* (1979) there is tendency for mothers to repeat gestational age and birth weight in successive births providing further evidence of the increased observed SGA risk given a previous low birth weight infant.

The fitted CG models provided an adequate fit to the data as can be seen from the size of the model deviances on a comparable number of degrees of freedom. Clearly, accurate classification of SGA births is rather hard. However, shifting the classification boundary by attaching a suitable loss factor (by between 15-30, say) does enable more accurate SGA prediction. The weighted logistic regression approach does prove to give more consistent results in terms of raw and smoothed error rates and should be preferred to CG classification with these data.

Chapter 5

Conclusion

In this thesis we have considered the use of conditional independence graphs in describing the inter-relationships among a set of mixed discrete (mainly binary) and continuous random variables. The framework we adopted was described by Lauritzen & Wermuth (1989) and is based on the CG interaction parametrization. We have not attempted to extend this framework but have illustrated how one might use it to model more complex datasets.

The FORTRAN program CGM was developed as a research aid to understanding the structure of CG models. CGM does overlap to some extent with Edwards' (1995) PC software MIM. However, our aim was not to develop an identical tool but to exploit the CG likelihood properties not available in MIM. It allowed us to illustrate how a quasi-Newton procedure may be used to estimate the parameters of a CG model and to obtain an estimated parameter covariance matrix. CGM also provided us with a flexible tool for studying the predictive approach to classification in the CG framework. Scope for improvement of CGM exists by implementing the analytic formulae of Frydenberg & Lauritzen (1989) for CG models with decomposable conditional independence graphs. This would allow computer intensive bootstrap computations to be performed, e.g. to bootstrap the model selection process. However, identification of CG models with analytic maximum likelihood based on its graphical structure does involve a good deal of additional programming.

5.1 Further work

Scope exists to use conditional independence graphs as prior belief structures so defining a *Bayesian* or *belief network*. Belief networks often incorporate directed edges denoting causal relationships. A 'real' example is given by Spiegelhalter *et al.* (1993) who construct a belief network for birth asphyxia. The topology of a discrete data network is defined using expert opinion before any data modelling is performed. The network is modified to ensure that it is decomposable and when new data become available. Local computations are used which exploit the graphical structure and simplify the parameter estimation problem. Lauritzen (1992) looks at the case of a belief network with mixed discrete and continuous data. The basis for model building is the CG distribution and again local computations may be used to simplify the parameter estimation problem. Gammerman *et al.* (1995) review exact and approximate algorithms currently available for handling mixed data in Bayesian belief networks.

As is often stated (e.g. Draper, 1995) performing model selection leading to a single 'best' model and then making inferences from this one model as if it were the true model ignores model uncertainty. Bayesian model averaging appears to provide an answer to this problem. This averages over a set of m models $\mathcal{M}_1, \ldots, \mathcal{M}_m$ given data, D (see

Madigan & York, 1995). In this situation, we have

$$p(\mathcal{M}_k \mid D) = \frac{p(D \mid \mathcal{M}_k)p(\mathcal{M}_k)}{\sum_{l=1}^{m} p(D \mid \mathcal{M}_l)p(\mathcal{M}_l)}, \tag{5.1}$$

where $p(\mathcal{M}_k)$ is a prior model probability that $\mathcal{M}_k$ is the true model and $p(D \mid \mathcal{M}_k)$ is the marginal likelihood of model $\mathcal{M}_k$ given by

$$p(D \mid M_k) = \int p(D \mid \mathcal{M}_k, \theta_k)p(\theta_k)d\theta_k. \tag{5.2}$$

In (5.2) $p(\theta_k)$ is the prior probability for θ_k under model $\mathcal{M}_k$ and $p(D \mid \mathcal{M}_k, \theta_k)$ is the likelihood. We might also average over the topologies of any specified graphical structures. The main difficulty is in evaluating (5.2), which typically involves integrating over a large number of dimensions. Stewart (1987) solves this problem by employing a Monte Carlo procedure for computing the integrals. Stewart's approach compares $2^8 = 256$ simple logistic regression models. One disadvantage with the method employed by Stewart is that it requires a prior distribution to be defined on the maximal model. For sub-models this involves assigning positive probability to a parameter being zero.

With large numbers of models averaging over all models generally becomes impractical. Madigan & Raftery (1994) adopt a model selection procedure based on Bayes factors in retaining only those models that appear to be reasonably useful. In addition, by appealing to the principle of Occam's razor more complicated models are penalized in favour of simpler ones. Madigan & York (1995) look at Bayesian model averaging in the discrete data case and implement a Markov chain Monte Carlo in order to explore the space of models. As with most practical Bayesian applications model averaging in both cases is based on a subset of selected models.

Appendix A

Miscellaneous matrix results

Some results concerning mainly vector and matrix differentiation are stated below without proofs.

DEFINITION A.1 *The Kronecker product.*

Let A be an $m \times n$ matrix and B a $p \times q$ matrix. The *Kronecker product* of A and B, written $A \otimes B$, is defined as the $mp \times nq$ matrix

$$\begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}.$$

DEFINITION A.2 *The vec operator.*

Let A be an $m \times n$ matrix and a_i its ith column. The *vec operator* is defined as the $mn \times 1$ vector

$$\text{vec}A = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}.$$

Thus the vec operator is defined by stacking the columns of A one underneath the other.

DEFINITION A.3 *The svec operator.*

Let A be an square symmetric $m \times m$ matrix and a_i its ith column containing only lower-triangular elements of A. Therefore a_i is of length $m - (i-1)$ for column i. The *svec operator* is defined as the $n = m(m+1)/2 \times 1$ vector

$$\text{svec}A = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}.$$

Thus the svec operator is defined by stacking the columnwise lower-triangular elements of A one underneath the other.

DEFINITION A.4 *The commutation matrix.*

Let A be an $m \times n$ matrix. The *commutation matrix* is the unique $mn \times mn$ permutation matrix, K_{mn}, which transforms vec A into vec A', i.e.

$$K_{mn}\text{vec}A = \text{vec}A'.$$

(If $m = n$, then we simply write K_n instead of K_{nn}.)

Table A.1 Vector derivatives $(x : p \times 1)$.

$\phi(x)$	$\partial\phi(x)/\partial x$
$a'x$	a
$x'x$	$2x$
$x'Ax$ $(A : p \times p)$	$(A + A')x$
$x'Ax$ (symmetric A)	$2Ax$

Table A.2 Matrix derivatives $(X : n \times n)$; X_s = symmetric X.

$\phi(X)$	$\partial\phi(X)/\partial X$	$\partial\phi(X)/\partial X_s$
$\log\|X\|$	$(X^{-1})'$	$(2X^{-1})' - \text{diag}(X^{-1})$
$\|X\|$	$\|X\|(X^{-1})'$	$\|X\|\{(2X^{-1})' - \text{diag}(X^{-1})\}$
$\|X\|^r$	$r\|X\|^r(X^{-1})'$	$r\|X\|^r\{(2X^{-1})' - \text{diag}(X^{-1})\}$

Table A.3 Matrix derivatives $(X : m \times n)$; X_s = symmetric X $(m = n)$.

$\phi(X)$	$\partial\phi(X)/\partial X$	$\partial\phi(X)/\partial X_s$
$\text{tr}(AX)$ $(A : n \times m)$	A'	$2A' - \text{diag}(A)$
$\text{tr}(A'X)$ $(A : n \times m)$	A	$2A - \text{diag}(A)$
$\text{tr}(XX')$	$2X$	$2\{X + X' - \text{diag}(X)\}$
$\text{tr}(X^r)$	rX^{r-1}	$r\{X^{r-1} + (X^{r-1})' - \text{diag}(X^{r-1})\}$
$\text{tr}(AX^{-1})$ $(A = A')$	$-(X^{-1}AX^{-1})$	$-2(X^{-1}AX^{-1}) + \text{diag}(X^{-1}AX^{-1})$

RESULT A.1 *The differential of a matrix inverse.*

If X is a non-singular $n \times n$ matrix, then the matrix function

$$\phi(X) = X^{-1}$$

has differential

$$\partial\phi(X) = -X^{-1}(\partial X)X^{-1}.$$

DEFINITION A.5 *The Hessian matrix of a scalar function.*

For a scalar function ϕ of a $p \times 1$ vector x, the Hessian matrix, H, of ϕ at x is given by the $p \times p$ matrix of second-order partial derivatives

$$\frac{\partial^2 \phi(x)}{\partial x \partial x'}.$$

RESULT A.2 *The Hessian matrix of a log determinant.*

For a scalar function $\phi(X) = \log |X|$, where X is an $n \times n$ matrix with positive determinant, its Hessian matrix is given by

$$\mathsf{H}\phi(X) = -K_n\{(X')^{-1} \otimes X^{-1}\}.$$

RESULT A.3 *The Hessian matrix of the trace of a quadratic function.*

For a scalar function $\phi(X) = \mathrm{tr}(X'AX)$, where X is an $n \times m$ matrix and A is $n \times n$, its Hessian matrix is defined as

$$\mathsf{H}\phi(X) = I \otimes (A + A').$$

RESULT A.4 *The Hessian matrix of the trace of an inverse (1).*

For a scalar function $\phi(X) = \mathrm{tr}(X^{-1})$, where X is an $n \times n$ matrix, its Hessian matrix is defined as

$$\mathsf{H}\phi(X) = K_n(X'^{-2} \otimes X^{-1} + X'^{-1} \otimes X^{-2}).$$

RESULT A.5 *The Hessian matrix of the trace of an inverse (2).*

For a scalar function $\phi(X) = \mathrm{tr}(AX^{-1})$, where X is an $n \times n$ matrix and A is $n \times n$, its Hessian matrix is defined as

$$\mathsf{H}\phi(X) = K_n(X'^{-1}AX'^{-1} \otimes X^{-1} + X'^{-1} \otimes X^{-1}AX^{-1}).$$

Appendix B

Quasi-Newton minimization

Quasi-Newton algorithms are a general class of methods for solving unconstrained minimization problems. The algorithms work by assuming a quadratic form as a local approximation to a multidimensional function. This approximation requires first and second partial derivatives to be evaluated, but instead of evaluating the Hessian analytically it is built up using some iterative scheme. Here we provide a brief sketch of one particular quasi-Newton algorithm, which is used to solve the maximum likelihood parameter estimation problem for an arbitrary Conditional Gaussian model.

B.1 Iterative descent

Quasi-Newton algorithms follow an iterative descent path that seeks the minimum of some arbitrary d-dimensional objective function, $f(\theta)$. This is accomplished by accumulating information from successive line minimizations, so that d such line minimizations lead to the exact minimum of a d-dimensional quadratic form. If $f(\theta)$ is not exactly a quadratic form then repeated cycles of d line minimizations will usually be needed to ensure that the method converges to a local minimum.

If first and second derivatives of $f(\theta)$ are available, then a local approximation of the objective function may be obtained by taking the first three terms in a Taylor-series expansion about the current point, θ, i.e.

$$f(\theta^*) \approx f(\theta) + (\theta^* - \theta)g(\theta) + \frac{1}{2}(\theta^* - \theta)'H(\theta)(\theta^* - \theta), \tag{B.1}$$

where $g(\theta)$ is the gradient of $f(\theta)$, $H(\theta)$ is the Hessian of $f(\theta)$ and θ^* is the predicted minimum point. A linear approximation to the derivative of $f(\theta^*)$ is given by

$$g(\theta^*) = g(\theta) + H(\theta)(\theta^* - \theta), \tag{B.2}$$

since $g(\theta^*) = 0$ we obtain

$$\theta^* = \theta - H^{-1}(\theta)g(\theta). \tag{B.3}$$

A quasi-Newton iterative descent path is given by the sequence of steps

$$\theta_{j+1} = \theta_j - kB(\theta_j)g(\theta_j), \tag{B.4}$$

where $H^{-1}(\theta)$, the inverse Hessian, in (B.3) is replaced by some approximation $B(\theta)$ and k defines a step length. The direction of the step is given by $-B(\theta)g(\theta)$.

If the unit matrix I instead of $B(\theta)$ is used in (B.3) together with any step length k that ensures a reduction in the objective function then we obtain the method of steepest descents. The method of steepest descents is guaranteed to converge to a local

minimum. However, its convergence rate in even mildly ill-conditioned problems can become unacceptably slow. The main problem is that the search directions generated are not conjugate to one another. If $\{x_i\}$ is a set of search direction vectors and A is a symmetric positive definite matrix then the $\{x_i\}$ are *mutually conjugate* with respect to A if

$$x_i' A x_j = 0 \qquad (i \neq j),$$

where $i = 1, \ldots, d$ and $j = 1, \ldots, d$. Non-conjugate search directions result in the method tending to 'hem-stitch' towards the minimum, rather than follow the floor of the function directly to its minimum. (For details of the problems inherent in the method of steepest descents see Gill *et al.*, 1981, p. 103.)

B.2 Quasi-Newton algorithms

Quasi-Newton methods overcome the problem of poor convergence in steepest descents by ensuring that the search directions generated are conjugate to one another with respect to $B(\theta)$, the approximation to the inverse Hessian. The approximation to the inverse Hessian is constructed using a sequence of matrices $B(\theta_j)$. A positive definite approximation (usually the unit matrix) is chosen at step $j = 1$. Subsequent $B(\theta_j)$ matrices are computed in such a way that they remain symmetric positive definite.

The approximation to the inverse Hessian, $B(\theta)$, is formed using the following updating rule

$$B(\theta_{j+1}) = B(\theta_j) + C(\theta_j), \tag{B.5}$$

where $C(\theta)$ is some correction term. The differences in quasi-Newton algorithms depend on how the correction term is computed. The method we use here is known as Broyden–Fletcher–Goldfarb–Shanno (BFGS). The BFGS updating procedure is known to give superior results, in terms of computational accuracy, to some of the other well known quasi-Newton algorithms (for details concerning its history see Gill *et al.*, 1981, p. 119). The BFGS correction term is given by

$$C(\theta_j) = a_2 s s' - [s\{B(\theta_j)u\}' + \{B(\theta_j)u\}s']/a_1, \tag{B.6}$$

where s is the step taken, i.e.

$$s = \theta_{j+1} - \theta_j = -kB(\theta_j)g(\theta_j),$$

u is the gradient difference

$$u = \{g(\theta_{j+1}) - g(\theta_j)\},$$

and the coefficients a_1 and a_2 are given by

$$a_1 = s'u,$$

and

$$a_2 = (1 + u'B(\theta_j)u/a_1)/a_1.$$

Further details about the implementation of this algorithm may be found in Nash (1990).

Appendix C

CGM programming details

CGM is a FORTRAN 77 program consisting of roughly 9,500 lines of code originally developed on a Sun SPARCstation 1. It has subsequently been ported to an Intel-compatible PC running RedHat Linux version 4. The FORTRAN code has been compiled using `fort77`, which converts the source-code to C and then automatically calls the GNU C source compiler, `gcc` version 2.7. The amount of memory taken up by the program ultimately depends on the storage size of arrays used. As a guide CGM runs in approximately 1MB of memory if array bounds are set for a maximum of 500 observations, 5 discrete variables, 5 continuous variables, 32 cells and 640 fitted parameters. Array bounds are set in the source code prior to compilation.

C.1 Interaction expansions

One of the most awkward tasks, from a computational point of view, is to set up interaction expansions for CG sub-models over the full set of factor levels, $i \in \mathcal{I}$. The example lower-triangular matrix shown overleaf gives generic interactions $\tau_{\cdot}$ for a maximum of four factors, which might represent $\omega(i)$, $\beta(i)$ or $\varphi(i)$ (for $\varphi(i)$ set the column labelled τ equal to zero). The matrix reflects the estimation of interaction expansions for all higher-order interactions via main effects or single-factors only. Interaction expansions for CG sub-models are described in Section 2.3.3 on page 23. Using the example matrix shown overleaf we can define interaction expansions using the following set of general rules:

1. Create a new matrix and fill it with zeros (the size of the matrix is determined by the number of interactions present);
2. label the columns and rows of the matrix with the full set of interactions using standard order;
3. set the diagonal elements equal to one for those interactions that are to be estimated;
4. set the constant τ column equal to one;
5. for each element on the leading diagonal that is zero copy the corresponding row from the interaction matrix;
6. for each zero element on the leading diagonal, zero its column;
7. for each zeroed off-diagonal element (zeroed at the last step), take its row and add to it the values in the row that had a zeroed diagonal element.

(For the last two steps start with the lowest-order interactions first.)

	τ	τ_A	τ_B	τ_{AB}	τ_C	τ_{AC}	τ_{BC}	τ_{ABC}	τ_D	τ_{AD}	τ_{BD}	τ_{ABD}	τ_{CD}	τ_{ACD}	τ_{BCD}	τ_{ABCD}
τ	1															
τ_A	1	1														
τ_B	1	0	1													
τ_{AB}	−1	1	1	0												
τ_C	1	0	0	0	1											
τ_{AC}	−1	1	0	0	1	0										
τ_{BC}	−1	0	1	0	1	0	0									
τ_{ABC}	1	−1	−1	1	−1	1	1	0								
τ_D	1	0	0	0	0	0	0	0	1							
τ_{AD}	−1	1	0	0	0	0	0	0	1	0						
τ_{BD}	−1	0	1	0	0	0	0	0	1	0	0					
τ_{ABD}	1	−1	−1	1	0	0	0	0	−1	1	1	0				
τ_{CD}	−1	0	0	0	1	0	0	0	1	0	0	0	0			
τ_{ACD}	1	−1	0	0	−1	1	0	0	−1	1	0	0	1	0		
τ_{BCD}	1	0	−1	0	−1	0	1	0	−1	0	1	0	1	0	0	
τ_{ABCD}	−1	1	1	−1	1	−1	−1	1	1	−1	−1	1	−1	1	1	0

For example, given three factors A, B, and C with conditional independence relationship $A \perp B \mid C$ we estimate interactions τ, τ_A, τ_B, τ_C, τ_{AC} and τ_{BC} thus after applying rules 1–4, we obtain the following matrix:

	τ	τ_A	τ_B	τ_{AB}	τ_C	τ_{AC}	τ_{BC}	τ_{ABC}
τ	1							
τ_A	1	1						
τ_B	1	0	1					
τ_{AB}	1	0	0	0				
τ_C	1	0	0	0	1			
τ_{AC}	1	0	0	0	0	1		
τ_{BC}	1	0	0	0	0	0	1	
τ_{ABC}	1	0	0	0	0	0	0	0

Now, apply rule 5 so that τ_{AB} and τ_{ABC} are given by

	τ	τ_A	τ_B	τ_{AB}	τ_C	τ_{AC}	τ_{BC}	τ_{ABC}
τ_{AB}	−1	1	1	0				
τ_{ABC}	1	−1	−1	1	−1	1	1	0

then 6

	τ	τ_A	τ_B	τ_{AB}	τ_C	τ_{AC}	τ_{BC}	τ_{ABC}
τ_{AB}	−1	1	1	0				
τ_{ABC}	1	−1	−1	0	−1	1	1	0

and finally rule 7 gives

	τ	τ_A	τ_B	τ_{AB}	τ_C	τ_{AC}	τ_{BC}	τ_{ABC}
τ_{AB}	−1	1	1	0				
τ_{ABC}	0	0	0	0	−1	1	1	0

so that $\tau_{ABC} = \tau_{AC} + \tau_{BC} - \tau_C$ and $\tau_{AB} = \tau_A + \tau_B - \tau$. (Note that the entries other than τ_{ABC} and τ_{AB} do not change after step 4.)

CGM generalizes this procedure for more interactions and interactions with different numbers of levels. It is also worth noting that the gradient vector for sub-models may be built-up by summation over columns.

C.2 Accumulating the log-likelihood

Double precision arithmetic was used to accumulate the value of the log-likelihood for saturated CG, homogeneous saturated CG and all CG sub-models described in Section 2.3. The negative of the log-likelihood is minimized using the quasi-Newton routine described in Appendix B. The elements of the gradient vector given by (2.26) in Section 2.3 were calculated using double precision arithmetic. A check that the gradient has been accumulated correctly was initially made using the central difference approximation

$$\frac{f(x+h)-f(x-h)}{2h} = f'(x) + O(h^2),$$

where $f'(x)$ is the true derivative of $f(x)$ (see Gill *et al.*, 1981, p. 54).

C.2.1 A starting point for the algorithm

Recall from Chapter 2 (page 21) that we need to estimate the values of the parameters $\{\varphi(i)\}_{i\in\mathcal{I}}$, $\{\beta(i)\}_{i\in\mathcal{I}}$ and $\{\Omega(i)\}_{i\in\mathcal{I}}$. A suitable starting point for the algorithm is to set the $\{\varphi(i)\}_{i\in\mathcal{I}}$ and $\{\beta(i)\}_{i\in\mathcal{I}}$ equal to zero, and set the $\{\Omega(i)\}_{i\in\mathcal{I}}$ equal to the identity matrix.

C.3 Matrix storage and inversion

Symmetric $n \times n$ matrices are stored by rows using one-dimensional arrays of length $n(n+1)/2$, i.e.

$$\text{Matrix} \rightarrow \text{Array} = \{a_{11} \mid a_{21}\, a_{22} \mid a_{31}\, a_{32}\, a_{33} \mid \cdots a_{nn}\}. \qquad \text{(C.1)}$$

The (i,j)th element of a matrix, a_{ij}, in row i and column j is indexed by $j + i(i-1)/2$ for $j = 1, \ldots, n$ and $i = j, \ldots, n$.

C.3.1 The Cholesky decomposition

Matrix inversion is computed using the Cholesky decomposition. For a symmetric positive definite $n \times n$ matrix A the Cholesky decomposition constructs a lower triangular matrix L such that

$$A = LL'.$$

The elements of L are given by $l_{ij} = l'_{ji}$, i.e.

$$\begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ l_{(n-1)1} & \cdots & l_{(n-1)(n-1)} & 0 \\ l_{n1} & \cdots & l_{n(n-1)} & l_{nn} \end{bmatrix}.$$

The general formula for the Cholesky decomposition is given by

$$l_{ii} = \left(a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2 \right)^{1/2},$$

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk}}{l_{jj}} \quad (j < i).$$

(see Jennings & McKeown, 1992, p. 101). Both A and L are stored using the matrix storage scheme defined by (C.1).

The inverse of A may be computed from its Cholesky decomposition using

$$A^{-1} = L^{-1}(L')^{-1}.$$

The determinant of A may be obtained from the elements of L using

$$|A| = \left(\prod_{i=1}^{n} l_{ii} \right)^2.$$

C.4 Supporting routines

Additional FORTRAN 77 routines used in CGM were obtained from two sources: *The Journal of the Royal Statistical Society C* (*Applied Statistics*) and the book *Numerical Recipes* by Press *et al.* (1992). *Applied Statistics* algorithms are available from the Statlib electronic archive via the Internet (`http://lib.stat.cmu.edu`). Brief details of the additional routines used in CGM are given below.

Function Name	Description
`AS66`:	calculation of the tail area under a normal curve. *Applied Statistics* (1973) **22**, #3. [Used in the predictive logistic calculation (see page 57).]
`brent`:	one-dimensional minimizer using Brent's method. *Numerical Recipes*, §10.2. [Used to obtain the value of the smoothing parameter; see Appendix E.]
`choldc`:	Cholesky decomposition of a matrix. *Numerical Recipes*, §2.9. [The procedure was re-written to work with a lower-triangular matrix supplied on input.]
`choldsl`:	solves the set equations $A\mathbf{x} = \mathbf{b}$, where A is a $p \times p$ positive-definite symmetric matrix and $\mathbf{x}$ and $\mathbf{b}$ are real-valued $p \times 1$ vectors. *Numerical Recipes*, §2.9. [Used in conjunction with `choldc` for matrix inversion. The procedure was re-written to work with a lower-triangular matrix supplied on input.]
`gammln`:	the logarithm of the gamma function. *Numerical Recipes*, §6.1. [Used in the normalization of the predictive densities given in Chapter 3. It is useful to note that $n! = \Gamma(n+1)$.]
`gammq`:	chi-square probability function via the incomplete gamma function. *Numerical Recipes*, §6.2 & §15.1. [Gives $p(\chi^2_{\text{calc}} \mid \nu)$ for integer number of degrees of freedom ν.]

C.5 NIC model selection

A more accurate correction term for the expected difference between the true and estimated log-likelihoods for θ is given by NIC (known as the Network Information Criterion), Murata *et al.* (1991) also derived by Stone (1977). NIC is defined as

$$\text{NIC} = D_{\mathcal{M}_k} + 2\text{tr}[IK^{-1}], \tag{C.2}$$

where $D_{\mathcal{M}_k}$ is the deviance for model k, I is the usual definition of the Fisher information and K is the expectation of the observed information, i.e.

$$I = I(\theta) = \text{Var}\left[\frac{\partial f(x_i;\theta)}{\partial\theta}\right] \quad \text{and} \quad K = K(\theta) = \text{E}\left[-\frac{\partial^2 f(x_i;\theta)}{\partial\theta^2}\right],$$

where $f(x_i, \theta)$ is the density function corresponding to the parameter value θ on the space of a single observation. Model selection is based on minimizing NIC. The basic idea here is that if the parametric family contains the true density then $\text{tr}[IK^{-1}]$ equals d (the number of parameters) yielding the AIC criterion. Thus, NIC may be viewed as being more appropriate in dealing with model uncertainty.

An estimate of I may be obtained by calculating the variance of the gradient summands of each of the n data points at $\hat{\theta}$. An estimate of K^{-1} may be obtained from program CGM by multiplying by n the BFGS approximation to the inverse Hessian of the negative log-likelihood. Further details regarding the derivation and practical estimation of NIC may be found in Ripley (1996, pp. 31–35) and the references therein.

Appendix D

CGM command syntax

The following commands may be specified in the parameter file `params.dat` before running program CGM. Commands may be typed in full, or abbreviated to the first few letters printed in upper case. Commands may be listed in any order. The majority of the output from CGM is written to file `CGM.log`. Additional output files created depend on some of the commands listed below.

The arguments for each of the commands listed below are enclosed in square brackets `[ ]`. (Note that square brackets are not included around arguments in the parameter file.) Arguments must be separated from commands by at least one space. The type of argument is denoted by either `int` for integer, `real` for real, `dble` for double precision or `char*i` for character arguments (where `i` indicates the maximum length of the character string). An `array` indicates that the argument of a command takes multiple values. Where appropriate, default parameter values are indicated after an equals sign.

`#`

Any text following the hash symbol, #, as the first non-blank character on a single line is assumed to be a comment and is ignored.

`Bootstrap [int=200]`

Set the number of bootstrap samples to be used in calculating '.632' classification error rates.

`CLassify [char*1]`

Specify a discrete classification variable. This classification variable must also be declared using `factor`. Each specific level of the classification variable determines a class.

`CONTinuous [char*9]`

Specify the names of the continuous variables or factors. Each continuous variable is declared as a single letter in the range `A - Z` (lower case input is converted to upper case on output). A maximum of nine continuous variables may be declared. Do not separate variable names with spaces.

`CONVerge [dble=0.00001]`

Set the convergence tolerance for the quasi-Newton procedure.

`CPutime [dble=5.0]`

Set the maximum amount of cpu time (in minutes) in which the quasi-Newton procedure has to converge. The procedure is stopped if this limit is reached before convergence.

`CRitical [dble=0.05]`

Set the critical value when doing stepwise model selection. (Used only when computing chi-squared tests of deviance.)

`CV [int=v]`

Specify the number of pieces to divide the dataset into when calculating v-fold cross-validated classification error rates. (The default v is taken to be the smaller of $\sqrt{n}$ or 10.)

`Doubt [real=1.0]`

Specify the probability, *dpr*, when computing the classification confusion matrix. An observation is declared as being of doubtful origin if its maximum posterior probability is less than $1 - dpr$.

`Evals [int=2000]`

Set the maximum number of function evaluations that the quasi-Newton procedure may perform. The procedure is stopped if this limit is reached before convergence.

`FACtor [char*9]`

Specify the names of the discrete variables or factors. Each discrete variable is declared as a single letter in the range `A - Z` (lower case input is converted to upper case on output). A maximum of nine discrete variables may be declared. Do not separate variable names with spaces.

`FILe [char*80]`

Specify a datafile from which the raw data are to be read.

`Ksmooth [real=1.0]`

Smooth observed cell probabilities using a binary data kernel. Specify the value of the smoothing parameter, h ($0.5 \leqslant h \leqslant 1.0$). When $h = 0.5$ uniform weight is given to all cells. Alternatively, when $h = 1.0$ the original observed cell probabilities are reproduced. Setting $h = 0.0$ forces CGM to estimate h from the observed data (see Appendix E).

`LEvels [int array]`

Specify the maximum level for each discrete variable. Each discrete variable is allowed up to nine levels. Do not separate variable levels with spaces.

`LOgistic`

Use the quasi-Newton minimizer to fit a logistic regression to the binary variables when computing posterior probabilities. The estimated probabilities are used as part of the Bayes classification rule. The structure of the logistic regression model is assumed to the same as that specified for the discrete part of the stated CG model with the class indicator treated as a response variable. (Runtime defaults, such as maximum number of function evaluations, will be the same as those specified for the quasi-Newton CG model fit.)

`Model [char*256]`

Define a model to be fitted. If more than one model is defined each model is fitted in turn. A maximum of 70 models may be defined at any one time. (Note that only

the first model specified will be used to generate sub-models when performing stepwise model selection.)

`PRINt [char*26="a"]`

Print options:

`a` - run information;

`b` - estimated .632 bootstrap error rates;

`c` - estimated classification rule;

`d` - raw data;

`e` - estimated standard errors for classification rule parameters;

`f` - fitted counts, means and covariances;

`g` - fitted discrete, linear and precision parameters;

`h` - estimated Hessian (BFGS update);

`k` - value of the kernel smoothing parameter;

`s` - empirical counts, means and covariances;

`t` - track progress of fitting algorithm output in `fort.99`;

`v` - estimated cross validated error rates.

Do not separate print options with spaces.

`PRIOr [real array]`

Specify class prior probabilities when doing classification. Delimit each probability with one or more spaces. (By default equal class prior probabilities are assumed.)

`REad [char*18]`

List the order in which the declared variables are to be read from the raw datafile.

`RPar`

Read parameter estimates from file `b.dat` (no model fitting is done).

`SAtmod [char*80]`

Specify a maximal model to fit. By default the maximal model is assumed to be the first model specified using `model`.

`SEed [int=234984]`

Set the seed value for choosing random samples. (Used only in conjunction with `bootstrap`.)

`SKip [int=0]`

Skip over a specified number of lines when reading a datafile.

`STEpwise [char*3="chi"]`

Perform backwards model selection using either chi-squared tests of deviance (`chi`), AIC (`aic`) or NIC (`nic`) as model selection criterion.

`STOp [int=99]`

Specify a step at which to halt stepwise model selection procedure. Note that the procedure is stopped at the start of the step. (Setting zero for `stop` fits just the first model read.)

`Title [char*132]`

Specify a title to be included in the output file `CGM.log`.

`WPar`

Write parameter estimates to file `b.dat`.

Appendix E

Smoothed cell probabilities

This Appendix outlines a method of data smoothing that overcomes the problem of observed cell frequencies of zero. We concentrate on the case where the discrete random variables are all binary. However, the method is easily extended to work with discrete variables of more than two levels. Let $i = (i_1, \ldots, i_p)'$ be a $p \times 1$ binary variable vector. There are clearly 2^p possible values of i. Let the full set of all possible values of i be given by $\mathcal{I}$. Given vectors $i, j \in \mathcal{I}$ let $d(i,j)$ be the number of disagreements in the corresponding components of i and j, i.e.

$$d(i,j) = (i-j)'(i-j). \tag{E.1}$$

Aitchison & Aitken (1976) define a kernel, K, for a p-dimensional binary space as

$$K(i \mid j, h) = h^{p-d(i,j)}(1-h)^{d(i,j)}, \tag{E.2}$$

where h is a smoothing parameter $1/2 \leqslant h \leqslant 1$. This gives the density at a point i based on the value of j. The more disagreements there are between i and j the more weight is placed on $h^{p-d(i,j)}$. Note that when $h = 1/2$ we get a uniform distribution over $\mathcal{I}$ and when $h = 1$ we get the observed relative frequencies, i.e.

$$\begin{aligned} K(i \mid j, \tfrac{1}{2}) &= (\tfrac{1}{2})^p, \\ K(i \mid j, 1) &= \begin{cases} 1 & (i = j), \\ 0 & (i \neq j). \end{cases} \end{aligned}$$

The kernel K is shown to satisfy

$$\sum_i K(i \mid j, h) = 1 \quad \text{for all } j \text{ and } h. \tag{E.3}$$

(For details about suitable kernels for ordered and unordered discrete data of more than two levels see Aitchison & Aitken (1976).)

Given a sample of n independent observations $I \subset \mathcal{I}$ we wish to estimate the density function of an unknown distribution over the sample space $\mathcal{I}$. An estimate $\check{p}(i)$ of the underlying density function is given by the kernel estimator

$$\check{p}(i \mid I, h) = 1/n \sum_{j \in I} K(i \mid j, h). \tag{E.4}$$

The value of the smoothing parameter in the interval $1/2 \leqslant h \leqslant 1$ may be estimated in a number of different ways. A pseudo-Bayesian approach described by Titterington (1980) places a beta prior on h combined with the likelihood to obtain a posterior

density for $p(i)$. An estimate of h may be chosen as the posterior mode or mean, or the posterior density may be used to estimate $p(i)$ directly. An approach that maximized a pseudo-likelihood using cross-validation in order to obtain an estimate for h was used by Habbema *et al.* (1974) with continuous data. Aitchison & Aitken (1976) adopt a similar approach in the case of multivariate binary data. Aitchison & Aitken's pseudo-likelihood for multivariate binary data is defined to be

$$W(h \mid I) = \prod_{j \in I} p(j \mid I - j, h). \tag{E.5}$$

where $I - j$ denotes the set I with the binary variable vector j excluded. (See Titterington (1980) for a comparative review of discrete kernel density estimation procedures.)

Here we adopt the approach of Aitchison & Aitken to estimate h. For speed of computation we use a v-fold method of cross-validation to estimate h by dividing the dataset randomly into $v = \min(10, \sqrt{n})$ parts and estimating each unique $p(j)$ in one of the v parts left out using the remaining $v - 1$ parts of the data. An estimate of h is then found by maximization of the logarithm of a pseudo-likelihood, which we define to be

$$W^{\dagger}(h \mid I) = \sum_{m=1}^{v} \sum_{j \in I_m} n(j) \log p(j \mid I_{v \setminus m}, h), \tag{E.6}$$

where I_v denotes the full dataset randomly partitioned into v parts, $I_{v \setminus m}$ denotes the dataset with part m removed, I_m denotes part m of the dataset, j indexes each unique cell vector in I_m and $n(j)$ is a count of the cell vectors taking value j. Maximization of (E.6) is performed by equivalently minimizing the negative of $W^{\dagger}(h \mid I)$ using NAG routine E04ABF (a one dimensional minimization routine using quadratic interpolation).

Having estimated h using (E.6) and calculated the smoothed cell probabilities, $\check{p}(i)$, using (E.4) we adjust the observed cell counts, $n(i)$, by taking

$$\check{n}(i) = \check{p}(i)n \qquad \text{for all } i \in \mathcal{I},$$

where $n = \sum_{i \in \mathcal{I}} n(i)$ and $\check{n}(i)$ is the adjusted cell count. Non-zero cell totals are adjusted by taking

$$\check{t}(i) = \frac{\check{n}(i)t(i)}{n(i)} \qquad \text{for all } n(i) > 0$$

and zero cell totals are imputed using

$$\check{t}(i) = \frac{\check{n}(i)t}{n} \qquad \text{for all } n(i) = 0,$$

where $t = \sum_{i \in \mathcal{I}} t(i)$ is a $q \times 1$ vector of grand totals and $\check{t}(i)$ is an updated $q \times 1$ vector of cell i totals for the continuous variables. The effect of this adjustment is to leave the empirical non-zero cell means for each continuous variables unchanged and to impute grand means for the zero cells. Having computed this adjustment, maximum likelihood fitting of HCG models is performed in exactly the same way as described in Section 2.3.

Appendix F

Estimation of error rates

A classification rule used to predict class membership is normally trained on observed data. It is then typically used to assign a class to some new observation whose group membership is unknown. The true classification error rate is the probability that this new observation will be assigned incorrectly. In this section we outline at a few of the commonly used methods for estimating the true error rate. We start by defining one parametric measure of the true error rate. However, we are mainly concerned with the estimation of error rates using more generally applicable non-parametric methods.

F.1 Mahalanobis distance

Given observations from two multivariate normal populations with differing means μ_1 and μ_2, common covariance Σ the minimum attainable misclassification rate is

$$\text{error} = \Phi(-\tfrac{1}{2}\Delta),$$

where Δ is the true distance between the two populations given by

$$\Delta = \{(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)\}^{1/2},$$

known as the Mahalanobis distance and Φ denotes the standard normal distribution function. An obvious plug-in estimator for Δ is given by

$$D = \{(\hat{\mu}_1 - \hat{\mu}_2)'\widehat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_2)\}^{1/2}.$$

McLachlan (1992, §10.6) points out that D is a biased estimate of Δ, which tends to overestimate Δ. This then gives an optimistic estimate of the true classification error rate. McLachlan discusses bias correcting D and the application of Mahalanobis distance as the basis for more general parametric error rate estimators.

It is worth noting that error rates based on D critically depend on the assumption of normality. In addition, they estimate optimal classification rather than the error associated with the actual classification method.

F.2 Non-parametric error rates

Consider a sample pair (y, c) drawn randomly from a population, where y is an observation vector and c is its class. The probability of incorrect classification is given by

$$p(\text{incorrect} \mid Y = y) = 1 - p(c = \arg\max_l p(l \mid y) \mid Y = y) = 1 - \max_l p(l \mid y), \tag{F.1}$$

where $l = 1, \ldots, g$ for g classes. The true overall classification error rate may be obtained by averaging over the population, Y, i.e.

$$\text{err} = p(\text{incorrect}) = E[1 - \max_l p(l \mid Y)].$$

An estimate of (F.1) is given by $1 - \max_l \hat{p}(l \mid y)$ based on any calculated classification rule that yields estimated posterior probabilities. An alternative estimate of (F.1) may be computed using the indicator function

$$\begin{aligned} I[c, \hat{c} = \arg\max_l \hat{p}(l \mid y)] &= 0 \quad \text{if} \quad c = \hat{c}, \\ &= 1 \quad \text{if} \quad c \neq \hat{c}. \end{aligned} \tag{F.2}$$

In (F.2) the (assumed) true class is given by c and the assigned class by $\hat{c}$. The value of the indicator function is 0 if an observation from true class c is assigned to the same class and 1 otherwise. An estimate of the true error rate is calculated by averaging over the values of the indicator function. Averaging over the estimated posterior probabilities yields a smoothed estimate of the true error rate. Ripley (1994b) notes that the smoothed estimate has lower variance than the raw estimate since it averages over the distribution of the actual posterior probabilities given $Y = y$. Another advantage of the smoothed estimate is that it does not depend directly on the supplied classifications. This is a useful feature if the supplied classifications are thought to be unreliable.

Smoothed estimates of the class-conditional error rates may be obtained using

$$\widehat{\text{err}}_c = \frac{1}{n_c} \sum_{\nu=1}^{n} [\hat{p}(c \mid y^{(\nu)})] I[c, \hat{c} = \arg\max_l \hat{p}(l \mid y)], \tag{F.3}$$

where $y^{(\nu)}$ indexes each observation in a sample of size n and I is the indicator function defined in (F.2). Here we average over the estimated probability of being in class c given a predicted class different from c. If the number of observations in each class, n_c, is unknown then Basford & McLachlan (1985) suggest the following alternative estimator for the class-conditional error rates

$$\widetilde{\text{err}}_c = \sum_{\nu=1}^{n} [\hat{p}(c \mid y^{(\nu)})] I[c, \hat{c} = \arg\max_l \hat{p}(l \mid y)] \Big/ \sum_{\nu=1}^{n} [\hat{p}(c \mid y^{(\nu)})]. \tag{F.4}$$

Raw error rates may be estimated by taking account of the supplied classifications and averaging over the number of the incorrectly assigned observations in each class.

Typically, we try to estimate error rates using a training set, T. The *apparent error rate* is then defined as the average over $y \in T$ based on either the smoothed or raw errors. Clearly, the apparent error rate, $\widehat{\text{err}}$, is likely to underestimate the true error rate since the same data are used to both train and test the classification rule. A better estimate of the true classification error rate is given by splitting a sample in test and training sets. This uses new data to assess a previously computed classification rule, i.e. the training set is used to compute a classification rule and the test set used to validate the derived rule. If the original dataset is too small to be split into reliable training and test sets, a v-fold method of cross-validation may be used. This method partitions the dataset into v roughly equal-sized groups. Each data partition is (in turn) removed from the training set and a classification rule computed. The data previously removed from the training set is then classified. The estimated overall error rate is then calculated

by averaging over n. By choosing v to be equal to n we obtain the leave one out estimate of misclassification error (Lachenbruch & Mickey, 1968). Ripley (1996, Ch. 2) suggests that choosing $v = n$ should give the least biased assessment of classification error as the true size of the training set is most closely modelled. However, Ripley justifies using a smaller value for v by noting that dropping one observation assesses the classification rule using $O(1/n)$ perturbations, whereas, the sampling variations in the parameter estimates are (usually) $O(1/\sqrt{n})$. Ripley advocates a sensible choice of v as $v = \min(10, \sqrt{n})$ giving larger bias with reduced variance.

Another estimate of classification error is based on the *bootstrap* (see Efron & Gong (1982), Efron (1983) and also Efron & Tibshirani (1993)). This uses b bootstrap samples drawn from T to estimate the true error rate. (A bootstrap sample is a random sample of size n drawn with replacement from an initial set of n observations.) Let T^* denote a bootstrap sample, then an apparent estimate of classification error, err*, is computed for T^*. The idea is that the behaviour of $(\text{err}^* - \widehat{\text{err}})$ should mimic that of $(\widehat{\text{err}} - \text{err})$ and so $(\text{err}^* - \widehat{\text{err}})$ should provide a sensible basis for bias correcting $\widehat{\text{err}}$. An average of $(\text{err}^* - \widehat{\text{err}})$ over the b samples is one way of estimating the bias correction for $\widehat{\text{err}}$.

A number of variants on the simple bootstrap exist including the randomized and double bootstrap (Efron, 1983). Here we concentrate on Efron's (1983) '.632' bootstrap estimate of classification error, which Efron & Tibshirani (1993, p. 255) claim gives superior results to other non-parametric methods of error rate estimation in samples of fixed size. Let $T^{*1}, \ldots, T^{*b}$ denote b bootstrap samples. For each of the b samples calculate the apparent error rate and average over b to form $\text{err}^{(*)}$, i.e. an estimate of apparent error averaged over all n observations and b bootstrap samples. Now for each observation $y \in T$, calculate its out-of-sample average classification error based only on those bootstrap samples that y did not appear in. Form an overall estimate of out-of-sample average classification error, $\text{err}^{(\text{avg})}$, by averaging over n. The final .632 bootstrap estimate of classification error is given by

$$\begin{aligned} \text{err}^{(.632)} &= \text{err}^{(*)} + 0.632[\text{err}^{(\text{avg})} - \text{err}^{(*)}] \\ &= 0.368\text{err}^{(*)} + 0.632\text{err}^{(\text{avg})} \end{aligned} \tag{F.5}$$

The factor of $0.632 = 1 - e^{-1}$ in (F.5) is derived by Efron (1983) as the limit for large n of the probability that an observation from T appears in T^*.

Either raw or smoothed overall and class-conditional error rates may be estimated using v-fold cross-validation or by employing the .632 bootstrap estimator. Choice of b in computing bootstrap statistical estimates is discussed in Efron & Tibshirani (1993, p. 50–53). For the purposes of computing bootstrap estimates of classification error $b = 200$ is viewed by Efron & Tibshirani as being adequate.

Appendix G

CGM model fitting results for the SGA births study

This appendix lists the last stage in the model selection procedure for each of the SGA models given in Chapter 4. The column labelled '$-2L$' gives minus twice the maximized log-likelihood, 'd' gives the number of independently adjusted parameters and 'DF' the number of degrees of freedom. The 'p-value' is determined using a chi-squared test of deviance on the specified number of degrees of freedom. AIC values are given when appropriate, i.e. when AIC model selection is employed. Note that the presence of p-values for AIC fitted models are given for information purposes only and are not used in the model selection procedure. The final model is specified before each table but its description is given in Chapter 4. The first line of each table gives the value of $-2L$ for the final model.

SGA15: Model 1 (p. 70)

$$BI, CI, FI, BC, BF, DF/BIW, FIW, BFW, DFW, BIY, CIY, FIY, BCY, BFY, DFY/$$
$$BIWY, CIY, FIWY, BCY, BFWY, DFWY$$

Table G.1 Edge exclusion deviances for each edge in the conditional independence graph associated with the above model. Model selection was performed using chi-squared tests of deviance (both 10% and 5% critical values giving the same results).

Edge Excluded	$-2L$	d	Deviance Difference	DF	p-value
[-]	−11000.54	62	.00	0	-
[BI]	−10911.70	56	88.83	6	.0000
[CI]	−10982.52	59	18.02	3	.0004
[FI]	−10932.29	56	68.24	6	.0000
[BC]	−10069.84	59	930.70	3	.0000
[BF]	−10980.30	56	20.24	6	.0025
[DF]	−10981.71	56	18.83	6	.0045
[BW]	−10983.17	53	17.36	9	.0433
[IW]	−10961.80	53	38.73	9	.0000
[DW]	−10960.56	56	39.97	6	.0000
[FW]	−10942.28	50	58.26	12	.0000
[BY]	−10971.93	51	28.61	11	.0026
[IY]	−10980.07	51	20.47	11	.0393
[CY]	−10968.45	56	32.09	6	.0000
[DY]	−10987.74	56	12.80	6	.0464
[FY]	−10965.83	50	34.70	12	.0005
[WY]	− 9751.85	53	1248.69	9	.0000

SGA15: Model 2 (p. 70)

$$BI, CI, FI, BC, BF, CD, CF, DF / BIW, FIW, BFW, DFW, BCY, BFY, CFY / BIW, FIW, BCY, BFWY, CFY$$

Table G.2 Edge exclusion deviances for each edge in the conditional independence graph associated with the above model. Model selection was performed by minimizing AIC.

Edge	$-2L$	d	Deviance	DF	p-value	AIC
[-]	−10978.29	49	.00	0	-	98.00
[BI]	−10889.77	46	88.52	3	.0000	180.52
[CI]	−10959.60	48	18.69	1	.0000	114.69
[FI]	−10911.50	46	66.80	3	.0000	158.80
[BC]	−10056.90	46	921.40	3	.0000	1013.40
[BF]	−10953.25	43	25.04	6	.0003	111.04
[CD]	−10974.75	48	3.55	1	.0596	99.55
[CF]	−10970.65	46	7.64	3	.0540	99.64
[BW]	−10962.29	41	16.01	8	.0423	98.01
[IW]	−10914.58	43	63.72	6	.0000	149.72
[DW]	−10929.74	45	48.56	4	.0000	138.56
[FW]	−10925.20	39	53.10	10	.0000	131.10
[BY]	−10943.73	41	34.56	8	.0000	116.56
[CY]	−10949.47	43	28.82	6	.0001	114.82
[FY]	−10946.60	41	31.70	8	.0001	113.70
[WY]	−9879.67	45	1098.62	4	.0000	1188.62

SGA10: Model 1 (p. 73)

$$BF, BH, DF, EF, EH, FH / BFW, BHW, DFW, FHW, BFY, BHY, DFY, EFY, EHY, FHY / BFWY, BHWY, DFWY, EFY, EHY, FHWY$$

Table G.3 Edge exclusion deviances for each edge in the conditional independence graph associated with the above model. Model selection was performed using chi-squared tests of deviance with a 10% critical level.

Edge	$-2L$	d	Deviance	DF	p-value
[-]	−8605.46	62	.00	0	-
[BF]	−8585.23	56	20.23	6	.0025
[BH]	−8529.61	56	75.86	6	.0000
[DF]	−8593.23	56	12.23	6	.0569
[EF]	−8594.96	59	10.51	3	.0147
[EH]	−8597.38	59	8.08	3	.0443
[FH]	−8553.62	56	51.85	6	.0000
[BW]	−8589.76	53	15.71	9	.0733
[DW]	−8574.67	56	30.79	6	.0000
[FW]	−8550.58	50	54.88	12	.0000
[HW]	−8585.80	53	19.67	9	.0201
[BY]	−8586.22	53	19.24	9	.0232
[EY]	−8589.71	56	15.75	6	.0152
[FY]	−8554.54	48	50.93	14	.0000
[HY]	−8582.83	51	22.63	11	.0199
[WY]	−7337.11	53	1268.35	9	.0000

SGA10: Model 2 (p. 73)

$$BF, BH, D, EF, FH/DW, FHW, BFY, BHY, EFY, FHY/BFY, BHY, DW, EFY, FHWY$$

Table G.4 Edge exclusion deviances for each edge in the conditional independence graph associated with the above model. Model selection was performed using chi-squared tests of deviance with a 5% critical level.

Edge	$-2L$	d	Deviance	DF	p-value
[-]	−8565.32	41	.00	0	-
$[BF]$	−8548.45	38	16.87	3	.0008
$[BH]$	−8491.71	38	73.60	3	.0000
$[EF]$	−8554.31	38	11.00	3	.0117
$[FH]$	−8512.74	35	52.58	6	.0000
$[DW]$	−8533.02	39	32.29	2	.0000
$[FW]$	−8520.51	35	44.80	6	.0000
$[HW]$	−8547.90	35	17.42	6	.0079
$[BY]$	−8545.63	35	19.68	6	.0032
$[EY]$	−8553.63	37	11.68	4	.0199
$[FY]$	−8526.69	31	38.63	10	.0000
$[HY]$	−8547.42	33	17.89	8	.0220
$[WY]$	−7298.82	37	1266.50	4	.0000

SGA10: Model 3 (p. 73)

$$BF, BH, DH, EF, EH, FH/DW, BFW, BH, DH, EFY, EHY, FHY/$$
$$DW, BFW, EFY, EHY, FHY, FWY$$

Table G.5 Edge exclusion deviances for each edge in the conditional independence graph associated with the above model. Model selection was performed by minimizing AIC.

Edge	$-2L$	d	Deviance	DF	p-value	AIC
[-]	−8566.80	37	.00	0	-	74.00
$[BF]$	−8549.30	34	17.50	3	.0006	85.50
$[BH]$	−8491.70	36	75.10	1	.0000	147.10
$[DH]$	−8559.28	36	7.52	1	.0061	79.52
$[EF]$	−8543.18	34	23.62	3	.0000	91.62
$[EH]$	−8558.70	34	8.10	3	.0440	76.10
$[FH]$	−8512.15	34	54.65	3	.0000	122.65
$[DW]$	−8492.38	35	74.43	2	.0000	144.43
$[BW]$	−8513.27	33	53.54	4	.0000	119.54
$[FW]$	−8520.26	32	46.54	5	.0000	110.54
$[EY]$	−8551.01	31	15.79	6	.0149	77.79
$[FY]$	−8531.51	30	35.29	7	.0000	95.29
$[HY]$	−8532.15	31	34.65	6	.0000	96.65
$[WY]$	−7285.64	35	1281.16	2	.0000	1351.16

Bwt: Model 1 (p. 76)

$$ABC, BCF/BW, BCFX, BCFY, ABCZ, BCFZ/WX, WYZ$$

Table G.6 Edge exclusion deviances for each edge in the conditional independence graph associated with the above model. Model selection was performed using chi-squared tests of deviance with a 10% critical level. However, AIC model selection produced the same final model.

Edge	$-2L$	d	Deviance	DF	p-value	AIC
$[\ -\ ]$	−22647.29	49	.00	0	-	98.00
$[AB]$	−22616.46	45	30.83	4	.0000	120.83
$[AC]$	−22597.29	45	50.00	4	.0000	140.00
$[BC]$	−21888.02	39	759.27	10	.0000	837.27
$[BF]$	−22613.85	41	33.45	8	.0001	115.45
$[CF]$	−22624.84	41	22.45	8	.0041	104.45
$[BW]$	−22644.54	48	2.75	1	.0970	98.75
$[BX]$	−22632.79	45	14.50	4	.0059	104.50
$[CX]$	−22635.41	45	11.88	4	.0183	101.88
$[FX]$	−22514.78	45	132.51	4	.0000	222.51
$[BY]$	−22628.79	45	18.50	4	.0010	108.50
$[CY]$	−22632.67	45	14.62	4	.0055	104.62
$[FY]$	−22625.43	45	21.86	4	.0002	111.86
$[AZ]$	−22632.97	45	14.33	4	.0063	104.33
$[BZ]$	−22534.75	43	112.54	6	.0000	198.54
$[CZ]$	−22634.35	43	12.94	6	.0440	98.94
$[FZ]$	−22572.75	45	74.54	4	.0000	164.54
$[WX]$	−22641.69	48	5.60	1	.0179	101.60
$[WY]$	−21451.62	48	1195.67	1	.0000	1291.67
$[WZ]$	−22560.34	48	86.96	1	.0000	182.96
$[YZ]$	−22633.06	48	14.23	1	.0001	110.23

Bwt: Model 2 (p. 76)

$$ABC, BCF/W, BCFX, BCFY, ABCZ, BCFZ/WX, WYZ$$

Table G.7 Edge exclusion deviances for each edge in the conditional independence graph associated with the above model. Model selection was performed using chi-squared tests of deviance with a 5% critical level.

Edge	$-2L$	d	Deviance	DF	p-value
$[\ -\ ]$	−22644.54	48	.00	0	-
$[AB]$	−22613.71	44	30.83	4	.0000
$[AC]$	−22594.54	44	50.00	4	.0000
$[BC]$	−21885.27	38	759.27	10	.0000
$[BF]$	−22611.09	40	33.45	8	.0001
$[CF]$	−22622.09	40	22.45	8	.0041
$[BX]$	−22630.28	44	14.26	4	.0065
$[CX]$	−22632.66	44	11.88	4	.0183
$[FX]$	−22512.03	44	132.51	4	.0000
$[BY]$	−22621.90	44	22.64	4	.0001
$[CY]$	−22629.91	44	14.62	4	.0056
$[FY]$	−22604.08	44	40.45	4	.0000
$[AZ]$	−22630.21	44	14.33	4	.0063
$[BZ]$	−22527.42	42	117.12	6	.0000
$[CZ]$	−22631.60	42	12.94	6	.0440
$[FZ]$	−22570.01	44	74.53	4	.0000
$[WX]$	−22639.08	47	5.46	1	.0195
$[WY]$	−21443.05	47	1201.49	1	.0000
$[WZ]$	−22549.13	47	95.41	1	.0000
$[YZ]$	−22631.24	47	13.30	1	.0003

References

Aitchison, J. and Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, **63**, 413–420.

Aitchison, J. and Begg, C. B. (1976). Statistical diagnosis when basic cases are not classified with certainty. *Biometrika*, **63**, 1–12.

Aitchison, J. and Dunsmore, I. (1975). *Statistical Prediction Analysis*. Cambridge University Press.

Aitken, C. G. G. (1978). Methods of discrimination in multivariate binary data. In *Proceedings of COMPSTAT 1978* (editors L. C. A. Corsten and J. Hermans), pages 155–161. Physica-Verlag, Vienna.

Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1989). *Statistical Modelling in GLIM*. Oxford University Press, Oxford.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory* (editors B. N. Petrov and F. Cáski), pages 267–281. Akademiai Kaidó.

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control*, **AC-19**, 716–722.

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression analysis. *Biometrika*, **71**, 1–10.

Anderson, E. (1935). The irises of the Gaspe Peninsula. *Bulletin of the American Iris Society*, **59**, 2–5.

Anderson, J. A. (1975). Quadratic logistic discrimination. *Biometrika*, **62**, 149–154.

Anderson, J. A. (1982). Logistic discrimination. In *Handbook of Statistics, Volume 2* (editors P. R. Krishnaiah and L. N. Kanal), pages 169–191. North-Holland, Amsterdam.

Armitage, P. and Berry, G. (1987). *Statistical Methods in Medical Research*. Blackwell Scientific Publications, Oxford.

Bakketeig, L. S., Hoffman, H. J. and Harley, E. E. (1979). The tendency to repeat gestational age and birth weight in successive births. *American Journal of Obstetrics and Gynecology*, **135**, 1086–1103.

Bakketeig, L. S., Jacobsen, G., Hoffman, H. J., Lindmark, G., Bergsjø, P., Molne, K. and Rødstein, J. (1993). Pre-pregnancy risk factors of small for gestational age births among parous women in Scandinavia. *Acta Obstetricia et Gynecologica Scandinavica*, **72**, 273–279.

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, New York.

Bartlett (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society A*, **160**, 268–82.

Basford, K. E. and McLachlan, G. J. (1985). Estimation of allocation rates in a cluster analysis context. *Journal of the American Statistical Association*, **80**, 286–293.

Berge, C. (1973). *Graphs and Hypergraphs*. North-Holland, Amsterdam.

Bergsjø, P., Hoffman, H. J., Davis, R. O., Goldenberg, R. L., Lindmark, G., Jacobsen, G., Cutter, G., Markestad, T., Nelson, K. G. and Bakketeig, L. S. (1989). Preliminary results from the collaborative Alabama and Scandinavian study of small for gestational age births. *Acta Obstetricia et Gynecologica Scandinavica*, **68**, 19–25.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* The MIT Press, Cambridge, Massachusetts.

Bjerkedal, T. and Skjærven, R. (1980). Percentiles of birth weight and crown-heel length for single live births. *Tidsskrift for den Norsk Lægeforening*, **100**, 1088–91.

Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika*, **36**, 317–346.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistcal Analysis*. Wiley, New York. (Originally published by Addison-Wesley, Reading, Massachusetts.)

Breiman, L. and Ihaka, R. (1984). Nonlinear discriminant analysis via ACE and scaling. Technical Report 40, Department of Statistics, University of California, Berkeley.

Bunday, B. D. and Kiri, V. A. (1987). Maximum likelihood estimation — practical merits of variable metric optimisation methods. *The Statistician*, **36**, 349–355.

Campbell, N. A. (1980). Shrunken estimators in discriminant and canonical variate analysis. *Journal of the Royal Statistical Society C*, **29**, 5–13, 24.

Cornfield, J. (1962). Joint dependence of risk of coronary heart disease on serum cholestrol and systolic blood pressure: a discriminant function approach. *Federation Proceedings. Federation of American Societies for Experimental Biology*, **11**, 58–61.

Cox, D. R. (1958). Two further applications of a model for binary regression. *Biometrika*, **45**, 562–565.

Cox, D. R. (1966). Some procedures associated with the logistic qualitative response curve. In *Research Papers in Statistics: Festschrift for J. Neyman* (editor F. N. David), pages 51–71. Wiley, New York.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data*. Second Edition. Chapman & Hall, London.

Cox, D. R. and Wermuth, N. (1996). *Multivariate Dependencies*. Chapman & Hall, London.

Darroch, J. N., Lauritzen, S. L. and Speed, T. P. (1980). Markov fields and log-linear interaction models for contingency tables. *Annals of Statistics*, **8**, 522–539.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society B*, **41**, 1–31.

Dawid, A. P. (1980). Conditional independence for statistical operations. *Annals of Statistics*, **8**, 598–617.

Day, N. E. and Kerridge, D. F. (1967). A general maximum likelihood discriminant. *Biometrics*, **23**, 313–23.

Deming, M. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, **11**, 427–444.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, **28**, 157–175.

Devijver, P. A. and Kittler, J. V. (1982). *Pattern Recognition: A Statistical Approach.* Prentice Hall: Englewood Cliffs, NJ.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society B*, **57**, 45–70, 71–97.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*. Second Edition. Wiley, New York.

Edwards, D. (1995). *Introduction to Graphical Modelling*. Springer-Verlag, New York.

Edwards, D. and Havránek, T. (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrika*, **72**, 339–351.

Edwards, D. and Havránek, T. (1987). A fast model selection procedure for large families of models. *Journal of the American Statistical Association*, **82**, 205–211.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, **70**, 892–898.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, **78**, 316–331.

Efron, B. and Gong, G. (1982). A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician*, **78**, 36–48.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics (London)*, **7**, 179–188.

Fisher, R. A. (1959). *Statistical Methods and Scientific Inference*. Hafner Press, New York.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, **84**, 165–175.

Frydenberg, M. and Lauritzen, S. L. (1989). Decomposition of maximum likelihood in mixed graphical interaction models. *Biometrika*, **76**, 539–555.

Gammerman, A., Luo, Z., Aitken, C. G. G. and Brewer, M. J. (1995). Exact and approximate algorithms and their implementations in mixed graphical models. In *Probabilistic Reasoning and Bayesian Belief Networks* (editor A. Gammerman), pages 33–53. Alfred Waller, Henley-on-Thames.

Geisser, S. (1993). *Predictive Inference: An Inroduction*. Chapman & Hall, New York.

Geisser, S. and Cornfield, J. (1963). Posterior distributions for multivariate normal parameters. *Journal of the Royal Statistical Society B*, **25**, 368–376.

Gill, P. E., Murray, W. and Wright, M. H. (1981). *Practical Optimization*. Academic Press, New York.

Goodman, L. A. (1970). The multivariate analysis of qualitative data: Interaction among multiple classifications. *Journal of the American Statistical Association*, **65**, 226–256.

Goodman, L. A. (1971). Partitioning of chi-square, analysis of marginal contingency tables, and estimation of expected frequencies in multidimensional contingency tables. *Journal of the American Statistical Association*, **66**, 339–344.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, **27**, 857–74.

Habbema, J. D. F., Hermans, J. and van der Broeck, K. (1974). A stepwise discriminant analysis program using density estimation. In *Compstat 1974* (editor G. Bruckmannn), pages 101–110. Physica Verlag, Vienna.

Hammersley, J. M. and Clifford, P. (1971). Markov fields on finite graphs and lattices. Unpublished manuscript.

Hastie, T., Buja, A. and Tibshirani, R. (1994). Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, **89**, 1255–1270.

Hastie, T. J. and Pregibon, D. (1992). Generalized linear models. In *Statistical Models in S* (editors J. M. Chambers and T. J. Hastie). Chapman & Hall, New York.

Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression.* Wiley, New York.

Jacobsen, G. (1992). *Detection of Intrauterine Growth Deviations: A Comparison Between Serial Symphysis Fundus Height and Ultrasound Measurements.* Yale University, USA and Trondheim University, Norway. (Report available from the Department of Community Medicine and General Practice, University of Trondheim, Norway.)

James, W. and Stein, C. (1961). Estimation with quadratic loss, In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability (Volume 1)*, pages 361–379. University of California Press, Berkeley.

Jennings, A. and McKeown, J. J. (1992). *Matrix Computation.* Second Edition. Wiley, Chichester.

Kimura, F., Takashina, K., Tsuruoka, S. and Miyake, Y. (1987). Modified quadratic discriminant functions and the application to Chinese character sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-9**, 149–153.

Krusińska, E. (1991). Suitable location model selection in the terminology of graphical models. *Biom. J.*, **32**, 817–826.

Krzanowski, W. J. (1975). Discrimination and classification using both binary and continuous variables. *Journal of the American Statistical Association*, **70**, 782–790.

Krzanowski, W. J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics*, **36**, 493–499.

Lachenbruch, P. A. and Mickey, M. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, **10**, 1–11.

Lauritzen, S. L. (1992). Propagation of probabilities, means, and variances in mixed graphical association models. *Journal of the American Statistical Association*, **87**, 1098–1108.

Lauritzen, S. L. (1996). *Graphical Models.* Oxford University Press, Oxford.

Lauritzen, S. L. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, **17**, 31–57.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, **89**, 1535–1546.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, **63**, 215–232.

Magnus, J. R. and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Wiley, Chichester.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis.* Academic Press, London.

McKay, R. J. and Campbell, N. A. (1982a). Variable selection techniques in discriminant analysis I. Description. *British Journal of Mathematical and Statistical Psychology*, **35**, 1–29.

McKay, R. J. and Campbell, N. A. (1982b). Variable selection techniques in discriminant analysis II. Allocation. *British Journal of Mathematical and Statistical Psychology*, **35**, 30–41.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.

Minsky, M. (1961). Steps towards artifical intelligence. *Proceedings of the IRE*, **49**, 8–30.

Morrison, D. F. (1967). *Multivariate Statistical Methods*. McGraw-Hill, New York.

Murata, N., Yoshizawa, S. and Amari, S. (1991). A criterion for determining the number of parameters in an artificial neural networks model. In *Artificial Neural Networks Volume I* (editors T. Kohonen, K. Mäkisara, O. Simula and J. Kangas), pages 9–14. North-Holland, Amsterdam.

Nash, J. C. (1990). *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation*. Second Edition. Adam Hilger, Bristol.

Olkin, I. and Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics*, **32**, 448–465.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery (1992). *Numerical Recipes in FORTRAN: The Art of Scientific Computing*. Second Edition. Cambridge University Press, Cambridge.

Rao, C. (1948). The utilization of multiple measurements in problems of biological classification (with discussion). *Journal of the Royal Statistical Society B*, **10**, 159–203.

Rice, J. A. (1988). *Mathematical Statistics and Data Analysis*. Wadsworth and Brooks/Cole, Pacific Grove, California.

Ripley, B. D. (1994a). Flexible non-linear approaches to classification. In *From Statistics to Neural Networks. Theory and Pattern Recognition Applications* (editors V. Cherkassky, J. H. Friedman and H. Wechsler), pages 105–126. ASI Proceedings, subseries F, Computer and Systems Sciences. Springer-Verlag.

Ripley, B. D. (1994b). Neural networks and related methods for classification (with discussion). *Journal of the Royal Statistical Society B*, **56**, 409–456.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

Roverato, A. and Whittaker, J. (1996). Standard errors for the parameters of graphical Gaussian models. *Statistics and Computing*, **6**, 297–302.

Schoener, T. W. (1970). Nonsynchronous spatial overlap of lizards in patchy habitats. *Ecology*, **51**, 408–418.

Speed, T. P. and Kiiveri, H. T. (1986). Gaussian Markov distributions over finite graphs. *Annals of Statistics*, **14**, 138–150.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L. and Cowell, R. G. (1993). Bayesian analysis in expert systems. *Statistical Science*, **8**, 219–283.

Stewart, L. (1987). Hierarchical Bayesian analysis using Monte Carlo integration: computing posterior distributions when there are many possible models. *The Statistician*, **36**, 211–219.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society B*, **39**, 44–47.

Tate, R. F. (1954). Correlation between a discrete and continuous variable. *Annals of Mathematical Statistics*, **25**, 603–607.

Titterington, D. M. (1980). A comparative study of kernel-based density estimates for categorical data. *Technometrics*, **22**, 259–268.

Truett, J., Cornfield, J. and Kannel, W. B. (1967). A multivariate analaysis of the risk of coronary heart disease in Framingham. *Journal of Chronic Diseases*, **20**, 511–524.

Venables, W. N. and Ripley, B. D. (1997). *Modern Applied Statistics with S-Plus*. Second Edition. Springer-Verlag, New York.

Vlachonikolis, I. G. (1990). Predictive discrimination and classification with mixed binary and continuous variables. *Biometrika*, **77**, 657–662.

Wermuth, N. and Lauritzen, S. L. (1990). On substantive research hypotheses, conditional independence and graphical chain models (with discussion). *Journal of the Royal Statistical Society B*, **52**, 21–50, 51–72.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester.

Wilson, R. J. (1985). *Introduction to Graph Theory*. Longman, Harlow.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, **20**, 557–585.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, **5**, 161–215.