Lectures on Contingency Tables

Electronic edition

Steffen L. Lauritzen
Aalborg University

October 2002

Preface

The present set of lecture notes are prepared for the course "Statistik 2" at the University of Copenhagen. It is a revised version of notes prepared in connection with a series of lectures at the Swedish summerschool in Särö, June 11–17, 1979.

The notes do by no means give a complete account of the theory of contingency tables. They are based on the idea that the graph theoretic methods in Darroch, Lauritzen and Speed (1978) can be used directly to develop this theory and, hopefully, with some pedagogical advantages.

My thanks are due to the audience at the Swedish summerschool for patiently listening to the first version of these lectures, to Joseph Verducci, Stanford, who read the manuscript and suggested many improvements and corrections, and to Ursula Hansen, who typed the manuscript.

Copenhagen, September 1979

Steffen L. Lauritzen

Preface to the second edition

The second edition is different from the first basically by the fact that a number of errors have been removed, certain simplifications have been achieved and some recent results about collapsibility have been included.

I am indebted to Søren Asmussen, Copenhagen and Ole Barndorff-Nielsen, Aarhus for most of these improvements.

Aalborg, August 1982

Steffen L. Lauritzen

Preface to the third edition

In the third edition more errors have been corrected and other minor changes have been made. There has been some effort to bring the list of references up to date.

Finally a section on other sampling schemes (Poisson models and sampling with one fixed marginal) has been added. This section is based upon notes by Inge Henningsen and I am grateful for her permission to include this material.

Again Søren Asmussen and Søren Tolver Jensen have found errors and made many detailed suggestions for improvements in the presentation. Helle Andersen and Helle Westmark have helped transforming the typescript into a typeset document. Søren L. Buhl has given me very helpful comments on this manuscript. Thank you.

Aalborg, August 1989

Steffen L. Lauritzen

Preface to the electronic edition

The only essential changes from the third edition to this one is that some figures and tables have been redrawn and some changes have been made to the typesetting, to accomodate the electronic medium. Some references have also been updated.

Klitgaarden, Skagen, October 2002

Steffen L. Lauritzen

Contents

1 Introduction

Before we proceed to develop the theory of multidimensional contingency tables we shall briefly consider the basics of tables having dimension 2 or 3.

Suppose n objects are classified according to two criteria A and B, these having levels $A_1, \ldots, A_r$ and $B_1, \ldots, B_s$. This gives rise to a table of the following form:

	B_1	$\cdots$	B_j	$\cdots$	B_s	totals
A_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1s}	$n_{1\cdot}$
.	.	$\cdots$	.	$\cdots$	.	.
.	.	$\cdots$	.	$\cdots$	.	.
.	.	$\cdots$	.	$\cdots$	.	.
A_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{is}	$n_{i\cdot}$
.	.	$\cdots$	.	$\cdots$	.	.
.	.	$\cdots$	.	$\cdots$	.	.
.	.	$\cdots$	.	$\cdots$	.	.
A_r	n_{r1}	$\cdots$	n_{rj}	$\cdots$	n_{rs}	$n_{r\cdot}$
totals	$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot s}$	n

Here n_{ij} are the *counts*, n_{ij} being the number of objects classified to belong to the *cell* (A_i, B_j). The symbols $n_{\cdot 1}, \ldots, n_{\cdot s}$ and $n_{1\cdot}, \ldots, n_{r\cdot}$ denote the *marginal counts*.

The statistical models we shall consider for observations of this type will describe relations (or rather lack of relations) between such criteria. It is important to notice that the models treated here are based on very little structure indeed. Thus the models do not take into account e.g.

- an *ordering* on the levels $A_1, \ldots, A_r$ (like low, medium, high etc.)

- a *scaling* on the levels (like when the criteria in fact are artificially discretised continuous variables)

- an *asymmetry* of the type that, say, A is a treatment and B a response or that A_1 is a 'control' and $A_2, \ldots, A_r$ corresponds to various 'treatments'

4

- a *theoretical knowledge* about the probability structure. In e.g. genetics where one can be interested in various linkage problems etc. there might be a good *a priori* knowledge about the possible probability models.

It is a serious limitation on the usefulness of the models that structures of the above kind disappear in the analysis. Nevertheless, any theory has a starting point and the methods developed might be useful for a rough, preliminary statistical analysis of many problems.

We shall use probability models based on the assumption that p_{ij} is the probability that any object belongs to cell (A_i, B_j) and that the cells of all the n objects are stochastically independent such that the joint probability distribution of the entire table of counts is multinomial:

$$P\{N_{ij} = n_{ij}, i = 1, \ldots, r, j = 1, \ldots, s\} = \frac{n!}{\prod_{ij} n_{ij}!} \prod_{ij} p_{ij}^{n_{ij}}$$

In general the only restriction on p_{ij} is that $p_{ij} \geq 0$ and $\sum_{ij} p_{ij} = 1$. The models we consider in a two-way table are given by restricting the probability p_{ij} as follows:

M_0: p_{ij} *unrestricted*

M_1: *independence:*

$$p_{ij} = p_{i\cdot} \times p_{\cdot j}$$

M_2: *irrelevance of B:*

$$p_{ij} = p_{i\cdot} \times \frac{1}{s}$$

M_2^*: *irrelevance of A:*

$$p_{ij} = \frac{1}{r} \times p_{\cdot j}$$

M_3: *complete irrelevance:*

$$p_{ij} = \frac{1}{rs}$$

The models are clearly related as follows:

$$M_0 \supset M_1 \supset \begin{matrix} M_2 \\ M_2^* \end{matrix} \supset M_3$$

and the maximum likelihood estimate of the p_{ij}'s under assumptions of the various models are given as:

M_0:

$$\hat{p}_{ij} = \frac{n_{ij}}{n}$$

M_1:

$$\hat{p}_{ij} = \frac{n_{i\cdot}}{n} \times \frac{n_{\cdot j}}{n}$$

M_2:

$$\hat{p}_{ij} = \frac{n_{i\cdot}}{n} \times \frac{1}{s}$$

M_2^*:

$$\hat{p}_{ij} = \frac{1}{r} \times \frac{n_{\cdot j}}{n}$$

M_3:

$$\hat{p}_{ij} = \frac{1}{rs}$$

The models can be tested by the likelihood ratio statistic which is given as:

$$-2\log Q = 2\sum_{ij} n_{ij}(\log \hat{p}_{ij} - \log \hat{\hat{p}}_{ij}),$$

where $\hat{p}_{ij}$ is the estimate of p_{ij} under the model assumed and $\hat{\hat{p}}_{ij}$ under that to be tested. Of course the latter has to be a submodel of the former.

The likelihood ratio statistic can be approximated by the Pearson χ^2-statistic:

$$-2\log Q \approx \chi^2 = \sum_{ij} \frac{(n\hat{p}_{ij} - n\hat{\hat{p}}_{ij})^2}{n\hat{\hat{p}}_{ij}}$$

6

The exact distribution of any of these two statistics is in general intractable and any practical inference has to be based on the asymptotic distribution which can be shown to be χ^2 with an appropriate number of degrees of freedom. It is a delicate question to judge whether the asymptotic results can be used in any practical situation and no clear results of practical use are known. A guideline is that $n\hat{p}_{ij}$ should be large, which seldom is the case for practical investigations with tables of high dimension. It is sometimes possible to collapse tables so as to reduce the dimensionality. We shall return to that later.

Three–way tables are treated mostly the same way. If n_{ijk} denote the counts and p_{ijk} the cell probabilities, we can as before consider models as e.g.

- *independence* between (A, B) and C:

$$p_{ijk} = p_{ij\cdot} \times p_{\cdot\cdot k}$$

- *irrelevance* of C

$$p_{ijk} = p_{ij\cdot} \times \frac{1}{t}$$

  where $C \sim C_1, \ldots, C_t$

- *complete independence*

$$p_{ijk} = p_{i\cdot\cdot} \times p_{\cdot j\cdot} \times p_{\cdot\cdot k}$$

  etc. etc.

But apart from these a new hypothesis naturally emerges stating that A and C are *conditionally independent* given B:

$$p_{ik|j} = p_{i|j} \times p_{k|j}$$

or equivalently

$$p_{ijk} = \frac{p_{ij\cdot} \times p_{\cdot jk}}{p_{\cdot j\cdot}}.$$

7

If we look at the three–way table as a box, the condition $B = B_j$ determines a slice, i.e. a two–dimensional table in A and C. The above hypothesis corresponds to the assumption of independence in all these s two–way tables.

Under the assumption above, one can show that the maximum likelihood estimate is given as

$$\hat{p}_{ijk} = \frac{n_{ij.} \times n_{.jk}}{n_{.j.} \times n}.$$

Also in a three–way table the statistical inference can be based on the analogous statistics

$$-2 \log Q = \sum_{ijk} n_{ijk} (\log \hat{p}_{ijk} - \log \hat{\hat{p}}_{ijk}) \approx \sum_{ijk} \frac{(n\hat{p}_{ijk} - n\hat{\hat{p}}_{ijk})^2}{n\hat{\hat{p}}_{ijk}}.$$

Going from dimension 2 to 3, we saw that an essentially new kind of model, that of conditional independence occurred. Generalizing to many dimensions makes everything rather complex and such a generalization can be made in different ways. We shall primarily be concerned with models given by conditional independence, independence and irrelevance and try to give tools that makes it possible to deal with all these such that the general overview and insight does not disappear. The approach is based on the ideas and results in Darroch, Lauritzen and Speed (1980).

A more usual way of developing the theory is to assume $p_{ijk} > 0$ and make an expansion as

$$\log p_{ijk} = \alpha + \beta_i^1 + \beta_j^2 + \beta_k^3 + \gamma_{ij}^{12} + \gamma_{jk}^{23} + \gamma_{ik}^{13} + \delta_{ijk}^{123}$$

and then assume certain of these terms not to be present. The model of conditional independence corresponds to assuming

$$\delta_{ijk}^{123} \equiv 0 \quad \text{and} \quad \gamma_{ik}^{13} \equiv 0.$$

Models of this type are denoted *log–linear interaction models* and the terms in the expansion are denoted *interactions*. The method has the advantage that it immediately generalizes to n dimensions but the disadvantage

that it becomes difficult to understand what the models in fact mean for the problem that has to be analysed. Log-linear models and many other aspects of the analysis of contingency tables are treated in detail in the classic book by Bishop, Fienberg and Holland (1975).

One of the most difficult problems when the generalization has to be made is to get a convenient notation. Therefore the next section will introduce a new notation.

Secondly, since conditional independence becomes fundamental, we shall briefly discuss the basic properties of this notion.

Finally, it shall turn out that the problem of dealing with many conditional independencies is most easily taken care of by means of a little bit of graph theory which also is contained in the following.

2   Preliminaries and notation

Let Γ be the set of *classification criteria*. Γ is supposed to be finite and $|\Gamma|$ is the number of elements in Γ. For each $\gamma \in \Gamma$ there are *levels* $\mathcal{I}_\gamma$ which again is a finite set. A *cell* in the table is a point $i = (i_\gamma)_{\gamma \in \Gamma}$ where

$$i \in \mathcal{I} = \times_{\gamma \in \Gamma} \mathcal{I}_\gamma$$

We assume that n objects are classified and the set of *counts*

$$n(i) = \text{the number of objects in cell } i, \quad i \in \mathcal{I}$$

constitute the contingency table, such that

$$n = \sum_{i \in \mathcal{I}} n(i).$$

The number of criteria $|\Gamma|$ is the *dimension* of the table. The *marginal tables* are those given by only classifying the objects according to a subset $a \subseteq \Gamma$ of the criteria. We have the *marginal cells* $i_a = (i_\gamma)_{\gamma \in a}$

$$i_a \in \mathcal{I}_a = \times_{\gamma \in a} \mathcal{I}_\gamma$$

and the corresponding *marginal counts*

$$n(i_a) = \text{the number of objects in } i_a = \sum_{j:j_a=i_a} n(j).$$

As before we have the probabilities

$$p(i) = \text{probability that an object belongs to cell } i,$$

where $p(i) \geq 0$ and $\sum_{i \in \mathcal{I}} p(i) = 1$. And the joint distribution of the entire table is multinomial:

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

We are interested in *marginal probabilities*

$$\begin{aligned} p(i_a) &= \text{probability that an object belongs to the marginal cell } i_a \\ &= \sum_{j:j_a=i_a} p(j). \end{aligned}$$

The notion of conditional independence is of great importance to us and we say for three discrete–valued random variables X, Y and Z that X and Y are *conditionally independent* given Z if

$$p(x, y \,|\, z) = p(x \,|\, z)p(y \,|\, z) \quad \text{whenever} \quad p(z) > 0$$

where $p(x, y \,|\, z)$ is an imprecise but hopefully clear notation for

$$p(x, y \,|\, z) = P\{X = x, Y = y \,|\, Z = z\}.$$

If X and Y are conditionally independent given Z, we write

$$X \perp\!\!\!\perp Y \,|\, Z.$$

Note that the following statements are true, where these all should be read with the sentence following: 'whenever all quantities are well–defined':

CI0: $X \perp\!\!\!\perp Y \,|\, Z \Leftrightarrow p(x, y, z) = p(x, z)p(y, z)/p(z)$

10

CI1:	$X \perp\!\!\!\perp Y \mid Z \Leftrightarrow p(x \mid y, z) = p(x \mid z)$
CI2:	$X \perp\!\!\!\perp Y \mid Z \Leftrightarrow p(x, y \mid z) = p(x \mid z)p(y \mid z)$
CI3:	$X \perp\!\!\!\perp Y \mid Z \Leftrightarrow \exists f, g : p(x, y, z) = f(x, z)g(y, z)$
CI4:	$X \perp\!\!\!\perp Y \mid Z \Leftrightarrow p(x, y, z) = p(x \mid z)p(y, z)$

The statements below give rules for deducing conditional independence statements from others:

CI5:	$X \perp\!\!\!\perp Y \mid Z \Leftrightarrow Y \perp\!\!\!\perp X \mid Z$
CI6:	$X \perp\!\!\!\perp Y \mid Z \Rightarrow X \perp\!\!\!\perp Y \mid (Z, f(Y))$
CI7:	$X \perp\!\!\!\perp Y \mid Z \Rightarrow f(X, Z) \perp\!\!\!\perp Y \mid Z$
CI8:	$[X \perp\!\!\!\perp Y \mid Z \text{ and } X \perp\!\!\!\perp W \mid (Y, Z)] \Rightarrow X \perp\!\!\!\perp (Y, W) \mid Z$

The proof of all these assertions is an exercise in elementary probability and omitted. In fact the rules CI5–CI8 can be used as a system of axioms for probabilistic independence, such as pointed out by Dawid (1979, 1980). This has also been further exploited to sketch a theory of *irrelevance* in connection with the logic of reasoning cf. Pearl (1988).

We shall in our contingency table be especially interested in the random variables corresponding to *marginals* such that we let $X_a(i) = i_a$ for $a \subseteq \Gamma$ and $X_\gamma(i) = i_\gamma$ for $\gamma \in \Gamma$. Instead of writing
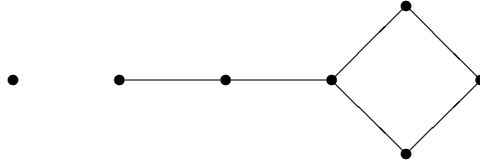
$$X_a \perp\!\!\!\perp X_b \mid X_c$$

we shall simply write

$$a \perp\!\!\!\perp b \mid c.$$

Only conditionally independencies of this type shall be dealt with.

We shall use the notion of a (simple, undirected) *graph*, i.e. a pair (V, E) where V is a finite set, the *vertex set*, and E is a subset of the set of all unordered pairs of V, the *edge set*. We display such a graph as e.g.

where the 'dots' correspond to elements of V and 'bars' to elements of E.

We say that two vertices $\alpha, \beta \in V$ are *adjacent* or *neighbours* and write $\alpha \sim \beta$, if there is an edge between them, i.e. if $\{\alpha, \beta\} \in E$.

The *boundary* of a subset $v \subseteq V$ is denoted by ∂v and consists of those vertices that are not in v but adjacent to some vertex in v:

$$\partial v = \{\beta \in V \setminus v \mid \exists \alpha \in v : \alpha \sim \beta\}.$$

Instead of $\partial\{\alpha\}$ we simply write $\partial\alpha$. The *closure* of a set $v \subseteq V$ is denoted as $\overline{v}$ and defined as $\overline{v} = v \cup \partial v$.

A subset $v \subseteq V$ with all vertices being mutual neighbours is called a *complete* subset, i.e.

$$v \text{ complete} \Leftrightarrow \forall \alpha, \beta \in v : \{\alpha, \beta\} \in E.$$

If a complete subset is maximal (w.r.t. inclusion) with this property, it is called a *clique:*

$$c \text{ is a clique} \Leftrightarrow [\, c \text{ is complete and } c^* \supset c \Rightarrow c^* \text{ not complete} \,].$$

The set of cliques of a graph is denoted by $\mathcal{C}$. A graph with at most one clique is a *complete graph*. (The empty graph is complete).

A graph (W, F) is a *subgraph* of (V, E) if $W \subseteq V$ and $F \subseteq E$. If $A \subseteq V$ it induces a subgraph

$$(A, E_A) = (A, \{\{\alpha, \beta\} \in E : \alpha \in A \wedge \beta \in A\}).$$

We can make unions and intersections of graphs by

$$\begin{aligned} (V, E) \cup (W, F) &= (V \cup W, E \cup F) \\ (V, E) \cap (W, F) &= (V \cap W, E \cap F). \end{aligned}$$

12

If $(V \cap W, E \cap F)$ is a complete graph we say that the above union is *direct* and write

$$(V \cup W, E \cup F) = (V, E) \mathbin{\dot{\cup}} (W, F).$$

We then say that we have a *decomposition* of $(V \cup W, E \cup F)$.

A *path* of length n in the graph (V, E) is a string of vertices $\alpha_0, \ldots, \alpha_n$ such that $\alpha_{i-1} \sim \alpha_i$ for $i = 1, \ldots, n$. Note that a path can have length zero. An *n–cycle* is a path of length n with $\alpha_0 = \alpha_n$.

We can define an equivalence relation on V as

$$\alpha \sim_p \beta \Leftrightarrow \text{there is a path } \alpha_0, \ldots, \alpha_n \text{ with } \alpha_0 = \alpha, \alpha_n = \beta.$$

The subgraphs induced by the equivalence classes are the *connected components* of (V, E). If there is only one equivalence class, we say that (V, E) is *connected.*

Two subsets a, b of Γ are said to be *separated* by the subset c with $a \cap b \subseteq c$ if all paths from a to b go via c, i.e. intersect c at some vertex. Note that if

$$(V \cup W, E \cup F) = (V, E) \cup (W, F)$$

then V and W are separated by $V \cap W$ in the graph $(V \cup W, E \cup F)$. In particular this holds for any decomposition. Thus two induced subgraphs (a, E_a) and (b, E_b) form a decomposition of (Γ, E) if and only if

(i) $a \cup b = \Gamma$

(ii) $a \cap b$ is complete

(iii) $a \cap b$ separates $a \setminus b$ from $b \setminus a$.

We then say that a and b *decompose* (Γ, E).

3 The Markov property

The models shall be given by conditional independence restrictions of Markov type: So let there be given a graph (V, E) with $V \subseteq \Gamma$.

Definition 1 A probability p is said to be *Markov* w.r.t. (V, E) if

M1: $p(i) > 0$ for all $i \in \mathcal{I}$

M2: $p(i) = p(i_V) \prod_{\gamma \notin V} |\mathcal{I}_\gamma|^{-1}$

M3: If $\alpha, \beta \in \Gamma, \alpha \not\sim \beta$ then $\alpha \perp\!\!\!\perp \beta \mid \Gamma \setminus \{\alpha, \beta\}$.

A probability is *extended Markov* if $p(i) = \lim_{n \to \infty} p^{(n)}(i)$ where $p_{(n)}$ are all Markov.

Note that M2 and M3 also hold for extended Markov probabilities whereas M1 might be violated.

M2 corresponds to *irrelevance* of the criteria in $\Gamma \setminus V$. This implies that these criteria can be ignored for statistical purposes and that the analysis can be performed in the marginal table corresponding to V. It is thus no severe restriction to assume $V = \Gamma$ which is henceforth done.

A famous result due to many authors: Hammersley, Clifford, Spitzer, Averintsev, Grimmett and others, asserts the equivalence between the Markov property and the existence of an expansion of the logarithm of the probability of a special structure. We shall let the functions for $a \subseteq \Gamma, b \subseteq \Gamma$

$$\phi_a(i), V_b(i)$$

denote functions that depend on i through i_a respectively i_b only. Thus $\phi_\emptyset, V_\emptyset$ are simply constants etc.

The theorem mentioned was once considered hard but now a simple proof is available which is based on the following lemma.

Lemma 1 (Möbius Inversion) *Let V and Φ be functions defined on the set of all subsets of a finite set Γ. Then the following two statements are equivalent:*

(1) $\forall a \subseteq \Gamma : V(a) = \sum_{b \subseteq a} \Phi(b)$

14

(2) $\forall a \subseteq \Gamma : \Phi(a) = \sum_{b \subseteq a} (-1)^{|a \backslash b|} V(b)$

Proof We show $(2) \Rightarrow (1)$:

$$
\begin{aligned}
\sum_{b \subseteq a} \Phi(b) &= \sum_{b \subseteq a} \sum_{c \subseteq b} (-1)^{|b \backslash c|} V(c) \\
&= \sum_{c \subseteq a} V(c) \Big(\sum_{c \subseteq b \subseteq a} (-1)^{|b \backslash c|} \Big) \\
&= \sum_{c \subseteq a} V(c) \Big(\sum_{h \subseteq a \backslash c} (-1)^{|h|} \Big).
\end{aligned}
$$

The latter sum is equal to zero unless $a \backslash c = \emptyset$ i.e. $c = a$, because any finite, non–empty set has the same number of subsets of even as of odd cardinality. The proof of $(1) \Rightarrow (2)$ is performed analogously. $\qquad \square$

We are then ready to formulate and prove the following

Theorem 1 *A probability p is Markov w.r.t. (Γ, E) if and only if there are functions ϕ_a, $a \subseteq \Gamma$ such that $\phi_a \equiv 0$ unless a is a complete subset of Γ and such that*

$$
\log p(i) = \sum_{a \subseteq \Gamma} \phi_a(i). \tag{1}
$$

Proof Suppose we have the representation (1). The condition M1 is then clearly fulfilled. Suppose now that α and β are not adjacent. This implies that no complete subset can contain both α and β. We can then write

$$
\log p(i) = \sum_{a \in A_1} \phi_a(i) + \sum_{a \in A_2} \phi_a(i)
$$

$$
= \psi_1(i) + \psi_2(i)
$$

with $A_1 = \{a \subseteq \Gamma \mid \alpha \notin a \wedge \beta \in a\}$ and $A_2 = \{a \subseteq \Gamma \mid \beta \notin a\}$ so that ψ_1 does not depend on i_α and ψ_2 does not depend on i_β. Thus

$$
p(i) = e^{\psi_1(i)} e^{\psi_2(i)}
$$

which gives us a factorisation as in CI3. Thus

$$\alpha \perp\!\!\!\perp \beta \mid \Gamma \setminus \{\alpha, \beta\}$$

i.e. we have shown that p is Markov.

Assume now p to be Markov. Choose a fixed but arbitrary cell $i^* \in \mathcal{I}$. Define

$$V_a(i) = \log p(i_a, i^*_{a^c}),$$

where $(i_a, i^*_{a^c})$ is the cell j with $j_\gamma = i_\gamma$ for $\gamma \in a$ and $j_\gamma = i^*_\gamma$ for $\gamma \notin a$. Let further

$$\phi_a(i) = \sum_{b \subseteq a} (-1)^{|a \setminus b|} V_b(i).$$

Clearly, $\phi_a(i)$ depends only on i through i_a. By the Möbius inversion lemma we have

$$\log p(i) = V_\Gamma(i) = \sum_{a \subseteq \Gamma} \phi_a(i)$$

such that we have proved the theorem if we can show that $\phi_a \equiv 0$ whenever a is not a complete subset of Γ. So let us assume that $\alpha, \beta \in a$ and $\alpha \not\sim \beta$. Let further $c = a \setminus \{\alpha, \beta\}$.

Then we can write

$$\phi_a(i) = \sum_{b \subseteq c} (-1)^{|c \setminus b|} \left[V_b(i) - V_{b \cup \{\alpha\}}(i) - V_{b \cup \{\beta\}}(i) + V_{b \cup \{\alpha, \beta\}}(i) \right].$$

By M3 and CI4 we have, if we let $d = b \cup \{\alpha, \beta\}$

$$
\begin{aligned}
V_{b \cup \{\alpha, \beta\}}(i) - V_{b \cup \{\alpha\}}(i) &= \log \frac{p(i_b, i_\alpha, i_\beta, i^*_{d^c})}{p(i_b, i_\alpha, i^*_\beta, i^*_{d^c})} \\
&= \log \frac{p(i_\alpha \mid i_b, i^*_{d^c}) p(i_\beta, i_b, i^*_{d^c})}{p(i_\alpha \mid i_b, i^*_{d^c}) p(i^*_\beta, i_b, i^*_{d^c})} \\
&= \log \frac{p(i^*_\alpha \mid i_b, i^*_{d^c}) p(i_\beta, i_b, i^*_{d^c})}{p(i^*_\alpha \mid i_b, i^*_{d^c}) p(i^*_\beta, i_b, i^*_{d^c})} \\
&= \log \frac{p(i_b, i^*_\alpha, i_\beta, i^*_{d^c})}{p(i_b, i^*_\alpha, i^*_\beta, i^*_{d^c})} = V_{b \cup \{\beta\}}(i) - V_b(i).
\end{aligned}
$$

Thus all terms in the square brackets are zero and henceforth the sum is zero. This completes the proof. $\qquad\square$

The functions $\phi_a(i)$ in the above theorem are called *interactions*. If a has only one element, ϕ_a is called a *main effect*. Since we have assumned $\Gamma = V$ all main effects are included in our representation (1). If a has m elements, we say that ϕ_a is the *interaction of $(m-1)$'st order* among the criteria in a. In terms of interactions, the above theorem says that in a Markov model, *interactions are permitted exactly among neighbours. Main effects are permitted exactly for those criteria that are not irrelevant.*

Note that the representation (1) is not at all unique without imposing further restrictions on the functions $\phi_a, a \subseteq \Gamma$. Restrictions could be of the symmetric type

$$b \subset a \Rightarrow \sum_{j:j_b = i_b} \phi_a(j) = 0 \qquad (2)$$

for all $a \neq \emptyset$ and $i_b \in \mathcal{I}_b$, or of the type given by a reference cell

$$\forall a \neq \emptyset : \phi_a(i) = 0 \quad \text{if } i_\gamma = i_\gamma^* \text{ for some } \gamma \in a \qquad (3)$$

where $i^* \in \mathcal{I}$ is fixed but arbitrary.

The interactions constructed in the proof of Theorem 1 are the unique interactions satisfying (3). For a systematic discussion of interaction representations the reader is referred to Darroch and Speed (1983).

On the other hand, if (2) or (3) do not have to be satisfied, we can construct a Markov probability by choosing *any* system of functions $\phi_a, a \neq \emptyset$ with $\phi_a \equiv 0$ if a is not complete, and then adjust $\phi_\emptyset$ such that the probability given by (1) is properly normalised.

Using the rules of conditional independence given in Section 1, one can of course derive a number of other conditional independence relations. This can, however, be done once and for all and this is what we shall do by the following:

Proposition 1 *For a probability p with $p(i) > 0$, the following statements are equivalent:*

(1) *p is Markov*

(2) *$a \perp\!\!\!\perp b \mid c$, whenever c separates a and b*

(3) *$\forall a \subseteq \Gamma : a \perp\!\!\!\perp a^c \mid \partial a$*

(4) *$\forall \alpha \in \Gamma : \alpha \perp\!\!\!\perp \alpha^c \mid \partial \alpha$*

Proof

$(1) \Rightarrow (2)$: Consider the subgraph induced by $\Gamma \setminus c$. That c separates a and b means that the vertices in $a \setminus c$ and $b \setminus c$ lie in different connected components of this graph. Let A denote the vertices in components containing vertices from $a \setminus c$. Then A, c and $\Gamma \setminus (A \cup c)$ are disjoint and a subset d of Γ can only be complete if either $d \subseteq A \cup c$ or $d \subseteq \Gamma \setminus A$. By the exponential representation in Theorem 1 we have

$$p(i) = \exp \sum_d \phi_d(i) = \exp \left(\sum_{d \subseteq A \cup c} \phi_d(i) + \sum_{d \subseteq \Gamma \setminus A, d \not\subseteq c} \phi_d(i) \right)$$

since $\phi_d \equiv 0$ unless $d \subseteq A \cup c$ or $d \subseteq \Gamma \setminus A$. By CI3 we get that $(A \cup c) \perp\!\!\!\perp (\Gamma \setminus A) \mid c$ and using CI7 twice gives $a \perp\!\!\!\perp b \mid c$.

$(2) \Rightarrow (3)$: ∂a separates a from a^c.

$(3) \Rightarrow (4)$: (4) is a special case of (3) if we let $a = \{\alpha\}$

$(4) \Rightarrow (1)$: This follows from CI6 and CI7 since

$$\alpha \perp\!\!\!\perp \alpha^c \mid \partial \alpha \Rightarrow \alpha \perp\!\!\!\perp \alpha^c \mid \Gamma \setminus \{\alpha, \beta\} \Rightarrow \alpha \perp\!\!\!\perp \beta \mid \Gamma \setminus \{\alpha, \beta\}$$

because $\beta \in \alpha^c$ and $\beta \notin \partial \alpha$ since $\alpha \not\sim \beta$.

This completes the proof. □

The property (M3) is known as the *pairwise* Markov property, (2) is the *global* Markov property and (4) is the *local* Markov property. They are not equivalent in general without the positivity assumption (M1).

The global Markov property is strongest possible in the sense that if c does not separate a and b, there is a Markov probability p such that, according to p, a and b are *not* conditionally independent given c. Such a p can easily be constructed using the exponential representation.

Thus (2) enables the researcher to read off all conditional independencies implied by the graph.

In the case where the graph is of the type

we get from Proposition 1 that $\{k\} \perp\!\!\!\perp \{1, \ldots, k-2\} \mid \{k-1\}$ for $3 \leq k \leq n$ which exactly is the usual Markov property on the line. In other words, the Markov property on a graph is an extension of the usual one.

If we consider extended Markov probabilities we have

$$p \text{ is extended Markov} \Rightarrow (M3) \wedge (2) \wedge (3) \wedge (4).$$

This can be seen by arguing that (2) is satisfied, which we shall now do: Let a, b, c be disjoint and assume that c separates a from b and that $\Gamma = a \cup b \cup c$. This can be done without loss of generality as in the proof of $(1) \Rightarrow (2)$ in Proposition 1. Since p is extended Markov we have $p = \lim p_n$ and thus if $p(i_c) > 0$ we get from CI0 that

$$p(i_a, i_b, i_c) = \lim p_n(i_a, i_b, i_c)$$

$$= \lim \frac{p_n(i_a, i_c) p_n(i_b, i_c)}{p_n(i_c)}$$

$$= \frac{p(i_a, i_c)p(i_b, i_c)}{p(i_c)}$$

whereby $a \perp\!\!\!\perp b \mid c$ as desired.

Thus the *extended* Markov probabilities have the same conditional independence interpretations as the Markov ones, but *no exponential representation* as in Theorem 1. In other words, it is not meaningful in general to discuss interactions for extended Markov probabilities.

When a and b form a decomposition of (Γ, E) the Markov property is decomposed accordingly. More precisely we have, if we let $0/0=0$

Proposition 2 *Assume that a and b decompose (Γ, E). Then $p(i)$ is (extended) Markov if and only if both $p(i_a)$ and $p(i_b)$ are (extended) Markov w.r.t. (a, E_a) and (b, E_b) respectively and*

$$p(i) = \frac{p(i_a)p(i_b)}{p(i_{a \cap b})}. \tag{4}$$

Proof   Suppose that p is Markov w.r.t. (Γ, E). Then (4) follows because $a \cap b$ separates $a \setminus b$ from $b \setminus a$. To show that $p(i_a)$ is Markov w.r.t. (a, E_a) we show that if $\alpha \not\sim \beta$ and $\alpha, \beta \in a$, then $a \setminus \{\alpha, \beta\}$ separates α from β in (Γ, E). So let α, β be nonadjacent in vertices in a. Because $a \cap b$ is complete, at least one of them, say α, is in $a \setminus b$. If there were a path from α to β avoiding $a \setminus \{\alpha, \beta\}$ it would go via $b \setminus a$ contradicting that $a \setminus b$ is separated from $b \setminus a$ by $a \cap b$.

The result for extended Markov probabilities follows by taking limits and the converse is an immediate consequence of the exponential representation (1). □

For further results about Markov random fields see e.g. Kemeny, Snell and Knapp (1976), Speed (1979) or Isham (1981).

4 Estimation in graphical models

We shall now consider the problem of finding the maximum likelihood estimate of the probability of any given cell under the assumption that this probability belongs to the class of extended Markov probabilities with respect to a given graph (Γ, E).

So let (Γ, E) be such a graph and let $\mathcal{P}$ be the set of Markov probabilities w.r.t. (Γ, E). Let $\overline{\mathcal{P}}$ denote the set of extended Markov probabilities. A model of the type $p \in \overline{\mathcal{P}}$, is called a *graphical model*. Let $\mathcal{C}$ denote the class of cliques of the graph (Γ, E). The likelihood function is proportional to

$$L(p) \propto \prod_i p(i)^{n(i)},$$

and we have to maximise this continuous function over the compact set of extended Markov probabilities. Because we are considering $\overline{\mathcal{P}}$ rather than $\mathcal{P}$, this maximum is always attained. We can further show that it is unique and that it is given by a simple set of equations that roughly corresponds to 'fitting the marginals'. Note that a probability p is an element of $\mathcal{P}$ if and only if it can be expanded as

$$p(i) = \exp \sum_{c \in \mathcal{C}} \psi_c(i_c)$$

where $\mathcal{C}$ is the cliques of the graph (Γ, E). This is easily derived from the exponential representation used in the main theorem of the previous section.

Note further that if $\hat{p}$ satisfies

$$L(\hat{p}) = \sup_{p \in \overline{\mathcal{P}}} L(p)$$

then $\hat{p}(i) = 0 \Rightarrow n(i) = 0$ i.e. $n(i) > 0 \Rightarrow \hat{p}(i) > 0$. If this was not the case, then $L(\hat{p})$ would be equal to zero.

Let $\overline{\mathcal{P}}^*$ be the set of p's in $\overline{\mathcal{P}}$ such that $n(i) > 0 \Rightarrow p(i) > 0$. Note that $\overline{\mathcal{P}}^*$ depends on the observed counts $n(i), i \in \mathcal{I}$. Define for $p \in \overline{\mathcal{P}}^*$ the operation

of 'adjusting a marginal' by

$$(T_c p)(i) = p(i) \frac{n(i_c)/n}{p(i_c)}, \quad \text{where } 0/0 = 0.$$

Note that $T_c p$ is a probability and that $(T_c p)(i_c) = n(i_c)/n$. We have then

Lemma 2 *The transformation T_c satisfies for all $c \in \mathcal{C}$*

i) *T_c is continuous on $\overline{\mathcal{P}}^*$;*

ii) *$T_c(\overline{\mathcal{P}}^*) \subseteq \overline{\mathcal{P}}^*$ for all $c \in \mathcal{C}$;*

iii) *$L(T_c p) \geq L(p)$ with equality if and only if $p(i_c) = n(i_c)/n, \forall i_c \in \mathcal{I}_c$ which happens if and only if $T_c p = p$.*

Proof   Let $\epsilon > 0$ and define $\pi_\epsilon(i) = (n(i) + \epsilon)/(n + \epsilon|\mathcal{I}|)$. Then π_ϵ is a strictly positive probability and for $p \in \mathcal{P}$

$$(T_c p)(i) = \lim_{\epsilon \to 0} p(i) \frac{\pi_\epsilon(i_c)}{p(i_c)}.$$

It then follows from the exponential representation that

$$T_c(\mathcal{P}) \subseteq \overline{\mathcal{P}}. \tag{5}$$

If $p_m \to p^*$ with $p_m, p^* \in \overline{\mathcal{P}}^*$ we have

$$\lim_{m \to \infty} (T_c p_m)(i) = \lim_{m \to \infty} p_m(i) \frac{n(i_c)/n}{p_m(i_c)} = (T_c p^*)(i)$$

since this is evident if $p^*(i_c) > 0$ and if $p^*(i_c) = 0$, we must have $n(i_c) = 0$ and the above relation is true as well. Thus i) is proved.

The continuity of T_c on $\overline{\mathcal{P}}^*$ together with (5) implies now

$$T_c(\overline{\mathcal{P}}^*) \subseteq \overline{\mathcal{P}}.$$

But in fact we must have $T_c(\overline{\mathcal{P}}^*) \subseteq \overline{\mathcal{P}}^*$ since for $p^* \in \overline{\mathcal{P}}^*$

$$(T_c p^*)(i) = p^*(i) \frac{n(i_c)/n}{p^*(i_c)}$$

and if $(T_c p^*)(i) = 0$, $p^*(i) = 0$ or $n(i_c) = 0$, both implying $n(i) = 0$ since $p^* \in \overline{\mathcal{P}}^*$ and $n(i) \leq n(i_c)$. Thus ii) is proved.

The last assertion of the lemma follows from

$$
\begin{aligned}
L(T_c p) &= L(p) \prod_{i \in \mathcal{I}} \left(\frac{n(i_c)/n}{p(i_c)} \right)^{n(i)} \\
&= L(p) \prod_{i_c \in \mathcal{I}_c} \left(\frac{n(i_c)/n}{p(i_c)} \right)^{n(i_c)}
\end{aligned}
$$

and the latter factor is ≥ 1 and $= 1$ if and only if $p(i_c) = n(i_c)/n$ by the information inequality. $\qquad\square$

We are then ready to prove the main result.

Theorem 2 *The maximum likelihood estimate $\hat{p}$ is the unique element of $\overline{\mathcal{P}}$ satisfying the system of equations*

$$
\hat{p}(i_c) = n(i_c)/n, \quad c \in \mathcal{C}, i_c \in \mathcal{I}_c.
$$

Proof The likelihood function is continuous on the compact set $\overline{\mathcal{P}}$ such that there is at least one $\hat{p} \in \overline{\mathcal{P}}$ such that

$$
L(\hat{p}) = \sup_{p \in \overline{\mathcal{P}}} L(p). \tag{6}
$$

But then $\hat{p} \in \overline{\mathcal{P}}^*$ as earlier noted and we get by Lemma 2 that

$$
L(\hat{p}) \leq L(T_c \hat{p}).
$$

Further (6) gives the reverse inequality which implies $L(\hat{p}) = L(T_c \hat{p})$. and iii) of Lemma 2 implies that $\hat{p}(i_c) = n(i_c)/n$ for all $c \in \mathcal{C}$, $i_c \in \mathcal{I}_c$.

Suppose conversely that $p^* \in \overline{\mathcal{P}}$ satisfies

$$
p^*(i_c) = n(i_c)/n \quad \text{for all } c \in \mathcal{C}, i_c \in \mathcal{I}_c.
$$

23

Then, for any $p \in \mathcal{P}$, we have

$$
\begin{aligned}
\log L(p) &= \sum_i n(i) \log p(i) \\
&= \sum_i \sum_c n(i) \psi_c(i_c) = \sum_c \sum_{i_c} n(i_c) \psi_c(i_c) \\
&= n \sum_c \sum_{i_c} p^*(i_c) \psi_c(i_c) = n \sum_i p^*(i) \log p(i),
\end{aligned}
$$

such that we by a continuity argument get that for all $p \in \overline{\mathcal{P}}$ we have

$$
L(p) = \prod_i p(i)^{np^*(i)}.
$$

But then we get by the information inequality

$$
L(p^*) = \prod_i p^*(i)^{np^*(i)} \geq \prod_i \hat{p}(i)^{np^*(i)} = L(\hat{p}).
$$

By definition of $\hat{p}$ we have then equality and therefore $p^*(i) = \hat{p}(i)$ which proves the uniqueness. $\qquad\square$

In general the maximum likelihood equations have to be solved iteratively. The iterative procedure we shall give is known as the IPS-algorithm (**I**terative **P**roportional **S**caling) and consists of successively fitting the marginals by using the operations $T_c, c \in \mathcal{C}$. More precisely, choose an ordering $c_1, \ldots, c_k$ of the cliques and let

$$
S = T_{c_k} \cdots T_{c_1}.
$$

Choose further any $p_0 \in \overline{\mathcal{P}}^*$ (e.g. $p_0(i) = |\mathcal{I}|^{-1}$) and let

$$
p_m = S^m p_0.
$$

Then we have

Theorem 3

$$
\hat{p} = \lim_{m \to \infty} p_m.
$$

Proof Let p_{m_k} be a convergent subsequence, $p_{m_k} \to p^*$. We have by the continuity of L that

$$L(p^*) \geq L(p_{m_k}) \geq L(p_0) > 0$$

such that $p^* \in \overline{\mathcal{P}}^*$. Since also S is continuous on $\overline{\mathcal{P}}^*$ by Lemma 2 we have

$$
\begin{aligned}
L(Sp^*) &= \lim_{k \to \infty} L(S(S^{m_k} p_0)) \\
&\leq \lim_{k \to \infty} L(S^{m_k+1} p_0) = \lim_{k \to \infty} L(p_{m_{k+1}}) = L(p^*).
\end{aligned}
$$

Again by Lemma 2, $L(Sp^*) \geq L(p^*)$ such that we must have equality

$$L(Sp^*) = L(T_{c_{k-1}} \cdots T_{c_1} p^*) = \cdots = L(T_{c_1} p^*) = L(p^*).$$

Reading these from right to left and using iii) of Lemma 2 gives $T_c p^* = p^*$ for all $c \in \mathcal{C}$, i.e. $p^*(i_c) = n(i_c)/n, \forall c \in \mathcal{C}, \forall i_c \in \mathcal{I}_c$. By Theorem 2 this implies $p^* = \hat{p}$. Then we have shown that any convergent subsequence of p_m converges to $\hat{p}$, which, by the compactness, implies $p_m \to \hat{p}$. $\square$

Apart from using this algorithm it is sometimes possible to obtain the maximum likelihood estimate in closed form. This is based on the fact that sometimes some marginal probabilities can be estimated from a marginal table, whereas it in general is necessary to estimate in the entire table and afterwards calculate marginal probabilities. For example we have that, if c is a clique

$$\hat{p}(i_c) = n(i_c)/n$$

and this is obviously an explicit estimate of a marginal probability based on a marginal table only. This trivially implies for a being any *complete* subset of (Γ, E)

$$\hat{p}(i_a) = n(i_a)/n.$$

But we can say much more about this. The important notion is that of a decomposition. So let a, b be subsets of Γ such that the induced subgraphs

25

(a, E_a) and (b, E_b) define a decomposition of (Γ, E). This implies that $a \cap b$ separates $a \setminus b$ from $b \setminus a$ and thus, by Proposition 1, we know that for all $p \in \overline{\mathcal{P}}$ we have

$$p(i) = \frac{p(i_a)p(i_b)}{p(i_{a \cap b})} \quad \forall p \in \overline{\mathcal{P}}$$

where $0/0 = 0$. But it also follows that the estimate decomposes in a similar fashion, as we shall now see.

If we let $\hat{p}_a(i_a)$ denote the maximum likelihood estimate of $p(i_a)$ based on the marginal counts $n(i_a)$ only and under the assumption that this marginal probability is Markov (extended) w.r.t. (a, E_a) we have

Proposition 3 *If a and b define a decomposition as above, then*

$$\hat{p}(i) = \frac{\hat{p}_a(i_a)\hat{p}_b(i_b)}{n(i_{a \cap b})/n}.$$

Proof Since a and b are separated by $a \cap b$ we have for all cliques that either $c \subseteq a$ or $c \subseteq b$. Let $\mathcal{C}_a$ be the cliques of (a, E_a) and $\mathcal{C}_b$ the cliques of (b, E_b). Then, clearly

$$\mathcal{C} \subseteq \mathcal{C}_a \cup \mathcal{C}_b.$$

The estimate $\hat{p}_a$ is given by the fact that it is (extended) Markov w.r.t. (a, E_a) and it satisfies

$$\hat{p}_a(i_c) = n(i_c)/n \quad \forall c \in \mathcal{C}_a,$$

and $\hat{p}_b$ analogously. By Proposition 1 we can let

$$p^*(i) = \frac{\hat{p}_a(i_a)\hat{p}_b(i_b)}{n(i_{a \cap b})/n}$$

and just show that p^* satisfies the likelihood equations. But

$$p^*(i_a) = \sum_{j:j_a=i_a} p^*(j) = \frac{\hat{p}_a(i_a)\hat{p}_b(i_{a \cap b})}{n(i_{a \cap b})/n} = \hat{p}_a(i_a)$$

since $a \cap b$ is complete. Thus for $c \in \mathcal{C}_a$

$$p^*(i_c) = \hat{p}_a(i_c) = n(i_c)/n$$

26

and similarly for $c \in \mathcal{C}_b$, which proves the result. $\qquad\square$

Formulating this in a slightly different way, we get:

Corollary 1 *If $\partial(a^c)$ is a complete subset then*

$$\hat{p}_a(i_a) = \hat{p}(i_a). \qquad (7)$$

Proof If we let $b = a^c \cup \partial(a^c) = \overline{a^c}$ then a and b define a decomposition exactly when $a \cap b = \partial(a^c)$ is a complete subset. Thus $\hat{p}_a(i_a) = \hat{p}(i_a)$. $\qquad\square$

The above corollary says in words that inference (or rather, *estimation*) of relations among criteria in a can be performed in the marginal table when $\partial(a^c)$ is complete. This can be a useful tool in reducing the dimensionality of any given problem.

The result in Corollary 1 can be considerably improved as pointed out to me by Søren Asmussen. In fact we have

Proposition 4 *If ∂b is complete for any connected component b of a^c, then*

$$\hat{p}_a(i_a) = \hat{p}(i_a).$$

Proof Let $a^c = b_1 \cup \cdots \cup b_k$ where b_i are the connected components of a^c. Defining $d = a \cup b_1 \cup \cdots \cup b_{k-1}$ and $e = b_k \cup \partial b_k$, d and e define a decomposition since

$$d \cap e = (a \cup b_1 \cup \cdots \cup b_{k-1}) \cap (b_k \cup \partial b_k) = a \cap \partial b_k = \partial b_k$$

which is complete by assumption. Thus

$$\hat{p}_d(i_d) = \hat{p}(i_d).$$

By repeating the argument until all b_i have been removed, we reach the conclusion. $\qquad\square$

If a fulfills the condition of Proposition 4 we say that the graphical model given by (Γ, E) is *collapsible* onto the variables in a.

A systematic investigation of this notion has been given by Asmussen and Edwards (1983), where it is also shown that the condition in Proposition 4 is necessary for (7) to hold for any set of counts.

5   Decomposable models

In the previous section we discussed the notion of a decomposition and showed that when a and b gave rise to a decomposition of (Γ, E), we could reduce the estimation problem to two marginal problems and a multiplication. Sometimes we can now decompose further such that e.g. (a, E_a) decompose into (a_1, E_{a_1}) and (a_2, E_{a_2}). If we can continue with this procedure such that finally all the marginal problems a_k are *complete* graphs, these marginal models correspond to simple estimates, since we have

$$\hat{p}_{a_k}(i_{a_k}) = n(i_{a_k})/n$$

and the entire estimate for the probability is then obtained by a multiplication of all these relative frequencies and divisions with relative frequencies corresponding to the intersections. In such cases we say that a model is *decomposable.* More formally we define

Definition 2 The graph (Γ, E) is *decomposable* if either (Γ, E) is complete or there exists $a, b \subseteq \Gamma$ with $|a| < |\Gamma|$ and $|b| < \Gamma$ such that a and b decompose (Γ, E) and such that (a, E_1) and (b, E_b) are both decomposable.

An alternative to this recursive definition is to say that there exists an ordering $c_1, \ldots, c_n$ of the cliques of the graph such that for all i, c_i and $\cup_{j<i}c_j$ decompose the subgraph induced by $\cup_{j \leq i}c_j$.

Graphs having this property have been studied by graph-theorists for many years and are known as *triangulated graphs, rigid circuit graphs, chordal graphs* and other names as well. We shall state the following result without proof but refer to e.g. Lauritzen, Speed and Vijayan (1984).

Proposition 5 *A graph is decomposable if and only if it contains no cycles of length ≥ 4 without a chord.*

Here a cycle is a sequence of vertices $\alpha_0, \alpha_1, \ldots, \alpha_n$ with $\alpha_0 = \alpha_n$ and $\alpha_i \sim \alpha_{i+1}$. Then n is the *length*. A *chord* is two vertices α_i, α_j such that $\alpha_i \sim \alpha_j$ but $j \neq i - 1, i + 1$ (modulo n).

Thus the smallest non-decomposable graph is the 4-cycle:
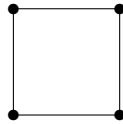
wehereas e.g. the graph
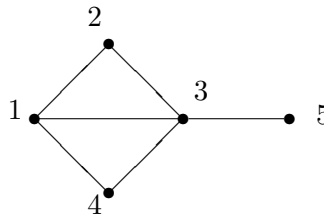
is decomposable because the 4-cycle (2,3,4,1,2) has the chord (1,3).

As earlier mentioned, the maximum likelihood estimate can then be computed by an explicit formula, multiplying together relative frequencies. We shall here just mention that for a *connected* graph this has the form:

$$\hat{p}(i) = \frac{1}{n} \prod_{d \in \mathcal{D}} n(i_d)^{\nu(d)}$$

where $\nu(d)$, $d \in \mathcal{D}$ is an index defined on $\mathcal{D}$, the set of all complete subsets of the graph. This index is related to the number of connected components of the graph when d is removed, see Lauritzen, Speed and Vijayan (1984).

If a graph is not connected, its components are independent and we get the estimate by multiplying together the estimates from different components.

Decomposability can be exploited for efficient computation of estimates but it can also be used as a basis for efficient computation of probabilities in expert systems, see Lauritzen and Spiegelhalter (1988).

6 The likelihood ratio and χ^2-tests

Consider now a subgraph $(\Gamma_0, E_0) \subseteq (\Gamma, E)$ and let $\mathcal{P} = \mathcal{P}_{(\Gamma,E)}$ and $\mathcal{P}_0 = \mathcal{P}_{(\Gamma_0,E_0)}$. Clearly we have

$$\overline{\mathcal{P}}_0 \subseteq \overline{\mathcal{P}}$$

such that we can ask for the likelihood ratio test for the hypothesis that $p \in \overline{\mathcal{P}}_0$ under the hypothesis that $p \in \overline{\mathcal{P}}$. In principle, given the estimation results, it is no problem to compute the likelihood ratio test statistic as

$$-2\log Q = 2\sum_i n(i) \log \frac{\hat{p}(i)}{\hat{p}_0(i)}$$

where $\hat{p}, \hat{p}_0$ are the maximum likelihood estimates of p under $\overline{\mathcal{P}}$ and $\overline{\mathcal{P}}_0$ respectively. This can be approximated by the Pearson χ^2-statistic

$$-2\log Q \approx \chi^2 = \sum_i \frac{(n\hat{p}(i) - n\hat{p}_0(i))^2}{n\hat{p}_0(i)}.$$

Under the assumption that $p \in \mathcal{P}_0$, both the $-2\log Q$ and χ^2-statistics have an asymptotic χ^2-distribution with degrees of freedom equal to

$$\dim(\mathcal{P}) - \dim(\mathcal{P}_0).$$

The dimensions can be obtained by the following rules

(i) If (Γ, E) is complete

$$\dim(\mathcal{P}_{(\Gamma,E)}) = \prod_{\gamma \in \Gamma} |\mathcal{I}_\gamma| - 1.$$

(ii) If $\Gamma = a \cup b$ and $E = E_a \cup E_b$ then

$$\dim(\mathcal{P}_{(\Gamma,E)}) = \dim(\mathcal{P}_{(a,E_a)}) + \dim(\mathcal{P}_{(b,E_b)}) - \dim(\mathcal{P}_{(a\cap b, E_{a\cap b})}).$$

The correctness of these rules is a consequence of the following facts. First, let for any $c \subseteq \Gamma$ the symbol L_c denote the vector space of functions of i that only depend on the c-marginal, i.e. on i through i_c only. These obviously satisfy

$$L_c \cap L_d = L_{c \cap d}$$

and

$$\dim(L_c) = \prod_{\gamma \in c} |\mathcal{I}_\gamma|.$$

Letting $\mathcal{C}$ denote the cliques of (Γ, E) we have for an arbitrary positive probability p that

$$p \in \mathcal{P}_{(\Gamma, E)} \Leftrightarrow \log p \in \sum_{c \in \mathcal{C}} L_c = H_{\mathcal{C}}$$

i.e. the probabilities in $\mathcal{P}_{(\Gamma, E)}$ can be injectively parametrised by the set of vectors $\theta \in H_{\mathcal{C}}$ satisfying

$$\sum_i e^{\theta(i)} = 1.$$

The set $\mathcal{P}_{(\Gamma, E)}$ is therefore a smooth surface of dimension $\dim(H_{\mathcal{C}}) - 1$. This reduces to (i) if $\mathcal{C}$ has only one element, i.e. if the graph is complete.

To prove the recursion formula we first note that

$$H_{\mathcal{C}} = H_{\mathcal{C}_a} + H_{\mathcal{C}_b},$$

where $\mathcal{C}_a$ and $\mathcal{C}_b$ are the cliques of (a, E_a) and (b, E_b) respectively. It remains to be shown that

$$H_{\mathcal{C}_{a \cap b}} = H_{\mathcal{C}_a} \cap H_{\mathcal{C}_b} \tag{8}$$

and the recursion becomes a consequence of the formula for dimension of the sum of vector spaces. The inclusion $\subseteq$ in (8) is trivial whereas the converse inclusion demands a bit of work. We sketch the arguments below. Let Π_c denote the orthogonal projection (usual inner product) onto L_c. By direct verification we get that these commute and that

$$\Pi_c \Pi_d = \Pi_d \Pi_c = \Pi_{c \cap d}. \tag{9}$$

The projections $\Pi_\mathcal{C}$ onto $H_\mathcal{C}$ must thus be of the form

$$\Pi_\mathcal{C} = \sum_{c \in \mathcal{C}} \sum_{b \subseteq c} \lambda_b \Pi_b \tag{10}$$

for some real constants λ_b. This can be seen by an induction argument of which we omit the details. From (9) and (10) we now obtain

$$\Pi_{\mathcal{C}_a} \Pi_{\mathcal{C}_b} = \Pi_{\mathcal{C}_b} \Pi_{\mathcal{C}_a}.$$

This implies that $\Pi_{\mathcal{C}_a} \Pi_{\mathcal{C}_b}$ is an orthogonal projection. Now (9) and (10) gives that the image of this projection must be contained in $H_{\mathcal{C}_{a \cap b}}$ whereby $\supseteq$ of (8) follows.

In practical use it is in general necessary to base the judgement of these test statistics on the above mentioned asymptotic results. For these to be usable as good approximations to exact results, it is important that cell frequencies are large. In tables of high dimension this will frequently not be the case and it is therefore important to use the results about decompositions to reduce the dimensionality of the problems.

So, suppose now that we have a decomposition (a, E_a), (b, E_b) of (Γ, E) and similarly (a_0, E_{a_0}), (b_0, E_{b_0}) of (Γ_0, E_0) such that $a_0 = a, E_{a_0} \subseteq E_a, b_0 = b, E_{b_0} \subseteq E_b$ and such that $a_0 \cap b_0 = a \cap b$ is complete in both (Γ, E) and (Γ_0, E_0). Then by the results in Section 3,

$$\hat{p}(i) = \frac{\hat{p}_a(i_a)\hat{p}_b(i_b)}{n(i_{a \cap b})/n}$$

and

$$\hat{p}_0(i) = \frac{\hat{p}_{a_0}(i_a)\hat{p}_{b_0}(i_b)}{n(i_{a \cap b})/n}$$

such that

$$-2 \log Q = (-2 \log Q_a) + (-2 \log Q_b)$$

where Q_a and Q_b are the likelihood ratio statistics for the tests of the corresponding hypotheses in the marginal tables given by a and b. Similarly we

have the corresponding *approximate* relationship for the χ^2-statistic:

$$\chi^2 \approx \chi_a^2 + \chi_b^2.$$

This gives us thus a *partitioning* of our test-statistic which has at least two advantages: we get the possibility of localising the term in the χ^2 giving rise to a bad fit and the cell frequencies in the marginal tables will be larger such that the use of asymptotic results is less likely to be dangerous.

The type of hypotheses most naturally formulated in the absence of any other particular knowledge are such where exactly one edge is removed. Thus, let $\Gamma_0 = \Gamma$ and $E_0 = E \setminus \{\alpha, \beta\}$. Using the definition of the Markov property, this is the hypothesis that α and β are conditionally independent given $\Gamma \setminus \{\alpha, \beta\}$, assuming all the conditional independencies given by (Γ, E). Using the results from Section 3, we see that if we find $a \subseteq \Gamma$ such that $\alpha, \beta \in a$ and a satisfies the condition in Proposition 4 in (Γ_0, E_0) it also does in (Γ, E) and it follows by successive decompositions that the test for the hypothesis that we can reduce from (Γ, E) to (Γ, E_0) can be carried out in the a-marginal as a test for the reduction (a, E_a) to (a, E_{a_0}). This can frequently reduce technical and conceptual problems considerably.

In particular if (Γ, E_0) has only one edge less than (Γ, E) and this edge is a member of one clique in (Γ, E) only, say c, the likelihood ratio test can be performed as a test of conditional independence in the c-marginal. This will automatically be the case if both of the graphs (Γ, E) and (Γ, E_0) are decomposable. Show this as an exercise.

7 Hierarchical models

To specify a hierarchical model, we begin by giving a set $\mathcal{C}$ of pairwise incomparable subsets of Γ, a so-called *generating class*. We then specify

$$p \in \mathcal{P}_{\mathcal{C}} \Leftrightarrow \log p(i) = \sum_{a \subseteq \Gamma} \phi_a(i)$$

where

$$\phi_a \equiv 0 \text{ unless } a \subseteq c \text{ for some } c \in \mathcal{C}.$$

The symbol $\overline{\mathcal{P}}_\mathcal{C}$ is $\mathcal{P}_\mathcal{C}$ extended with limits of such probabilities.

As we see from the definition, $\mathcal{C}$ is the set of *maximal permissible interactions* and if $\mathcal{C}$ is the cliques of a graph (Γ, E), the hierarchical model with generating class $\mathcal{C}$ is identical to the graphical model given by (Γ, E). But in general $\mathcal{C}$ does not correspond to the set of cliques of any graph.

The simplest example of a hierarchical, non-graphical model exists in 3 dimensions and is given by

$$\Gamma = \{1, 2, 3\}, \quad \mathcal{C} = \{\{1, 2\}, \{2, 3\}, \{1, 3\}\}.$$

This model is that of vanishing second order interaction in a three-way table. If it had been graphical, we see from $\mathcal{C}$ that $1 \sim 2, 2 \sim 3, 1 \sim 3$ such that its graph had to be
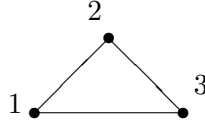
But then we should also allow interaction among $(1, 2, 3)$ since this is a complete subset of the graph.

In general, the hierarchical models are more difficult to interpret, but also for these models, graphs can be of some use. For a given generating class $\mathcal{C}$, define the graph $(\Gamma, E(\mathcal{C}))$ as

$$\{\alpha, \beta\} \in E(\mathcal{C}) \Leftrightarrow \exists c \in \mathcal{C} : \{\alpha, \beta\} \subseteq c.$$

It is not difficult to see, using the exponential representation, that this graph has the property that $\mathcal{P}_{(\Gamma, E(\mathcal{C}))}$ is the smallest graphical model containing $\mathcal{P}_\mathcal{C}$. Thus

$$p \in \mathcal{P}_\mathcal{C} \Rightarrow p \text{ is Markov w.r.t. } (\Gamma, E(\mathcal{C})).$$

Using the graph $(\Gamma, E(\mathcal{C}))$ we can again read off all the conditional in-dependencies and the interpretation of $\mathcal{P}_\mathcal{C}$ can then be formulated by conditional independence statements *combined* with a list of missing interactions relative to the graphical model.

The notion of a decomposable model was orginally defined for hierarchical models. For $\mathcal{C}_1, \mathcal{C}_2$ generating classes define

$$\begin{aligned} \mathcal{C}_1 \vee \mathcal{C}_2 &= \operatorname{red}\{\mathcal{C}_1 \cup \mathcal{C}_2\} \\ \mathcal{C}_1 \wedge \mathcal{C}_2 &= \operatorname{red}\{c_1 \cap c_2 : c_1 \in \mathcal{C}_1, c_2 \in \mathcal{C}_2\}, \end{aligned}$$

where 'red' stands for the operation of deleting the smaller of any two sets a_1, a_2 with $a_1 \subseteq a_2$, such as to make the subsets pairwise incomparable.

Then, we can define a decomposition of a generating class as $\mathcal{C}$ is *decomposed* into $\mathcal{C}_1$ and $\mathcal{C}_2$ if

$$\mathcal{C} = \mathcal{C}_1 \vee \mathcal{C}_2, \quad \mathcal{C}_1 \wedge \mathcal{C}_2 = \{c\},$$

i.e. if the minimum of $\mathcal{C}_1$ and $\mathcal{C}_2$ is a generating class with just one set. Note that $\mathcal{C}_1 \wedge \mathcal{C}_2 = \{c\}$ if and only if both conditions below are satisfied:

(i) for all $c_1 \in \mathcal{C}_1, c_2 \in \mathcal{C}_2$ we have $c_1 \cap c_2 \subseteq c$

(ii) there are $c_1^* \in \mathcal{C}_1, c_2^* \in \mathcal{C}_2$ such that $c_1^* \cap c_2^* = c$.

We proceed to define $\mathcal{C}$ to be *decomposable* if either $\mathcal{C} = \{c\}$ or $\mathcal{C} = \mathcal{C}_1 \vee \mathcal{C}_2$ with this being a decomposition and $\mathcal{C}_1$ and $\mathcal{C}_2$ decomposable. It turns out (Lauritzen, Speed and Vijayan, 1984) that

$$\mathcal{C} \text{ is decomposable} \Leftrightarrow \mathcal{C} = \text{cliques of } (\Gamma, E(\mathcal{C})) \text{ and } (\Gamma, E(\mathcal{C})) \text{ is}$$
$$\text{decomposable},$$

such that all decomposable hierarchical models are graphical as well.

It is convenient to realise that two subsets a and b of Γ define a decomposition if and only if $a \cap b = \Gamma$ and

(i) $a \cap b \subseteq c$ for some $c \in \mathcal{C}$

(ii) $a \cap b$ separates $a \setminus b$ from $b \setminus a$ in $(\Gamma, E(\mathcal{C}))$.

The induced decompositon has then

$$\mathcal{C}_a = \text{red}\{c \cap a : c \in \mathcal{C}\}, \quad \mathcal{C}_b = \text{red}\{c \cap b : c \in \mathcal{C}\}.$$

What concerns estimation in hierarchical models, most results for graphical models still hold. For example

Theorem 4 *The maximum likelihood estimate $\hat{p}(i)$ in a hierarchical model (extended) is given by the unique element $\hat{p}$ of $\overline{\mathcal{P}}_{\mathcal{C}}$ satisfying*

$$\hat{p}(i_c) = n(i_c)/n \quad \forall c \in \mathcal{C}, i_c \in \mathcal{I}_c.$$

Proof Just note that the proof given for graphical models did not use the fact that $\mathcal{C}$ was the cliques of a graph. $\square$

And further, the IPS algorithm, where $\mathcal{C} = \{c_1, \ldots, c_k\}$.

Proposition 6 *Let $p_0 \in \mathcal{P}_{\mathcal{C}}$. Define*

$$p_m = (T_{c_1} T_{c_2} \cdots T_{c_k})^m p_0.$$

Then

$$\hat{p}(i) = \lim_{m \to \infty} p_m(i).$$

Proof As for graphical models. $\square$

And the result for decompositions is obtained as follows. Let $\mathcal{C}$ be decomposed into $\mathcal{C}_1$ and $\mathcal{C}_2$ with $\mathcal{C}_1 \wedge \mathcal{C}_2 = \{c\}$. Let $C_1 = \cup_{c_1 \in \mathcal{C}_2} c_1, C_2 = \cup_{c_2 \in \mathcal{C}_2} c_2$. It follows that $C_1 \cap C_2 = c$. We then first obtain

Lemma 3 $p \in \overline{\mathcal{P}}_{\mathcal{C}}$ *if and only if $p(i_{C_1}) \in \overline{\mathcal{P}}_{\mathcal{C}_1}$ and $p(i_{C_2}) \in \overline{\mathcal{P}}_{\mathcal{C}_2}$ and*

$$p(i) = \frac{p(i_{C_1})p(i_{C_2})}{p(i_c)}.$$

Proof Since c separates C_1 from C_2 in $(\Gamma, E(\mathcal{C}))$, the conditional independence gives the factorisation for any $p \in \overline{\mathcal{P}}_\mathcal{C}$. That $p \in \mathcal{P}_\mathcal{C}$ implies $p(i_{C_1}) \in \mathcal{P}_{C_1}$ is most easily seen by using the exponential representation. It remains to take limits and to observe that the converse is trivial. $\qquad\square$

Proposition 7 *Let $\mathcal{C} = \mathcal{C}_1 \vee \mathcal{C}_2$ as above, then*

$$\hat{p}(i) = \frac{\hat{p}_{\mathcal{C}_1}(i_{C_1})\hat{p}_{\mathcal{C}_2}(i_{C_2})}{n(i_c)/n}$$

where $\hat{p}_{\mathcal{C}_1}, \hat{p}_{\mathcal{C}_2}$ are the estimates based on the marginal tables C_1, C_2 and generating classes $\mathcal{C}_1$ and $\mathcal{C}_2$.

Proof The proof is analogous to that of Proposition 3. $\qquad\square$

If we let $\dim(\mathcal{C})$ denote the dimension of the model given by $\mathcal{P}_\mathcal{C}$ we have

$$\dim(\mathcal{C}_1 \vee \mathcal{C}_2) = \dim(\mathcal{C}_1) + \dim(\mathcal{C}_2) - \dim(\mathcal{C}_1 \wedge \mathcal{C}_2),$$

which is seen as in the case of graphical models. If we combine this with the fact that

$$\dim(\{c\}) = \prod_{\gamma \in c} |\mathcal{I}_\gamma| - 1$$

we get a recursion for computing the dimension. As an example let $\Gamma = \{1, 2, 3, 4\}$ and $\mathcal{C} = \{\{1, 2\}, \{2, 3\}, \{1, 3, 4\}\}$. We get

$$
\begin{aligned}
\dim(\mathcal{C}) \;=\; & \dim(\{1, 3, 4\}) + \dim(\{1, 2\}, \{2, 3\}) - \dim(\{1\}, \{3\}) \\
=\; & |\mathcal{I}_1||\mathcal{I}_3||\mathcal{I}_4| - 1 + \dim(\{1, 2\}) + \dim(\{2, 3\}) \\
& - \dim(\{2\}) - (\dim(\{1\}) + \dim(\{3\})) \\
=\; & |\mathcal{I}_1||\mathcal{I}_3||\mathcal{I}_4| - 1 + |\mathcal{I}_1||\mathcal{I}_2| - 1 + |\mathcal{I}_2||\mathcal{I}_3| - 1 \\
& - |\mathcal{I}_2| + 1 - |\mathcal{I}_1| + 1 - |\mathcal{I}_3| + 1 \\
=\; & |\mathcal{I}_1||\mathcal{I}_3||\mathcal{I}_4| + |\mathcal{I}_1||\mathcal{I}_2| + |\mathcal{I}_2||\mathcal{I}_3| - |\mathcal{I}_1| - |\mathcal{I}_2| - |\mathcal{I}_3|.
\end{aligned}
$$

This determines in an obvious way the degrees of freedom of likelihood ratio and χ^2-tests.

The condition for collapsibility is obtained in the following way: First we form the graph $(\Gamma, E(\mathcal{C}))$. Let now $b_1, \ldots, b_k$ be the connected components of the subgraph of the above graph induced by a^c. If for all $i = 1, \ldots, k$ there is a $c_i \in \mathcal{C}$ such that

$$\partial b_i \subseteq c_i \tag{11}$$

then

$$\hat{p}(i_a) = \hat{p}_a(i_a).$$

And the condition (11) is also necessary, see Asmussen and Edwards (1983).

8 Other sampling schemes

In many applications, for example in analysis of traffic accidents and other spontaneous events, in clinical and epidemiological research, data for contingency tables are collected differently than assumed in the previous sections, where a fixed number n of objects were classified according to certain criteria.

For example, in some situations it is reasonable to assume the cell counts $n(i)$ to be independent and Poisson distributed, and in other situations data are collected in such a way that certain marginal counts are held fixed, determined by the experimental design.

We shall in the following briefly indicate how the results of the previous section can be carried over to deal with these cases.

We first consider the Poisson models, i.e. we assume that the counts $N(i)$ are independent and identically distributed with $E(N(i)) = \lambda(i)$, i.e.

$$P\{N(i) = n(i), i \in \mathcal{I}\} = \prod_{i \in \mathcal{I}} \frac{\lambda(i)^{n(i)}}{n(i)!} e^{-\lambda(i)}$$

where $\lambda(i) \geq 0$. In analogy with the previous section we define the *hierarchical model* with generating class $\mathcal{C}$ to be determined by the set of λ's with

$\lambda(i) > 0$ and such that

$$\log \lambda(i) = \sum_{a \subseteq \Gamma} \phi_a(i)$$

with $\phi_a \equiv 0$ unless $a \subseteq c$ for some $c \in \mathcal{C}$. The *extended hierarchical models* are obtained by taking weak limits. We denote the set of admissible λ's by $\Lambda_{\mathcal{C}}$ and $\overline{\Lambda}_{\mathcal{C}}$ respectively.

The likelihood function becomes

$$L(\lambda) \propto \prod_i \lambda(i)^{n(i)} e^{-\lambda(i)}$$

and if we let

$$\lambda_{\cdot} = \sum_{i \in \mathcal{I}} \lambda(i) = \lambda(i_\emptyset)$$

we get, (assuming that $\lambda_{\cdot} \neq 0$ for $\lambda \in \overline{\Lambda}_{\mathcal{C}}$),

$$L \propto \lambda_{\cdot}^n e^{-\lambda_{\cdot}} \prod_{i \in \mathcal{I}} (\frac{\lambda(i)}{\lambda_{\cdot}})^{n(i)}. \tag{12}$$

Letting $p(i) = \lambda(i)/\lambda_{\cdot}$ we find that

$$\lambda \in \overline{\Lambda}_{\mathcal{C}} \Leftrightarrow \lambda_{\cdot} > 0 \quad \text{and } p \in \overline{\mathcal{P}}_{\mathcal{C}}.$$

This means that the likelihood function (12) can be maximized by maximizing each factor separately such that the unique maximum likelihood estimate is given by the system of equations

$$\lambda_{\cdot} = n, \qquad p(i_c) = n(i_c)/n, \quad i_c \in \mathcal{I}_c, c \in \mathcal{C} \tag{13}$$

which is clearly equivalent to the system of equations

$$\lambda(i_c) = n(i_c), \quad i_c \in \mathcal{I}_c, c \in \mathcal{C}. \tag{14}$$

This system of equations has as before a unique solution if just $n > 0$ and the solution can be found by, for example, the IPS-algorithm.

But the equivalence of the equation systems (13) and (14) can be used in the other direction. The hierarchical log-linear Poisson model given by $\Lambda_{\mathcal{C}}$ is an example of a so-called generalised linear model see McCullagh and Nelder (1983) with Poisson error and log as link function. It follows that the maximum likelihood estimates can be calculated using the program GLIM (or GENSTAT).

This can then be exploited for estimation in the models with n fixed, i.e. those with multinomial sampling. This can sometimes, but not always, be advantageous.

Consider next the sampling situation when the experiment by design collects data in such a way that the numbers $(n(i_b), i_b \in \mathcal{I}_b)$ are fixed for a particular set of criteria $b \subseteq \Gamma$.

Then the sampling distribution is appropriately described by a *product of multinomial* distributions, i.e.

$$P\{N(i) = n(i), i \in \mathcal{I}\} = \prod_{i_b \in \mathcal{I}_b} \frac{n(i_b)!}{\prod_{j:j_b=i_b} n(j)!} \prod_{j:j_b=i_b} \left(\frac{p(j)}{p(i_b)} \right)^{n(j)}. \tag{15}$$

Let now $\eta(i) = p(i)n(i_b)/(p(i_b)n)$. The likelihood function then becomes

$$L(p) \propto \prod_{i \in \mathcal{I}} \eta(i)^{n(i)} \tag{16}$$

and η satisfies

$$\eta(i) \geq 0, \quad \sum_{j:j_b=i_b} \eta(j) = \eta(i_b) = \frac{n(i_b)}{n}$$

whereby we get $\sum_i \eta(i) = 1$.

Suppose now that we assume that $p \in \overline{\mathcal{P}}_{\mathcal{C}}$ with $b \subseteq c_0$ for some $c_0 \in \mathcal{C}$. We then have that the maximum likelihood estimate for the conditional probability based upon the model (15) and thus the likelihood (16), is given by

$$\hat{p}(i_{b^c} \,|\, i_b) = p^*(i)/p^*(i_b),$$

where p^* is the unique element of $\overline{\mathcal{P}}_\mathcal{C}$ that solves the equations

$$p^*(i_c) = n(i_c)/n, \quad i_c \in \mathcal{I}_c, c \in \mathcal{C}.$$

To see this we proceed as follows. If $\eta > 0$ we have that

$$\log \eta(i) = \log p(i) + \log n(i_b) - \log p(i_b) - \log n$$

and, because $b \subseteq c_0$, that

$$\log \eta \in H_\mathcal{C} \Leftrightarrow \log p \in H_\mathcal{C}.$$

Repeating the argument in the proof of Theorem 2 we thus obtain that the value of η maximizing the likelihood is given as

$$\hat{\eta}(i_c) = n(i_c)/n, \quad i_c \in \mathcal{I}_c, c \in \mathcal{C},$$

and, since $b \subseteq c_0$ for some $c_0 \in \mathcal{C}$, this automatically satisfies the restriction

$$\hat{\eta}(i_b) = n(i_b)/n, \quad i_b \in \mathcal{I}_b.$$

Thus we may let

$$\hat{p}(i_{b^c} \,|\, i_b) = \hat{\eta}(i)n/n(i_b) = \hat{\eta}(i)/\hat{\eta}(i_b) = p^*(i)/p^*(i_b).$$

We warn the reader that the similar result is false in general if we fix more than one marginal or if the fixed marginal b is *not* contained in a generating set $c_0 \in \mathcal{C}$.

Finally we mention that if we consider the likelihood ratio statistic for testing one model with generator $\mathcal{C}_0 \subseteq \mathcal{C}$ assuming the model with generator $\mathcal{C}$ in *all three sampling situations* is equal to

$$-2\log Q = 2\sum_{i \in \mathcal{I}} n(i) \log \frac{\hat{\lambda}(i)}{\hat{\lambda}_0(i)},$$

where $\hat{\lambda}$ and $\hat{\lambda}_0$ are the estimates under the Poisson model. Thus

$$\frac{\hat{\lambda}(i)}{\hat{\lambda}_0(i)} = \frac{\hat{\lambda}(i)/n}{\hat{\lambda}_0(i)/n} = \frac{\hat{p}(i)}{\hat{p}_0(i)} = \frac{\hat{p}(i)/n(i_b)}{\hat{p}_0(i)/n(i_b)} = \frac{\hat{p}(i)/\hat{p}(i_b)}{\hat{p}_0(i)/\hat{p}(i_b)},$$

41

provided, as always, that $b \subseteq c_0$ for some $c_0 \in \mathcal{C}_0$.

Although the exact distribution of the likelihood ratio statistic is different in the three cases, their *asymptotic distribution is the same* χ^2-distribution with degrees of freedom equal to $\dim(\mathcal{C}) - \dim(\mathcal{C}_0)$. This is due to the fact that although the dimensions of the models are different, the *difference* of the dimensions remain unchanged by the conditioning.

9   Miscellaneous

Apart from the hierarchical models it is sometimes convenient to consider the larger class of *general interaction models.* These are obtained by expanding the logarithm to the probability $p(i)$ as

$$\log p(i) = \sum_{a \subseteq \Gamma} \phi_a(i)$$

and then specify a list of subsets $\mathcal{A} = \{a_1, ..., a_k\}$ such that

$$\phi_a \equiv 0 \quad \text{if } a \notin \mathcal{A}$$

and ϕ_a is arbitrary otherwise. This demands a choice of how to make the representation unique and the model will depend effectively on that choice, cf. (2) and (3) in Section 3.

We shall not discuss inference in these models in detail but just refer to the general theory of exponential families, see e.g. Barndorff-Nielsen (1978), Johansen (1979) or Brown (1987).

To each general interaction model there is a smallest hierarchical model containing it and then a smallest graphical model containing the hierarchical one.

As it seemed convenient to interpret a hierarchical model by referring to those interactions missing to make it graphical and listing the graph given the conditional independencies, it is probably useful to interpret a general interaction model by referring to the smallest hierarchical extension.
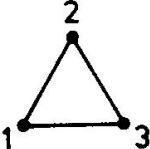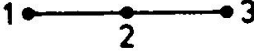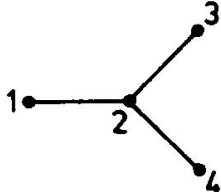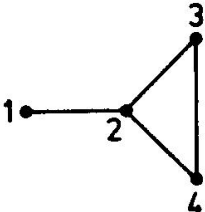
The various types of models: decomposable, graphical, hierarchical and general, represent increasing levels of complexity both in terms of interpretation and inference procedures. The notion of a decomposition is probably more important than that of a decomposable model, because it allows the possibility of collapsing tables of high dimension thus reducing inference problems, sometimes even quite drastically.

As mentioned in the introduction, the models described here are not very refined and probably most useful in a preliminary statistical analysis, with the purpose of discovering very basic structures in a given data set.

It seems therefore recommendable to start a model search procedure by fitting a graphical model, which then finds the interesting conditional independencies. Depending on the amount of knowledge one could then formulate relevant hierarchical hypotheses of vanishing interactions or even non-hierarchical models. Such an approach ensures that the final model has a reasonable interpretation which is far from unimportant.

On the following pages we have reproduced a list of all graphical models of dimension ≤ 5 from Darroch, Lauritzen and Speed (1980).

TABLE 1
Decomposable models of dimension less than or equal to five.

graph	interpretation
	unrestricted
	unrestricted
	unrestricted
	[1 ⊗ 3|2]
	[1 ⊗ 3, 4|2] ∩ [1, 2 ⊗ 4|3]
	[1 ⊗ 3 ⊗ 4|2]
	[1 ⊗ 3, 4|2]

44
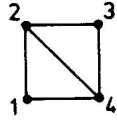
$[1 \otimes 3 | 2, 4]$
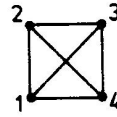
unrestricted
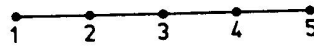
$[1 \otimes 3, 4, 5 | 2], \text{ etc.}$
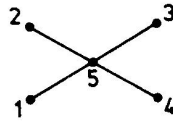
$[1 \otimes 2 \otimes 3 \otimes 4 | 5]$
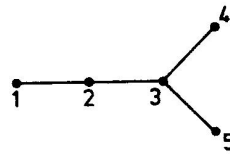
$[1 \otimes 3, 4, 5 | 2] \cap [1, 2 \otimes 4 \otimes 5 | 3]$
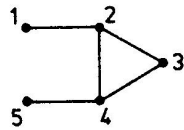
$[1 \otimes 5 \otimes 3 | 2, 4] \cap [1 \otimes 3, 4, 5 | 2]$
$\cap [5 \otimes 1, 2, 3 | 4]$
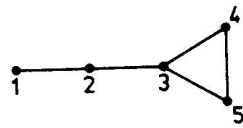
$[1, 2 \otimes 4, 5 | 3] \cap [1 \otimes 3, 4, 5 | 2]$
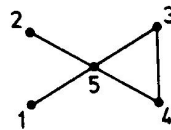
$[1 \otimes 2 \otimes 3, 4 | 5]$
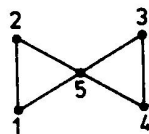
$[1, 2 \otimes 3, 4 | 5]$
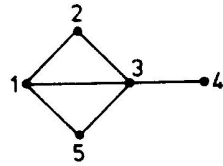
$[2 \otimes 5 \otimes 4 | 1, 3] \cap [1, 2, 5 \otimes 4 | 3]$
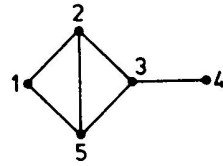
$[1, 2, 5 \otimes 4 | 3] \cap [1 \otimes 3, 4 | 2, 5]$
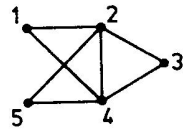
$[1 \otimes 3 \otimes 5 | 2, 4]$
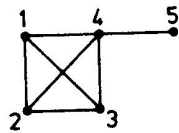
$[1, 2, 3 \otimes 5 | 4]$
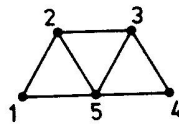
$[1 \otimes 3, 4 | 2, 5] \cap [1, 2 \otimes 4 | 3, 5]$
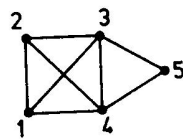
$[1, 2 \otimes 5 | 3, 4]$
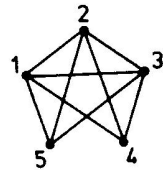
$[4 \otimes 5 | 1, 2, 3]$
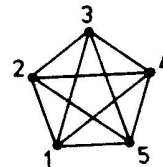
unrestricted.

TABLE 2

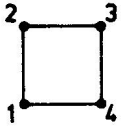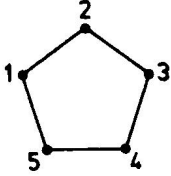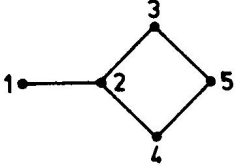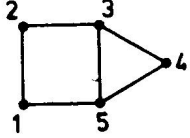*Nondecomposable models that are graphical of dimension less than or equal to five.*
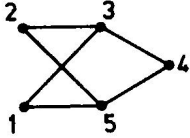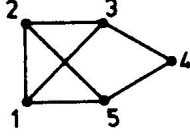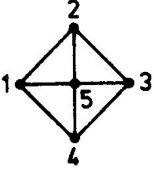
graph	> 4-cycle	interpretation
	{1, 2, 3, 4}	[1 ⊗ 3|2, 4] ∩ [2 ⊗ 4|1, 3]
	{1, 2, 3, 4, 5}	[1, 2 ⊗ 4|3, 5], etc.
	{2, 3, 4, 5}	[1, 2 ⊗ 5|3, 4] ∩ [1 ⊗ 3, 4, 5|2] ∩[3 ⊗ 1 ⊗ 4|2, 5]
	{1, 2, 3, 5}	[1, 2 ⊗ 4|3, 5] ∩ [1 ⊗ 3, 4|2, 5] ∩[2 ⊗ 4, 5|1, 3]
	{1, 3, 4, 5} {2, 3, 4, 5} and {1, 2, 3, 5}	[1 ⊗ 2 ⊗ 4|3, 5] ∩[3 ⊗ 5|1, 2, 4]
	{1, 3, 4, 5} and {2, 3, 4, 5}	[1, 2 ⊗ 4|3, 5] ∩[3 ⊗ 5|1, 2, 4]
	{1, 2, 3, 4}	[1 ⊗ 3|2, 4, 5] ∩[2 ⊗ 4|1, 3, 5].

47

Table 3: Number of models of given type

			Dimension		
Type	1	2	3	4	5
Interaction	2	8	128	32,768	2,147,483,648
Hierarchical	2	5	19	167	7,580
Graphical	2	5	18	113	1,450
Decomposable	2	5	18	110	1,233

To illustrate the complexity of the various types of models, we also do the combinatorial exercise of computing the number of models of a given type for a given dimension.

The number of general interaction models of dimension $|\Gamma| = n$ is equal to $2^{2^n - 1}$ since there are 2^n subsets of Γ. For all non-empty subsets a we can choose ϕ_a to vanish or not.

The number of graphical models is given as

$$\sum_{i=0}^{n} \binom{n}{i} 2^{\binom{i}{2}}$$

since we can choose i criteria to be relevant in $\binom{n}{i}$ ways $i = 0, \ldots, n$ and then choose each of the $\binom{i}{2}$ edges to be present or not.

The number of decomposable and hierarchical models cannot be determined by a closed expression, but Table 3 gives all the numbers for $n = 2, 3, 4, 5$.

The number 2^{31} should scare anybody from trying to fit "the right" general interaction model for a 5-dimensional contingency table without specific knowledge of the relevance of such a model. Still, it is an interesting problem is to design procedures for model selection within these classes. This has, for example, been adressed by Wermuth (1976b), Edwards and Havránek (1985, 1987) and Whittaker (1984).

48

10   Analogies to covariance selection models

Models analogous to the graphical models exist also for the multivariate normal distribution. These are the so-called *covariance selection* models introduced by Dempster (1972).

Here $X = X_\Gamma$ is a vector of real-valued random variables indexed by Γ. We assume that

$$X \sim N(0, \Sigma), \quad \text{with } \Sigma \in \mathcal{S}_{(\Gamma, E)}$$

where $\mathcal{S}_{(\Gamma, E)}$ is the set of positive definite symmetric matrices such that

$$\sigma^{\alpha\beta} = 0 \quad \text{whenever } \alpha \not\sim \beta.$$

Here

$$\Sigma = \{\sigma_{\alpha\beta}\}, \quad \Sigma^{-1} = \{\sigma^{\alpha\beta}\}.$$

Since we have for a normal distribution that

$$\sigma^{\alpha\beta} = 0 \Leftrightarrow X_\alpha \perp\!\!\!\perp X_\beta \mid X_{\Gamma \setminus \{\alpha, \beta\}}$$

we see that this model is given by exactly the same type of conditional independencies as our graphical models.

Further, since the multivariate distribution of a normal vector is completely specified by its 'first-order interactions' $\sigma^{\alpha\beta}$, there exist no 'hierarchical' models that are not graphical for this type.

If we have independent observations $X^{(1)}, \ldots, X^{(n)}$, and form the empirical covariance matrix S, we further have that the maximum likelihood estimate of Σ is given by the equations:

$$\Sigma \in \mathcal{S}_{(\Gamma, E)}, \qquad \Sigma_c = S_c \quad \forall c \in \mathcal{C},$$

where $\mathcal{C}$ is the cliques of (Γ, E) and Σ_c, S_c are the marginals corresponding to $(X_\gamma, \gamma \in c)$. So for these models, the question of estimation is the question of 'fitting the marginals'. An IPS-algorithm exists cf. Speed and Kiiveri (1986) etc.

If the graph is decomposable, one can show that we have

$$\hat{\Sigma}^{-1} = \sum_d \nu(d)[S_d^{-1}]^0$$

and that

$$\det \Sigma = \prod_d (\det \Sigma_d)^{\nu(d)} \quad \forall \Sigma \in \mathcal{S}_{(\Gamma, E)},$$

where $[S_d^{-1}]^0$ is the matrix obtained by inverting S_d, filling in the corresponding elements and letting other elements be equal to zero and $\nu(d)$ is the index mentioned in Section 5. A more detailed discussion of this can be found in Wermuth (1976a).

Thus we see that the analogy to these models is much more direct than that to models for the analysis of variance, where we make an expansion of the *expectation* of a normal variate into its 'interactions'.

Recently, the theory of contingency tables and covariance selection models have been unified and extended to graphical association models for *mixed* quantitative and categorical data. The basic theory for graphical models is developed in Lauritzen and Wermuth (1989) and extended to hierarchical models by Edwards (1990).

The models have also been extended to deal with situations where some variables are responses and some explanatory. Wermuth and Lauritzen (1983, 1990) discuss for example this aspect.

A comprehensive exposition of these developments as well as further references can be found in Lauritzen (1996).

References

Andersen, A. H. (1974). Multidimensional contingency tables. *Scand. J. Statist.* **1**, 115–127

Asmussen, S. and Edwards, D. (1983). Collapsibility and response variables in contingency tables. *Biometrika* **70**, 567–578

Bishop, Y. M. M., Fienberg, S.E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* MIT Press, Cambridge, Mass.

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory.* Wiley, New York.

Brown, L. D. (1987). *Fundamentals of Exponential Families with Applications in Decision Theory.* IMS Monographs, Vol. IX, California.

Darroch, J. N., Lauritzen S. L. and Speed, T. P. (1980). Markov-fields and log-linear models for contingency tables. *Ann. Statist.* **8**, 522–539

Darroch, J. N. and Speed, T. P. (1983). Additive and multiplicative models and interactions. *Ann. Statist.* **11**, 724–738

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc. Ser. B* **41**, 1–31

Dawid, A. P. (1980). Conditional independence for statistical operations. *Ann. Statist.* **8**, 598–617

Dempster, A. P. (1972). Covariance selection. *Biometrics* **28**, 157–175

Edwards, D. (1990). Hierarchical interaction models (with discussion). *J. Roy. Statist. Soc. Ser. B* **52**, 3–20 and 51–72

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. *J. Amer. Statist. Assoc.* **82**, 205–211

Haberman, S. J. (1974). *The Analysis of Frequency Data.* Univ. of Chicago Press.

Isham, V. (1981). An introduction to spatial point processes and Markov random fields. *Int. Statist. Review* **49**, 21–43

Jensen, S. T. (1978). *Flersidede kontingenstabeller.* (Danish). Univ. Cop. Inst. Math. Stat.

Johansen S. (1979). *An Introduction to the Theory of Regular Exponential Families.* Lecture Notes 3, Univ. Cop. Inst. Math. Stat.

Kemeny, J. G., Snell, J. L., Knapp, A. W. and Griffeath, D. (1976). *Denumerable Markov Chains. 2nd ed.* Springer, Heidelberg.

Lauritzen, S. L. (1996). *Graphical Models.* Clarendon Press, Oxford.

Lauritzen, S. L., Speed, T. P. and Vijayan, K. (1984). Decomposable graphs and hypergraphs. *J. Austral. Math. Soc. A* **36**, 12–29

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **50**, 157–224

Lauritzen, S. L. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.* **17**, 31–57

Pearl, J. (1988). *Probabilistic Inference in Intelligent Systems.* Morgan Kaufmann, San Mateo.

Speed, T. P. (1978). Graph-theoretic methods in the analysis of interactions. Mimeographed lecture notes. Univ. Cop. Inst. Math. Stat.

Speed, T. P. (1979). A note on nearest-neighbour Gibbs and Markov probabilities. *Sankhya A* **41**, 184–197

Speed, T. P. and Kiiveri, H. (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist.* **14**, 138–150

Wermuth, N. (1976a). Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics* **32**, 95–108

Wermuth, N. (1976b). Model search among multiplicative models. *Biometrics* **32**, 253–263

Wermuth, N. and Lauritzen, S. L. (1983). Graphical and recursive models for contingency tables. *Biometrika* **70**, 537–552

Wermuth, N. and Lauritzen, S. L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion). *J. Roy. Statist. Soc. Ser. B* **52**, 21–72

Whittaker, J. (1984). Fitting all possible decomposable and graphical models to multiway contingency tables. In Havránek, T. (ed.). COMPSTAT 84, 98–108. Physica Verlag, Vienna.