AALBORG UNIVERSITY

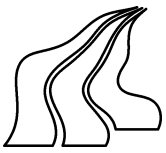
Decomposable log-linear models

by

P. Svante Eriksen

R-2005-16 Maj 2005

Department of Mathematical Sciences
Aalborg University
Fredrik Bajers Vej 7 G ▪ DK - 9220 Aalborg Øst ▪ Denmark
Phone: +45 96 35 80 80 ▪ Telefax: +45 98 15 81 29
URL: http://www.math.aau.dk



ISSN 1399–2503 ▪ On-line version ISSN 1601–7811

Decomposable log-linear models.

P. Svante Eriksen
Department of Mathematical Sciences
Aalborg University, Denmark

2005-05-06

Abstract

The present paper considers discrete probability models with exact computational properties. In relation to contingency tables this means closed form expressions of the maksimum likelihood estimate and its distribution. The model class includes what is known as decomposable graphical models, which can be characterized by a structured set of conditional independencies between some variables given some other variables.

We term the new model class decomposable log-linear models, which is illustrated to be a much richer class than decomposable graphical models. It covers a wide range of non-hierarchical models, models with structural zeroes, models described by quasi independence and models for level merging. Also, they have a very natural interpretation as they may be formulated by a structured set of conditional independencies between two events given some other event. In relation to contingency tables we term such independencies as context specific independencies.

Key words: decomposable model, log-linear model, exact inference, context specific independence, quasi independence, level merging, incomplete contingency table.

1 Introduction

The central issue of this paper is to study discrete probability models with exact computational properties. We explore the structure of what we will call decomposable log-linear models and describe how exact inference can be performed.

Decomposable graphical models for contingency tables is a well studied class of models and was originally introduced by Haberman (1970) in connexion with hierarchical log-linear models. The interpretation in terms of Markov properties attached to undirected graphs is due to Darroch *et al.* (1980), whereas exact results on distributional properties of estimators were given by Sundberg (1975). A rigourous and comprehensive treatment of the subject may be found in Lauritzen (1996).

In recent years there has been a growing interest in non-hierarchical modelling for contingency tables, where we allow that interactions may vanish in certain contexts. A subclass of these models focus on Markov properties and are often called context specific independence models. Some recent papers are Wermuth and Cox (1998), Teugels and Van Horebeek (1998), Højsgaard (2004) and Corander (2003), where the latter uses a labelled graph to display context specific independence. Also within the computer science community, the ideas have gathered substantial interest, e.g. illustrated by Poole and Zhang (2003) and Rosen *et al.* (2004).

Another area of relevance to context specific independence in contingency tables is level merging, which have recently been considered in Dellaportas and Forster (1999).

Finally, we mention the analysis of contingency tables with structural zeroes, also known as incomplete contingency tables. The socalled block-stairway tables introduced by Bishop *et al.* (1975) is an example of a decomposable log-linear model. The structure is intimately related to models for quasi independence, when quasi independence actually can be interpreted in terms of conditional independencies of events. A recent work on such structures is given by Rapallo (2003), where the algebraic structure is used for developing simulation algorithms.

Most applications of decomposable log-linear models are related to contingency tables, but the formulation is free of "coordinates", as we do not explicitly connect the outcome structure to some given factors.

The framework is intimately related to the concept of conditional independence - but not in terms of some predefined factors. Rather, the models can be fully specified by a set of conditional independence statements between two events given another event.

2 Preliminaries and notation

We start by establishing some notation and definitions of fundamental concepts. In order to motivate these we give a simple but hopefully illustrative example.

Example 2.1. Consider 3 binary variables A, B, C taking values $\{-, +\}$. Define

- $v^* = \{V = *\}$ when $V \in \{A, B, C\}$ and $* \in \{-, +\}$.
- $V = \{v^-, v^+\}$ when $V \in \{A, B, C\}$, i.e. we allow that the variable is identified with its corresponding events.

We study the model where B and C are independent given $\{A = +\} = a^+$. If $a \wedge b$ denotes intersection of events a and b then we may state the assumption as

$$P(a^+ \wedge b^* \wedge c^*) = \frac{P(a^+ \wedge b^*)P(a^+ \wedge c^*)}{P(a^+)} \qquad b^* \in B,\ c^* \in C$$

with the convention that $\frac{0 \cdot 0}{0} = 0$. The crucial point to note is the factorization in terms of probabilities of events, and we want to explore the structure of these events.

Define

- $a^+V = \{a^+ \wedge v | v \in V\}, \quad V \in \{B, C\}$.
- $a^*BC = \{a^* \wedge b^* \wedge c^* | b^* \in B,\ c^* \in C\},\ \ a^* \in A$.
- $ABC = a^-BC \cup a^+BC$.

Then the outcome space ABC is the union of mutually disjoint events in a^-BC and a^+BC. Furthermore, a^+B and a^+C both form a decomposition of a^+ into disjoint subevents, and

$$a^+BC = \{b \wedge c | b \in a^+B,\ c \in a^+C\}$$

We may say that a^+B, a^+C form a cartesian product that spans a^+BC.

■

The crucial points to emphasize is the decomposition of ABC into a^-BC and a^+BC and the representation of a^+BC as a cartesian product of a^+B and a^+C. This should motivate the following definitions.

Definition 2.1.
Subsequently, we consider a finite set Ω, which we denote the **outcome space**. Any subset of Ω is called an **event**. When a and b are events in Ω we use $a \wedge b$ to denote their intersection.

A non empty set $\mathcal{A}$ of events is defined to be **a paving on** Ω.

If $\mathcal{A}$ and $\mathcal{B}$ are pavings on Ω, then we define the **wedge product of pavings** $\mathcal{A} \wedge \mathcal{B}$ to be all non empty events of the form $a \wedge b$, where $a \in \mathcal{A}$ and $b \in \mathcal{B}$.

If $\mathcal{A}$ and $\mathcal{B}$ are pavings on Ω such that $\mathcal{A} \wedge \mathcal{B} = \emptyset$, then we say that $\mathcal{A} \cup \mathcal{B}$ is **a direct sum of pavings** and we use the notation $\mathcal{A} \oplus \mathcal{B}$ to denote the union of pavings with empty wedge product.

A **factor** f on Ω is a set of disjoint events in Ω.

When f is a factor, we use $\mathrm{s}(f)$ to denote the union of the events in f.

Two factors f and g on Ω are **cartesian factors** if

$$a \wedge b \neq \emptyset, \quad a \in f, \quad b \in g$$

and we let $f \otimes g$ denote the wedge product of cartesian factors. □

Example 2.2. Turning back to example 2.1 we consider the factors

- $f_0 = a^-BC$, $f_1 = a^+B$ and $f_2 = a^+C$.

and we may represent the outcome space $\Omega = ABC$ as

$$\Omega = f_0 \oplus (f_1 \otimes f_2)$$

Next let $a \subseteq \Omega$ and define

$$P(a)^\omega = \begin{cases} P(a) & \text{if } \omega \in a \\ 1 & \text{otherwise} \end{cases}$$

If we define the pavings $\mathcal{A} = f_0 \cup f_1 \cup f_2$ and $\mathcal{S} = \{a^+\}$ then it is easily seen that we may represent the distribution of (A, B, C) as

$$P(\{\omega\}) = \frac{\prod_{a \in \mathcal{A}} P(a)^\omega}{\prod_{s \in \mathcal{S}} P(s)^\omega} \quad \omega \in \Omega$$

We call a^+ a separating event and note that it is attached to the cartesian product $f_1 \otimes f_2$ and fulfills

$$P(a^+) = P(\mathrm{s}(f_1)) = \sum_{a \in f_1} P(a) = \sum_{a \in f_2} P(a) = P(\mathrm{s}(f_2))$$

i.e. the probability of the separator is easily derived from events in $\mathcal{A}$. ■

Example 2.3. Let us consider a slightly more complicated example with 4 binary variables A, B, C, D where we suppose

- D is independent of C given (A, B) and D is independent of B given a^+.
- C is independent of B given a^+ and C is independent of A given b^+.

Define events and factors

- $e^- = a^- \wedge b^-$ and e^+ as the complementary event to e^-.
- $f_1 = e^- C$ and $f_2 = e^- D$ are cartesian with separator e^-.
- $f_3 = a^+ B$ and $f_4 = a^+ D$ are cartesian with separator a^+.
- If $f_5 = a^- b^+ D$ and $f_6 = e^+ C$ then $(f_3 \otimes f_4) \oplus f_5$ and f_6 are cartesian with separator e^+.

Also we may represent $\Omega = ABCD$ as

$$\Omega = (f_1 \otimes f_2) \oplus (((f_3 \otimes f_4) \oplus f_5) \otimes f_6)$$

Furthermore, if $\mathcal{A}$ is the union of $f_1, \ldots, f_6$ and $\mathcal{S} = \{a^+, e^-, e^+\}$ then one may verify that

$$P(\{\omega\}) = \frac{\prod_{a \in \mathcal{A}} P(a)^\omega}{\prod_{s \in \mathcal{S}} P(s)^\omega} \quad \omega \in \Omega$$

E.g. when $A = +$ this reads

$$P(a^+b^*c^*d^*) = \frac{P(a^+b^*)P(a^+d^*)P(e^+c^*)}{P(a^+)P(e^+)}$$

■

The crucial point to note is that the complete distribution is easily recovered from knowing the probabilities of the events in $\mathcal{A}$. Evidently, the distribution can be characterized by allowing a factorization in terms of numbers attached to events in $\mathcal{A}$. So we aim at studying distributions with factorization properties.

2.1 Factorization of distributions

In order to describe the general structure we introduce some more notation and concepts.

Definition 2.2.
Let $\mathcal{A}$ be a paving on Ω.

By $\sigma(\mathcal{A})$ we denote the smallest σ-algebra containing $\mathcal{A}$. Since $\mathcal{A}$ is finite, then $\sigma(\mathcal{A})$ is actually an algebra, which we call **the algebra of** $\mathcal{A}$.

By the **atoms of** $\mathcal{A}$ denoted $\alpha(\mathcal{A})$, we mean the minimal elements of $\sigma(\mathcal{A}) \setminus \{\emptyset\}$, when we consider the partial order given by set inclusion. □

The object of study is distributions - or equivalently - probability measures on $\sigma(\mathcal{A})$. We note that a distribution P can be characterized by the numbers $P(c),\ c \in \alpha(\mathcal{A})$, i.e. the probabilities on atoms. We continue to tell what we mean by factorization.

Definition 2.3.
Let $\mathcal{A}$ be a paving on Ω.

If a is an event in $\mathcal{A} \cup \{\Omega\}$ and λ_a is a non negative number we define the quantity λ_a^c, $c \in \alpha(\mathcal{A})$ by

$$\lambda_a^c = \begin{cases} \lambda_a & \text{if } c \subseteq a \\ 1 & \text{otherwise} \end{cases}$$

i.e. we assign the number λ_a to the atoms of the event a.

A distribution P on $\sigma(\mathcal{A})$ is said to **factorize w.r.t.** $\mathcal{A}$ if we can find positive numbers $\{\lambda_a | a \in \mathcal{A} \cup \{\Omega\}\}$ such that

$$P(c) = \prod_{a \in \mathcal{A} \cup \{\Omega\}} \lambda_a^c, \quad c \in \alpha(\mathcal{A})$$

$M(\mathcal{A})$ denotes the set of distributions that factorize w.r.t. $\mathcal{A}$. We call $M(\mathcal{A})$ a **log linear model** on Ω since

$$\log(P(c)) = \sum_{a \in \mathcal{A} \cup \{\Omega\}} \beta_a 1_a(c) \quad c \in \alpha(\mathcal{A})$$

where $\beta_a = \log(\lambda_a)$ and $1_a(c)$ is the function indicating whether or not c is a subset of the event a. So we can think of $\log(P)$ as a linear combination of indicator functions on $\alpha(\mathcal{A})$.

We let $L(\mathcal{A})$ denote the vector space spanned by the above indicator functions, i.e.

$$L(\mathcal{A}) = \text{span}\{1_c | c \in \mathcal{A} \cup \{\Omega\}\}$$

By $\Sigma(\mathcal{A})$ we denote all events, which has an indicator function belonging to $L(\mathcal{A})$, i.e.

$$\Sigma(\mathcal{A}) = \{c \in \sigma(\mathcal{A}) | 1_c \in L(\mathcal{A})\}$$

By the **generator corresponding to** $\mathcal{A}$ denoted $\gamma(\mathcal{A})$, we mean the minimal elements of $\Sigma(\mathcal{A}) \setminus \{\emptyset\}$, when we consider the partial order given by set-inclusion. □

Intuitively, the following proposition should be no surprise. The proof is deferred to appendix A.2.

Proposition 2.1. *Let $\mathcal{A}$ be a paving on Ω.*
Then $M(\mathcal{A}) = M(\Sigma(\mathcal{A})) = M(\gamma(\mathcal{A}))$.

We do not consider the problem to actually construct a generator. The problem is not trivial as illustrated by the next example.

Example 2.4. Let $\Omega = \{1, 2, 3, 4\}$ and consider the paving $\mathcal{A} = \{12, 13, 14\}$, where we allow 12 to denote the event $\{1, 2\}$ etc.
Obviously

$$1_{23} = 1_\Omega - 1_{14}$$

such that $23 \in \Sigma(\mathcal{A})$. Furthermore

$$1_1 = (1_{12} + 1_{13} - 1_{23})/2$$

We may conclude that $\gamma(\mathcal{A}) = \alpha(\mathcal{A})$, i.e. the generator is actually the atomic event. ■

About generators we note the following, which should be fairly evident from the definition.

Proposition 2.2. *Let $\mathcal{A}$ be a generator. Then $a \in \Sigma(\mathcal{A})$ is a union of disjoint events in $\mathcal{A}$. Furthermore we have*

- $\Omega \in \Sigma(\mathcal{A})$*, i.e.* Ω *is the union of a set of disjoint events in* $\mathcal{A}$*.*
- *If* $a, b \in \Sigma(\mathcal{A})$ *are disjoint, then* $a \cup b \in \Sigma(\mathcal{A})$*.*
- *If* $a \in \Sigma(\mathcal{A})$ *then the complement* $\Omega - a \in \Sigma(\mathcal{A})$*.*

We use the term **the set of marginal events in** $\mathcal{A}$ to denote $\Sigma(\mathcal{A})$.

Subsequently, we allways assume that $\mathcal{A} = \gamma(\mathcal{A})$, when we consider a model $M(\mathcal{A})$ and we will speak of a **log linear model** $M(\mathcal{A})$ **with generator** $\mathcal{A}$.

Actually, given a generator $\mathcal{A}$ we will focus on the extended model, where we allow the representation to contain zeroes, i.e. we define

Definition 2.4.
Let $\mathcal{A}$ be a generator. We then define $\overline{M}(\mathcal{A})$ to be all distributions on $\sigma(\mathcal{A})$ satisfying

$$P(c) = \prod_{a \in \mathcal{A}} \lambda_a^c, \quad c \in \alpha(\mathcal{A})$$

for some $\lambda(P) = \{\lambda_a \geq 0 | a \in \mathcal{A}\}$, which we call **a representation of** P. □

Remark that λ_Ω is no longer included as a normalizing constant. When $\mathcal{A}$ is a generator, we can by proposition 2.2 find disjoint events $\mathcal{A}_0 \subseteq \mathcal{A}$, where the union of these events is Ω. Hence we may absorb λ_Ω into $\{\lambda_a | a \in \mathcal{A}_0\}$.

A fundamental property of a distribution P that factorizes w.r.t. $\mathcal{A}$ is that the distribution is uniquely determined, when we know the probabilities $P(a),\ a \in \mathcal{A}$. This may represent a dramatic dimension reduction compared to specifying $P(c)$, $c \in \alpha(\mathcal{A})$. This property is summarized by

Proposition 2.3. *Let* $P, Q \in \overline{M}(\mathcal{A})$*. If* $P(a) = Q(a),\ a \in \mathcal{A}$ *then* $P = Q$*.*

The proof is given in appendix A in a slightly extended version.

The above proposition does not describe how to construct P from $P(a),\ a \in \mathcal{A}$, and in general this requires what is known as iterative proportional scaling. We do not pursue this issue, which in a more general setting requires extension of $\overline{M}(\mathcal{A})$ such that it is closed, i.e. contains the limits of distributions in $\overline{M}(\mathcal{A})$. Instead we consider situations, where exact reconstruction is possible, as this is the actual focus of this paper.

3 Decomposable models

We are going to describe the setup, which allows an easy reconstruction of $P \in \overline{M}(\mathcal{A})$ from knowing the probabilities of the events in the generator $\mathcal{A}$. We start by returning to example 2.3.

Example 3.1. Reconsider the paving $\mathcal{A}$ being the union of factors $f_1, \ldots, f_6$ such that

$$\alpha(\mathcal{A}) = (f_1 \otimes f_2) \oplus (((f_3 \otimes f_4) \oplus f_5) \otimes f_6)$$

with separators $\mathcal{S} = \{s_1, s_2, s_3\}$ fulfilling

- $s_1 = \mathrm{s}(f_1) = \mathrm{s}(f_2)$ and $s_2 = \mathrm{s}(f_3) = \mathrm{s}(f_4)$.
- $s_3 = \mathrm{s}(f_6) = s_2 + \mathrm{s}(f_5)$ and $s_1 + s_3 = \Omega$.

The crucial point to note is that if we remove any one of f_1, f_2, f_3, f_4 or f_6, then we obtain a model with a similar structure. E.g. removing f_4 we obtain a model with atomic structure

$$\alpha(\mathcal{A} \setminus f_4) = (f_1 \otimes f_2) \oplus ((f_3 \oplus f_5) \otimes f_6)$$

and separators $\mathcal{S} = \{s_1, s_3\}$. If $P \in \overline{M}(\mathcal{A})$ then we may consider the removal of f_4 as a marginalization of P in the universe s_2 and - as we shall see - we obtain a distribution in $\overline{M}(\mathcal{A} \setminus f_4)$. ■

The example above calls for some more definitions.

Definition 3.1.
A factor f on Ω is said to be **a complete factor** on Ω if $\mathrm{s}(f) = \Omega$.

A generator $\mathcal{A}$ on Ω is said to be **a complete generator** if $\mathcal{A}$ is a complete factor.

Let $\mathcal{A}$ be a paving and α a complete factor. If $\mathcal{A} \subseteq \sigma(\alpha)$ this is denoted by $\mathcal{A} \leq \alpha$.

A factor $f \subseteq \mathcal{A}$ is said to be **simplicial in** $\mathcal{A}$ if $\mathrm{s}(f) \in \Sigma(\mathcal{A} \setminus f)$ and we have the atomic structure

$$\alpha(\mathcal{A} \setminus f) = \alpha_0 \oplus \alpha_1 \tag{1}$$
$$\alpha(\mathcal{A}) = \alpha_0 \oplus f \otimes \alpha_1 \tag{2}$$

The paving $\mathcal{A}$ is said to be **decomposable** if either $\mathcal{A}$ is complete or $\mathcal{A}$ has a simplicial factor f such that $\mathcal{A} \setminus f$ is a decomposable paving. The factor f is said to be **a terminal factor** in $\mathcal{A}$.

Suppose that $\mathcal{A}$ is a decomposable paving. Clearly, this means that $\mathcal{A}$ can be split into a sequence $(f_1, \ldots f_k)$ of subsets such that f_i is a terminal factor in $\bigcup_{j=1}^{i} f_j$, $i = 2, \ldots, k$.

We denote the sequence to be **a perfect sequence of factors** in $\mathcal{A}$.

Suppose that $\mathcal{A}$ is a decomposable paving with $(f_1, \ldots f_k)$ as a perfect sequence of factors. Let $s_i = s(f_{i+1})$, $i = 1, \ldots, k-1$.

These events are called **separators**.

We denote $(s_1, \ldots s_{k-1})$ to be **a numbering of the separators** of $\mathcal{A}$. □

Example 3.2. Reconsider example 3.1 with atomic structure

$$\alpha(\mathcal{A}) = (f_1 \otimes f_2) \oplus (((f_3 \otimes f_4) \oplus f_5) \otimes f_6)$$

and separators $\mathcal{S} = \{s_1, s_2, s_3\}$ fulfilling

- $s_1 = \mathrm{s}(f_1) = \mathrm{s}(f_2)$ and $s_2 = \mathrm{s}(f_3) = \mathrm{s}(f_4)$.
- $s_3 = \mathrm{s}(f_6) = s_2 + \mathrm{s}(f_5)$ and $s_1 + s_3 = \Omega$.

One example of a perfect sequence is $(f_1 \oplus f_3 \oplus f_5, f_2, f_4, f_6)$ with (s_1, s_2, s_3) as a numbering of separators.
An alternative is $(f_2 \oplus f_6, f_4 \oplus f_5, f_3, f_1)$ with (s_3, s_2, s_1) as a numbering of separators. ■

Example 3.3. Let $f = \oplus_{i=1}^{3} f_i$ and $g = \oplus_{j=1}^{2} g_j$ be complete and cartesian factors on Ω. Suppose that events in

$$N = (f_1 \otimes g_1) \oplus (f_3 \otimes g_2)$$

have probalitity zero. This is an example of the socalled block-stairway incomplete table in Bishop *et al.* (1975). To avoid trivialities, we assume that all of the f_i, g_j factors contain more that one event. Define

- $\Omega_* = \Omega \setminus N$.
- $f_i^* = f_i \wedge \Omega_*$, $i = 1, 2, 3$ and $g_j^* = g_j \wedge \Omega_*$, $j = 1, 2$.
- $\mathcal{A}_0 = (\cup_{i=1}^{3} f_i^*) \cup (\cup_{j=1}^{2} g_j^*)$.

One might be tempted to think that $\mathcal{A}_0$ is a generator, but remark that

$$\mathrm{s}(f_1^*) = \mathrm{s}(f_1) \wedge \mathrm{s}(g_2)$$
$$\mathrm{s}(g_2^*) = \mathrm{s}(f_1) \wedge \mathrm{s}(g_2) + \mathrm{s}(f_2) \wedge \mathrm{s}(g_2)$$

and hence we need to include $a_{22} = \mathrm{s}(f_2) \wedge \mathrm{s}(g_2)$ in a generator. Similarly, since $a_{22} \subset \mathrm{s}(f_2)$ we include $a_{21} = \mathrm{s}(f_2) - a_{22}$ to get the generator $\mathcal{A} = \mathcal{A}_0 \cup \{a_{21}, a_{22}\}$. The atoms may be represented as

$$\alpha(\mathcal{A}) = (g_1^* \otimes (\{a_{21}\} \otimes f_2^*)) \oplus f_3^*) \oplus (g_2^* \otimes (f_1^* \oplus (\{a_{22}\} \otimes f_2^*)))$$

Both g_1^* and g_2^* are terminal with separators

- $s_1 = \mathrm{s}(g_1^*) = a_{21} + \mathrm{s}(f_3^*)$ and $s_2 = \mathrm{s}(g_2^*) = \mathrm{s}(f_1^*) + a_{22}$.

Once these are removed $\{a_{21}, a_{22}\}$ is terminal with separator $s_2 = a_{21} + a_{22} = \mathrm{s}(f_2^*)$ and we end up with the complete factor $f_1^* \oplus f_2^* \oplus f_3^*$. ■

Example 3.4. Let f, g and h be complete and mutually cartesian factors on Ω. Suppose that $f_0 \leq f$ where each event in f_0 corresponds to the union of two or more events in f. Assume similarly about $f_1 \leq f$ and consider the decompositions, definition and paving

- $f = \oplus_{a \in f_0} f_0^a$ where $\mathrm{s}(f_0^a) = a$, $a \in f_0$.
- $f = \oplus_{b \in f_1} f_1^b$ where $\mathrm{s}(f_1^b) = b$, $b \in f_1$.
- Define $g_a = \{a\} \otimes g$, $a \in f_0$ and $h_b = \{b\} \otimes h$, $b \in f_1$.
- $\mathcal{A} = f \cup (f_0 \otimes g) \cup (f_1 \otimes h) = f \cup (\oplus_{a \in f_0} g_a) \cup (\oplus_{b \in f_1} h_b)$.

Then the atomic structure is given by

- $\alpha(\mathcal{A}) = f \otimes g \otimes h = \sum_{b \in f_1} (f_1^b \otimes g) \otimes h_b$, where h_b is terminal with separator $b = \mathrm{s}(h_b) = \mathrm{s}(f_1^b \otimes h)$, $b \in f_1$.
- $\alpha(\mathcal{A} \setminus h) = f \otimes g = \sum_{a \in f_0} (f_0^a \otimes g_a)$, where g_a is terminal with separator $a = \mathrm{s}(g_a) = \mathrm{s}(f_0^a \otimes h)$, $a \in f_1$.

We conclude that $\mathcal{A}$ is decomposable with separators $\mathcal{S} = f_0 \cup f_1$.

One possible interpretation is that f, g, h represent categorical variables. And that g is independent of (f, h) conditionally on the aggregation f_0 of f, whereas h is independent of (f, g) conditionally on the aggregation f_1 of f.

Clearly, the model is a simple example, but anyhow it represents a structure, which is more general than Dellaportas and Forster (1999), since we allow that level merging is depending on local characteristics.

■

The preceeding examples indicate that decomposable log-linear models is a very general class. Integrating local modelling of context specific independence, structural zeroes and level merging demonstrates the applicability of decomposable log-linear models to a wide class of models, which - as we shall demonstrate - allow exact inference.

In order to study models associated with decomposable pavings we need the following, which is proved in appendix A.3.

Theorem 3.1. *If $\mathcal{A}$ is a decomposable paving then $\mathcal{A}$ is a generator.*

We have now set the scene for presenting the main theorem about the structure of decomposable models. The analogous result for decomposable graphical models is originally due to Haberman (1970), but can also be found in Andersen (1974). The proof is deferred to appendix C.

Theorem 3.2. *Suppose that $\mathcal{A}$ is decomposable and let $(s_1, \ldots, s_{k-1})$ be a numbering of the separators in $\mathcal{A}$.*

Let $P \in \overline{M}(\mathcal{A})$ *and* $I = \{i | P(s_i) > 0\}$. *Then*

$$P(c) = \frac{\prod_{a \in \mathcal{A}} P(a)^c}{\prod_{i \in I} P(s_i)^c} \quad c \in \alpha(\mathcal{A})$$

where

$$P(b)^c = \begin{cases} P(b) & \text{if } c \subseteq b \\ 1 & \text{otherwise} \end{cases}$$

This finishes our description of decomposable models. We continue to study inferential aspects.

4 Inference for decomposable models.

The kind of data at hand is represented by a vector $n_\alpha = \{n(a)\}_{a \in \alpha(\mathcal{A})}$. Our basic assumption is that these data represent a sample, which is modelled by a multinomial distribution.

Definition 4.1.
Let α be a complete generator. The stochastic vector $N_\alpha = \{N(a)\}_{a \in \alpha}$ is said to have a multinomial distribution with probabilities $p_\alpha = \{p(a)\}_{a \in \alpha}$ and sample size $n_+ = \sum_{a \in \alpha} n(a)$ if

$$P(N_\alpha = n_\alpha) = \binom{n_+}{n_\alpha} p_\alpha^{n_\alpha} = \frac{n_+!}{\prod_{a \in \alpha} n(a)!} \prod_{a \in \alpha} p(a)^{n(a)}$$

We denote this distribution $N_\alpha \sim \text{mult}_\alpha(n_+, p_\alpha)$. □

When $n_\alpha \in \mathbb{R}^\alpha$ we extend this to $\sigma(\alpha)$ by

$$n_\alpha(a) = \sum_{c \in a \wedge \alpha} n_\alpha(c) \quad a \in \sigma(\alpha)$$

Suppose $\mathcal{A}$ is a paving such that $\mathcal{A} \leq \alpha$. Then the $\mathcal{A}$ marginal of n_α is welldefined as

$$n_\mathcal{A} = \Pi_\mathcal{A}(n_\alpha) = \{n(a)\}_{a \in \mathcal{A}}$$

Remark that we allow the shorthand notation $n_\mathcal{A}$ when no ambiguities are present.

Concerning estimation in decomposable models we have the following result, which is proven in appendix C.1.

Theorem 4.1. *Suppose* $\mathcal{A}$ *is decomposable with separators* $(s_1, \ldots, s_k)$.
Let n_α *be an observation of* $N_\alpha \sim \text{mult}_\alpha(n^+, p_\alpha)$ *where* $p_\alpha \in \overline{M}(\mathcal{A})$ *and let* $I = \{1 \leq i \leq k | n(s_i) > 0\}$.

Then the maximum likelihood estimate of p_α is given by

$$\hat{p}(c) = \frac{\prod_{a\in\mathcal{A}} n(a)^c}{n_+ \prod_{i\in I} n(s_i)^c} \quad c \in \alpha(\mathcal{A}) \tag{3}$$

Subsequently, we describe the distribution of the estimator by characterizing the distribution of the sufficient "marginals" in a decomposable model. The proof is given in appendix C.2.

Theorem 4.2. *Let $\mathcal{A}$ be a decomposable paving with separators $s_1, \ldots, s_k$ and let $\alpha = \alpha(\mathcal{A})$. Suppose $N_\alpha \sim \text{mult}_\alpha(n_+, p_\alpha)$, where $p_\alpha \in \overline{M}(\mathcal{A})$. Then*

$$P(N_\mathcal{A} = n_\mathcal{A}) = \prod_{a\in\mathcal{A}} \frac{p(a)^{n(a)}}{n(a)!} \prod_{i=1}^{k} \frac{n(s_i)!}{p(s_i)^{n(s_i)}}$$

This generalizes the analogous result for decomposable graphical models as presented in (Sundberg 1975).

Example 4.1. Reconsider example 3.2 with the atomic structure

$$\alpha(\mathcal{A}) = (f_1 \otimes f_2) \oplus (((f_3 \otimes f_4) \oplus f_5) \otimes f_6)$$

Let $f_i^* \leq f_i$ be a binary subfactor of f_i, $i = 3, 4$ and suppose that we want to test the model extended by including the factor $f_3^* \otimes f_4^*$ against the model given by $\mathcal{A} = \cup_{i=1}^6 f_i$.

By theorem 4.1 it is verified that the likelihood ratio test statistic correponds to testing independence between the "variables" f_3^* and f_4^* conditionally on the "universe" $s = \text{s}(f_3^*) = \text{s}(f_4^*)$.

Also - within the framework of full exponential families - the exact test would be based on the distribution of $N_{f_3^* \otimes f_4^*}$ conditionally on $N_\mathcal{A}$, and referring to theorem 4.2, it should be obvious that under the model $\overline{M}(\mathcal{A})$ we have

$$P(N_{f_3^* \otimes f_4^*} = n_{f_3^* \otimes f_4^*} | N_\mathcal{A} = n_\mathcal{A}) = \frac{\prod_{a\in f_3^* \cup f_4^*} n(a)!}{n(s)! \prod_{a\in f_3^* \otimes f_4^*} n(a)!} \tag{4}$$

This is identified as the hypergeometric distribution for Fisher's usual exact test of independence between the binary variables f_3^* and f_4^* within the universe given by s. Furthermore, we note that $N_{f_3^* \otimes f_4^*}$ and $N_\mathcal{A}$ are conditionally independent given $N(s)$. ■

5 Concluding remarks

We have demonstrated that log-linear decomposable models provide a very flexible framework for integrating knowledge about context specific independence, structural zeroes, and level merging. And that this integration allows exact inference in terms of estimation and distribution of estimates.

Another, very substantial point, in connexion with decomposable graphical models, is effective algorithms for calculating posterior probabilities, i.e. conditional probabilities of a marginal event given an event specified by the intersection of a set of generating events. In relation to graphical models for discrete variables, this problem was originally treated in Lauritzen and Spiegelhalter (1988), and the algorithms have gathered widespread use in diverse areas, e.g. illustrated by Xiang (2002) and Storkey (2004). Remarking the similarity between decomposable log-linear models and decomposable graphical models, it seems to be straight forward to generalize these results.

Finally, one should mention the challenge to do model selection within the universe of decomposable log-linear models. Example 4.1 illustrates a basic step in a forward model selection algorithm, but evidently there is a need for clever search strategies to limit a search space, which can be prohibitively large.

References

Andersen, A. H. (1974). Multidimensional contingency tables. *Scand. J. Statist.*, **1**, (3), 115–27.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete multivariate analysis: theory and practice*. The MIT Press, Cambridge, Mass.-London.

Corander, J. (2003). Labelled graphical models. *Scand. J. Statist.*, **30**, (3), 493–508.

Darroch, J. N., Lauritzen, S. L., and Speed, T. P. (1980). Markov fields and log-linear interaction models for contingency tables. *Ann. Statist.*, **8**, (3), 522–39.

Dellaportas, P. and Forster, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika*, **86**, (3), 615–33.

Haberman, S. J. (1970). *The general log-linear model*. PhD thesis, Department of Statistics, University of Chicago.

Højsgaard, S. (2004). Statistical inference in context specific interaction models for contingency tables. *Scand. J. Statist.*, **31**, (1), 143–58.

Lauritzen, S. L. (1996). *Graphical models*, Oxford statistical science series, Vol. 17. The Clarendon Press Oxford University Press, New York.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *J. Roy. Statist. Soc. Ser. B*, **50**, (2), 157–224.

Poole, D. and Zhang, N. L. (2003). Exploiting contextual independence in probabilistic inference. *J. Artificial Intelligence Res.*, **18**, 263–313 (electronic).

Rapallo, F. (2003). Algebraic Markov bases and MCMC for two-way contingency tables. *Scand. J. Statist.*, **30**, (2), 385–97.

Rosen, T., Shimony, S. E., and Santos, Jr., E. (2004). Reasoning with BKBs—algorithms and complexity. *Ann. Math. Artif. Intell.*, **40**, (3-4), 403–25.

Storkey, A. J. (2004). Generalised propagation for fast fourier transforms with partial or missing data. In *Advances in Neural Information Processing Systems 16*, (ed. S. Thrun, L. Saul, and B. Schölkopf). MIT Press, Cambridge, MA.

Sundberg, R. (1975). Some results about decomposable (or Markov-type) models for multidimensional contingency tables: distribution of marginals and partitioning of tests. *Scand. J. Statist.*, **2**, (2), 71–9.

Teugels, J. L. and Van Horebeek, J. (1998). Generalized graphical models for discrete data. *Statist. Probab. Lett.*, **38**, (1), 41–7.

Wermuth, N. and Cox, D. (1998). On the application of conditional independence to ordinal data. *International Statistical Review*.

Xiang, Y. (2002). *Probabilistic reasoning in multiagent systems*. Cambridge University Press, Cambridge.

A Properties of log linear models

A.1 Proof of proposition 2.3

It turns out that proofs may me simplified by relazing the assumption, that we are dealing with probability distributions. Instead we consider set functions representing abstract measures on $\sigma(\mathcal{A})$ that factorize on $\alpha(\mathcal{A})$, where $\mathcal{A}$ is a generator. We denote these measures by $\overline{H}(\mathcal{A})$.

Definition A.1.
An element $h \in \overline{H}(\mathcal{A})$ is a function on $\sigma(\mathcal{A})$ defined by a representation

$$\mu(h, \mathcal{A}) = \{\mu_a \geq 0 | a \in \mathcal{A}\}$$

such that

$$h(c) = \prod_{a \in \mathcal{A}} \mu_a^c \quad c \in \alpha(\mathcal{A})$$

$$h(a) = \sum_{c \in a \wedge \alpha(\mathcal{A})} h(c) \quad a \in \sigma(\mathcal{A})$$

□

Suppose that $h, g \in \overline{H}(\mathcal{A})$ i.e.

$$h(c) = \prod_{a \in \mathcal{A}} \lambda_a^c \quad c \in \alpha(\mathcal{A})$$

$$g(c) = \prod_{a \in \mathcal{A}} \mu_a^c \quad c \in \alpha(\mathcal{A})$$

for suitable representations of h and g.

We now consider the situation $h(a) = g(a)$ when $a \in \mathcal{A}$ and want to show that $h = g$.

First of all we show that h and g have common support. Let $c_0 \in \alpha(\mathcal{A})$ and assume $h(c_0) = 0$. By the representation of h we conclude that $\lambda_a = 0$ for some $a \in \mathcal{A}$, where $c_0 \subseteq a$. Hence $h(c) = 0$ for any $c \subseteq a$, $c \in \alpha(\mathcal{A})$ whereby

$$\sum_{c \in a \wedge \alpha(\mathcal{A})} g(c) = g(a) = h(a) = \sum_{c \in a \wedge \alpha(\mathcal{A})} h(c) = 0$$

such that

$$h(c) = 0 \Leftrightarrow g(c) = 0 \quad c \in \alpha(\mathcal{A}) \tag{5}$$

By analogous considerations, it is clear that we may change the representation such that we obtain

$$\begin{gathered} \lambda_a = \mu_a = 0 \Leftrightarrow h(a) = g(a) = 0 \Leftrightarrow \\ \{g(c) = h(c) = 0,\ c \in \alpha(\mathcal{A}),\ c \subseteq a\} \quad a \in \mathcal{A} \end{gathered} \tag{6}$$

Next, note that since $\mathcal{A}$ is a generator, then - by proposition 2.2 - it has a factor f such that $\mathrm{s}(f) = \Omega$ and hence $h(\Omega) = g(\Omega)$. So actually, we may by normalization suppose that g and h are probabilities. Let $I(h||g)$ denote the Kullback Leibler distance between h and g considered as probability densities on $\alpha(\mathcal{A})$. Then with the usual convention that $0 \log(\frac{0}{0}) = 0$ and (5), (6) we obtain

$$\begin{aligned} I(h||g) = \sum_{c \in \alpha(\mathcal{A})} h(c) \log(\frac{h(c)}{g(c)}) = \\ \sum_{c \in \alpha(\mathcal{A})} h(c) \log(\prod_{a \in \mathcal{A}} \frac{\lambda_a^c}{\mu_a^c}) = \sum_{c \in \alpha(\mathcal{A})} h(c) \sum_{a \in \mathcal{A}} \log(\frac{\lambda_a^c}{\mu_a^c}) \\ = \sum_{a \in \mathcal{A}} \sum_{c \in \alpha(\mathcal{A})} h(c) \log(\frac{\lambda_a^c}{\mu_a^c}) = \sum_{a \in \mathcal{A}} h(a) \log(\frac{\lambda_a}{\mu_a}) = \\ \sum_{a \in \mathcal{A}} g(a) \log(\frac{\lambda_a}{\mu_a}) = \sum_{a \in \mathcal{A}} \sum_{c \in \alpha(\mathcal{A})} g(c) \log(\frac{\lambda_a^c}{\mu_a^c}) = \\ \sum_{c \in \alpha(\mathcal{A})} g(c) \sum_{a \in \mathcal{A}} \log(\frac{\lambda_a^c}{\mu_a^c}) = \sum_{c \in \alpha(\mathcal{A})} g(c) \log(\prod_{a \in \mathcal{A}} \frac{\lambda_a^c}{\mu_a^c}) = \\ \sum_{c \in \alpha(\mathcal{A})} g(c) \log(\frac{h(c)}{g(c)}) = -I(g||h) \leq 0 \end{aligned}$$

such that $I(h||g) = 0$ and thereby $h(c) = g(c)$, $c \in \alpha(\mathcal{A})$. □

A.2 Proof of proposition 2.1

It should be fairly obvious that

$$L(\gamma(\mathcal{A})) \subseteq L(\Sigma(\mathcal{A})) = L(\mathcal{A})$$

i.e. we only need to show that if $1_c \in L(\Sigma(\mathcal{A}))$ then $1_c \in L(\gamma(\mathcal{A}))$, which we do by induction on the number $n = |c|$ of elements in c.

If $n = 1$ then clearly c is minimal, i.e. $c \in \gamma(\mathcal{A})$. So suppose that $n > 1$ and $c \notin \gamma(\mathcal{A})$, i.e. we can find $b \in \gamma(\mathcal{A})$ such that $b \subset c$, whereby $1_{c-b} = 1_c - 1_b \in L(\Sigma(\mathcal{A}))$. By induction $1_{c-b} \in L(\gamma(\mathcal{A}))$ and hence $1_c = 1_b + 1_{c-b} \in L(\gamma(\mathcal{A}))$.

A.3 Proof of theorem 3.1

Proof. If $\mathcal{A}$ is complete, then the statement should be obvious. So, let f be a terminal factor in $\mathcal{A}$, $\mathcal{A}_0 = \mathcal{A} \setminus f$ and suppose that $c \in \gamma(\mathcal{A})$, i.e. c is a minimal set fulfilling

$$1_c = \sum_{a \in \mathcal{A}_0} \lambda_a 1_a + \sum_{b \in f} \lambda_b 1_b \tag{7}$$

We then need to show that $c \in \mathcal{A}$.

By induction on the number of separators we have $\gamma(\mathcal{A}_0) = \mathcal{A}_0$. Combining this with the terminality of f means that we can find a factor g in $\mathcal{A}_0$ such that $\mathrm{s}(g) = \mathrm{s}(f)$. Remark that we can add any constant to $\{\lambda_a\}_{a \in f}$ by subtracting the same constant from $\{\lambda_a\}_{a \in g}$. This means that we may assume that

$$\lambda_b \geq 0, \quad b \in f \tag{8}$$

$$\lambda_{b_0} = 0 \quad \text{for some } b_0 \in f \tag{9}$$

Terminality of f also means that

$$\alpha(\mathcal{A}_0) = \alpha_0 \oplus \alpha_1$$
$$\alpha(\mathcal{A}) = \alpha_0 \oplus (f \otimes \alpha_1)$$
$$\mathrm{s}(f) = \mathrm{s}(\alpha_1)$$

In general we have that c is the sum of its atoms in $\alpha(\mathcal{A})$, i.e.

$$1_c = \sum_{a \in \alpha_0} \mu_a 1_a + \sum_{a \in \alpha_1, b \in f} \mu_{ab} 1_{a \wedge b} \quad \mu_a \in \{0,1\}, \quad \mu_{ab} \in \{0,1\} \tag{10}$$

Similarly, any $a \in \mathcal{A}_0$ is the sum of its atoms in $\alpha(\mathcal{A}_0)$ whereby (7) writes

$$1_c = \sum_{a \in \alpha_0 \cup \alpha_1} \psi_a 1_a + \sum_{b \in f} \lambda_b 1_b \tag{11}$$

Now, comparing (10) and (11) yields

$$\psi_a = \mu_a \in \{0,1\}, \quad a \in \alpha_0$$
$$\psi_a + \lambda_b = \mu_{ab} \in \{0,1\}, \quad a \in \alpha_1,\ b \in f$$

By the last equation and (9) we obtain that $\psi_a \in \{0,1\},\ a \in \alpha_1$, ie. we actually have that $\psi_a \in \{0,1\},\ a \in \alpha_0 \cup \alpha_1$. Hence we may rewrite (11) as

$$1_c = 1_d + \sum_{b \in f} \lambda_b 1_b, \quad d \in \sigma(\mathcal{A}_0),\ \lambda_b \geq 0$$

Obviously, this means that $d \subseteq c$. If $d = c$ we must have $\lambda_b = 0,\ b \in f$ whereby $c \in \Sigma(\mathcal{A}_0)$ and by the induction hypothesis this means that $c \in \mathcal{A}_0$. If $d \subset c$ then we infer that $c - d \in \Sigma(\mathcal{A})$ and minimality of c then means that $d = \emptyset$. Removing 1_d from the above equation, it should be obvious from the minimality of c that $c \in f$. □

B Proof of factorization theorem

In order to prove theorem 3.2 we first establish a lemma, which shows that the factorization property is carried over, when we "marginalize out" terminal factors. The lemma is formulated for decomposable pavings, but actually only exploits the terminality. The set-up is slightly generalized as we extend it to $\overline{H}(\mathcal{A})$, i.e. non negative functions that factorize without necessarily summing to unity.

Lemma B.1. *Suppose that $\mathcal{A}$ is decomposable and incomplete, let f be a terminal factor and $\mathcal{A}_0 = \mathcal{A} \setminus f$.*
Let $h \in \overline{H}(\mathcal{A})$ and h_0 the restriction of h to $\sigma(\mathcal{A}_0)$.
Then $h_0 \in \overline{H}(\mathcal{A}_0)$ and

$$h(a \wedge b)h(s) = h(a)h_0(b) \quad a \in f,\ b \in \mathrm{s}(f) \wedge \alpha(\mathcal{A}_0)$$

Proof. Terminality of f is summarized by the structure

$$\alpha(\mathcal{A}_0) = \alpha_0 \oplus \alpha_1 \tag{12}$$
$$\alpha(\mathcal{A}) = \alpha_0 \oplus (f \otimes \alpha_1) \tag{13}$$
$$s = \mathrm{s}(f) = \mathrm{s}(\alpha_1) \in \Sigma(\mathcal{A}_0) \tag{14}$$

Let $\{\mu_a | a \in \mathcal{A}\}$ be a representation of h.
By (13) we have the factorization

$$h(c \wedge d) = \prod_{a \in f} \mu_a^{c \wedge d} \prod_{b \in \mathcal{A}_0} \mu_b^{c \wedge d} \quad c \in f, \quad d \in \alpha_1 \tag{15}$$

Consider the number $\mu_b^{c\wedge d}$, $b \in \mathcal{A}_0$, $c \in f$, $d \in \alpha_1$.
Now, if $d \not\subseteq b$ then $b \wedge d = \emptyset$ such that $\mu_b^{c\wedge d} = \mu_b^d = 1$. On the other hand, if $d \subseteq b$ then $\emptyset \subset c \wedge d \subseteq d$, such that we must have $\mu_b^{c\wedge d} = \mu_b^d = \mu_b$.
In summary, we conclude that $\mu_b^{c\wedge d} = \mu_b^d$, when $b \in \mathcal{A}_0$, $c \in f$, $d \in \alpha_1$.
Similarly, $\mu_a^{c\wedge d} = \mu_a^c$, $a, c \in f$, $d \in \alpha_1$ which allows us to rewrite (15) as

$$h(c \wedge d) = \prod_{a\in f} \mu_a^c \prod_{b\in\mathcal{A}_0} \mu_b^d \quad c \in f,\ d \in \alpha_1 \tag{16}$$

Summing out $c \in f$ in (16) and noting $d \subseteq \mathrm{s}(f) = s$ yields

$$h(d) = (\sum_{c\in f} \prod_{a\in f} \mu_a^c) \prod_{b\in\mathcal{A}_0} \mu_b^d \quad d \in \alpha_1 \tag{17}$$

So we need to include the term $\lambda_s = \sum_{c\in f} \prod_{a\in f} \mu_a^c$ in the representation of h_0. Since $s \in \Sigma(\mathcal{A}_0)$ and $\mathcal{A}_0$ is a generator by decomposability, then we can choose $g \in \mathcal{A}_0$ such that $\mathrm{s}(g) = s$. Define

$$\lambda_a = \begin{cases} \mu_a & a \in \mathcal{A}_0 \setminus g \\ \lambda_s \mu_a & a \in g \end{cases}$$

and observe that

$$h_0(d) = h(d) = \prod_{b\in\mathcal{A}_0} \lambda_b^d \quad d \in \alpha_1$$

By (13) it should also be evident that since $g \wedge \alpha_0 = \emptyset$ we have the factorization

$$h_0(c) = h(c) = \prod_{a\in\mathcal{A}_0\setminus g} \mu_a^c = \prod_{a\in\mathcal{A}_0\setminus g} \lambda_a^c = \prod_{a\in\mathcal{A}_0} \lambda_a^c \quad c \in \alpha_0$$

i.e. $h_0 \in \overline{H}(\mathcal{A}_0)$.
Next, summing out $d \in \alpha_1$ in (16) and noting $c \subseteq \mathrm{s}(\alpha_1) = s$ yields

$$h(c) = \prod_{a\in\mathcal{A}} \mu_a^c (\sum_{d\in\alpha_1} \prod_{b\in\mathcal{A}_0} \mu_b^d) \quad c \in f \tag{18}$$

Summing out $c \in f$ in (18) yields

$$h(s) = (\sum_{c\in f} \prod_{a\in f} \mu_a^c)(\sum_{d\in\alpha_1} \prod_{b\in\mathcal{A}_0} \mu_b^d) \tag{19}$$

It is now easy to combine the above equations to see that

$$h(c \wedge d)h(s) = h(c)h(d)$$

□

B.1 Proof of theorem 3.2

Again we give a proof in a slightly more general version, where we assume that $h \in \overline{H}(\mathcal{A})$ and $\mathcal{A}$ is decomposable with $(f_1, \ldots, f_k)$ as a perfect sequence.
If $I = \{1 < i \leq k | h(\mathrm{s}(f_i)) > 0\}$ then we intend to show

$$h(c) = \frac{\prod_{a \in \mathcal{A}} h(a)^c}{\prod_{i \in I} h(\mathrm{s}(f_i))^c} \quad c \in \alpha(\mathcal{A}) \tag{20}$$

The proof is induction on k. If $k = 1$ the theorem just states that

$$h(c) = \prod_{a \in f_1} h(a)^c, \quad c \in f_1$$

which is trivially true since f_1 is a complete factor.
So let $k > 1$. For notational convenience let $s = \mathrm{s}(f_k)$ and $\mathcal{A}_0 = \mathcal{A} \setminus f_k$.
Consider the preceeding lemma, by which we have

$$h(c) = h_0(c) \quad c \in \alpha_0 \tag{21}$$

$$h(c \wedge b)h(s) = h_0(c)h(b) \quad c \in \alpha_1,\ b \in f_k \tag{22}$$

where $\alpha(\mathcal{A}) = \alpha_0 \oplus (\alpha_1 \otimes f_k)$ and $h_0 \in \overline{H}(\mathcal{A}_0)$ coinsides with h on $\sigma(\mathcal{A}_0)$.
If $I_0 = I \setminus \{k\}$ then by the induction hypothesis we obtain

$$h_0(c) = h(c) = \frac{\prod_{a \in \mathcal{A}_0} h(a)^c}{\prod_{i \in I_0} h(\mathrm{s}(f_i))^c} \quad c \in \alpha_0 \oplus \alpha_1$$

where we have exploited that $h(a) = h_0(a)$, $a \in \mathcal{A}_0$.
If $c \in \alpha_0$ the result in (20) is obvious since then $c \wedge s = \emptyset$, which means that the contributions attached to f_k and $\mathrm{s}(f_k)$ disappear, so that we have the above representation of h_0.
So let us consider $c \wedge b \in \alpha_1 \otimes f_k$. If $h(s) = 0$ then $h(b \wedge c) = h(b) = 0$ and the result follows by noting that $h(b)$ enters the product on the right hand side of (20). Otherwise, we just need to divide by $h(s)$ in (22) and use the representation of h_0. □

C Properties of estimation

C.1 Proof of theorem 4.1

Let $p_\alpha \in \overline{M}(\mathcal{A})$ and $I(p_\alpha) = \{1 \leq i \leq k | p(s_i) > 0\}$.
Then the likelihood is by theorem 3.2 proportional to

$$L(p_\alpha | n_\alpha) = \prod_{c \in \alpha(\mathcal{A})} p(c)^{n(c)} = \frac{\prod_{a \in \mathcal{A}} p(a)^{n(a)}}{\prod_{i \in I(p_\alpha)} p(s_i)^{n(s_i)}} \quad p_\alpha \in \overline{M}(\mathcal{A})$$

Define the function n^*_α on $\sigma(\mathcal{A})$ by

$$n^*(c) = \frac{\prod_{a\in\mathcal{A}} n(a)^c}{\prod_{i\in I} n(\mathrm{s}_i)^c} \quad c \in \alpha(\mathcal{A}) \tag{23}$$

where $I = \{1 \leq i \leq k | n(s_i) > 0\}$. We note by section B.1 that

$$n^*(a) = n(a) \quad a \in \mathcal{A} \tag{24}$$

and in particular that $n^*(\Omega) = n_+$ since $\Omega \in \Sigma(\mathcal{A})$ whereby

$$p^*_\alpha = \frac{n^*_\alpha}{n_+} \in \overline{M}(\mathcal{A})$$

Furthermore, it is easy to verify that

$$L(p_\alpha|n_\alpha) = L(p_\alpha|n^*_\alpha)$$

and as $L(p^*_\alpha|n^*_\alpha) > 0$, this allows us to consider the alternative likelihood

$$L^*(p_\alpha) = \frac{L(p_\alpha|n^*_\alpha)}{L(p^*_\alpha|n^*_\alpha)} \qquad p_\alpha \in \overline{M}(\mathcal{A}) \cap \{p_\alpha | L^*((p_\alpha) \geq 1\}$$

Hence

$$\log(L^*(p_\alpha)) = n_+ \sum_{c\in\alpha(\mathcal{A})} p^*(c) \log(\frac{p(c)}{p^*(c)}) = -n_+ I(p^*_\alpha || p_\alpha) \tag{25}$$

where the Kullback-Leibler distance is minimized when $p^*_\alpha = p_\alpha$.

C.2 Proof of theorem 4.2

Induction on k. If $k = 1$ the result is obvious. Next consider $k = 2$, i.e. we must have that $\mathcal{A} = f_0 \cup f_1 \cup f_2$ and

$$\alpha = \alpha(\mathcal{A}) = f_0 \oplus (f_1 \otimes f_2)$$

If we let $s_0 = \mathrm{s}(f_0)$ and $s_1 = \mathrm{s}(f_1) = \mathrm{s}(f_2)$ then $Z = (N(s_0), N(s_1))$ is binomial. If $Y_i = \{N(c)_{c\in f_i}\}$, $i = 0, 1, 2$ and $X = \{N(c)_{c\in f_1\otimes f_2}\}$ then (Y_0, X) are independent given Z, and Y_0 given Z is multinomial.

Similarly X given Z is multinomial corresponding to a twoway table, where the model specifies independence, which allows us to conclude independence and multinomial distributions of (Y_1, Y_2) conditionally on Z. Piecing this together gives the prescribed distribution of (Y_0, Y_1, Y_2).

So let $k > 2$ and f be terminal in $\mathcal{A}$, $\mathcal{A}_0 = \mathcal{A} \setminus f$, $\beta = \alpha(\mathcal{A}_0) = \alpha_0 \oplus \alpha_1$ and $\alpha = \alpha(\mathcal{A}) = \alpha_0 \oplus f \otimes \alpha_1$. We then refer to the case $k = 2$ to obtain that

$$P(N_\beta = n_\beta, N_f = n_f) = P(N_\beta = n_\beta) \frac{\prod_{a\in f} p(a)^{n(a)}}{p(\mathrm{s}(f))^{n(\mathrm{s}(f))}} \tag{26}$$

As N_β is marginal to N_α it is multinomial and by lemma B.1 we obtain that $p_\beta \in \overline{M}(\mathcal{A}_0)$.

By the induction assumption we obtain a factorization of $P(N_{\mathcal{A}_0})$ which combined with (26) yields the result, when we note that $\{N_{A_0}, N_f\} = N_{\mathcal{A}}$.