## Aalborg Universitet

## AALBORG UNIVERSITY

## Decomposable log-linear models

Eriksen, Poul Svante

Publication date:
2005

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Eriksen, P. S. (2005). Decomposable log-linear models. Department of Mathematical Sciences, Aalborg
University. Research Report Series No. R-2005-16

## General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

## AALBORG UNIVERSITY

## Decomposable log-linear models

by<br>P. Svante Eriksen



# Decomposable log-linear models. 

P. Svante Eriksen<br>Department of Mathematical Sciences<br>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 $* \in\{-,+\}$.
- $V=\left\{v^{-}, v^{+}\right\}$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

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

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

- $a^{+} V=\left\{a^{+} \wedge v \mid v \in V\right\}, \quad V \in\{B, C\}$.
- $a^{*} B C=\left\{a^{*} \wedge b^{*} \wedge c^{*} \mid b^{*} \in B, c^{*} \in C\right\}, a^{*} \in A$.
- $A B C=a^{-} B C \cup a^{+} B C$.

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

$$
a^{+} B C=\left\{b \wedge c \mid b \in a^{+} B, c \in a^{+} C\right\}
$$

We may say that $a^{+} B, a^{+} C$ form a cartesian product that spans $a^{+} B C$.

The crucial points to emphasize is the decomposition of $A B C$ into $a^{-} B C$ and $a^{+} B C$ and the representation of $a^{+} B C$ 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 $\Omega$, which we denote the outcome space. Any subset of $\Omega$ is called an event. When $a$ and $b$ are events in $\Omega$ we use $a \wedge b$ to denote their intersection.
A non empty set $\mathcal{A}$ of events is defined to be a paving on $\Omega$.
If $\mathcal{A}$ and $\mathcal{B}$ are pavings on $\Omega$, 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 $\Omega$ 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 $\Omega$ is a set of disjoint events in $\Omega$.
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 $\Omega$ 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

- $f_{0}=a^{-} B C, f_{1}=a^{+} B$ and $f_{2}=a^{+} C$.
and we may represent the outcome space $\Omega=A B C$ as

$$
\Omega=f_{0} \oplus\left(f_{1} \otimes f_{2}\right)
$$

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 $\mathcal{S}=\left\{a^{+}\right\}$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

$$
P\left(a^{+}\right)=P\left(\mathrm{~s}\left(f_{1}\right)\right)=\sum_{a \in f_{1}} P(a)=\sum_{a \in f_{2}} P(a)=P\left(\mathrm{~s}\left(f_{2}\right)\right)
$$

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 $\left(f_{3} \otimes f_{4}\right) \oplus f_{5}$ and $f_{6}$ are cartesian with separator $e^{+}$.

Also we may represent $\Omega=A B C D$ as

$$
\Omega=\left(f_{1} \otimes f_{2}\right) \oplus\left(\left(\left(f_{3} \otimes f_{4}\right) \oplus f_{5}\right) \otimes f_{6}\right)
$$

Furthermore, if $\mathcal{A}$ is the union of $f_{1}, \ldots, f_{6}$ and $\mathcal{S}=\left\{a^{+}, e^{-}, e^{+}\right\}$then one may verify that

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

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

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

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 $\Omega$.
By $\sigma(\mathcal{A})$ we denote the smallest $\sigma$-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 $\sigma(\mathcal{A}) \backslash\{\emptyset\}$, 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 $\Omega$.
If $a$ is an event in $\mathcal{A} \cup\{\Omega\}$ and $\lambda_{a}$ is a non negative number we define the quantity $\lambda_{a}^{c}, c \in \alpha(\mathcal{A})$ by

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

i.e. we assign the number $\lambda_{a}$ 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 $\left\{\lambda_{a} \mid a \in \mathcal{A} \cup\{\Omega\}\right\}$ 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 $\Omega$ since

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

where $\beta_{a}=\log \left(\lambda_{a}\right)$ 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.

$$
L(\mathcal{A})=\operatorname{span}\left\{1_{c} \mid c \in \mathcal{A} \cup\{\Omega\}\right\}
$$

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

$$
\Sigma(\mathcal{A})=\left\{c \in \sigma(\mathcal{A}) \mid 1_{c} \in L(\mathcal{A})\right\}
$$

By the generator corresponding to $\mathcal{A}$ denoted $\gamma(\mathcal{A})$, we mean the minimal elements of $\Sigma(\mathcal{A}) \backslash\{\emptyset\}$, 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 $\Omega$.
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

$$
1_{1}=\left(1_{12}+1_{13}-1_{23}\right) / 2
$$

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. $\Omega$ 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 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 $\bar{M}(\mathcal{A})$ 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 $\lambda(P)=\left\{\lambda_{a} \geq 0 \mid a \in \mathcal{A}\right\}$, which we call a representation of $P$.
Remark that $\lambda_{\Omega}$ 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 $\Omega$. Hence we may absorb $\lambda_{\Omega}$ into $\left\{\lambda_{a} \mid a \in \mathcal{A}_{0}\right\}$.
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 $P, Q \in \bar{M}(\mathcal{A})$. 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 $\bar{M}(\mathcal{A})$ such that it is closed, i.e. contains the limits of distributions in $\bar{M}(\mathcal{A})$. 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 $P \in \bar{M}(\mathcal{A})$ 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

$$
\alpha(\mathcal{A})=\left(f_{1} \otimes f_{2}\right) \oplus\left(\left(\left(f_{3} \otimes f_{4}\right) \oplus f_{5}\right) \otimes f_{6}\right)
$$

with separators $\mathcal{S}=\left\{s_{1}, s_{2}, s_{3}\right\}$ fulfilling

- $s_{1}=\mathrm{s}\left(f_{1}\right)=\mathrm{s}\left(f_{2}\right)$ and $s_{2}=\mathrm{s}\left(f_{3}\right)=\mathrm{s}\left(f_{4}\right)$.
- $s_{3}=\mathrm{s}\left(f_{6}\right)=s_{2}+\mathrm{s}\left(f_{5}\right)$ 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

$$
\alpha\left(\mathcal{A} \backslash f_{4}\right)=\left(f_{1} \otimes f_{2}\right) \oplus\left(\left(f_{3} \oplus f_{5}\right) \otimes f_{6}\right)
$$

and separators $\mathcal{S}=\left\{s_{1}, s_{3}\right\}$. If $P \in \bar{M}(\mathcal{A})$ 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 $\bar{M}\left(\mathcal{A} \backslash f_{4}\right)$.

The example above calls for some more definitions.
Definition 3.1.
A factor $f$ on $\Omega$ is said to be a complete factor on $\Omega$ if $\mathrm{s}(f)=\Omega$.
A generator $\mathcal{A}$ on $\Omega$ is said to be a complete generator if $\mathcal{A}$ is a complete factor.
Let $\mathcal{A}$ be a paving and $\alpha$ 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 $\mathrm{s}(f) \in \Sigma(\mathcal{A} \backslash f)$ and we have the atomic structure

$$
\begin{align*}
& \alpha(\mathcal{A} \backslash f)=\alpha_{0} \oplus \alpha_{1}  \tag{1}\\
& \alpha(\mathcal{A})=\alpha_{0} \oplus f \otimes \alpha_{1} \tag{2}
\end{align*}
$$

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 $\mathcal{A} \backslash f$ 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 $\left(f_{1}, \ldots f_{k}\right)$ 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 $\left(f_{1}, \ldots f_{k}\right)$ as a perfect sequence of factors. Let $s_{i}=s\left(f_{i+1}\right), i=1, \ldots, k-1$.
These events are called separators.
We denote $\left(s_{1}, \ldots s_{k-1}\right)$ to be a numbering of the separators of $\mathcal{A}$.

Example 3.2. Reconsider example 3.1 with atomic structure

$$
\alpha(\mathcal{A})=\left(f_{1} \otimes f_{2}\right) \oplus\left(\left(\left(f_{3} \otimes f_{4}\right) \oplus f_{5}\right) \otimes f_{6}\right)
$$

and separators $\mathcal{S}=\left\{s_{1}, s_{2}, s_{3}\right\}$ fulfilling

- $s_{1}=\mathrm{s}\left(f_{1}\right)=\mathrm{s}\left(f_{2}\right)$ and $s_{2}=\mathrm{s}\left(f_{3}\right)=\mathrm{s}\left(f_{4}\right)$.
- $s_{3}=\mathrm{s}\left(f_{6}\right)=s_{2}+\mathrm{s}\left(f_{5}\right)$ and $s_{1}+s_{3}=\Omega$.

One example of a perfect sequence is $\left(f_{1} \oplus f_{3} \oplus f_{5}, f_{2}, f_{4}, f_{6}\right)$ with $\left(s_{1}, s_{2}, s_{3}\right)$ as a numbering of separators.
An alternative is $\left(f_{2} \oplus f_{6}, f_{4} \oplus f_{5}, f_{3}, f_{1}\right)$ with $\left(s_{3}, s_{2}, s_{1}\right)$ 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 $\Omega$. Suppose that events in

$$
N=\left(f_{1} \otimes g_{1}\right) \oplus\left(f_{3} \otimes g_{2}\right)
$$

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

- $\Omega_{*}=\Omega \backslash N$.
- $f_{i}^{*}=f_{i} \wedge \Omega_{*}, i=1,2,3$ and $g_{j}^{*}=g_{j} \wedge \Omega_{*}, j=1,2$.
- $\mathcal{A}_{0}=\left(\cup_{i=1}^{3} f_{i}^{*}\right) \cup\left(\cup_{j=1}^{2} g_{j}^{*}\right)$.

One might be tempted to think that $\mathcal{A}_{0}$ is a generator, but remark that

$$
\begin{gathered}
\mathrm{s}\left(f_{1}^{*}\right)=\mathrm{s}\left(f_{1}\right) \wedge \mathrm{s}\left(g_{2}\right) \\
\mathrm{s}\left(g_{2}^{*}\right)=\mathrm{s}\left(f_{1}\right) \wedge \mathrm{s}\left(g_{2}\right)+\mathrm{s}\left(f_{2}\right) \wedge \mathrm{s}\left(g_{2}\right)
\end{gathered}
$$

and hence we need to include $a_{22}=\mathrm{s}\left(f_{2}\right) \wedge \mathrm{s}\left(g_{2}\right)$ in a generator. Similarly, since $a_{22} \subset \mathrm{~s}\left(f_{2}\right)$ we include $a_{21}=\mathrm{s}\left(f_{2}\right)-a_{22}$ to get the generator $\mathcal{A}=\mathcal{A}_{0} \cup\left\{a_{21}, a_{22}\right\}$. The atoms may be represented as

$$
\alpha(\mathcal{A})=\left(g_{1}^{*} \otimes\left(\left(\left\{a_{21}\right\} \otimes f_{2}^{*}\right)\right) \oplus f_{3}^{*}\right) \oplus\left(g_{2}^{*} \otimes\left(f_{1}^{*} \oplus\left(\left\{a_{22}\right\} \otimes f_{2}^{*}\right)\right)\right)
$$

Both $g_{1}^{*}$ and $g_{2}^{*}$ are terminal with separators

- $s_{1}=\mathrm{s}\left(g_{1}^{*}\right)=a_{21}+\mathrm{s}\left(f_{3}^{*}\right)$ and $s_{2}=\mathrm{s}\left(g_{2}^{*}\right)=\mathrm{s}\left(f_{1}^{*}\right)+a_{22}$.

Once these are removed $\left\{a_{21}, a_{22}\right\}$ is terminal with separator $s_{2}=a_{21}+a_{22}=$ $\mathrm{s}\left(f_{2}^{*}\right)$ 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 $\Omega$. 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 $\mathrm{s}\left(f_{0}^{a}\right)=a, a \in f_{0}$.
- $f=\oplus_{b \in f_{1}} f_{1}^{b}$ where $\mathrm{s}\left(f_{1}^{b}\right)=b, b \in f_{1}$.
- Define $g_{a}=\{a\} \otimes g, a \in f_{0}$ and $h_{b}=\{b\} \otimes h, b \in f_{1}$.
- $\mathcal{A}=f \cup\left(f_{0} \otimes g\right) \cup\left(f_{1} \otimes h\right)=f \cup\left(\oplus_{a \in f_{0}} g_{a}\right) \cup\left(\oplus_{b \in f_{1}} h_{b}\right)$.

Then the atomic structure is given by

- $\alpha(\mathcal{A})=f \otimes g \otimes h=\sum_{b \in f_{1}}\left(f_{1}^{b} \otimes g\right) \otimes h_{b}$, where $h_{b}$ is terminal with separator $b=\mathrm{s}\left(h_{b}\right)=\mathrm{s}\left(f_{1}^{b} \otimes h\right), b \in f_{1}$.
- $\alpha(\mathcal{A} \backslash h)=f \otimes g=\sum_{a \in f_{0}}\left(f_{0}^{a} \otimes g_{a}\right)$, where $g_{a}$ is terminal with separator $a=\mathrm{s}\left(g_{a}\right)=\mathrm{s}\left(f_{0}^{a} \otimes h\right), a \in f_{1}$.

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 $\left(s_{1}, \ldots, s_{k-1}\right)$ be a numbering of the separators in $\mathcal{A}$.

Let $P \in \bar{M}(\mathcal{A})$ and $I=\left\{i \mid P\left(s_{i}\right)>0\right\}$. Then

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

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 $\alpha$ 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

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

We denote this distribution $N_{\alpha} \sim \operatorname{mult}_{\alpha}\left(n_{+}, p_{\alpha}\right)$.
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 $n_{\alpha}$ is welldefined as

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

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 $\left(s_{1}, \ldots, s_{k}\right)$.
Let $n_{\alpha}$ be an observation of $N_{\alpha} \sim \operatorname{mult}_{\alpha}\left(n^{+}, p_{\alpha}\right)$ where $p_{\alpha} \in \bar{M}(\mathcal{A})$ and let $I=\left\{1 \leq i \leq k \mid n\left(s_{i}\right)>0\right\}$.

Then the maximum likelihood estimate of $p_{\alpha}$ is given by

$$
\begin{equation*}
\hat{p}(c)=\frac{\prod_{a \in \mathcal{A}} n(a)^{c}}{n_{+} \prod_{i \in I} n\left(s_{i}\right)^{c}} \quad c \in \alpha(\mathcal{A}) \tag{3}
\end{equation*}
$$

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 $N_{\alpha} \sim \operatorname{mult}_{\alpha}\left(n_{+}, p_{\alpha}\right)$, where $p_{\alpha} \in \bar{M}(\mathcal{A})$. Then

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

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

$$
\alpha(\mathcal{A})=\left(f_{1} \otimes f_{2}\right) \oplus\left(\left(\left(f_{3} \otimes f_{4}\right) \oplus f_{5}\right) \otimes f_{6}\right)
$$

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" $s=\mathrm{s}\left(f_{3}^{*}\right)=\mathrm{s}\left(f_{4}^{*}\right)$.
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 $\bar{M}(\mathcal{A})$ we have

$$
\begin{equation*}
P\left(N_{f_{3}^{*} \otimes f_{4}^{*}}=n_{f_{3}^{*} \otimes f_{4}^{*}} \mid N_{\mathcal{A}}=n_{\mathcal{A}}\right)=\frac{\prod_{a \in f_{3}^{*} \cup f_{4}^{*}} n(a)!}{n(s)!\prod_{a \in f_{3}^{*} \otimes f_{4}^{*}} n(a)!} \tag{4}
\end{equation*}
$$

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 BKBsalgorithms 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 $\bar{H}(\mathcal{A})$.

## Definition A.1.

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

$$
\mu(h, \mathcal{A})=\left\{\mu_{a} \geq 0 \mid a \in \mathcal{A}\right\}
$$

such that

$$
\begin{gathered}
h(c)=\prod_{a \in \mathcal{A}} \mu_{a}^{c} \quad c \in \alpha(\mathcal{A}) \\
h(a)=\sum_{c \in a \wedge \alpha(\mathcal{A})} h(c) \quad a \in \sigma(\mathcal{A})
\end{gathered}
$$

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

$$
\begin{array}{ll}
h(c)=\prod_{a \in \mathcal{A}} \lambda_{a}^{c} & c \in \alpha(\mathcal{A}) \\
g(c)=\prod_{a \in \mathcal{A}} \mu_{a}^{c} & c \in \alpha(\mathcal{A})
\end{array}
$$

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 $h\left(c_{0}\right)=0$. 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

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

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

$$
\begin{gather*}
\lambda_{a}=\mu_{a}=0 \Leftrightarrow h(a)=g(a)=0 \Leftrightarrow \\
\{g(c)=h(c)=0, c \in \alpha(\mathcal{A}), c \subseteq a\} \quad a \in \mathcal{A} \tag{6}
\end{gather*}
$$

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 $0 \log \left(\frac{0}{0}\right)=0$ and (5), (6) we obtain

$$
\begin{aligned}
& I(h \| g)=\sum_{c \in \alpha(\mathcal{A})} h(c) \log \left(\frac{h(c)}{g(c)}\right)= \\
& \quad \sum_{c \in \alpha(\mathcal{A})} h(c) \log \left(\prod_{a \in \mathcal{A}} \frac{\lambda_{a}^{c}}{\mu_{a}^{c}}\right)=\sum_{c \in \alpha(\mathcal{A})} h(c) \sum_{a \in \mathcal{A}} \log \left(\frac{\lambda_{a}^{c}}{\mu_{a}^{c}}\right) \\
& =\sum_{a \in \mathcal{A}} \sum_{c \in \alpha(\mathcal{A})} h(c) \log \left(\frac{\lambda_{a}^{c}}{\mu_{a}^{c}}\right)=\sum_{a \in \mathcal{A}} h(a) \log \left(\frac{\lambda_{a}}{\mu_{a}}\right)= \\
& \quad \sum_{a \in \mathcal{A}} g(a) \log \left(\frac{\lambda_{a}}{\mu_{a}}\right)=\sum_{a \in \mathcal{A}} \sum_{c \in \alpha(\mathcal{A})} g(c) \log \left(\frac{\lambda_{a}^{c}}{\mu_{a}^{c}}\right)= \\
& \sum_{c \in \alpha(\mathcal{A})} g(c) \sum_{a \in \mathcal{A}} \log \left(\frac{\lambda_{a}^{c}}{\mu_{a}^{c}}\right)=\sum_{c \in \alpha(\mathcal{A})} g(c) \log \left(\prod_{a \in \mathcal{A}} \frac{\lambda_{a}^{c}}{\mu_{a}^{c}}\right)= \\
& \sum_{c \in \alpha(\mathcal{A})} g(c) \log \left(\frac{h(c)}{g(c)}\right)=-I(g \| h) \leq 0
\end{aligned}
$$

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 $\mathcal{A}, \mathcal{A}_{0}=\mathcal{A} \backslash f$ and suppose that $c \in \gamma(\mathcal{A})$, i.e. $c$ is a minimal set fulfilling

$$
\begin{equation*}
1_{c}=\sum_{a \in \mathcal{A}_{0}} \lambda_{a} 1_{a}+\sum_{b \in f} \lambda_{b} 1_{b} \tag{7}
\end{equation*}
$$

We then need to show that $c \in \mathcal{A}$.
By induction on the number of separators we have $\gamma\left(\mathcal{A}_{0}\right)=\mathcal{A}_{0}$. 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 $\left\{\lambda_{a}\right\}_{a \in f}$ by subtracting the same constant from $\left\{\lambda_{a}\right\}_{a \in g}$. This means that we may assume that

$$
\begin{gather*}
\lambda_{b} \geq 0, \quad b \in f  \tag{8}\\
\lambda_{b_{0}}=0 \quad \text { for some } b_{0} \in f \tag{9}
\end{gather*}
$$

Terminality of $f$ also means that

$$
\begin{gathered}
\alpha\left(\mathcal{A}_{0}\right)=\alpha_{0} \oplus \alpha_{1} \\
\alpha(\mathcal{A})=\alpha_{0} \oplus\left(f \otimes \alpha_{1}\right) \\
\mathrm{s}(f)=\mathrm{s}\left(\alpha_{1}\right)
\end{gathered}
$$

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

$$
\begin{equation*}
1_{c}=\sum_{a \in \alpha_{0}} \mu_{a} 1_{a}+\sum_{a \in \alpha_{1}, b \in f} \mu_{a b} 1_{a \wedge b} \quad \mu_{a} \in\{0,1\}, \quad \mu_{a b} \in\{0,1\} \tag{10}
\end{equation*}
$$

Similarly, any $a \in \mathcal{A}_{0}$ is the sum of its atoms in $\alpha\left(\mathcal{A}_{0}\right)$ whereby (7) writes

$$
\begin{equation*}
1_{c}=\sum_{a \in \alpha_{0} \cup \alpha_{1}} \psi_{a} 1_{a}+\sum_{b \in f} \lambda_{b} 1_{b} \tag{11}
\end{equation*}
$$

Now, comparing (10) and (11) yields

$$
\begin{gathered}
\psi_{a}=\mu_{a} \in\{0,1\}, \quad a \in \alpha_{0} \\
\psi_{a}+\lambda_{b}=\mu_{a b} \in\{0,1\}, \quad a \in \alpha_{1}, \quad b \in f
\end{gathered}
$$

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

$$
1_{c}=1_{d}+\sum_{b \in f} \lambda_{b} 1_{b}, \quad d \in \sigma\left(\mathcal{A}_{0}\right), \quad \lambda_{b} \geq 0
$$

Obviously, this means that $d \subseteq c$. If $d=c$ we must have $\lambda_{b}=0, b \in f$ whereby $c \in \Sigma\left(\mathcal{A}_{0}\right)$ 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 $\bar{H}(\mathcal{A})$, 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 $\mathcal{A}_{0}=\mathcal{A} \backslash f$.
Let $h \in \bar{H}(\mathcal{A})$ and $h_{0}$ the restriction of $h$ to $\sigma\left(\mathcal{A}_{0}\right)$.
Then $h_{0} \in \bar{H}\left(\mathcal{A}_{0}\right)$ and

$$
h(a \wedge b) h(s)=h(a) h_{0}(b) \quad a \in f, b \in \mathrm{~s}(f) \wedge \alpha\left(\mathcal{A}_{0}\right)
$$

Proof. Terminality of $f$ is summarized by the structure

$$
\begin{gather*}
\alpha\left(\mathcal{A}_{0}\right)=\alpha_{0} \oplus \alpha_{1}  \tag{12}\\
\alpha(\mathcal{A})=\alpha_{0} \oplus\left(f \otimes \alpha_{1}\right)  \tag{13}\\
s=\mathrm{s}(f)=\mathrm{s}\left(\alpha_{1}\right) \in \Sigma\left(\mathcal{A}_{0}\right) \tag{14}
\end{gather*}
$$

Let $\left\{\mu_{a} \mid a \in \mathcal{A}\right\}$ be a representation of $h$.
By (13) we have the factorization

$$
\begin{equation*}
h(c \wedge d)=\prod_{a \in f} \mu_{a}^{c \wedge d} \prod_{b \in \mathcal{A}_{0}} \mu_{b}^{c \wedge d} \quad c \in f, \quad d \in \alpha_{1} \tag{15}
\end{equation*}
$$

Consider the number $\mu_{b}^{c \wedge d}, b \in \mathcal{A}_{0}, c \in f, d \in \alpha_{1}$.
Now, if $d \nsubseteq b$ 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

$$
\begin{equation*}
h(c \wedge d)=\prod_{a \in f} \mu_{a}^{c} \prod_{b \in \mathcal{A}_{0}} \mu_{b}^{d} \quad c \in f, d \in \alpha_{1} \tag{16}
\end{equation*}
$$

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

$$
\begin{equation*}
h(d)=\left(\sum_{c \in f} \prod_{a \in f} \mu_{a}^{c}\right) \prod_{b \in \mathcal{A}_{0}} \mu_{b}^{d} \quad d \in \alpha_{1} \tag{17}
\end{equation*}
$$

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 $s \in \Sigma\left(\mathcal{A}_{0}\right)$ 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

$$
\lambda_{a}= \begin{cases}\mu_{a} & a \in \mathcal{A}_{0} \backslash g \\ \lambda_{s} \mu_{a} & a \in g\end{cases}
$$

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

$$
h_{0}(c)=h(c)=\prod_{a \in \mathcal{A}_{0} \backslash g} \mu_{a}^{c}=\prod_{a \in \mathcal{A}_{0} \backslash g} \lambda_{a}^{c}=\prod_{a \in \mathcal{A}_{0}} \lambda_{a}^{c} \quad c \in \alpha_{0}
$$

i.e. $h_{0} \in \bar{H}\left(\mathcal{A}_{0}\right)$.

Next, summing out $d \in \alpha_{1}$ in (16) and noting $c \subseteq \mathrm{~s}\left(\alpha_{1}\right)=s$ yields

$$
\begin{equation*}
h(c)=\prod_{a \in \mathcal{A}} \mu_{a}^{c}\left(\sum_{d \in \alpha_{1}} \prod_{b \in \mathcal{A}_{0}} \mu_{b}^{d}\right) \quad c \in f \tag{18}
\end{equation*}
$$

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

$$
\begin{equation*}
h(s)=\left(\sum_{c \in f} \prod_{a \in f} \mu_{a}^{c}\right)\left(\sum_{d \in \alpha_{1}} \prod_{b \in \mathcal{A}_{0}} \mu_{b}^{d}\right) \tag{19}
\end{equation*}
$$

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 $h \in \bar{H}(\mathcal{A})$ and $\mathcal{A}$ is decomposable with $\left(f_{1}, \ldots, f_{k}\right)$ as a perfect sequence.
If $I=\left\{1<i \leq k \mid h\left(\mathrm{~s}\left(f_{i}\right)\right)>0\right\}$ then we intend to show

$$
\begin{equation*}
h(c)=\frac{\prod_{a \in \mathcal{A}} h(a)^{c}}{\prod_{i \in I} h\left(\mathrm{~s}\left(f_{i}\right)\right)^{c}} \quad c \in \alpha(\mathcal{A}) \tag{20}
\end{equation*}
$$

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 $s=\mathrm{s}\left(f_{k}\right)$ and $\mathcal{A}_{0}=\mathcal{A} \backslash f_{k}$.
Consider the preceeding lemma, by which we have

$$
\begin{align*}
& h(c)=h_{0}(c) \quad c \in \alpha_{0}  \tag{21}\\
& h(c \wedge b) h(s)=h_{0}(c) h(b) \quad c \in \alpha_{1}, b \in f_{k} \tag{22}
\end{align*}
$$

where $\alpha(\mathcal{A})=\alpha_{0} \oplus\left(\alpha_{1} \otimes f_{k}\right)$ and $h_{0} \in \bar{H}\left(\mathcal{A}_{0}\right)$ coinsides with $h$ on $\sigma\left(\mathcal{A}_{0}\right)$.
If $I_{0}=I \backslash\{k\}$ then by the induction hypothesis we obtain

$$
h_{0}(c)=h(c)=\frac{\prod_{a \in \mathcal{A}_{0}} h(a)^{c}}{\prod_{i \in I_{0}} h\left(\mathrm{~s}\left(f_{i}\right)\right)^{c}} \quad c \in \alpha_{0} \oplus \alpha_{1}
$$

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 $\mathrm{s}\left(f_{k}\right)$ 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 $p_{\alpha} \in \bar{M}(\mathcal{A})$ and $I\left(p_{\alpha}\right)=\left\{1 \leq i \leq k \mid p\left(s_{i}\right)>0\right\}$.
Then the likelihood is by theorem 3.2 proportional to

$$
L\left(p_{\alpha} \mid n_{\alpha}\right)=\prod_{c \in \alpha(\mathcal{A})} p(c)^{n(c)}=\frac{\prod_{a \in \mathcal{A}} p(a)^{n(a)}}{\prod_{i \in I\left(p_{\alpha}\right)} p\left(s_{i}\right)^{n\left(s_{i}\right)}} \quad p_{\alpha} \in \bar{M}(\mathcal{A})
$$

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

$$
\begin{equation*}
n^{*}(c)=\frac{\prod_{a \in \mathcal{A}} n(a)^{c}}{\prod_{i \in I} n\left(\mathrm{~s}_{i}\right)^{c}} \quad c \in \alpha(\mathcal{A}) \tag{23}
\end{equation*}
$$

where $I=\left\{1 \leq i \leq k \mid n\left(s_{i}\right)>0\right\}$. We note by section B. 1 that

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

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

$$
p_{\alpha}^{*}=\frac{n_{\alpha}^{*}}{n_{+}} \in \bar{M}(\mathcal{A})
$$

Furthermore, it is easy to verify that

$$
L\left(p_{\alpha} \mid n_{\alpha}\right)=L\left(p_{\alpha} \mid n_{\alpha}^{*}\right)
$$

and as $L\left(p_{\alpha}^{*} \mid n_{\alpha}^{*}\right)>0$, this allows us to consider the alternative likelihood

$$
L^{*}\left(p_{\alpha}\right)=\frac{L\left(p_{\alpha} \mid n_{\alpha}^{*}\right)}{L\left(p_{\alpha}^{*} \mid n_{\alpha}^{*}\right)} \quad p_{\alpha} \in \bar{M}(\mathcal{A}) \cap\left\{p_{\alpha} \mid L^{*}\left(\left(p_{\alpha}\right) \geq 1\right\}\right.
$$

Hence

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

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

$$
\alpha=\alpha(\mathcal{A})=f_{0} \oplus\left(f_{1} \otimes f_{2}\right)
$$

If we let $s_{0}=\mathrm{s}\left(f_{0}\right)$ and $s_{1}=\mathrm{s}\left(f_{1}\right)=\mathrm{s}\left(f_{2}\right)$ then $Z=\left(N\left(s_{0}\right), N\left(s_{1}\right)\right)$ is binomial. If $Y_{i}=\left\{N(c)_{c \in f_{i}}\right\}, i=0,1,2$ and $X=\left\{N(c)_{c \in f_{1} \otimes f_{2}}\right\}$ then $\left(Y_{0}, X\right)$ 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 $\left(Y_{1}, Y_{2}\right)$ conditionally on $Z$. Piecing this together gives the prescribed distribution of $\left(Y_{0}, Y_{1}, Y_{2}\right)$.
So let $k>2$ and $f$ be terminal in $\mathcal{A}, \mathcal{A}_{0}=\mathcal{A} \backslash f, \beta=\alpha\left(\mathcal{A}_{0}\right)=\alpha_{0} \oplus \alpha_{1}$ and $\alpha=\alpha(\mathcal{A})=\alpha_{0} \oplus f \otimes \alpha_{1}$. We then refer to the case $k=2$ to obtain that

$$
\begin{equation*}
P\left(N_{\beta}=n_{\beta}, N_{f}=n_{f}\right)=P\left(N_{\beta}=n_{\beta}\right) \frac{\prod_{a \in f} p(a)^{n(a)}}{p(\mathrm{~s}(f))^{n(\mathrm{~s}(f))}} \tag{26}
\end{equation*}
$$

As $N_{\beta}$ is marginal to $N_{\alpha}$ it is multinomial and by lemma B. 1 we obtain that $p_{\beta} \in \bar{M}\left(\mathcal{A}_{0}\right)$.
By the induction assumption we obtain a factorization of $P\left(N_{\mathcal{A}_{0}}\right)$ which combined with (26) yields the result, when we note that $\left\{N_{A_{0}}, N_{f}\right\}=N_{\mathcal{A}}$.

