Borrower:

Location MATH LIB.
SHELVED BY ~~TITLE~~
LIB. HAS 1(1974)-30(2003)

**ILL.library.orst**

Record 31 of 46

Record 33 of 46 3/4

ILL pe
CAN YOU SUPPLY ? YES NO COND FUTUREDATE
:ILL: 3155759 :Borrower: ORE :ReqDate: 20050301 :NeedBefore: 20050331
:Status: IN PROCESS 20050301 :RecDate: :RenewalReq:
:OCLC: 2243570 :Source: OCLCILL :DueDate: N/A :NewDueDate:
:Lender: *ORU,IYU,TXA,OKS,LDL
:CALLNO: *Lender's OCLC LDR: 1- 1974-
:TITLE: Scandinavian journal of statistics, theory and applications.
:IMPRINT: Stockholm, Almqvist & Wiksell Periodical Co.
:ARTICLE: Jukka Corander: Labelled Graphical Models
:VOL: 30 :NO: 3 :DATE: September 2003 :PAGES: 493-508
:VERIFIED: <TN:119824>OCLC ISSN: 0303-6898 [Format: Serial]
:PATRON: Bulatov, Yaroslav
:SHIP TO:
Library-ILL
Oregon State University
121 The Valley Library
Cross Streets Jefferson Way & Waldo Pl
Corvallis, OR 97331-4501

:BILL TO: same.0005911 ** GWLA MEMBER **

:SHIP VIA: IF POSSIBLE, PLEASE USE ARIEL: OSU-ILL.library.orst.edu
:MAXCOST: $21.25IFM :COPYRT COMPLIANCE: CCG
:FAX: 541-737-1328
:E-MAIL: valley.ill@oregonstate.edu
:BILLING NOTES: BRI usercode 51-0446. FEIN#93-600-1786 --ISI Acct #76636/TGA Acct #61659--CAI ACCT #DD684104--CAS CUSTOMER #1088297-- ****ATTENTION***** NON-USA LIBRARIES, we can not convert to your currency. We can pay via IFLA coupons, IRC coupons, IFM or Visa only. ALL LIBRARIES: Please let us know if you can accept Visa instead of a check for payments. Thanks.
:AFFILIATION: OCLC Western, GWLA (BTP), ORBIS Cascade Alliance

OCLC Number:

**Library-ILL**
**Oregon State University**
**121 The Valley Library**
**Cross Streets Jefferson Way & Waldo Pl**
**Corvallis, OR 97331-4501**

© Board of the Foundation of the Scandinavian Journal of Statistics 2003. Published by Blackwell Publishing Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA Vol 30: 493–508, 2003

# Labelled Graphical Models

JUKKA CORANDER

*University of Helsinki*

**ABSTRACT. A class of log-linear models, referred to as labelled graphical models (LGMs), is introduced for multinomial distributions. These models generalize graphical models (GMs) by employing partial conditional independence restrictions which are valid only in subsets of an outcome space. Theoretical results concerning model identifiability, decomposability and estimation are derived. A decision theoretical framework and a search algorithm for the identification of plausible models are described. Real data sets are used to illustrate that LGMs may provide a simpler interpretation of a dependence structure than GMs.**

*Key words:* Bayesian model determination, graphical models, log-linear models, reference analysis, utility

## 1. Introduction

A widely used approach to modelling multinomial data is the class of log-linear *graphical models* (GMs) introduced by Darroch *et al.* (1980). Theoretical aspects of graphical modelling are comprehensively treated in Whittaker (1990) and Lauritzen (1996). While GMs provide a simple interpretation of the dependence structure among a set of variables, they are restricted to the representation of independence statements that are valid across the complete outcome space. To overcome this restriction *labelled* graphical models (LGMs), which allow different dependence structures in subsets of an outcome space, are introduced here.

Various related approaches which allow dependence structures to vary over subsets of an outcome space have been discussed in the literature. Bock (1986, 1994, 1996) considered entropy clustering of contingency tables, leading to sliced tables with different interaction structures. Højsgaard (1998, 2003a) introduced a general class of context specific independence (CSI) models and Højsgaard (2003b) focused on a subclass of CSI models called split models. Decomposability of the CSI models has also been investigated by Eriksen (1999). The development of the class of CSI models illustrates that some non-hierarchical log-linear models may have a simple interpretation in terms of conditional independencies. Teugels & Van Horebeek (1998a) briefly discussed the idea of allowing the presence or absence of an association between two variables to depend on the values of all other variables under investigation. A related algebraic approach of parametrizing a multinomial distribution in terms of its moments was also introduced in Teugels & Van Horebeek (1998b).

Similar to split models, the class of LGMs is a subclass of the CSI models. While these two subclasses share some properties, they are distinct in an important respect. Namely, a split model is based on a nested separation (split) of the data, such that different models are separately fitted to different subsets. As pointed out by Teugels & Van Horebeek (1998a), this strategy has the disadvantage that when a variable is used in separation of the data, it can no longer be used in a subsequent independence statement. On the contrary, an LGM imposes restrictions to multinomial probabilities in a joint rather than nested fashion. This feature is achieved by labelling the edges of a GM with outcomes of their respective common neighbours, such that conditional on those outcomes the two variables represented by a particular edge are independent. Complementary to the labels imposed to the edges of a GM, the interaction structure corresponding to an LGM may be represented by a collection of coloured

independence graphs associated with mutually exclusive subsets of the outcome space, such that the complete outcome space is the union of these subsets.

There has been a considerable interest recently, from the perspective of the Bayesian paradigm, in the task of determining which GMs are plausible in the light of data (see Dawid & Lauritzen, 1993; Madigan & Raftery, 1994; Madigan & York, 1995; Dellaportas & Forster, 1999; Giudici & Green, 1999; Giudici *et al.*, 1999; Corander, 2003). The class of LGMs poses an additional challenge to model determination for several reasons. For instance, unlike for GMs, the aspect of model identifiability needs to be taken into account, and also, the number of possible models becomes astronomic already for a moderate number of variables. As a solution to the model determination problem we describe a heuristic search algorithm, which utilizes the decision theoretical approach of Corander (2003). When this approach is too demanding computationally, asymptotic criteria, for instance that of Schwarz (1978), may also be used.

The present paper is structured as follows. General properties of various log-linear models for multinomial data are discussed in section 2 and LGMs are introduced in section 3. Section 4 is concerned with model determination strategies, while empirical examples are considered in section 5. Some concluding remarks are given in the final section.

## 2. Log-linear models for multinomial distributions

Let $\Delta$ denote a set of $k$ random variables for which the outcomes are labelled by the integers in the finite set $\mathcal{I}_\delta = \{0,1,\ldots,r_\delta\}$, $\delta \in \Delta$. The joint outcome space of a subset $a \subseteq \Delta$ is denoted by $\mathcal{I}_a = \times_{\delta \in a}\mathcal{I}_\delta$. The models for the joint distribution of $\Delta$ considered in the present paper belong to the exponential family for which a detailed theory is developed in Barndorff-Nielsen (1978) and Brown (1986).

To obtain a minimal parametrization of the joint distribution of $\Delta$, we adopt the convention that the probability of the outcome labelled by 0 is determined by the remaining probabilities for all $\delta \in \Delta$. Let $\mathcal{I}_\delta^*$ denote the outcome set where 0 is excluded, and accordingly, let $\mathcal{I}_a^* = \times_{\delta \in a}\mathcal{I}_\delta^*$. Given an arbitrary element $i \in \mathcal{I}_\Delta$, we write $i_a$ for the outcome of the subset $a \subset \Delta$.

The joint distribution can be characterized either directly by the probabilities $\mathbf{p} = \{p(i), i \in \mathcal{I}_\Delta\}$ or by a parameter vector $\boldsymbol{\theta} \in \Theta$ with subvectors $\boldsymbol{\theta}_a$ corresponding to the $2^k - 1$ non-empty subsets $a$ of $\Delta$ in a lexicographical order. The elements of $\boldsymbol{\theta}_a$, labelled by $\theta_{a(i_a)}, i_a \in \mathcal{I}_a^*$, are coordinate projection functions (or interactions), such that

$$\log p(i) = \sum_{a \subseteq \Delta} \theta_{a(i_a)}, \tag{1}$$

where the elements of $i$ equal to 0 are ignored. This corresponds to the restriction $\boldsymbol{\theta}_{a(i_a)} = 0$, for all $i_a \in \mathcal{I}_a$ such that $i_\delta = 0$ for some $\delta \in a$ (see Whittaker, 1990). The value of $\log p(i)$, say $\boldsymbol{\theta}_\emptyset$, for the case with $i_\delta = 0$, for all $\delta \in \Delta$, is determined by the $|\mathcal{I}_\Delta| - 1$ elements in $\boldsymbol{\theta}$. It is assumed in the sequel that $p(i) > 0$, for all $i \in \mathcal{I}_\Delta$. For any such probability distribution, vector $\boldsymbol{\theta}$ attains finite values in $\mathbb{R}^{|\mathcal{I}_\Delta|-1}$. For a subset $a \subset \Delta$, the marginal probability of $i_a$, $i_a \in \mathcal{I}_a$, is denoted by $p_a(i_a)$. Given the probabilities $\mathbf{p}$, the values of $\boldsymbol{\theta}_{a(i_a)}$ are recursively determined using the elements of $\mathcal{I}_\Delta$ in lexicographical order (see Whittaker, 1990).

The class of models defined through $\boldsymbol{\theta}$ is very general, and one could, for instance, consider exponential models where $\boldsymbol{\theta}$ is restricted to an arbitrary *affine* subspace of $\Theta$ (see Lauritzen, 1996). However, the main problems with such generality are lack of an intuitive interpretation and the difficulty of empirically verifying whether such models are plausible. The class of graphical models $\mathcal{G}$, where the elements are labelled by the $2^{\binom{k}{2}}$ possible undirected graphs on





$\Delta$, is considerably m
is provided here; for
(1996). The terms GM
sequel.

An undirected gra
$E \subseteq \{\Delta \times \Delta\}$, i.e. pair
contains exactly those
is the set $a \subseteq \Delta\backslash\delta$ of
vertices in every succ
lapping subsets $(a, b,$
arbitrary vertices $\delta \in$
corresponding to max
broken into a sequen
*separators* $\mathcal{S}(G)$ of a d
For a decomposable g
distributions of the cl

Under the current a
of a GM are all equiv

Pairwise Markov
$\delta \perp \gamma|\Delta\backslash\{\delta,\gamma\} \Leftrightarrow$
Local Markov pr
$\delta \perp \Delta\backslash\mathrm{bd}(\delta)|\mathrm{bd}(\delta$
Global Markov p
$a \perp b|s$, for all di

The log-linear param
$\boldsymbol{\theta}_{a(i_a)} = 0$ whenever $\{\delta$
corresponding subspa
(1996). If $G$ is decomp
function of the margi

$$p(i) = \frac{\prod_{c \in \mathcal{C}(G)} p_c(}{\prod_{s \in \mathcal{S}(G)} p_s(}$$

Consequently, the ma
decomposable model
frequencies) with re
decomposable model
projection, in particu
subspace $\Theta_G$ can b
decomposable models

Although models i
restrictions on $\boldsymbol{\theta}$, as
statements in a mor
models with the fea
specific. Let $a$ and $b$
letting the pair $(i_b, a$
be written as

outcome space, such
e of the Bayesian pa-
of data (see Dawid &
ellaportas & Forster,
class of LGMs poses
r instance, unlike for
and also, the number
er of variables. As a
arch algorithm, which
this approach is too
warz (1978), may also
ous log-linear models
d in section 3. Section
nples are considered in
elled by the integers in
et $a \subseteq \Delta$ is denoted by
e present paper belong
rndorff-Nielsen (1978)
e adopt the convention
emaining probabilities
and accordingly, let
outcome of the subset
by the probabilities
corresponding to the
ents of $\boldsymbol{\theta}_a$, labelled by
h that
(1)
estriction $\boldsymbol{\theta}_{a(i_a)} = 0$, for
alue of $\log p(i)$, say $\boldsymbol{\theta}_\emptyset$,
nents in $\boldsymbol{\theta}$. It is assumed
distribution, vector $\boldsymbol{\theta}$
ability of $i_a$, $i_a \in \mathcal{I}_a$, is
sively determined using
., for instance, consider
ce of $\Theta$ (see Lauritzen,
intuitive interpretation
plausible. The class of
e undirected graphs on



$\Delta$, is considerably more tractable in both these respects. Only a brief treatment of GMs is provided here; for further details and explanations, see Whittaker (1990) or Lauritzen (1996). The terms GM and (conditional independence) graph are used interchangeably in the sequel.

An undirected graph $G = G(\Delta, E)$ contains the set of vertices $\Delta$ and the set of edges $E \subseteq \{\Delta \times \Delta\}$, i.e. pairs of elements from $\Delta$. For a subset $a \subset \Delta$, a *subgraph* $G_a = G(a, E_a)$ of $G$ contains exactly those edges $\{\delta, \gamma\} \in E$ for which $\{\delta, \gamma\} \subset a$. The *boundary* $\mathrm{bd}(\delta)$ of a vertex $\delta$ is the set $a \subseteq \Delta \backslash \delta$ of vertices adjacent to $\delta$. A *path* is a sequence of vertices of $G$ such that vertices in every successive pair are adjacent. A subset of vertices $s$ in a triple of non-overlapping subsets $(a, b, s)$ is said to *separate* subsets $a$ and $b$, when any path in $G$ between two arbitrary vertices $\delta \in a$, $\gamma \in b$ intersects $s$. The *cliques* $\mathcal{C}(G)$ of a graph $G$ are the subsets of $\Delta$ corresponding to maximal complete subgraphs of $G$. A graph is *decomposable* when it can be broken into a sequence of its cliques by certain basic operations. From this sequence, the *separators* $\mathcal{S}(G)$ of a decomposable graph can be obtained as intersections of successive cliques. For a decomposable graph $G$ there exists a unique factorization of $p(i)$ in terms of the marginal distributions of the cliques and the separators (see below).

Under the current assumption of the positivity of each $p(i)$, the following Markov properties of a GM are all equivalent (here $\perp$ denotes conditional independence).

Pairwise Markov property:
$\delta \perp \gamma | \Delta \backslash \{\delta, \gamma\} \Leftrightarrow \{\delta, \gamma\} \notin E$, for all $\{\delta, \gamma\} \subset \Delta$.
Local Markov property:
$\delta \perp \Delta \backslash \mathrm{bd}(\delta) | \mathrm{bd}(\delta)$, for all $\delta \in \Delta$.
Global Markov property:
$a \perp b | s$, for all disjoint subsets $(a, b, s)$ of $\Delta$ such that $s$ separates $a$ from $b$.

The log-linear parametrization for a GM is determined through the cliques $\mathcal{C}(G)$ by setting $\boldsymbol{\theta}_{a(i_a)} = 0$ whenever $\{\delta, \gamma\} \subseteq a$ and the edge $\{\delta, \gamma\}$ is absent in $G$. The dimensionality of the corresponding subspace $\Theta_G$ of $\Theta$ can be calculated using the formulas given in Lauritzen (1996). If $G$ is decomposable, the probability $p(i)$ can also be directly written as the following function of the marginal probabilities of the clique and separator subsets

$$p(i) = \frac{\prod_{c \in \mathcal{C}(G)} p_c(i_c)}{\prod_{s \in \mathcal{S}(G)} p_s(i_s)} \tag{2}$$

Consequently, the maximum likelihood estimate of $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}(G)$, is explicitly available for a decomposable model and it corresponds to a projection of the empirical probabilities (relative frequencies) with respect to the affine subspace $\Theta_G$ (see Lauritzen, 1996). For non-decomposable models there are several iterative methods available for obtaining the projection, in particular Rudas (1998) and Corander (2003) utilized the fact that the affine subspace $\Theta_G$ can be represented as an intersection of the affine subspaces of certain decomposable models in which $G$ is nested (see also Csiszar, 1975).

Although models in the class $\mathcal{G}$ have many useful properties, they impose rather strong restrictions on $\boldsymbol{\theta}$, as discussed in the previous section. To present conditional independence statements in a more flexible way, Højsgaard (1998) introduced the general class of CSI models with the feature that dependence (or interaction) of variables can be outcome specific. Let $a$ and $b$ be disjoint subsets of $\Delta$. Using the notation of Højsgaard (1998), by letting the pair $(i_b, a)$ to represent an element in the *generating class* $\mathcal{D}$, a CSI model can be written as

$$\log p(i) = \sum_{(i_b,a)\in\mathcal{D}} \psi^{i_b}_{a(i_a)} \tag{3}$$

where $\psi^{i_b}_{a(i_a)}$ is a real number, equal to zero for all $j \in \mathcal{I}_\Delta$ with $j_b \neq i_b$. For details on model representation through generating classes, see Lauritzen (1996). From (3) it is seen that the interaction terms are outcome (context) specific and vanish outside the subsets $i_b$. A GM is a special case of a CSI model where the index set $b$ is empty for all terms in $\mathcal{D}$. The class of CSI models is extremely general and enables specification of a wide range of conditional independence hypotheses. On the other hand, the price of this generality is that model identifiability and empirical plausibility are difficult to assess (by model non-identifiability we mean the existence of at least two different generator classes leading to the same affine parameter subspace).

Højsgaard (1998) also introduced a subclass of CSI models which he called split models. They admit a graphical representation of the interaction structure in terms of a collection of graphs embedded in a tree structure, where the absence of an edge corresponds to a CSI in each graph. The idea of imposing the CSI constraints on a GM rather than on an unrestricted distribution **p** considerably simplifies the empirical work with the models, and is utilized here as well. However, as split models are defined through deletion of edges of a GM and through splits of the data to subsets where the additional parameter restrictions are imposed separately, they do not always enable representation of the CSI restrictions in a joint manner. As noted earlier, when a variable is used in separation of the data in a split model, it can no longer be used in a subsequent independence statement. To overcome such restrictions while retaining as much simplicity of interpretation as possible, a subclass of CSI models referred to as LGMs is introduced in the next section.

## 3. Labelled graphical models

We start by specifying the class $\mathcal{G}_L$ of LGMs in the following definition.

**Definition 1**
*Let $G(\Delta, E)$ be a GM for $\Delta$. For all $\{\delta, \gamma\} \in E$, let $L_{\{\delta,\gamma\}}$ denote the set of vertices adjacent to both $\delta$ and $\gamma$. For a non-empty $L_{\{\delta,\gamma\}}$, define the label of the edge $\{\delta, \gamma\}$ as the set $\mathcal{L}_{\{\delta,\gamma\}} = \{i_{L_{\{\delta,\gamma\}}} \in \mathcal{I}_{L_{\{\delta,\gamma\}}} : \delta \perp\!\!\!\perp \gamma | L_{\{\delta,\gamma\}} = i_{L_{\{\delta,\gamma\}}}\}$ of all outcomes of $L_{\{\delta,\gamma\}}$ for which $\delta$ and $\gamma$ are conditionally independent. The graph $G$ where the edges are labelled with $\mathcal{L}_{\{\delta,\gamma\}}$ is called a labelled graphical model $G_L$ for $\Delta$.*

An LGM is a GM when the label sets are all empty, as no additional independence constraints are then imposed on the variables, or when the label sets equal $\mathcal{I}_{L_{\{\delta,\gamma\}}}$, in which case the conditional independence statements hold in the complete outcome space. The additional constraints imposed by a $G_L$ are called *partial* conditional independence ($\perp\!\!\!\perp$) constraints. The labelled models define a subclass of CSI models by first allowing the generator class to involve only cliques of a graph, and then allowing the exclusion of certain interactions.

To visualize the constraints imposed by an LGM, one can utilize two distinct approaches. First, the integer vectors in the set $\mathcal{L}_{\{\delta,\gamma\}}$ can be attached to the edges of $G_L$. Notice that given a fixed ordering of the variables, the label $i_{L_{\{\delta,\gamma\}}}$ need not contain the variable indices, as $L_{\{\delta,\gamma\}}$ contains all variables adjacent to both $\delta$ and $\gamma$ and is therefore visible in the graph. For large label sets this may be impractical, and alternatively, one can use *coloured* graphs to represent LGMs. In general, in coloured graphs any two vertices may be adjacent in several ways. One can consider the adjacency of a pair $\{\delta, \gamma\}$ in terms of a variable $Y_{\{\delta,\gamma\}}$ taking values labelled 0, ..., $r$, which are called colours. Typically, the colour with the zero-label corresponds to a





non-adjacent pair of ve
(corresponding to adja

For each $i \in \mathcal{I}_\Delta$, def
one of three possible c
that $\{\delta, \gamma\}$ is not in t
adjacent in $G^i_L$). For
$i_{L_{\{\delta,\gamma\}}} \in \mathcal{L}_{\{\delta,\gamma\}}$, the edg
notation, vertex pairs
ordinary GMs, where
straints. An LGM ma
corresponding to the

$$G^i_L = G^j_L, \text{ for all}$$

$$G^i_L \neq G^j_L, \text{ for all}$$

$$\mathcal{I}_\Delta = \bigcup_{l=1}^{m} \mathcal{I}^l \text{ and } \mathcal{I}$$

Interpretation of thes
Markov properties of
statements for edges
definition of an LGM
$p(i)$ as the underlying
independencies from
lead to such interpr
constraints imposed b

**Proposition 1**
*Affine subspaces of $\Theta$
variables with non-zer
imposed by the label a
outcome of $i_a$.*

*Proof.* See the appe

Maximum likelihoo
Højsgaard (1998, 200
However, an alternati
Corander (2003), such
subspaces correspond
conditional independe
responding $m$ subsets

To illustrate LGM
binary variables are g
coloured graphs are v
$G_L$ satisfies the condit
not jointly imply ma

non-adjacent pair of vertices. In this notation, ordinary graphs allow only two possible colours (corresponding to adjacency and non-adjacency) for each pair of vertices.

For each $i \in \mathcal{I}_\Delta$, define the coloured graph $G_L^i$ on $\Delta$ as the graph where each pair $\{\delta, \gamma\}$ has one of three possible colours (say $C_0$, $C_1$ and $C_2$) according to the following. All $\{\delta, \gamma\}$ such that $\{\delta, \gamma\}$ is not in the edge set of the underlying GM $G$, attain colour $C_0$ (i.e. are non-adjacent in $G_L^i$). For all $i$ and $\{\delta, \gamma\}$ with non-empty $L_{\{\delta,\gamma\}}$, such that $i$ contains a label $i_{L_{\{\delta,\gamma\}}} \in \mathcal{L}_{\{\delta,\gamma\}}$, the edge $\{\delta, \gamma\}$ in $G_L^i$ has the colour $C_1$, and colour $C_2$ otherwise. In this notation, vertex pairs associated with colours $C_0$ and $C_2$ have the same interpretation as in ordinary GMs, whereas colour $C_1$ represents the additional conditional independence constraints. An LGM may then be represented by the finite collection of distinct coloured graphs corresponding to the $m$ disjoint subsets $\mathcal{I}^1,\ldots,\mathcal{I}^m$ of $\mathcal{I}_\Delta$, where

$$G_L^i = G_L^j, \text{ for all } \{i,j\} \subset \mathcal{I}^l,$$

$$G_L^i \neq G_L^j, \text{ for all } i \in \mathcal{I}^l, j \in \mathcal{I}^{l^*},$$

$$\mathcal{I}_\Delta = \bigcup_{l=1}^{m} \mathcal{I}^l \text{ and } \mathcal{I}^l \cap \mathcal{I}^{l^*} = \emptyset, \text{ for all index pairs } \{l, l^*\}.$$

Interpretation of these coloured graphs is facilitated by the fact that they share the global Markov properties of the underlying GM and satisfy the additional conditional independence statements for edges having the colour $C_1$. The global Markov property follows from the definition of an LGM, as it induces the same *global* conditional independence constraints on $p(i)$ as the underlying GM. However, edge colouring is necessary for distinguishing conditional independencies from marginal independence, as the absence of the edges with colour $C_1$ might lead to such interpretations (see examples below). The form of additional parameter constraints imposed by an LGM is given in the following proposition.

**Proposition 1**
*Affine subspaces of $\Theta$ induced by LGMs. Let $a \subseteq (L_{\{\delta,\gamma\}} \cup \{\delta, \gamma\})$ contain $\{\delta, \gamma\}$ and all the variables with non-zero outcomes in a label $i_{L_{\{\delta,\gamma\}}}$ of $\{\delta, \gamma\}$. The $(r_\delta - 1)(r_\gamma - 1)$ restrictions imposed by the label are of the form $\sum_{b \subseteq a} \theta_{b(i_b)} = 0$, such that $\{\delta, \gamma\} \subseteq b$ and $i_b$ is a marginal outcome of $i_a$.*

*Proof.* See the appendix.

Maximum likelihood estimation of CSI model parameters has been considered in detail by Højsgaard (1998, 2003a) and Eriksen (1999), and the results may be utilized for LGMs. However, an alternative estimation scheme can be based on the approach of Rudas (1998) and Corander (2003), such that the affine subspace $\Theta_{G_L}$ is represented as an intersection of affine subspaces corresponding to the $m$ graphs with coloured edges. In practice this means that the conditional independence restrictions given by each graph $G_L^i$ are cyclically fitted to the corresponding $m$ subsets of $\mathcal{I}_\Delta$ until a convergence criterion is satisfied.

To illustrate LGMs, a variety of different models based on the complete graph for three binary variables are given in Table 1. The corresponding graphs with labelled edges and the coloured graphs are visualized in Figs 1 and 2, respectively. It is essential to notice that while $G_L$ satisfies the conditional independence restrictions according to all $G_L^i$, these restrictions do not jointly imply marginal independence. That is, the partial conditional independencies

Table 1. *Examples of LGMs based on the complete graph for three binary variables, here # $G_L$ denotes the number of distinct models with a particular number of labelled edges*

| # Labelled edges | Labelled edges | Labels | # $G_L$ | Constraints on $\boldsymbol{\theta}$ |
|---|---|---|---|---|
| 1 | {2, 3} | $i_1 = 0$ | 6 | $\boldsymbol{\theta}_{23(i_{23})} = 0$ |
| 2 | {1, 2} | $i_3 = 1$ | 12 | $\boldsymbol{\theta}_{123(i_{123})} = -\boldsymbol{\theta}_{12(i_{12})}$ |
| | {2, 3} | $i_1 = 0$ | | $\boldsymbol{\theta}_{23(i_{23})} = 0$ |
| 3 | {1, 2} | $i_3 = 1$ | 8 | $\boldsymbol{\theta}_{123(i_{123})} = -\boldsymbol{\theta}_{12(i_{12})}$ |
| | {1, 3} | $i_2 = 0$ | | $\boldsymbol{\theta}_{13(i_{13})} = 0$ |
| | {2, 3} | $i_1 = 0$ | | $\boldsymbol{\theta}_{23(i_{23})} = 0$ |

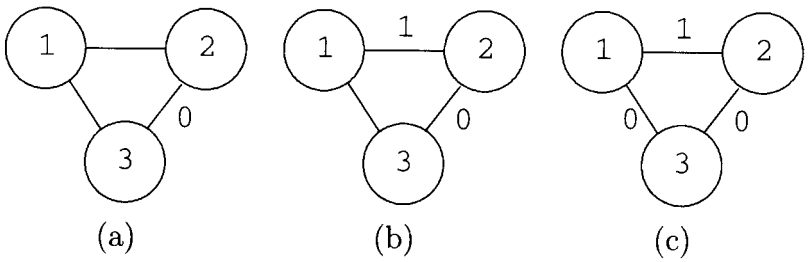



*Fig. 1.* LGMs for the example models in Table 1. Parts (a), (b) and (c) correspond to the cases with 1, 2 and 3 labels, respectively.

$\delta \perp \gamma|\sigma$ and $\delta \perp \sigma|\gamma$ do not imply $\delta \perp \{\gamma, \sigma\}$. For instance, the model in Fig. 1c where all edges are labelled, does not imply that $p(i_{123} = 001) = p(i_1 = 0)p(i_2 = 0)p(i_3 = 1)$, although all three partial conditional independencies are valid for the outcome $i_{123} = 001$. Deletion of the edges of the graphs $G_L^i$ (see Fig. 2c) instead of colouring would therefore occasionally lead to wrong conclusions about marginal independence.

The number of possible LGMs increases very rapidly with sizes of the cliques of a GM and the numbers of outcomes $r_\delta$. For instance, consider labellings of a complete graph on $k$ variables where each label set contains only a single outcome. Assuming further that the number of possible outcomes is equal for all variables (say $r$) the number of distinct labellings can be expressed as $\sum_{i=1}^{k(k-1)/2} \binom{k(k-1)/2}{i} r^{(k-2)i}$. However, in general, all labellings do not correspond to meaningful parameter constraints. Consider, for instance, the graph $G$ on a set of four binary variables, with the cliques {1, 2, 3} and {1, 3, 4}. The four potential labels of the edge {1, 3} and the corresponding parameter constraints are given in Table 2. As $\boldsymbol{\theta}_{1234(i_{1234})}$ is restricted to zero by $G$, the choice of labelling $\mathcal{L}_{\{1,3\}} = \{(00), (01), (10)\}$ leads to the same parametric model as the graph with the cliques {1,2}, {2,3}, {3,4} and {1,4}.

As there exist in general several generator classes for a CSI model leading to the same affine subspace, the interpretation of the conditional independence constraints may be ambiguous (as illustrated in the above example). To facilitate the interpretation of LGMs we introduce two conditions, which ensure that the affine subspace $\Theta_{G_L}$ does not coincide with the affine subspace according to any GM or another LGM with the same or a larger number of edge labels. The *regularity* condition ensures that none of the edges of a graph can be completely removed by the inclusion of the label constraints. In particular, this means that for a regular LGM $\mathcal{L}_{\{\delta,\gamma\}}$must always be distinct from $\mathcal{I}_{L_{\{\delta,\gamma\}}}$. The *maximality* condition in turn guarantees that no additional labels can be imposed to the edges without reducing the dimension of the affine subspace $\Theta_{G_L}$.





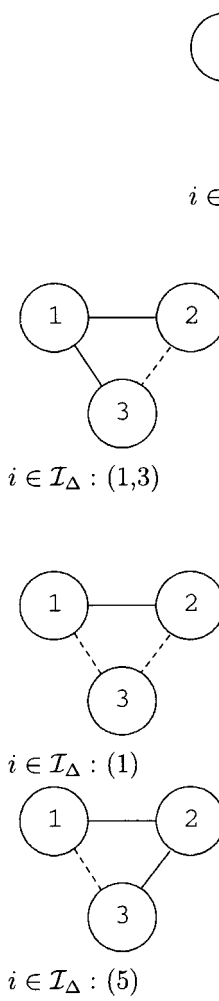


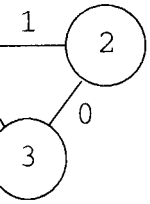

*Fig. 2.* Coloured indepen
to the cases with 1, 2 and
eight outcomes are labell
(000), (001),..., (111).

Table 2. *The f*
*constraints for t*
*variables*

| Label |
|---|
| $i_{24} = (00)$ |
| $i_{24} = (01)$ |
| $i_{24} = (10)$ |
| $i_{24} = (11)$ |

*ry variables, here*
*labelled edges*

| Constraints on $\boldsymbol{\theta}$ |
|---|
| $_{23(i_{23})} = 0$ |
| $_{123(i_{123})} = -\boldsymbol{\theta}_{12(i_{12})}$ |
| $_{23(i_{23})} = 0$ |
| $_{123(i_{123})} = -\boldsymbol{\theta}_{12(i_{12})}$ |
| $_{13(i_{13})} = 0$ |
| $_{23(i_{23})} = 0$ |




(c)

ond to the cases with 1, 2

Fig. 1c where all edges
= 1), although all three
. Deletion of the edges
asionally lead to wrong

ne cliques of a GM and
complete graph on $k$
uming further that the
ber of distinct labellings

ral, all labellings do not
e, the graph $G$ on a set
four potential labels of
in Table 2. As $\boldsymbol{\theta}_{1234(i_{1234})}$
(10)} leads to the same
1 {1,4}.
ading to the same affine
nts may be ambiguous
of LGMs we introduce
coincide with the affine
larger number of edge
raph can be completely
neans that for a regular
tion in turn guarantees
ng the dimension of the



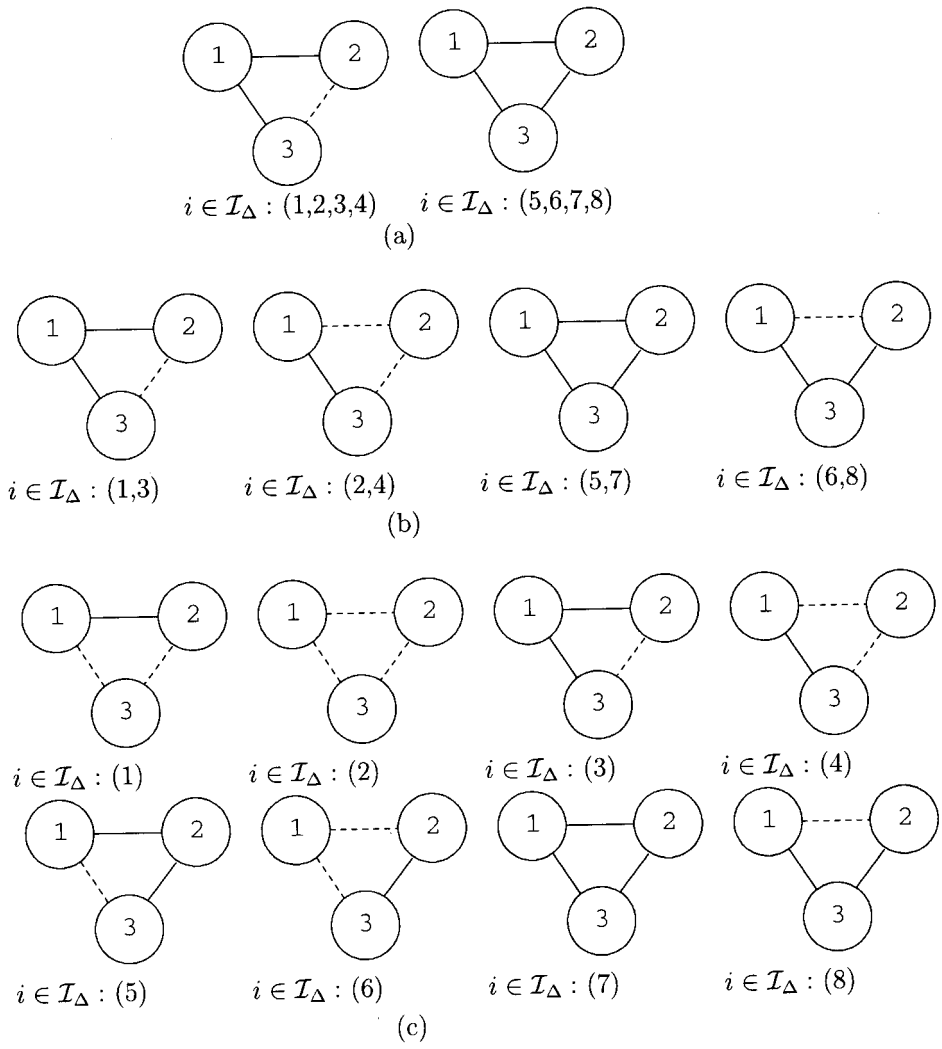



*Fig. 2.* Coloured independence graphs for the example models in Table 1. Parts (a), (b) and (c) correspond to the cases with 1, 2 and 3 labels, respectively. The dashed line indicates conditional independence and the eight outcomes are labelled according to a lexicographical order of the binary vectors $i = (i_1, i_2, i_3)$, i.e. (000), (001),..., (111).

Table 2. *The four possible labels of the edge* {1, 3} *and the corresponding parameter constraints for the graph G with cliques* {1, 2, 3} *and* {1, 3, 4}*, on a set of four binary variables*

| Label | Constraints on $\boldsymbol{\theta}$ |
|---|---|
| $i_{24} = (00)$ | $\boldsymbol{\theta}_{13(i_{13})} = 0$ |
| $i_{24} = (01)$ | $\boldsymbol{\theta}_{134(i_{134})} = -\boldsymbol{\theta}_{13(i_{13})}$ |
| $i_{24} = (10)$ | $\boldsymbol{\theta}_{123(i_{123})} = -\boldsymbol{\theta}_{13(i_{13})}$ |
| $i_{24} = (11)$ | $\boldsymbol{\theta}_{1234(i_{1234})} = -(\boldsymbol{\theta}_{123(i_{123})} + \boldsymbol{\theta}_{134(i_{134})} + \boldsymbol{\theta}_{13(i_{13})})$ |

**Definition 2**
*Regular LGMs. An LGM is called regular if, for no edge $\{\delta, \gamma\}$ in $G$, the edge labels induce an affine subspace $\Theta_{G_L}$ such that $\boldsymbol{\theta}_{a(x_a)} = 0$ whenever $\{\delta, \gamma\} \subseteq a$.*

**Definition 3**
*Maximal regular LGM. A regular LGM is called maximal if, for no edge $\{\delta, \gamma\}$ in $G$, additional labels can be imposed without reducing the dimension of the affine subspace $\Theta_{G_L}$.*

To illustrate the maximality condition, an example with five binary variables is given in Table 3. For instance, the LGM with all labels leads to the same parameter constraints as the two LGMs where all labels except the last one for either of the two edges are chosen. Similarly, the model with the first two labels for both edges corresponds to the two models where the first two labels are put on one edge and only the first label on the other edge.

Determination of the maximality of an LGM is simplified by noting that addition of the label $i_{L_{\{\delta,\gamma\}}}$ does not impose a new constraint on $\boldsymbol{\theta}$ if all terms in the sum $\sum_{b \subseteq a} \boldsymbol{\theta}_{b(i_b)} = 0$ are already constrained. This happens when for all $b \subseteq a$ there is an edge $\{\sigma, \tau\}$ in $G_L$ with a label including the elements of $i_{b\setminus\{\sigma,\tau\}}$, or when $b = \{\sigma, \tau\} = \{\delta, \gamma\}$, the label $i_{L_{\{\delta,\gamma\}}} = (0,\ldots,0)$ is in $\mathcal{L}_{\{\delta,\gamma\}}$. The dimension of the affine subspace $\Theta_{G_L}$ can be determined from the dimension of $\Theta_G$ by imposing the label constraints sequentially and controlling after each new label outcome whether further model reduction is achieved. Proposition 1 shows that a label can induce at most $(r_\delta - 1)(r_\gamma - 1)$ new restrictions. For instance, the maximal model with all labels in Table 3, imposes six constraints on $\boldsymbol{\theta}$. The next proposition formalizes the consequences of the regularity and maximality concepts, i.e. the affine subspaces of two maximal regular LGMs are always distinct.

**Proposition 2**
*The affine subspace of $\Theta$ induced by a maximal regular LGM is unique, such that $\Theta_{G_L} \neq \Theta_{G'_L}$ for all distinct $G$, $G'$ with arbitrary labellings $\mathcal{L}(G_L), \mathcal{L}(G'_L)$, and $\Theta_{G_L} \neq \Theta_{G_{L'}}$ for all distinct maximal labellings $\mathcal{L}(G_L), \mathcal{L}'(G_L)$ with an arbitrary $G$.*

*Proof.* Let $G$ and $G'$ differ in at least one edge, say $\{\delta, \gamma\}$, such that $\{\delta, \gamma\}$ is absent in $G'$. All $\boldsymbol{\theta}_{a(i_a)}$ where $\{\delta, \gamma\} \subseteq a$ are thereby restricted to zero in $\Theta_{G'_L}$, whereas at least one of these terms is by the regularity condition left unrestricted in $\Theta_{G_L}$ for an arbitrary labelling. To prove the latter part of the proposition, assume $\Theta_{G_L} = \Theta_{G_{L'}}$ for two distinct labellings $\mathcal{L}(G_L)$, $\mathcal{L}'(G_L)$ of an arbitrary $G$. Then, both models impose exactly the same restrictions on $\boldsymbol{\theta}$, but neither of the

Table 3. *Some edge labels and the corresponding parameter constraints for the complete graph on a set of five binary variables*

| | Constraints on $\boldsymbol{\theta}$ |
|---|---|
| Label for edge {1, 3} | |
| $i_{245} = (000)$ | $\boldsymbol{\theta}_{13(i_{13})} = 0$ |
| $i_{245} = (100)$ | $\boldsymbol{\theta}_{123(i_{123})} = -\boldsymbol{\theta}_{13(i_{13})}$ |
| $i_{245} = (010)$ | $\boldsymbol{\theta}_{134(i_{134})} = -\boldsymbol{\theta}_{13(i_{13})}$ |
| $i_{245} = (110)$ | $\boldsymbol{\theta}_{1234(i_{1234})} = -(\boldsymbol{\theta}_{123(i_{123})} + \boldsymbol{\theta}_{134(i_{134})} + \boldsymbol{\theta}_{13(i_{13})})$ |
| Label for edge {2, 3} | |
| $i_{145} = (000)$ | $\boldsymbol{\theta}_{23(i_{23})} = 0$ |
| $i_{145} = (100)$ | $\boldsymbol{\theta}_{123(i_{123})} = -\boldsymbol{\theta}_{23(i_{23})}$ |
| $i_{145} = (010)$ | $\boldsymbol{\theta}_{234(i_{234})} = -\boldsymbol{\theta}_{23(i_{23})}$ |
| $i_{145} = (110)$ | $\boldsymbol{\theta}_{1234(i_{1234})} = -(\boldsymbol{\theta}_{123(i_{123})} + \boldsymbol{\theta}_{234(i_{234})} + \boldsymbol{\theta}_{23(i_{23})})$ |





two is maximal since
dimension of the affin

**4. Model determinatio**

The Bayesian approac
the recent statistical lit
been proposed for ass
random variables. Ho
*et al.*, 2001) or a comp
may be inefficient or
algorithm in a decision
a similar approach, w
tional burden is prohi

To tackle the mod
(decomposability of L

1. Find the most plau
2. Given the models
among the most pl
3. For each of the m
plausible decompos
4. Given the models f

By definition 1, edge
the number of candid
suggest the use of a pr
found plausible, none

To apply the mode
plausibility is judged.
criteria based on the
according to

$$\log L(\hat{\boldsymbol{\theta}}(G_L)) - c($$

For instance, common
Hannan & Quinn (197
$|\Theta_{G_L}|(\log n)/2$ and $|\Theta_{G}$
in $\Theta_{G_L}$. As a more elab
criterion introduced in
$n$ exchangeable obser
expected logarithmic u

$$\bar{u}(G_L|\mathbf{x}) = n \int \left[ \sum_{i \in} \right.$$

where $q(i|G_L)$ is the pr
the posterior of $\mathbf{p}$. Wh
interpreted as an asyn
observed data.

Let $\boldsymbol{\alpha}_\Delta$ be a vector
denote the observed c

two is maximal since the labels of the other model can be added without reducing the dimension of the affine subspace.

## 4. Model determination

The Bayesian approach to empirical learning of model structures has attained vivid interest in the recent statistical literature. The Markov chain Monte Carlo (MCMC) strategy has widely been proposed for assessment of the graph structure underlying dependencies among a set of random variables. However, for large scale problems with many variables (see Myllymäki *et al.*, 2001) or a complicated model structure (as in the present case), the MCMC approach may be inefficient or very difficult to implement. Corander (2003) used a heuristic search algorithm in a decision theoretical context to learn GM structures from data. Here we describe a similar approach, where asymptotic approximations may also be used when the computational burden is prohibitive.

To tackle the model assessment problem we suggest the following four-step strategy (decomposability of LGMs is defined below).

1. Find the most plausible decomposable GMs.
2. Given the models from step 1, check whether any non-decomposable GMs should be among the most plausible models.
3. For each of the most plausible models according to steps 1 and 2, identify the most plausible decomposable LGMs.
4. Given the models from step 3, identify the most plausible non-decomposable LGMs.

By definition 1, edges can only be labelled in cliques with at least three variables, which reduces the number of candidate models considerably. To reduce further the number of models, we suggest the use of a principle of parsimony, such that whenever a model with label set $\mathcal{L}_{\{\delta,\gamma\}}$ is found plausible, none of the models with a label set $\mathcal{L}'_{\{\delta,\gamma\}} \subset \mathcal{L}_{\{\delta,\gamma\}}$ is considered plausible.

To apply the model search strategy sketched above, it has to be decided how a model's plausibility is judged. A relatively simple approach is to use asymptotic model determination criteria based on the maximum likelihood estimate $\hat{\boldsymbol{\theta}}(G_L)$ and a penalty term $c(G_L, n)$, according to

$$\log L(\hat{\boldsymbol{\theta}}(G_L)) - c(G_L, n) \tag{4}$$

For instance, commonly used criteria of this type are those introduced in Schwarz (1978), and Hannan & Quinn (1979). The penalty terms for the Schwarz and Hannan–Quinn criteria equal $|\Theta_{G_L}|(\log n)/2$ and $|\Theta_{G_L}| \log\log n$, respectively, where $|\Theta_{G_L}|$ is the number of free parameters in $\Theta_{G_L}$. As a more elaborate way of judging the plausibilities, we extend the GM determination criterion introduced in Corander (2003) to LGMs. According to this criterion, given a set $\mathbf{x}$ of $n$ exchangeable observations $i_1,\ldots,i_n$, the plausibility of $G_L$ is measured by the posterior expected logarithmic utility

$$\bar{u}(G_L|\mathbf{x}) = n \int \left[ \sum_{i \in \mathcal{I}_\Delta} p(i) \log q(i|G_L) \right] \pi(\mathbf{p}|\mathbf{x}) d\mathbf{p} - c(G_L, n) \tag{5}$$

where $q(i|G_L)$ is the projection of $p(i)$ according to the affine subspace $\Theta_{G_L}$, and $\pi(\mathbf{p}|\mathbf{x})$ denotes the posterior of $\mathbf{p}$. When the posterior shrinks towards $\hat{\boldsymbol{\theta}}$ for $n \to \infty$, the criterion (4) can be interpreted as an asymptotic approximation to the expected utility of a model to explain the observed data.

Let $\boldsymbol{\alpha}_\Delta$ be a vector of constants $(\alpha(i),\ i \in \mathcal{I}_\Delta)$, and let $n_\Delta = (n(i) = \sum_{j=1}^n I(i_j = i), i \in \mathcal{I}_\Delta)$ denote the observed counts of the different outcomes. Assume the prior distribution for the

probabilities of $i \in \mathcal{I}_\Delta$ is the Dirichlet ($\boldsymbol{\alpha}_\Delta$) distribution. The corresponding posterior is then Dirichlet ($\boldsymbol{\alpha}_\Delta + n_\Delta$). The choice of $\alpha(i)$ equal to $1/|\mathcal{I}_\Delta|$ for all $i \in \mathcal{I}_\Delta$, leads to a prior which is vague and symmetric with respect to marginalization. Analogous to Corander (2003), we use a reference approach with this particular prior, penalty term $|\Theta_{G_L}|$ log log $n$, and *relative* expected utilities of the models

$$\bar{u}(G_L|\mathbf{x})^* = \frac{\exp\{\bar{u}(G_L|\mathbf{x})\}}{\sum_{G_L \in \mathcal{G}_L^*} \exp\{\bar{u}(G_L|\mathbf{x})\}} \tag{6}$$

Here $\mathcal{G}_L^*$ is a subclass of $\mathcal{G}_L$ when $\mathcal{G}_L$ is too large to allow for exhaustive enumeration in model determination. An LGM may be considered plausible when (6) exceeds some boundary, such as 0.05 (cf. Madigan & Raftery, 1994).

Eriksen (1999) investigated decomposability conditions of the CSI models under which the maximum likelihood estimate is available in a closed form. We use the term *strong decomposability* (definition 4) to refer to LGMs for which the posterior expected utility is available analytically. For a strongly decomposable LGM, the formula (2) is modified by replacing the probability $p_c(i_c)$ by

$$\frac{p_{c\setminus\{\delta\}}(i_{c\setminus\{\delta\}})p_{c\setminus\{\gamma\}}(i_{c\setminus\{\gamma\}})}{p_{c\setminus\{\delta,\gamma\}}(i_{c\setminus\{\delta,\gamma\}})}$$

whenever $\{\delta, \gamma\} \subset c$ and $i_{c\setminus\{\delta,\gamma\}} \in \mathcal{L}_{\{\delta,\gamma\}}$.

**Definition 4**

*Strongly decomposable LGMs. An LGM is called strongly decomposable if (1) the underlying graph G is decomposable, and (2) no clique contains more than one labelled edge, and (3) no separator contains any labelled edges.*

The expected logarithmic utility of a decomposable GM has an analytic expression under the Dirichlet posterior, as shown in Corander (2003) using the result of Yuan & Kesavan (1997). The proposition below extends this result to strongly decomposable LGMs.

**Proposition 3**

*The expected logarithmic utility of a strongly decomposable LGM. Let the posterior $\pi(\mathbf{p}|\mathbf{x})$ of $\mathbf{p}$ be a Dirichlet distribution. The difference $\sum_{i \in \mathcal{I}_\Delta} p(i) \log q(i|G_L)$ between the negative Kullback–Leibler divergence and the entropy of $\mathbf{p}$ is proportional to the posterior expectation of*

$$\sum_{c \in \mathcal{C}(G)} \sum_{i_c \in \mathcal{I}_c} p_c(i_c) \log p_c^*(i_c) - \sum_{s \in \mathcal{S}(G)} \sum_{i_s \in \mathcal{I}_s} p_s(i_s) \log p_s(i_s)$$

*where*

$$p_c^*(i_c) = \frac{p_{c\setminus\{\delta\}}(i_{c\setminus\{\delta\}})p_{c\setminus\{\gamma\}}(i_{c\setminus\{\gamma\}})}{p_{c\setminus\{\delta,\gamma\}}(i_{c\setminus\{\delta,\gamma\}})}$$

*if $\{\delta,\gamma\} \subset c$, $\mathcal{L}_{\{\delta,\gamma\}} \neq \emptyset$, $i_{c\setminus\{\delta,\gamma\}} \in \mathcal{L}_{\{\delta,\gamma\}}$, and $p_c^*(i_c) = p_c(i_c)$ otherwise. Let $h(p_{b|a}(\cdot|i_a))$ denote the entropy of the conditional distribution $p_{b|a}(\cdot|i_a)$. The expression $\sum_{i_c \in \mathcal{I}_c} p_c(i_c) \log p_c^*(i_c)$ simplifies to*

$$\sum_{i_{c\setminus\{\delta,\gamma\}} \in \mathcal{L}_{\{\delta,\gamma\}}} [p(i_{c\setminus\{\delta,\gamma\}})[h(p_{\{\delta,\gamma\}|L_{\{\delta,\gamma\}}}(\cdot|i_{c\setminus\{\delta,\gamma\}})) - h(p_{\delta|L_{\{\delta,\gamma\}}}(\cdot|i_{c\setminus\{\delta,\gamma\}}))$$
$$- h(p_{\gamma|L_{\{\delta,\gamma\}}}(\cdot|i_{c\setminus\{\delta,\gamma\}}))]] - h(p_c(\cdot))$$

*which has an analytic expectation using the properties of the Dirichlet distribution.*





*Proof.* See the appe

When an LGM is n
calculated explicitly. I
random quantities fro
approximation. Notice
and therefore only ind
the projection of the p
replace the expected e
asymptotic formula (4

## 5. Empirical examples

We consider now exam
the Florida murderers
been considered by nu
are: (1) race of a victi
sentence (death or oth
Edwards (2000), where
partment (3), the last
denoted by the integers
data set reported in Fo
data set consists of 11

Graphical model de
plete graph with relati
ables. The conclusion
complete graph for th
meration in model det
with reasonably high
arithmic utility was de
the relevant affine subs

For the Florida mur
strongly supports the c
sentence are independe
in the conditional distr

Table 4. D

| Variable |
|---|
| Attendance |
| Gender |
| School type |
| Answer to mathema |
| Subject pre |
| Future plan |

*Proof.* See the appendix.

When an LGM is not strongly decomposable, the expected logarithmic utility cannot be calculated explicitly. In general, the expectation has to be calculated by generating a set of random quantities from the posterior, and projecting each of these to $\Theta_{G_L}$, to obtain an approximation. Notice that the posterior expectation is with respect to the complete graph, and therefore only independent realizations from a Dirichlet distribution are required. When the projection of the probabilities is computationally too intensive, it is more reasonable to replace the expected entropies with the maximum likelihood estimates, which leads to the asymptotic formula (4).

## 5. Empirical examples

We consider now examples of labelled graphical modelling using real data sets. The first one is the Florida murderers data containing 4764 observations on three binary variables, which has been considered by numerous authors (see, for instance, Whittaker, 1990, p. 47). The variables are: (1) race of a victim (black/white), (2) race of an alleged murderer (black/white) and (3) sentence (death or other). The second data set is the university admission data analysed in Edwards (2000), where 4544 applicants are classified by admittance (1), gender (2) and department (3), the last one having six different categories. In both cases the variables are denoted by the integers $\Delta = \{1, 2, 3\}$ in the given order. The final example is the mathematics data set reported in Fowlkes *et al.* (1988), and also analysed in Madigan & Raftery (1994). The data set consists of 1190 observations on variables described in Table 4.

Graphical model determination using the approach of Corander (2003) leads to the complete graph with relative utility approximately equal to 1 for both data sets with three variables. The conclusion is that none of the three edges is removable in these data. Given the complete graph for the data sets with three variables, it is feasible to use exhaustive enumeration in model determination. In the presentation of model search results, only models with reasonably high utilities are shown. For non-decomposable LGMs, the expected logarithmic utility was determined using 1000 draws from the posterior and projecting these into the relevant affine subspace of $\Theta$.

For the Florida murderers data, model search gave the results presented in Fig. 3. The data strongly supports the conclusion, that for black victims, the race of the alleged murderer and sentence are independent, while for white victims there is a considerable degree of dependence in the conditional distribution.

Table 4. *Description of the variables in the mathematics data set*

| Variable | Label | Outcomes | Label |
|---|---|---|---|
| Attendance in mathematics lectures | 1 | Attended | 0 |
| | | Did not attend | 1 |
| Gender | 2 | Female | 0 |
| | | Male | 1 |
| School type | 3 | Suburban | 0 |
| | | Urban | 1 |
| Answer to the statement 'I'll need mathematics in my future work' | 4 | Agree | 0 |
| | | Disagree | 1 |
| Subject preference | 5 | Math-science | 0 |
| | | Liberal arts | 1 |
| Future plans | 6 | College | 0 |
| | | Job | 1 |

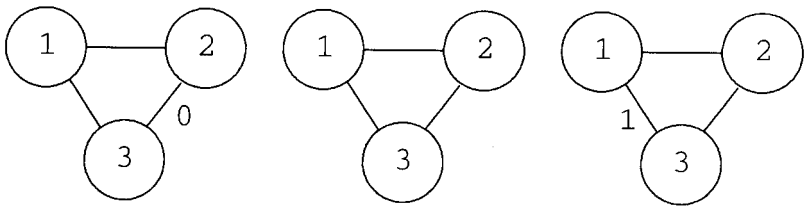



*Fig. 3.* LGMs with the highest relative utilities for Florida murderers data. Labels equal to 0 and 1 denote 'black' and 'white', respectively.

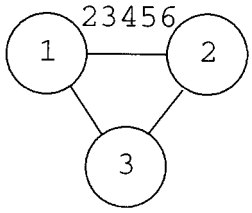



*Fig. 4.* The LGM with the highest relative utility for the university admission data (utility approximately equal to unity).

For the admission data, the single model given in Fig. 4 was superior to the others, leading to the conclusion that the admittance and gender of an applicant are dependent only within one of the six departments.

For the mathematics data, GM determination using the approach of Corander (2003) gave the following model the largest relative utility (0.62), which is approximately four times higher than utilities for some neighbouring models (Fig. 5).

To search among LGMs, labels were imposed on each edge and those resulting in a plausible model with respect to the GM (relative utility in a pairwise comparison higher than 0.05) were then combined in all possible ways. To avoid problems similar to the mass significance phenomenon, the subclass $\mathcal{G}_L^*$ in (6) was defined to include all models visited in the search among label combinations, i.e. also the non-plausible ones (Fig. 6).

The final utility of the GM with the cliques {1}, {245}, {345}, {346} used as the starting model, is so much lower ($< 0.01$) than the utilities of the labelled models, that it is not included

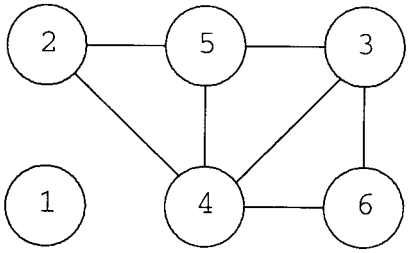



*Fig. 5.* The GM with the highest relative utility (0.62) for the mathematics data.





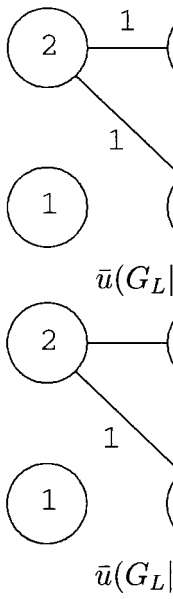



*Fig. 6.* LGMs with the h

in the final class of pla
the independencies st
mathematics are indep
are independent for th
preference are indepen

## 6. Remarks

The examples of secti
dependent interpretati
the relative expected u
models in the light of
develop a heuristic sea
the sense of Malvestut

A natural generaliz
where some of the vari
dependent parameter
pendence structures fo
conditioning set includ
the pure discrete case

## Acknowledgements

The author would like
structive comments a
manuscript.

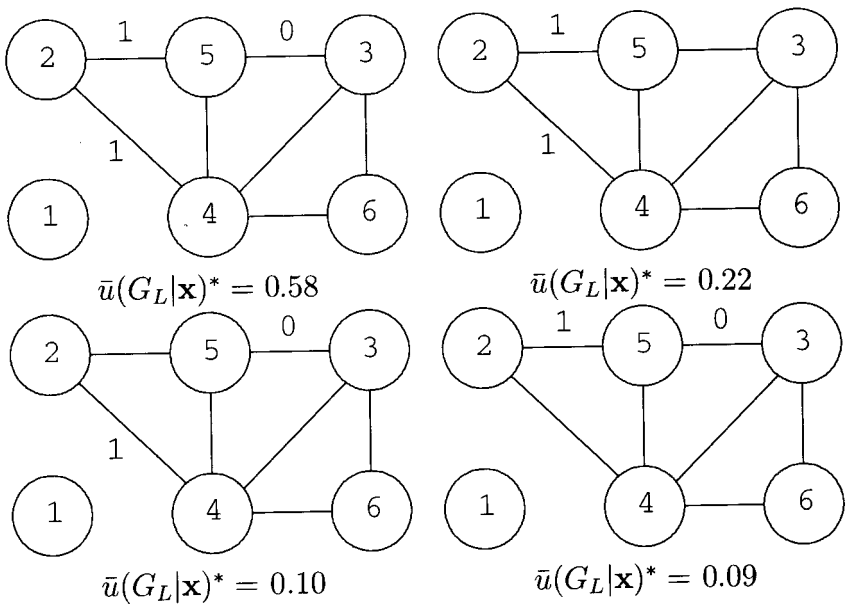



*Fig. 6.* LGMs with the highest relative utilities for the mathematics data.

in the final class of plausible models. Interpretation of the model search results, in addition to the independencies stated by the GM, gives the following: gender and attitude towards mathematics are independent for those who prefer liberal arts; gender and subject preference are independent for those with negative attitude towards mathematics; school type and subject preference are independent for those with positive attitude towards mathematics.

## 6. Remarks

The examples of section 5 illustrate the potential of LGMs to provide a simple, outcome-dependent interpretation of the dependence structure of a multinomial distribution. Moreover, the relative expected utilities provide a convenient measure of the plausibility of investigated models in the light of data. As the model space is typically very large, it might be useful to develop a heuristic search algorithm utilizing logical implications between different models in the sense of Malvestuto (1996), where hierarchical log-linear models were considered.

A natural generalization of the LGMs are labelled mixed graphical interaction models, where some of the variables are qualitative and some quantitative. In addition to the outcome-dependent parameter restrictions considered here, labelling would then allow varying dependence structures for pairs of quantitative and qualitative/quantitative variables when the conditioning set includes qualitative variables. Labelled directed acyclic and chain graphs for the pure discrete case could also provide interesting extensions.

## Acknowledgements

The author would like to thank Elja Arjas and the two anonymous reviewers for their constructive comments and suggestions that led to a significant improvement of the original manuscript.

## References


Barndorff-Nielsen, O. E. (1978). *Information and exponential families in statistical theory*. Wiley, New York.

Bock, H.-H. (1986). Loglinear models and entropy clustering methods for qualitative data. In *Classification as a tool of research* (eds W. Gaul & M. Schader), 19–26. North-Holland, Amsterdam.

Bock, H.-H. (1994). Information and entropy in cluster analysis. In *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: an informational approach* (ed. H. Bozdogan), 115–147. Kluwer, Amsterdam.

Bock, H.-H. (1996). Probability models and hypothesis testing in partitioning cluster analysis. In *Clustering and classification* (eds P. Arabie, L. Hubert & G. De Soete), 377–453. World Scientific, River Edge.

Brown, L. (1986). *Fundamentals of statistical exponential families*. Institute of Mathematical Statistics, Hayward.

Corander, J. (2003). Bayesian graphical model determination using decision theory. *J. Multivariate Anal.*, in press.

Csiszar, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **3**, 146–158.

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

Dawid, A. P. & Lauritzen, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21**, 1272–1317.

Dellaportas, P. & Forster, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika*, **86**, 615–633.

Edwards, D. (2000). *Introduction to graphical modelling*. Springer-Verlag, New York.

Eriksen, P. S. (1999). Context specific interaction models. Technical report, available at www.math.auc.dk/research/reports/.

Fowlkes, E. B., Freeny, A. E. & Landwehr, J. M. (1988). Evaluating logistic models for large contingency tables. *J. Amer. Statist. Assoc.* **83**, 611–622.

Giudici, P. & Green, P. J. (1999). Decomposable graphical Gaussian model determination. *Biometrika* **86**, 785–801.

Giudici, P., Green, P. J. & Tarantola, C. (1999). Efficient model determination for discrete graphical models. Technical report, available at www.statslab.cam.ac.uk/MCMC/.

Hannan, E. & Quinn, B. (1979). The determination of the order of an autoregression. *J. R. Stat. Soc. Ser B Stat. Methodol.* **41**, 190–195.

Højsgaard, S. (1998). Split models for contingency tables. PhD Thesis. Danish Institute of Agricultural Science, Biometry Research Unit, Tjele, Denmark.

Højsgaard, S. (2003a). Statistical inference in context specific interaction models for contingency tables. Accepted for publication in *Scand J Statist.*

Højsgaard, S. (2003b). Split models for contingency tables. *Comput. Statist. Data Anal.*, in press

Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press, Oxford.

Madigan, D. & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.* **89**, 1535–1546.

Madigan, D. & York, J. (1995). Bayesian graphical models for discrete data. *Int. Statist. Rev.*, **63**, 215–232.

Malvestuto, F. M. (1996). Testing implication of hierarchical log-linear models for probability distributions. *Statist. Comput.* **6**, 169–176.

Myllymäki, P., Silander, T., Tirri, H. & Uronen, P. (2001). Bayesian data mining on the Web with B-Course. In *Proceedings of the 2001 IEEE International Conference on Data Mining* (eds N. Cercone, T. Y. Lin & X. Wu), 626–629. IEEE Computer Society Press.

Rudas, T. (1998). A new algorithm for the maximum likelihood estimation of graphical log-linear models. *Comput. Statist.* **13**, 529–537.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

Teugels, J. & Van Horebeek, (1998a). Generalized graphical models for discrete data. *Statist. Probab. Lett.* **38**, 41–47.







Teugels, J. & Van Hore
*J. Multivariate Anal.* **67**
Whittaker, J. (1990). *Gra*
Yuan, L. & Kesavan, H. (
**26**, 139–148.


*Received September 2001,*


Jukka Corander, Rolf N
Finland.
E-mail: jukka.corander@


## Appendix

*Proof of proposition*
imposed by a GM whe
subset of such restricti
and $\mathcal{F}(G)^*$ the class of

$$\log p(i) = \sum_{b \in \mathcal{F}(G)^*} \theta$$

A GM where $\{\delta, \gamma\}$ is
for all $i \in \mathcal{I}_\Delta$. In partic
zero, the restriction be
interaction term $\boldsymbol{\theta}_{b(i_b)}$ is
and the lower order ter
$\boldsymbol{\theta}_{b(i_b)}$, $b \in \mathcal{F}(G)^*$, to ze
independence only for
LGM satisfies the const
variables in $\{\delta, \gamma\}$ and

**Lemma 1**
*The formula (3) in* Yu
*($\boldsymbol{\alpha}_\Delta + n_\Delta$) posterior*
*expression*

$$-\sum_{i \in \mathcal{I}_\Delta} \frac{\alpha(i) + n(i)}{n + \alpha(i)|\mathcal{I}_\Delta|}$$

*where $\psi(\cdot)$ is the digam*

**Lemma 2**
*Lemma 7.2 in Dawid &*
*Let $a \subseteq \Delta$, $b = \Delta \backslash a$, a*
*independent and distrib*

*Proof of proposition*

$$\sum_{i_c \in \mathcal{I}_c : i_{c\backslash\{\delta,\gamma\}} \in \mathcal{L}_{\{\delta,\gamma\}}} p_c(i_c$$

The second sum is equ

*tistical theory*. Wiley, New

s for qualitative data. In
-), 19–26. North-Holland,

*dings of the First US/Japan*
*ach* (ed. H. Bozdogan), 115–

cluster analysis. In *Clustering*
id Scientific, River Edge.
of Mathematical Statistics,

heory. *J. Multivariate Anal.*,

inimization problems. *Ann.*

near interaction models for

al analysis of decomposable

etermination for hierarchical

ew York.
ical report, available at

models for large contingency

etermination. *Biometrika* **86**,

nation for discrete graphical

ression. *J. R. Stat. Soc. Ser B*

ish Institute of Agricultural

odels for contingency tables.

*Data Anal.*, in press

odel uncertainty in graphical

a. *Int. Statist. Rev.*, **63**, 215–

ear models for probability

ta mining on the Web with
*ata Mining* (eds N. Cercone,

f graphical log-linear models.

61–464.
te data. *Statist. Probab. Lett.*



Teugels, J. & Van Horebeek, (1998b). Algebraic descriptions of nominal multivariate discrete data. *J. Multivariate Anal.* **67**, 203–226.
Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Wiley, Chichester.
Yuan, L. & Kesavan, H. (1997). Bayesian estimation of Shannon entropy. *Comm. Statist. Theory Methods* **26**, 139–148.

*Received September 2001, in final form November 2002*

Jukka Corander, Rolf Nevanlinna Institute, PO Box 4, FIN-00014, University of Helsinki, Helsinki, Finland.
E-mail: jukka.corander@rni.helsinki.fi

## Appendix

*Proof of proposition 1.* The proposition can be proved by considering the restrictions imposed by a GM where the edge $\{\delta, \gamma\}$ is absent and noting that an LGM imposes only a subset of such restrictions. Let $\mathcal{F}(G)$ denote the class of variable subsets *not* involving $\{\delta, \gamma\}$, and $\mathcal{F}(G)^*$ the class of remaining subsets. The log-probabilities may be written as

$$\log p(i) = \sum_{b \in \mathcal{F}(G)^*} \theta_{b(i_b)} + \sum_{d \in \mathcal{F}(G)} \theta_{d(i_d)}.$$

A GM where $\{\delta, \gamma\}$ is absent restricts the log-probabilities according to $\sum_{b \in \mathcal{F}(G)} \boldsymbol{\theta}_{b(i_b)} = 0$, for all $i \in \mathcal{I}_\Delta$. In particular, for the outcome $i$ where only $\delta$ and $\gamma$ have values distinct from zero, the restriction becomes $\boldsymbol{\theta}_{\{\delta,\gamma\}(i_{\{\delta,\gamma\}})} = 0$ (for each of the $(r_\delta - 1)(r_\gamma - 1)$ terms). As each interaction term $\boldsymbol{\theta}_{b(i_b)}$ is recursively defined in terms of the difference between a log-probability and the lower order terms involving all the subsets of $b$, the absence of $\{\delta, \gamma\}$ constraints *each* $\boldsymbol{\theta}_{b(i_b)}$, $b \in \mathcal{F}(G)^*$, to zero in a GM. On the contrary, an LGM imposes the conditional independence only for a subset of $\mathcal{I}_\Delta$, whereby the stated result follows. Notice that, as an LGM satisfies the constraints imposed by a GM, all the interaction terms involving both of the variables in $\{\delta, \gamma\}$ and at least one variable outside $L_{\{\delta,\gamma\}}$ are set to zero.

**Lemma 1**
*The formula (3) in* Yuan & Kesavan (1997) *states the following result. Under the Dirichlet* $(\boldsymbol{\alpha}_\Delta + n_\Delta)$ *posterior for* $\mathbf{p}$, *the expectation of the entropy* $-\sum_{i \in \mathcal{I}_\Delta} p(i) \log p(i)$ *has the expression*

$$-\sum_{i \in \mathcal{I}_\Delta} \frac{\alpha(i) + n(i)}{n + \alpha(i)|\mathcal{I}_\Delta|} [\psi(\alpha(i) + n(i) + 1) - \psi(n + \alpha(i)|\mathcal{I}_\Delta| + 1)]$$

*where* $\psi(\cdot)$ *is the digamma function.*

**Lemma 2**
*Lemma 7.2 in Dawid & Lauritzen (1993) states the following properties of Dirichlet distribution. Let* $a \subseteq \Delta$, $b = \Delta \backslash a$, *and let* $\mathbf{p}$ *have the Dirichlet* $(\boldsymbol{\alpha}_\Delta)$ *distribution. Then (1)* $p_{b|a}(\cdot|i_a)$ *are all independent and distributed as Dirichlet* $(\boldsymbol{\alpha}_\Delta(i_a))$, *and (2)* $p_{b|a} \perp p_a$.

*Proof of proposition 3.* The expression $\sum_{i_c \in \mathcal{I}_c} p_c(i_c) \log p_c^*(i_c)$ can be written as

$$\sum_{i_c \in \mathcal{I}_c : i_{c \backslash \{\delta,\gamma\}} \in \mathcal{L}_{\{\delta,\gamma\}}} p_c(i_c) \log p_c^*(i_c) + \sum_{i_c \in \mathcal{I}_c : i_{c \backslash \{\delta,\gamma\}} \in \mathcal{L}_{\{\delta,\gamma\}}} p_c(i_c) \log p_c^*(i_c)$$

The second sum is equal to

$$-h(p_c(\cdot)) - \sum_{i_c \in \mathcal{I}_c : i_{c\setminus\{\delta,\gamma\}} \in \mathcal{L}_{\{\delta,\gamma\}}} p_c(i_c) \log p_c(i_c)$$

and as the probability $p_c^*(i_c)$ in first sum factorizes, one obtains

$$\sum_{i_c \in \mathcal{I}_c : i_{c\setminus\{\delta,\gamma\}} \in \mathcal{L}_{\{\delta,\gamma\}}} p_c(i_c)[\log p_{c\setminus\{\delta\}}(i_{c\setminus\{\delta\}}) + \log p_{c\setminus\{\gamma\}}(i_{c\setminus\{\gamma\}}) - \log p_{c\setminus\{\delta,\gamma\}}(i_{c\setminus\{\delta,\gamma\}})]$$

The expression

$$\sum_{i_{c\setminus\{\delta,\gamma\}} \in \mathcal{L}_{\{\delta,\gamma\}}} [p(i_{c\setminus\{\delta,\gamma\}})[h(p_{\{\delta,\gamma\}|L_{\{\delta,\gamma\}}}(\cdot|i_{c\setminus\{\delta,\gamma\}})) - h(p_{\delta|L_{\{\delta,\gamma\}}}(\cdot|i_{c\setminus\{\delta,\gamma\}}))$$
$$- h(p_{\gamma|L_{\{\delta,\gamma\}}}(\cdot|i_{c\setminus\{\delta,\gamma\}}))] - h(p_c(\cdot))$$

follows then by rewriting the sums. Applied to the cliques of the current graph, lemma 2 shows that the expectation may be taken separately for each of the terms, and lemma 1 gives the explicit form for the expectation.





# A Likeliho for the Cl Marginal

CHRISTIAN BRE
*The Royal Veterinary*

**ABSTRACT. Multivaria failure times may be dep hazards model with estin consider the Clayton–Oa structure to improve on ef based estimating equation the model. We give the equation. Finally, we inv Carlo simulations.**

*Key words:* Clayton–Oak semiparametric likelihoo

## 1. Introduction

Multivariate or cluster where time to disease for various other biom statistical models and hazards models have b proportional hazards citly, estimation is fai 1992; Andersen *et al.*. Large sample properti 1995; Parner 1998). T comparing subjects w frailty variable (see Ar approach, however, is choice of frailty distri

The marginal propo ested in comparing su failure times $T_{1k},\ldots,$ $Z_{1k}(t),\ldots,Z_{nk}(t)$, the m functions for $T_{ik}$, $i =$

$$\lambda_{ik}(t) = \lambda_0(t)e^{\beta_0^T Z_{ik}}$$

where $\lambda_0(t)$ denotes t regression parameters considered the stratifi model (1).