Conditional Log-Linear Structures for Log-Linear Modelling

SUNG-HO KIM[1]

A log-linear modelling will take quite a long time if the data involves many variables and if we try to deal with all the variables at once. Fienberg and Kim (1999) investigated the relationship between log-linear model and its conditional, and we will show how this relationship is employed to make the log-linear modelling easier. The result of this article applies to all the hierarchical log-linear models and the proposed method is used successfully to real data.

KEY WORDS: hybrid of tree and log-linear structures; strong hierarchy principle; induced subgraph; variable-removal; zero-w phenomenon.

[1]Sung-Ho Kim is Associate Professor of Statistics, Division of Applied Mathematics, Korea Advanced Institute of Science and Technology, Daejeon, South Korea, 305-701 (E-mail: shkim@amath.kaist.ac.kr).

1 INTRODUCTION

Suppose that the log-linear model (LLM) of $\mathbf{X} = (X_1, \cdots, X_p)$ is hierarchical and that X_i is multinomial with category levels $1, \cdots, \lambda_i$ and let $\mathbf{X}_{(i)}$ be the $(p-1)$-vector obtained by removing X_i from $\mathbf{X}$. As is well known, a log-linear model of $\mathbf{X}$ is used to represent the logarithm of the joint probability distribution of $\mathbf{X}$ as a linear combination of the terms each of which is defined on a subset of $X_1, \cdots, X_p$. When the model is hierarchical, the model structure of $\mathbf{X}$ is determined by the terms whose domain subsets are maximal, i.e., each of the maximal domain subsets is not contained in any other domain subset in the model. This is why we can represent the model structure of a hierarchical log-linear model by a set of maximal domain subsets, and we will call the set the log-linear structure (LLS) of the model. The set used to be called *generating class* in literature.

Consider a conditional LLM of $\mathbf{X}_{(1)}$ given that $X_1 = x_1$ and denote its LLS by CS_{x_1}. Since X_1 takes on λ_1 different values, we can think of λ_1 such CS's, whose collection actually makes the LLS of $\mathbf{X}$. In other words, the collection is another form of model representation for $\mathbf{X}$. The conditional log-linear structures (CLLSs) are low-level model-structures for $\mathbf{X}_{(1)}$. We will depict the collection in a tree shape, with a node for X_1 and λ_1 arrows from it to the λ_1 nodes of CS's. Since this tree shape is a hybrid of tree and log-linear structures, we will call it a *hybrid*, represent it as $Hyb(X_1;\ CS_1, \cdots, CS_{\lambda_1})$, and call it a *one-node hybrid*.

Adapting our notation from Bishop, Fienberg, and Holland (1975), we will represent a LLS by $\{\theta_1, \cdots, \theta_k\}$, where θ_i's are maximal domain subsets. For notational convenience, we use the indexes of the X variables to represent the LLS. Fienberg and Kim (1999) found (Lemma 1 thereof) that if θ appears in at least one of the CLLSs of a given one-node hybrid, the original log-linear model, say M, must contain a maximal domain subset θ or the union of θ and the index, say c, of the conditional variable. In particular, it was found that if the θ is found in all the CLLSs, then either θ or $\theta \cup \{c\}$ is a maximal domain subset of the model M, and that otherwise, $\theta \cup \{c\}$ must be a maximal domain subset of M provided that θ is maximal in $\cup_{i=1}^{\lambda_c} CS_i$. A key point in this is that parts of the model structure of a LLM are found in the CLLSs of the LLM.

Conditional models are often used for model representation in AI. The Bayesian network (Pearl, 1988) is one of the most popular graphical models in AI that are depicted via directed acyclic graph. D'Ambrosio (1995), Geiger and Heckerman (1996), and Mahoney and Laskey (1999) consider a problem of representing Bayesian networks, using conditional probability

models and conditional variables, which are capable of explicitly capturing much of lower-level structural details. They use the hybrid type of model representation with the CLLS replaced with the conditional probability distribution. Chickering, Heckerman, and Meek (1997) and Friedman and Goldszmidt (1998) employ the detailed model representation in learning Bayesian networks. Conditional models are embedded in a whole model as local descriptions of the whole model, making the detailed model-representation possible. Our purpose of using the CLLS is quite different from them in that we aim to use the CLLS for searching model structures rather than using it as a way of model representation.

Our main goal of this article is to explore further the relationship between LLS and its CLLS and propose an easier method of log-linear modelling which makes use of the pieces of information on model structure that are obtained from CLLSs. The time of parameter-estimation for log-linear modelling increases exponentially in the number of variables involved. At an early stage of modelling, it is desirable that we have a reasonable model structure to begin with, which is in practice not an easy matter.

This paper consists of 7 sections. Section 2 briefly describes the relation between an LLM and its conditional LLM. We view this relation from another perspective in section 3 assuming that the log-linear model is graphical. This will be helpful in getting an insight into the relation. Section 4 then presents a theory which is instrumental for structure searching. While we assume single conditional variables in section 4, we consider CLLSs with multiple conditional variables in section 5 and derive generalized versions of the theorems in its preceding sections. The results in section 5 are applied successfully to real data in section 6, and finally we close the article in section 7 with some concluding remarks.

2 CONDITIONAL LOG-LINEAR STRUCTURE

We briefly review, through a couple of examples, a LLM and its conditionals. The examples are about how CLLSs are related with a given log-linear structure (LLS for short). After these examples we will describe in general terms the relation between a LLM and its CLLM.

Example 1 Consider a *LLS* for 5 variables, $X_1, \cdots, X_5$, as given by

$$\{\{1,2,3\},\{1,3,4\},\{3,4,5\}\}. \tag{1}$$

If we denote the three component sets in (1), respectively by $\{\theta_1, \theta_2, \theta_3\}$, we can write the

corresponding LLM as

$$\log p_{12345}(x_1, \cdots, x_5) = u_{\theta_1} + u_{\theta_2} + u_{\theta_3} + R,$$

where R includes a constant u-term plus the summation of the u_θ-terms, each of whose subscript set is a strict subset of θ_i for some $i \in \{1, 2, 3\}$.

Let

$$\begin{aligned} w_\theta^{(x_1)} &= u_\theta + u_{\theta \cup \{1\}}, \\ w^{(x_1)} &= u + u_1 - \log p_1(x_1), \end{aligned}$$

Then the logarithm of the conditional probability of $X_2 = x_2, \cdots, X_5 = x_5$ given $X_1 = x_1$ is given by

$$\begin{aligned} &\log p_{2345|1}(x_2, \cdots, x_5; x_1) \\ &\quad = w^{(x_1)} + w_2^{(x_1)} + w_3^{(x_1)} + w_4^{(x_1)} + u_5 + w_{23}^{(x_1)} + w_{34}^{(x_1)} + u_{35} + u_{45} + u_{345}. \end{aligned} \tag{2}$$

Note in (2) that the CLLS is determined by the terms $w_{23}^{(x_1)}$ and u_{345} if neither of them equals zero. Also note that the index sets of these terms, $\{2, 3\}$ and $\{3, 4, 5\}$, are included in the set

$$\{\theta_1 \setminus \{1\}, \theta_2 \setminus \{1\}, \theta_3 \setminus \{1\}\} = \{\{2, 3\},\ \{3, 4\},\ \{3, 4, 5\}\}. \qquad \square$$

As connoted in the last equation, we need know when to expect to see the full set of $u-$ or $w-$terms for a successful trip back to a LLM from a specific version of its conditional model (that depends on the value x_1). The lemma below plays an important role in searching for the set of the CLLSs that appear in a hybrid. Although the proof of the lemma is simple it is presented because it gives us an insight into the relation between a LLM and its CLLM.

Lemma 1 *Let $\theta \cap \{1\} = \emptyset$. Then, $u_\theta = u_{\{1\} \cup \theta} = 0$, iff $w_\theta^{(x_1)} = 0$ for all $x_1 = 1, \cdots, \lambda_1$.*

Proof: Suppose that $w_{\theta(\mathbf{x}_\theta)}^{(x_1)} = 0$, for all possible realizations $\mathbf{x}_\theta$, and for all $x_1 = 1, \cdots, \lambda_1$. Then it follows that, for $x_1 = 1, \cdots, \lambda_1$,

$$u_{\{1\} \cup \theta(x_1, \mathbf{x}_\theta)} = -u_{\theta(\mathbf{x}_\theta)}, \ \ \forall \mathbf{x}_\theta. \tag{3}$$

Since $\sum_{x_1=1}^{\lambda_1} u_{\{1\} \cup \theta(x_1, \mathbf{x}_\theta)} = 0$, equation (3) implies that $u_\theta = 0$. The proof for the other direction is straightforward. $\square$

According to Lemma 1, it is possible that $w_{34}^{(x_1)} = 0$ for some value x_1 of X_1 when $u_{134} \neq 0$. If $w_{34}^{(x_1)}$ in (2) equals 0 for all values of x_1, then the full LLM is not hierarchical since $u_{345} \neq 0$; similarly, it is also possible that $w_3^{(x_1)} = 0$ while $w_{34}^{(x_1)} \neq 0$. These "non-hierarchical" situations are difficult to interpret. We shall assume, throughout the remainder of the paper, that the hierarchy principle holds for both the $CLLM$ and the LLM to avoid such situations. We refer to this as the *strong hierarchy principle* or the SHP for short.

Under the SHP, the CLLS is subject to whether a particular w-term is zero or not. For instance, the $CLLS$ of the model represented by expression (2) is determined by u_{345} and $w_{23}^{(x_1)}$ when $w_{23}^{(x_1)} \neq 0$, determined by u_{345} and $w_2^{(x_1)}$ when $w_{23}^{(x_1)} = 0$ and $w_2^{(x_1)} \neq 0$, and determined by u_{345} only when both $w_{23}^{(x_1)}$ and $w_2^{(x_1)}$ are non-zero. We refer to such situations where $w^{(\cdot)} = 0$ as *zero-w* phenomena.

Taking the zero-w phenomena into consideration in (2), we come up with the possible $CLLS$s for $X_2, \cdots, X_5$ given X_1 as

$$\{\{2,3\},\{3,4,5\}\},\ \ \{\{2\},\{3,4,5\}\},\ \ \{\{3,4,5\}\}. \tag{4}$$

We can also have the same result directly from the LLS as in (1). Once we condition on X_1 in the model given in (1), $\{1,3,4\}\setminus\{1\} = \{3,4\}$ is a subset of $\{3,4,5\}$ while $\{1,2,3\}\setminus\{1\} = \{2,3\}$ is not. Applying the zero-w phenomenon to $\{2,3\}$ leads us to the list in (4).

By applying Lemma 1 to the list in (4), we can obtain hybrids for the 5 random variables in Example 1. According to the lemma, the CLLS $\{\{2,3\},\ \{3,4,5\}\}$ appears in every possible hybrid that corresponds to the LLM in expression (1), but not for the other CLLSs in (4). For example, if the model $Hyb(X_1; \{\{3,4,5\}\}, \{\{2\},\ \{3,4,5\}\})$ holds true, then by Lemma 1 the set $\{1,2,3\}$ can not show up in (1). Note that $\{3,4,5\}$ appears in all the structures in (4), but this is not guaranteed when a set in a LLS contains "1". To make it clearer, we consider a simple LLS which consists of the sets only that contains "1" each.

Example 2 Consider a submodel of the model in (1):

$$\{\{1,2,3\},\ \{1,3,4\}\}. \tag{5}$$

Since "1" is contained in both of the sets in expression (5), we represent the $CLLM$s in terms of w-terms only. Thus, as in Example 1, considering all the possible zero-w phenomena concerning both $\{1,2,3\}$ and $\{1,3,4\}$ gives rise to possible $CLLS$s of the form:

$$\{\varphi_1, \varphi_2\} \qquad \text{with} \qquad \varphi_1 \subseteq \{2,3\},\ \varphi_2 \subseteq \{3,4\},$$

where $\{2,3\}$ and $\{3,4\}$ must show up at least once in the CLLSs by Lemma 1. $\square$

It is worthwhile to note in both of the examples that for each set θ in a given LLS, $\theta \setminus \{1\}$ shows up in at least one of the CLLSs unless $\theta \setminus \{1\}$ is a subset of any other in LLS. Thus we may claim that for every set θ that is maximal in $\bigcup_{i=1}^{\lambda_1} CS_i$, either θ or $\theta \cup \{1\}$ must show up in the corresponding LLS. This relationship between LLS and its CLLS is a useful piece of information for finding the LLS.

We can apply the approach suggested by these two examples to get a list of the possible CLLSs for a given LLS involving p categorical variables, $X_1, X_2, \cdots, X_p$. We consider CLLSs of $X_2, X_3, \cdots, X_p$ conditional on the outcomes of X_1, and assume that X_1 takes on λ_1 values $1, \cdots, \lambda_1$ and that it is effective for the LLM of $X_1, X_2, \cdots, X_p$. Thus we may restrict ourselves to the LLMs that include the u_1 term, i.e.,

$$u_1 \neq 0. \tag{6}$$

Consider a LLS given by

$$\{\theta_1, \theta_2, \cdots, \theta_k\}, \tag{7}$$

where $\theta_1, \theta_2, \cdots, \theta_k$ are distinct subsets of $\{1, 2, \cdots, n\}$. Under the assumption in (6), there must exist at least one set in expression (7) which contains "1". Suppose that there are r of these, $\theta_1, \theta_2, \cdots, \theta_r$. When a LLM is conditioned by X_1, the "1" disappears into the $w^{(x_1)}$-terms in the CLLM and the terms affect the CLLS under the SHP. If the index set of a w-term (e.g., $w^{(x_1)}_{\{3,4\}}$) is a subset of the index set of some u-term, the w-term does not affect the CLLS; otherwise (e.g., $w^{(x_1)}_{\{2,3\}}$), the w-term affect the CLLS. We call a set such as $\{1,3,4\}$ in Example 1 as a *disappearing* set, a set such as $\{1,2,3\}$ as a *remaining* set, and a set such as $\{3,4,5\}$ as a *settled* set. If a set of a LLS is free of "1", it is *settled*; otherwise, it is either *remaining* or *disappearing*.

We may suppose, without loss of generality, that there are d disappearing sets, $\theta_1, \cdots, \theta_d$, $0 \leq d < r$, in expression (7). Thus the sets, $\theta_{d+1}, \cdots, \theta_r$, are the remaining sets in (7). After appropriate rearrangements, we have $d = 1,\ r = 2,\ k = 3$ in Example 1, and $d = 0,\ r = k = 2$ in Example 2. When all the sets in a LLS contain "1", $r = k$ and $d = 0$.

Many of the $CLLS$s associated with a LLS are the results of zero-w phenomena involving the remaining sets in the LLS. We define an operator $\langle \cdot \rangle$ as in that, for a collection A of sets, $\langle A \rangle$ is a collection of the maximal sets in A, where "maximal" is in the sense of set-inclusion.

When the LLS is given by (7), the generic form of the $CLLS$s for $X_2, \cdots, X_p$ given $X_1 = x_1$ is given, under the SHP, by

$$\langle \varphi_{d+1}, \varphi_{d+2}, \cdots, \varphi_r, \theta_{r+1}, \cdots, \theta_k \rangle, \quad \text{for } \varphi_j \in \mathcal{C}_j, \quad j = d+1, \cdots, r, \tag{8}$$

where $\mathcal{C}_j$ is a collection of subsets of the remaining set, $\theta_j \setminus \{1\}$ for $d+1 \leq j \leq r$. The largest of the $CLLS$s is given by $\{\theta_{d+1} \setminus \{1\}, \cdots, \theta_r \setminus \{1\}, \theta_{r+1}, \cdots, \theta_k\}$, and the smallest is given by $\{\theta_{r+1}, \cdots, \theta_k\}$. The settled sets show up in each of the $CLLS$s.

There must exist, by Lemma 1, $\theta_j \setminus \{1\}$ in at least one of the λ_1 CLLSs for $d+1 \leq j \leq r$. Thus if we take the maximal sets in $\cup_{i=1}^{\lambda_1} CS_i$ we have

$$\langle \cup_{i=1}^{\lambda_1} CS_i \rangle = \{\theta_{d+1} \setminus \{1\}, \cdots, \theta_r \setminus \{1\}, \theta_{r+1}, \cdots, \theta_k\}, \tag{9}$$

which we will denote by $\Lambda_{\{1\}}(M)$ where M denotes the LLM represented as in (7) and the subscript $\{1\}$ stands for that the conditional variable is X_1. If a set S_M represents the structure of a LLM M, we will write

$$\Lambda_{\{1\}}(S_M) = \Lambda_{\{1\}}(M).$$

Expressions (8) and (9) also imply that, for any subset $B \subseteq \{1, 2, \cdots, \lambda_1\}$,

$$\Big\langle \cup_{i \in B} CS_i \Big\rangle \subseteq \Big(\{\varphi_j;\ \varphi_j \subseteq (\theta_j \setminus \{1\}), j = d+1, \cdots, r\} \cup \{\theta_{r+1}, \cdots, \theta_k\} \Big)$$

and

$$\Big\langle \cap_{i \in B} CS_i \Big\rangle \supseteq \{\theta_{r+1}, \cdots, \theta_k\}.$$

In other words, $\Big\langle \cup_{i \in B} CS_i \Big\rangle$ contains the *settled* sets as well as subsets of the *remaining* sets.

Equation (9) is valid for the hierarchical log-linear models since Lemma 1 holds for every hierarchical LLM. Although our interest lies in the hierarchical log-linear models, we will make use of an attractive feature of graphical log-linear model (Darroch, Lauritzen, and Speed 1980; Fienberg 1980) to visualize the relation between LLS and its CLLS. This will help getting an insight into how CLLSs are useful for log-linear modelling.

3 CONDITIONAL LOG-LINEAR STRUCTURE AND INDUCED SUBGRAPH

A log-linear model is called a graphical log-linear model if its model structure is representable via an undirected graph of vertices and edges. The undirected graph is also called independence

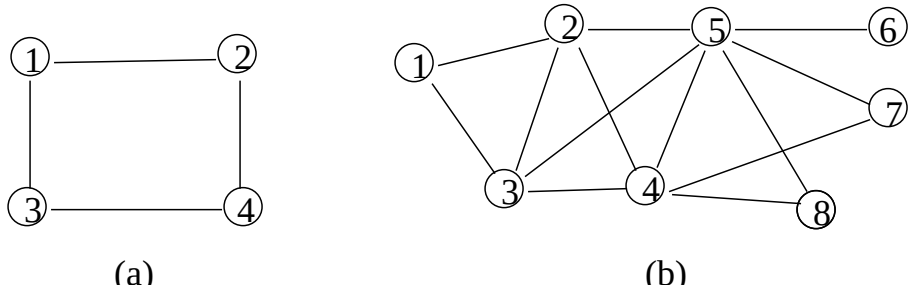

Figure 1: Examples of undirected graph

graph (Whittaker 1990) and in computer science it is called Markov network (Pearl 1988). We will call the graphical log-linear model simply by graphical model. As is well known, the conditional independence relationship among the variables involved in a graphical model is represented in a graph in such a way that the conditional independence of X and Y given Z is depicted in the corresponding graph of the model as that all the paths from X to Y are through Z (Whittaker 1990, The Separation Theorem).

We denote a graph by $\mathcal{G} = (V, E)$ with V and E as the set of the vertices in $\mathcal{G}$ and the set of the edges in $\mathcal{G}$, respectively. Since we consider only the undirected graph, $(\alpha, \beta) \in E$ implies $(\beta, \alpha) \in E$ and vice versa. So if an edge with end points α and β is in $\mathcal{G}$, we will simply write either $(\alpha, \beta) \in E$ or $(\beta, \alpha) \in E$.

An induced subgraph of $\mathcal{G}$ is defined as

$$\mathcal{G}_A = (A, E_A)$$

where the edge set $E_A = E \cap (A \times A)$ is obtained by keeping edges with both endpoints in A. From this definition, we can see that for a set $V_{(a)} = V \setminus \{a\}$ with a a vertex in $\mathcal{G}$, $\mathcal{G}_{V_{(a)}}$ is obtained by removing the vertex a along with the edges for which a is an endpoint.

A clique in $\mathcal{G}$ is a maximal complete subgraph of $\mathcal{G}$ where maximal is in the sense of set inclusion. If the LLS in (7) is graphical with graph $\mathcal{G}$, each θ_i, $i = 1, \cdots, k$, corresponds to a clique in $\mathcal{G}$. For example, the graph in panel (a) of Figure 1 is the graph of the LLS, $\{\{1,2\},\ \{1,3\},\ \{2,4\},\ \{3,4\}\}$ and that in panel (b) is the graph of the LLS, $\{\{1,2,3\},\ \{2,3,4,5\},\ \{5,6\},\ \{4,5,7\},\ \{4,5,8\}\}$. A good introduction to the graphical log-linear model is given in Darroch, Lauritzen, and Speed (1980).

Note that in panel (a) the induced subgraph $\mathcal{G}_{\{2,3,4\}}$ is expressed as

$$\{\{2,4\},\ \{3,4\}\}; \tag{10}$$

similarly, as for panel (b), the induced subgraph is expressed as

$$\{\{2,3,4,5\},\ \{5,6\},\ \{4,5,7\},\ \{4,5,8\}\}. \tag{11}$$

If we denote the LLSs in Figure 1 by M_a and M_b respectively, we can see from (7) that $\Lambda_{\{1\}}(M_a)$ and $\Lambda_{\{1\}}(M_b)$ are the same as those in (10) and (11), respectively. We can summarize this as in the theorem below. If a graphical model, M, has its corresponding graph $\mathcal{G}$, we will use $\Lambda_a(\mathcal{G})$ in the same sense as $\Lambda_a(M)$. Since a graph can be expressed in the form of set, we will use the symbol $\psi(\mathcal{G})$ for the set-expression of the graph $\mathcal{G}$.

Theorem 1 *Suppose that the LLS of* $\mathbf{X} = (X_1, \cdots, X_p)$ *is graphical with the corresponding graph* $\mathcal{G}$*. Then*

$$\Lambda_{\{1\}}(\mathcal{G}) = \psi(\mathcal{G}_{V\setminus\{1\}}).$$

Proof: Without loss of generality, we may assume the LLS as in (7) with d and r interpreted as they are therein. By definition, $\Lambda_{\{1\}}(\mathcal{G})$ is the same as the right-hand side of equation (9).

$\mathcal{G}_{V\setminus\{1\}}$ is obtained by removing node 1 along with the edges connected to node 1. There are two situations of the nodes that are at the other end of the to-be-removed edges. One is that the node, α say, belongs to only one clique; and the other is that node α belongs to multiple cliques. In the former situation, node α itself remains as a clique or belongs to a new clique; and in the latter situation, node α remains as a member of an existing clique of $\mathcal{G}$ and any clique that includes both nodes 1 and α disappears along with node 1. In the context of the LLS (7), $\{\theta_{d+1}, \cdots, \theta_r\}$ pertains to the former situation and $\{\theta_1, \cdots, \theta_d\}$ the latter situation. Therefore, we have

$$\begin{aligned} \psi(\mathcal{G}_{V\setminus\{1\}}) &= \langle\{\theta_1 \setminus \{1\}, \cdots, \theta_k \setminus \{1\}\}\rangle \\ &= \text{the right-hand side of (9).} \end{aligned}$$

This completes the proof. □

4 COMBINATION OF CLLSs WITH SINGLE CONDITIONAL VARIABLES

Consider a graph $\mathcal{G} = (V, E)$ and let $A \subseteq V$. Then, for any induced subgraph $\mathcal{G}_A$, $E_A \subseteq E$. In an undirected graph, we say that two vertices α and β are adjacent if the two are connected by an edge, i.e., $(\alpha, \beta) \in E$.

When two nodes, α and β, are not adjacent in $\mathcal{G}$, then it is obvious, by the definition of induced subgraph, that

$$\psi(\mathcal{G}_{V\setminus\alpha}) \cup \psi(\mathcal{G}_{V\setminus\beta}) = \psi(\mathcal{G}).$$

Thus, if there are at least one pair of non-adjacent nodes in a LLS, it is always true that

$$\cup_{\alpha\in V}\psi(\mathcal{G}_{V\setminus\alpha}) = \psi(\mathcal{G}). \tag{12}$$

We will prove this in a general setting.

Theorem 2 *Consider a LLM, M, represented as in (7). Then the following holds:*

(i) *If $\{\alpha,\beta\} \subseteq \theta'$ for some θ' in S_M, then $\langle \Lambda_{\{\alpha\}}(M) \cup \Lambda_{\{\beta\}}(M)\rangle \subset S_M$.*

(ii) *Otherwise, $\langle \Lambda_{\{\alpha\}}(M) \cup \Lambda_{\{\beta\}}(M)\rangle = S_M$.*

Proof: In case (i), it is obvious that θ' does not show up in $\Lambda_{\{\alpha\}}(M) \cup \Lambda_{\{\beta\}}(M)$ since it is not included in neither of $\Lambda_{\{\alpha\}}(M)$ and $\Lambda_{\{\beta\}}(M)$. In case (ii), we may suppose without loss of generality that α is included in $S_\alpha = \{\theta_1,\cdots,\theta_a\}$ and β in $S_\beta = \{\theta_{a+1},\cdots,\theta_{a+b}\}$ for some positive integers a and b with $a+b \le k$. Since the sets in S_α are all distinct from the sets in S_β, it follows that

$$(S_M \setminus S_\alpha) \cup (S_M \setminus S_\beta) = S_M.$$

Therefore, by applying (8), we have

$$\langle \Lambda_{\{\alpha\}}(M) \cup \Lambda_{\{\beta\}}(M)\rangle = S_M.$$

□

Note that by Theorem 1, equation (12) is immediate from result (ii) of Theorem 2. From result (i) of Theorem 2, we can see that (12) does not hold if a graph is complete. For instance, if $\mathcal{G}$ is a complete graph of nodes 1, 2, and 3, then

$$\cup_{\alpha\in\{1,2,3\}}\psi(\mathcal{G}_{V\setminus\alpha}) = \{\{1,2\},\ \{2,3\},\ \{1,3\}\}, \tag{13}$$

while $\psi(\mathcal{G}) = \{\{1,2,3\}\}$. As a matter of fact, the LLS as in the right-hand side of (13) is a good example of a hierarchical log-linear model which is not graphical (Darroch, Lauritzen, and Speed 1980).

5 COMBINATION OF CLLSs WITH MULTIPLE CONDITIONAL VARIABLES

So far, we have considered CLLMs with only one conditional variable. But we can extend the argument to the situations of multiple conditional variables. Let A be a subset of $V = \{1, 2, \cdots, p\}$ and suppose that $\mathbf{X}_A$ is the vector of the conditional variables. We define, $\mathbf{X}_A = \mathbf{x}_A$,

$$w_\theta^{(\mathbf{x}_A)} = \sum_{\varphi \subseteq A} u_{\theta \cup \varphi}.$$

Then, we can have an extended version of Lemma 1.

Theorem 3 *Let $\theta \cap A = \emptyset$ for a set $A \subset V$. Then*

$$u_{\varphi \cup \theta} = 0, \quad \textit{for all } \varphi \subseteq A$$

if and only if

$$w_\theta^{(\mathbf{x}_A)} = 0 \quad \textit{for all values of } \mathbf{x}_A.$$

Proof: See Appendix A.

From this theorem, we can have a generalized version of (8) and (9), which plays a key role along with Theorem 2 in searching for model structures. Consider the LLS in (7), let $A \subset V$, and assume that

$$\theta_i \cap A \neq \emptyset \quad \text{for} \quad i \in \{1, 2, \cdots, r\}$$

and

$$\theta_i \cap A = \emptyset \quad \text{for} \quad i \in \{r+1, \cdots, k\}.$$

Further assume that for $i \in \{1, \cdots, d\},\ d \leq r$,

$$(\theta_i \setminus A) \subset \theta_l, \quad \text{for some } l \in \{r+1, \cdots, k\},$$

but not for $i \in \{r+1, \cdots, r\}$. Then the CLLS of $\mathbf{X}_{V \setminus A}$ is given in the form of

$$\langle \varphi_{d+1}, \varphi_{d+2}, \cdots, \varphi_r, \theta_{r+1}, \cdots, \theta_k \rangle \ \text{ for } \ \varphi_j \in \mathcal{C}_j, \ \ j = d+1, \cdots, r,$$

where $\mathcal{C}_j$ is a collection of subsets of the remaining set, $\theta_j \setminus A$ for $d+1 \leq j \leq r$.

We denote by χ_A the set of all the possible values of $\mathbf{X}_A$. Then, by Theorem 3, the maximal sets in $\cup_{\mathbf{x}\in\chi_A} CS_{\mathbf{x}}$ is given by

$$\langle \cup_{\mathbf{x}\in\chi_A} CS_{\mathbf{x}} \rangle = \{\theta_{d+1} \setminus A, \cdots, \theta_r \setminus A, \theta_{r+1}, \cdots, \theta_k\},$$

which we will denote by $\Lambda_A(M)$.

The theorem below is an extension of Theorem 1 to a situation of multiple conditional variables.

Theorem 4 *Suppose that the LLM M of $\mathbf{X} = (X_1, \cdots, X_p)$ is graphical with the corresponding graph $\mathcal{G}$. Then, for $A \subset V$,*

$$\langle \cup_{\mathbf{x}\in\chi_A} CS_{\mathbf{x}} \rangle = \psi(\mathcal{G}_{V\setminus A}).$$

Proof: According to Theorem 3, only the set, θ say, that appears in $\cup_{\mathbf{x}\in\chi_A} CS_{\mathbf{x}}$ shows up in the LLS S_M in the form of $\varphi_\theta \cup \theta$ for some $\varphi_\theta \subseteq A$ which can be an empty set. And so, under the strong hierarchy principle, the LLS S_M is given by

$$\langle \{\theta \cup \varphi_\theta;\ \theta \in \cup_{\mathbf{x}\in\chi_A} CS_{\mathbf{x}}\} \rangle,$$

where φ_θ's correspond to the θ and are subsets of A.

Since M is graphical, the induced subgraph of M confined to $V \setminus A$ can be expressed in a set form by

$$\langle \{\theta \cup \varphi_\theta;\ \theta \in \cup_{\mathbf{x}\in\chi_A} CS_{\mathbf{x}}\} \rangle \setminus A,$$

which is the same as $\langle \cup_{\mathbf{x}\in\chi_A} CS_{\mathbf{x}} \rangle$ since all the φ_θ's are contained in A. This completes the proof. □

When multiple variables are conditional, whether all the conditional variables are contained in a component set of a LLS is an important issue in structure searching through CLLSs.

Theorem 5 *Consider a LLM M as given in (7). Let A and B be non-empty subsets of V. Then the following holds:*

(i) *If $A \cap B \neq \emptyset$, then $\Lambda_A(M) \cup \Lambda_B(M) \subset S_M$.*

(ii) *If $A \cap \theta' \neq \emptyset$ and $B \cap \theta' \neq \emptyset$ for some $\theta' \in S_M$, then $\Lambda_A(M) \cup \Lambda_B(M) \subset S_M$.*

(iii) *Otherwise, $\Lambda_A(M) \cup \Lambda_B(M) = S_M$.*

Table 1: The 8 labels of items and their meanings. The random variables are listed according to their contents. The first four are about lecturing, the next three concerning homework, and the last may be regarded as an overall level of evaluation.

Variables	key words	item contents
A	Class atmosphere	Was the lecture given in an interactive atmosphere?
T	Thought provoking	Was the lecture thought-provoking?
O	Organization	Was the lecture well organized throughout the course?
E	Explanation	Was the lecture given with an explanation good enough for a clear understanding?
D	Difficulty	Were the test and the homework problems at appropriate levels?
F	Feedback	Did you get a satisfactory feedback from the comments on your homework?
H	Homework	Was the homework assignments helpful in understanding the lecture and related subjects?
R	Recommendation	Would you recommend this course to your friends?

Proof: See Appendix A.

This theorem says that when using multiple conditional variables, it is desirable that at least one pair of sets of conditional variables be disjoint. This point is to be observed when we try to find a reasonable model structure in log-linear modelling.

6 AN APPLICATION: LOG-LINEAR MODELLING WITH REAL DATA

In this section a real data set of size 28,270 is used, which is collected for lecture evaluation for the courses lectured during in year 2000 at a university in South Korea. The survey items that are selected for use are 8 out of 23. The selected 8 items are about lecture and homework, and tests. We will label the random variables for the 8 items by A, T, O, E, D, F, H, and R, which are explained in Table 1. Item D has three options, easier, middle, and harder, and each of the other 7 items has three options, *negative, half-and-half, positive*, and so the random variables are all ternary taking on values 1 (for negative), 2 (for half-and-half), or 3 (for positive). The frequency table of the data is of $3^8 = 6,561$ cells. Because of the size of the table, the data set is not included in the article but is given as a file named "lec00.sel8.dat" in the web site, "http://amath.kaist.ac.kr/∼slki/research/data."

The eight variables look all related each other and so it is not easy to guess an initial model structure to begin with. Although only eight variables are involved, it takes too long a time to use a backward deletion method starting from the full model which is of 6,560 parameters to estimate. However, if we use the information from CLLSs of a subset of the eight variables, we can easily guess a possible model structure for the whole data.

The smaller the model, the easier the modelling. The log-linear modelling with five variables is feasible and takes much less time than dealing with six variables. So we tried to find a set of three conditioning variables which are closely related each other. CART (Breiman et al. 1984) and S-plus (Chambers and Hastie 1992) are useful in this respect since it yields a regression tree where the variables that are more highly associated with a given dependent variable tend to appear earlier in the tree topology.

The set of variables that was selected at the initial stage of modelling is $\{E, O, R\}$. It was obtained by tree-regressing R which indicates a level of course recommendation by students and is mostly interested by teachers. The CLLS of the rest five variables, A, T, D, F, and H, that is appropriate conditional on $\{E, O, R\}$ is [ATFH][TDFH]. The goodness-of-fit levels are listed in panel (a) of Table 2. In this CLLS, $\{T, F, H\}$ is contained in both of the sufficient sets of the model. So we selected the variables in the set as conditional and the model [AEO][DOR][AOR][EDR] is selected as reasonably good. The corresponding goodness-of-fit levels are listed in panel (b) of Table 2. Out of all the possible configurations of the conditional variables, only those configurations are used where the corresponding sample sizes (n) are near or larger than 5 times the cell count (i.e., $3^5 = 243$) of each conditional contingency table. This applies to all the panels of the table.

Note that the two conditional sets are disjoint and so it is possible by (i) of Theorem 2 that there be at least one sufficient set that involves some variables in both of the two conditional sets. To check this, two more conditional modelings were carried out, one conditional on $\{A, E, D\}$ and another conditional on $\{A, D, F\}$, and their results are summarized in panels (c) and (d) of Table 2, respectively. In panel (c), the P-value is small for two configurations, 222 and 332. But this may not undermine the modelling process since when combining the CLLSs toward the whole model, such local deficiencies may be patched up.

Using the notation for CLLS and denoting the LLM for the data by M, we can express the for CLLSs as follows:

$$\Lambda_{\{O,E,R\}}(M) \;=\; \{\{A, T, F, H\},\ \{T, D, F, H\}\},$$

Table 2: The goodness-of-fit levels of the four CLLMs corresponding to four sets of conditional variables. The values are obtained from the SAS package.

(a) [ATFH][TDFH]
conditional on $\{O, E, R\}$

configuration	n	P-value
222	2569	0.09
223	1696	0.72
233	1837	0.40
323	1669	0.31
332	1934	0.66
333	11760	0.08

(b) [AEO][DOR][AOR][EDR]
conditional on $\{T, H, F\}$

configuration	n	P-value
221	1181	0.99
222	2357	0.13
223	1145	0.99
231	1056	0.85
232	1751	0.69
233	2017	0.30
322	1370	0.99
323	1027	0.90
331	1558	0.21
332	3327	0.46
333	6519	0.60

(c) [THO][TFH][TOR][FHR]
conditional on $\{A, E, D\}$

configuration	n	P-value
222	2482	0.0015
223	1426	0.28
232	2583	0.30
322	1635	0.98
323	1291	0.995
332	7180	0.004
333	4689	0.64

(d) [EO][TEH][TOR][THR][HOR][ER]
conditional on $\{A, D, F\}$

configuration	n	P-value
221	1309	0.74
222	2502	0.02
223	1669	0.80
232	1293	0.77
233	1206	0.99
321	1683	0.999
322	3158	0.83
323	4201	0.05
331	1060	0.998
332	1781	0.42
333	3434	0.67

$$
\begin{aligned}
\Lambda_{\{T,H,F\}}(M) &= \{\{A,E,O\},\ \{D,O,R\},\ \{A,O,R\},\ \{E,D,R\}\},\\
\Lambda_{\{A,E,D\}}(M) &= \{\{T,H,O\},\ \{T,F,H\},\ \{T,O,R\},\ \{F,H,R\}\},\\
\Lambda_{\{A,D,F\}}(M) &= \{\{E,O\},\ \{T,E,H\},\{T,O,R\},\ \{T,H,R\},\ \{H,O,R\},\ \{E,R\}\}.
\end{aligned}
$$

This model does not fit well to the whole data, for which the goodness-of-fit level is 0.0012. So we put the four sets

$$\{\{T,H,O\},\ \{T,O,R\},\ \{T,H,R\},\ \{H,O,R\}\}$$

together into $\{T, H, O, R\}$ with the other sets in (15) remaining in the model, which ends up with the model $\{\{A,T,F,H\},\ \{T,D,F,H\},\ \{A,E,O\},\ \{D,O,R\},\ \{A,O,R\},\ \{E,D,R\},$ $\{T,H,O,R\},\ \{F,H,R\},\ \{T,E,H\}\}$. The goodness-of-fit level of this model is 0.0316, which

is yet to be improved. In searching for a larger model, we examined the third through sixth sets in the preceding model and found that out of the five variables, A, D, E, O, R, only the A-and-D pair is not two-way interactive. So we considered a largest model possible for the five variables that is given by $\{\{A, E, O, R\},\ \{D, E, O, R\}\}$. With this replacement made to the preceding model, we have the model

$$\begin{aligned}&\{\{A, T, F, H\},\ \{T, D, F, H\},\ \{A, E, O, R\},\ \{D, E, O, R\},\\ &\quad \{T, H, O, R\},\ \{F, H, R\},\ \{T, E, H\}\},\end{aligned} \tag{14}$$

whose goodness-of-fit level is 0.46 and whose ANOVA-like summary from the SAS package is given in Appendix B.

Combining these four CLLSs yields

$$\begin{aligned}&\langle \Lambda_{\{O,E,R\}}(M) \cup \Lambda_{\{T,H,F\}}(M) \cup \Lambda_{\{A,E,D\}}(M) \cup \Lambda_{\{A,D,F\}}(M) \rangle \\ &\quad = \{\{A, T, F, H\},\ \{T, D, F, H\},\ \{A, E, O\},\ \{D, O, R\},\ \{A, O, R\},\ \{E, D, R\},\\ &\qquad \{T, H, O\},\ \{T, O, R\},\ \{F, H, R\},\ \{T, E, H\}, \{T, H, R\},\ \{H, O, R\}\}\end{aligned} \tag{15}$$

It is interesting to see that in the model (14) variable T is in the same set as H, E and O are in the same set as R, and F with H. This may imply that the contribution of homework (H) has something to do with a thought-provoking lecture (T) and with the feedback from the homework (F) and that the course-recommendation level (R) is influenced by, among others, the clarity of explanation (E) and the organization of a course (O) by the lecturer. Since our purpose is to demonstrate the proposed method of structure searching, we will refrain from deviating that line of pursuit concerning the data.

7 CONCLUDING REMARKS

We will call a set A irreducible if $\langle A \rangle = A$. When a LLM M is graphical, we learned that the induced subgraph which is obtained by removing the nodes of the conditional variables from the graph of M can also be obtained as the irreducible set of the component sets of all the CLLSs of M. This node-removal from a graph is equivalent to the variable-elimination from the structure of a hierarchical LLM. A nice feature in conditionalizing a LLM M is that the conditional variables disappear from the LLS S_M with the other variables in M remaining untouched. Individual CLLSs may vary across the values of the conditional variables. However,

from the collection of the CLLSs, we can obtain a model structure which is an analogy of induced subgraph. We have shown that we can construct a model structure for M if two sets of conditional variables are disjoint and they do not share a component set of S_M.

When modelling with real data and if it is time-consuming to find a good-looking model structure for the data, it is relatively easy, as we have seen in the previous section, to find a model structure which is reasonable to begin with. It is usually the case that parameter estimation with conditional models is much more time-efficient than with a model of all the variables involved. What is worse when we deal with all the variables involved is that a wrong model that we begin with might cost us a long time until we could find a reasonable model at an early stage of modelling.

As we have seen in the real data example, we may not have to look at all the CLLSs. We may ignore the CLLSs for which the sample sizes are not large enough. Although the irreducible set, say S, of all the component sets of the CLLSs such as in (15) may not fit well to the data, it serves as a solid ground for building an appropriate model by trying, if necessary, some models larger than S.

APPENDIX A: PROOFS

Proof of of Theorems 3: Necessity is obvious by the definition of w. Suppose that $w_\theta^{(\mathbf{x}_A)} = 0$ for all possible values of $\mathbf{x}_A$ and $\mathbf{x}_\theta$. So we have

$$u_{A\cup\theta}(\mathbf{x}_A, \mathbf{x}_\theta) = -\sum_{\varphi\subset A} u_{\varphi\cup\theta}(\mathbf{x}_\varphi, \mathbf{x}_\theta) \quad \text{for all values of } \mathbf{x}_A \text{ and } \mathbf{x}_\theta. \tag{A.1}$$

By the constraint that

$$\sum_{x_i} u_{\varphi\cup\theta}(\mathbf{x}_\varphi, \mathbf{x}_\theta) = 0 \quad \text{for } i \in (\varphi\cup\theta) \ \text{ with } \ \varphi \subseteq A,$$

we have from (A.1), for $i \in A$, that

$$0 = u_\theta(\mathbf{x}_\theta) + u_{\{i\}\cup\theta}(x_i, \mathbf{x}_\theta).$$

So, it follows that

$$u_\theta = u_{\{i\}\cup\theta} = 0.$$

Similarly, we can show that

$$u_\theta = u_{\{\alpha\}\cup\theta} = 0 \quad \text{for every } \alpha \in A. \tag{A.2}$$

Now suppose that (A.2) holds for i and j in A. If $A = \{i, j\}$, then the following equation is immediate from (A.1) and the supposition:

$$u_{\{i,j\}\cup\theta} = 0. \tag{A.3}$$

Otherwise, we have

$$\begin{aligned} \sum_{x_k;\, k\in(A\setminus\{i,j\})} u_{A\cup\theta}(\mathbf{x}_A, \mathbf{x}_\theta) &= -\sum_{\varphi\subseteq\{i,j\}} u_{\varphi\cup\theta}(\mathbf{x}_\varphi, \mathbf{x}_\theta) && \text{(A.4)} \\ &= -u_{\{i,j\}\cup\theta}(x_i, x_j, \mathbf{x}_\theta) && \text{(A.5)} \\ &= 0. && \text{(A.6)} \end{aligned}$$

The left-hand side of (A.4) equals zero due to the constraint upon the u terms, and equation (A.5) follows from the result (A.2).

For a proper subset B of A, assume that

$$u_{\varphi\cup\theta}(\mathbf{x}_\varphi, \mathbf{x}_\theta) = 0, \quad \text{for every subset } \varphi \subseteq B.$$

By applying the same argument as above, we can derive

$$u_{B\cup\theta} = 0.$$

Thus by inductive reasoning, the sufficiency is proved. □

Proof of Theorem 5: If $A \cap B \neq \emptyset$, then any variable with its index $\alpha \in (A \cap B)$ does not show up in $\Lambda_A(M) \cup \Lambda_B(M)$. So (i) holds true. If the condition of statement (ii) holds, $\theta' \setminus A$ is a subset of a component set of $\Lambda_A(M)$ and $\theta' \setminus B$ is a subset of a component set of $\Lambda_B(M)$. So the set θ' can not show up in $\Lambda_A(M) \cup \Lambda_B(M)$, which proves (ii).

In statement (iii), A and B do not share variables with the same set in S_M. Suppose that A share variables with $\theta_{i_1}, \cdots, \theta_{i_a}$ in S_M and B with $\theta_{j_1}, \cdots, \theta_{j_b}$ in S_M for some positive integers a and b with $a + b \leq k$, where

$$(\cup_{k=1}^{a} \theta_{i_k}) \cap (\cup_{k=1}^{b} \theta_{j_k}) = \emptyset. \tag{A.7}$$

Then by applying Theorem 3, we have

$$\Lambda_A(M) \supseteq \langle (S_M \setminus (\cup_{k=1}^{a} \theta_{i_k})) \rangle \tag{A.8}$$

since $A \subseteq (\cup_{k=1}^{a} \theta_{i_k})$, and by the same reason

$$\Lambda_B(M) \supseteq \langle (S_M \setminus (\cup_{k=1}^{b} \theta_{j_k})) \rangle. \tag{A.9}$$

From (A.7) follows that

$$(S_M \setminus (\cup_{k=1}^{a} \theta_{i_k})) \cup (S_M \setminus (\cup_{k=1}^{b} \theta_{j_k})) = S_M.$$

So, we have from (A.8) and (A.9), that

$$\Lambda_A(M) \cup \Lambda_B(M) \supseteq S_M.$$

But by definition, each component set of $\Lambda_A(M)$ is a subset of a component set of S_M and similarly for $\Lambda_B(M)$. This leads to the desired result of (iii). □

APPENDIX B: SAS OUTPUT FOR THE MODEL IN EXPRESSION (14)

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
A	2	94.89	<.0001
T	2	137.60	<.0001
A*T	4	753.38	<.0001
F	2	26.46	<.0001
A*F	4	37.96	<.0001
T*F	4	20.70	0.0004
A*T*F	8	26.82	0.0008
H	2	84.52	<.0001
A*H	4	111.45	<.0001
T*H	4	14.86	0.0050
A*T*H	8	90.96	<.0001
H*F	4	61.15	<.0001
A*H*F	8	22.51	0.0041
T*H*F	8	4.15	0.8430
A*T*H*F	16	58.61	<.0001
D	2	308.39	<.0001
T*D	4	65.60	<.0001
D*F	4	26.72	<.0001
T*D*F	8	12.45	0.1323
H*D	4	182.51	<.0001
T*H*D	8	36.20	<.0001
H*D*F	8	29.66	0.0002
T*H*D*F	16	19.74	0.2323
O	2	172.28	<.0001
A*O	4	59.67	<.0001
E	2	63.55	<.0001
A*E	4	112.15	<.0001
O*E	4	288.55	<.0001
A*O*E	8	42.70	<.0001
R	2	85.54	<.0001
A*R	4	28.08	<.0001

O*R	4	72.22	<.0001
A*O*R	8	10.52	0.2306
E*R	4	82.91	<.0001
A*E*R	8	20.83	0.0076
O*E*R	8	16.76	0.0327
A*O*E*R	16	50.10	<.0001
O*D	4	41.82	<.0001
E*D	4	36.53	<.0001
O*E*D	8	50.24	<.0001
D*R	4	75.25	<.0001
O*D*R	8	22.06	0.0048
E*D*R	8	21.68	0.0056
O*E*D*R	16	27.68	0.0345
T*O	4	166.07	<.0001
O*H	4	167.19	<.0001
T*O*H	8	17.32	0.0270
T*R	4	117.39	<.0001
H*R	4	183.99	<.0001
T*H*R	8	9.02	0.3403
T*O*R	8	13.61	0.0924
O*H*R	8	24.37	0.0020
T*O*H*R	16	97.55	<.0001
F*R	4	97.89	<.0001
H*F*R	8	26.32	0.0009
T*E	4	417.46	<.0001
E*H	4	172.52	<.0001
T*E*H	8	35.79	<.0001
Likelihood Ratio	3E3	3044.42	0.4587

REFERENCES

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

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*. Belmont, CA: Wadsworth.

Chambers, J.M. and Hastie, T.J. (1992), *Statistical Models in S*. Pacific Grove, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Chickering, D.M., Heckerman, D., and Meek, C. (1997), "A Bayesian approach to learning Bayesian Networks with local structure," *Proceedings of Thirteenth Conference on Uncertainty in Artificial Intelligence*, eds. D. Geiger and P. Shenoy, Providence, RI: Morgan Kaufmann, 80-89.

D'Ambrosio, B. (1995), "Local expression languages for probabilistic dependence," *International Journal of Approximate Reasoning*, **13**, 61-81.

Darroch, J. N., Lauritzen, S. L. and Speed, T. P. (1980), "Markov fields and log-linear interaction models for contingency tables," *The Annals of Statistics*, **8**, 3, 522-539.

Fienberg, S. E. (1980), *The Analysis of Cross-Classified Categorical Data*. Cambridge, MA: MIT Press.

Fienberg, S.E. and Kim, S.-H. (1999), "Combining conditional log-linear structures," *Journal of the American Statistical Association*, **94**, 445, 229-239.

Friedman, N. and Goldszmidt, M. (1998), "Learning Bayesian networks with local structure," in *Learning in Graphical Models*, ed. M.I. Jordan. Dordrecht, Netherlands: Kluwer, pp. 421-460.

Geiger, D. and Heckerman, D. (1996), "Knowledge representation and inference in similarity networks and Bayesian multinets," *Artificial Intelligence*, **82**, 45-74.

Mahoney, S.M. and Laskey, K.B. (1999), "Representing and combining partially specified CPTs," in *Uncertainty in Artificial Intelligence*, eds. K.B. Laskey and H. Prade. San Francisco, CA: Morgan Kaufmann Publishers, pp. 391-400.

Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics*. New York: Wiley.