A note on collapsibility in recursive graphical models of contingency tables

By SUNG-HO KIM and SEONG-HO KIM
Division of Applied Mathematics, Korea Advanced Institute of Science and Technology
Daejeon, 305-701, South Korea
Sung-Ho.Kim@kaist.ac.kr and mathan72@hotmail.com

SUMMARY

Necessary and sufficient conditions for collapsibility of a recursive model for a contingency table are derived. By applying the conditions, we can easily check collapsibility over any variable in a given model either by using the joint probability distribution or by using the graph of the model structure.

Some key words: Graphical log-linear model; Markov equivalence; Model collapsibility; Maximum likelihood estimation; Node removal.

1 INTRODUCTION

The notion of collapsibility is important for the practical and conceptual analysis of contingency tables which includes hypothesis testing, model selection, and data reduction. Contingency tables of high dimension are difficult to deal with for detecting model structure, and cells are apt to be empty or to have very low frequencies when the dimension is large. Collapsing a large table into manageable marginal tables might save us from such a difficulty in statistical analysis. But some parameters describing the interactions among a set of variables may be affected by summing over the rest of the variables. Discussion of this issue of collapsibility dates back to the Yule-Simpson paradox (Simpson 1951) and is continued by many investigators including Darroch (1962), Lewis (1962), Birch (1963), and Plackett (1969), to name only the first few in this line of work.

The notion of collapsibility is refined into several types of collapsibility. They are graphical collapsibility (Whittaker 1990), parametric collapsibility (Bishop 1971, Whittemore 1978, Wermuth 1987, Geng 1992), model collapsibility (Asmussen and Edwards 1983, Frydenberg 1990a, Lauritzen 1996), and test collapsibility (Whittaker 1990). These types of collapsibility are related to each other in a somewhat hierarchical manner. Among these types, we will con-

fine our attention to model collapsibility, which may be called commutativity of model-fitting and marginalization.

We will say that a distribution $\mathcal{P}$ is Markov with respect to an undirected graph if, for three disjoint sets of random variables, A, B, S, A and B are conditionally independent given S under $\mathcal{P}$ whenever S separates A and B in the graph. Frydenberg (1990a) proposes a theorem (Theorem 5.4 therein) which lists several equivalent conditions of model collapsibility of a probability distribution $\mathcal{P}$ which is Markov with respect to an undirected graph under the positivity condition of P. According to the theorem, we can interpret collapsibility regarding an undirected graph in the context of a simple graphical chain model (Lauritzen and Wermuth 1984, 1989) which has two chain components. Didelez and Edwards (2004) propose a theorem concerning collapsibility with regard to a marked, undirected graph of mixed variables, which gives necessary and sufficient conditions for collapsibility for graphical CG-regression model.

Asmussen and Edwards (1983) define a causal chain model in terms of model collapsibility (Theorem 3.3 (a) therein). If every chain component is of a node, the causal chain model is a recursive model. Consider a recursive model of a set of nodes $V = \{v_1, \cdots, v_m\}$ where the arrows run from nodes with lower indexes to nodes with higher indexes. We define the sets $B_1 = \{v_1\}$, $B_i = \{v_i\} \cup B_{i-1}$ $(i = 1, \cdots, m)$. Asmussen and Edwards proved that a hierarchical log-linear model is a recursive model if and only if the model is collapsible onto B_i $(i = 1, \cdots, m-1)$. Since a recursive model is a particular form of a graphical chain model, we can say that Asmussen and Edwards (1983) and Frydenberg (1990a) described the model collapsibility with regard to a recursive model, in a sequential manner, in reversed order of the node indexes. Our goal in this paper is to extend the notion of model collapsibility with regard to a recursive model to any subset of V, a set of categorical random variables.

This paper consists of 5 sections. Section 2 presents notation and graph-theoretic terminologies. In section 3, we present a main result of the paper, Theorem 3.2, which describes two equivalent conditions of collapsibility over a variable in a recursive model. In section 4, we then compare collapsibility between a recursive model and the moral graph of the model structure, and discuss collapsibility over a set of variables. Section 5 concludes the paper with summarizing remarks.

2 GRAPHICAL TERMINOLOGIES AND NOTATION

A graphical model is a statistical model whose model structure can be represented by a graph, and we will denote the graph of a graphical model by $\mathcal{G} = (V, E)$, where V is the index set of the nodes involved in the model and E a set of edges between the nodes in V. E is given as a set of ordered pairs (u, v) such that $E \subseteq V \times V$ where (u, v) symbolizes a directed edge or an arrow from node u to node v in graph $\mathcal{G}$. If both (u, v) and (v, u) are included in E, this means that there is an undirected edge between nodes u and v. Thus if $\mathcal{G} = (V, E)$ is the model structure of a recursive model and $(u, v) \in E$, then $(v, u) \notin E$. A node in the graph of a graphical model corresponds to a variable of the model.

We will use a lowercase x to denote the cell location of a contingency table and use x_A for the contingency table of the variables indexed in A. If there is an arrow (u, v), then we will say that node u is a parent of node v and node v is a child of node u. We will denote by $pa(v)$ the set of the parents of node v and by $ch(v)$ the set of the children of v and let $fa(v) = \{v\} \cup pa(v)$. For $A \subseteq V$, we define

$$\begin{aligned} pa(A) &= \cup_{v \in A} pa(v) \setminus A, \\ ch(A) &= \cup_{v \in A} ch(v) \setminus A, \\ fa(A) &= A \cup pa(A), \end{aligned}$$

and

$$clan(v) = fa(ch(v) \cup \{v\}) \quad \text{for a node } v \in V.$$

For $A \subseteq V$, an *induced subgraph* of $\mathcal{G}$ confined to A is defined as $\mathcal{G}_A = (A, E_A)$, $E_A = E \cap (A \times A)$. For a set of edges E, we define $sym(E) = \{(b, a) | (a, b) \in E\} \cup E$. A graph $\mathcal{G} = (V, E)$ is *undirected* if $E = sym(E)$. The *associated undirected graph* of graph $\mathcal{G} = (V, E)$ is an undirected version of $\mathcal{G}$ and we will represent it by $\mathcal{G}^{\sim} = (V, sym(E))$.

A node which does not have any child node will be called a *terminal* node. If $(a, b) \in E$, we say that node a is *adjacent* to node b or vice versa. A sequence of nodes $u = v_1, \cdots, v_r = v$ is called a *chain* from u to v (or from v to u) if $(v_i, v_{i+1}) \in sym(E)$, $i = 1, 2, \cdots, r-1$. If $(v_i, v_{i+1}) \in E$, $i = 1, 2, \cdots, r-1$, the sequence is called a path of length r from u to v. If there is a path from u to v, we write $u \longmapsto v$. If, for $A \subseteq V$,

$$u \longmapsto v \quad \text{and } v \longmapsto u \quad \text{for every pair } u, v \in A,$$

we call A a *connectivity component* of $\mathcal{G}$.

We define the boundary of a node v in $\mathcal{G} = (V, E)$ as $bd(v) = \{u;\ \ (u, v) \in E\}$. A graph is said to be *complete* if all vertices are adjacent to each other. A complete subgraph is a subgraph which is complete. A complete subgraph that is maximal in the sense of set-inclusion in $\mathcal{G}$ is called a *clique* of $\mathcal{G}$.

For $A, B \subseteq V$, we let $P_{B|A}(x_B|x_A) = P(X_B = x_B|X_A = x_A)$. When $A = V$, we will write $P_A(x_A) = P(x)$. We denote by $n_A(x_A)$ the cell frequency at the cell-entry x_A for a set A of variables and by n the total frequency. We will denote the collection of all the cell locations x_A, for an index set A, by $\mathcal{X}_A$. If confusion is not likely, we will ignore the argument 'x' in $n_A(x_A)$. The cardinality of a set A will be denoted by $|A|$.

3 COLLAPSING OVER A VARIABLE

In this section, we will present a theorem which lists a couple of conditions that are equivalent to collapsibility over a variable in a recursive model. We say that a probability distribution $\mathcal{P}$ admits a recursive factorization according to a recursive model $\mathcal{G}$ of discrete variables (Lauritzen 1996), if there exist conditional probabilities $P_{v|pa(v)}(\cdot|\cdot),\ \ v \in V$, such that

$$P(x) = \prod_{v \in V} P_{v|pa(v)}(x_v|x_{pa(v)}).$$

Let a probability distribution $\mathcal{P}$ admit a recursive factorization according to $\mathcal{G}$. Then, the *maximum likelihood estimate* (MLE) $\widehat{P}_{\mathcal{G}}(x)$ of $P(x)$ which is obtained under the model $\mathcal{G}$ is given by

$$\widehat{P}_{\mathcal{G}}(x) = \prod_{v \in V} \widehat{P}_{v|pa(v)}(x_v|x_{pa(v)}) = \prod_{v \in V} \frac{n_{fa(v)}(x_{fa(v)})}{n_{pa(v)}(x_{pa(v)})} \tag{1}$$

where $n_\emptyset(x_\emptyset) = n$. If confusion is not likely, we will simply write $\widehat{P}(x)$ instead of $\widehat{P}_{\mathcal{G}}(x)$. It is possible that

$$pa(v) = fa(v') \tag{2}$$

for some nodes v and v'. Thus, cancelling the common terms in the numerator and the denominator of the right hand side of (1) yields

$$\widehat{P}(x) = \prod_{i=1}^{\rho} \frac{n_{C_i}(x_{C_i})}{n_{S_i}(x_{S_i})} \tag{3}$$

where $C_i, S_i \subseteq V$, $1 \leq i \leq \rho \leq |V|$ and $C_i \neq S_j$ for all i, j with $1 \leq i, j \leq \rho$ and it is possible that $S_i = \emptyset$ for some i. Obviously,

$$\{C_i \ ; \ i = 1, 2, \cdots, \rho\} \subseteq \{fa(v) \ ; \ v \in V\} \text{ and } \{S_i \ ; \ i = 1, 2, \cdots, \rho\} \subseteq \{pa(v) \ ; \ v \in V\}. \quad (4)$$

We will call S_i and C_i, respectively, *f-separator* and *f-clique* (think of factorization for "f") of the recursive model. For convenience's sake, we will denote

$$\begin{array}{rcl} n_{fa(v)}(x_{fa(v)})/n & \text{by} & [fa(v)](x_{fa(v)}), \\ n_{pa(v)}(x_{pa(v)})/n & \text{by} & [pa(v)](x_{pa(v)}), \\ n_{C_i}(x_{C_i})/n & \text{by} & [C_i](x_{C_i}), \\ \text{and } n_{S_i}(x_{S_i})/n & \text{by} & [S_i](x_{S_i}). \end{array} \quad (5)$$

When confusion is not likely, we will ignore the arguments in the above bracketed expressions $[\cdot](x_\cdot)$.

Let $x_{(v)}$ be the subvector of x with $x_{\{v\}}$ only removed; analogously for a subset A of V, we will denote by $x_{(A)}$ the subvector of x with x_A only removed. Furthermore, we will write $\mathcal{G}_{V \setminus \{v\}} = \mathcal{G}_{(v)}$.

Definition 3.1. *Let $\mathcal{G} = (V, E)$ be a recursive model. If, at all $x_{(v)} \in \mathcal{X}_{V \setminus \{v\}}$,*

$$\widehat{P}(x_{(v)}) = \widehat{P}_{\mathcal{G}_{(v)}}(x_{(v)}), \quad (6)$$

i.e., the marginal, $\widehat{P}(x_{(v)})$, of $\widehat{P}$ onto $V \setminus \{v\}$ is the same as the MLE, $\widehat{P}_{\mathcal{G}_{(v)}}(x_{(v)})$, of the marginal of P onto $\mathcal{G}_{(v)}$, then we say that P is collapsible over v and call the node v a removable *node of $\mathcal{G}$.*

We will use both of the terms, collapsibility and node removability, for convenience and clarity of expression. For instance, we will simply say that a node is removable from a recursive model $\mathcal{G}$ instead of saying that a distribution $\mathcal{P}$ which is global $\mathcal{G}$-Markov (see Appendix A) is collapsible over a variable (or a node).

Theorem 3.2. *Consider a node v^* in a recursive model $\mathcal{G} = (V, E)$. Then the following statements are equivalent.*

(i) The node v^* is contained in one and only one f-clique of $\mathcal{G}$.

(ii) The nodes in $\mathcal{G}_{clan(v^*)}$ are all adjacent to each other possibly except for the nodes in $pa(v^*)$.

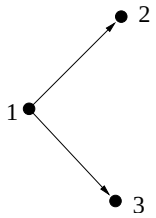

Figure 1: A recursive model of three nodes where nodes 2 and 3 are removable but node 1 is not.

(iii) The node v^* is removable from $\mathcal{G}$.

Proof: See Appendix B.

Theorem 3.2 says that every terminal node is removable and that, as for a non-terminal node, we can check easily, by applying condition (ii) of the theorem, whether the node is removable. We will see two simple examples of removable nodes, the former being simpler than the latter.

Example 3.3. *Consider a recursive model $\mathcal{G} = (V, E)$ in Figure 1. We have that*

$$\widehat{P}(x_V) = \frac{[\{1,2\}][\{1,3\}]}{n \cdot [\{1\}]}.$$

In this equation, we can see that nodes 2 and 3 are each contained in one and only one f-clique, but node 1 is contained in both of f-cliques $\{1,2\}$ and $\{1,3\}$. Furthermore, nodes 2 and 3 are terminal, and $clan(1) = \{1,2,3\}$ cannot be made into a clique in the context of condition (ii) of Theorem 3.2 as is obvious in Figure 1. Thus node 1 is not removable while nodes 2 and 3 are. □

Example 3.4. *Consider a recursive model $\mathcal{G}$ in panel (a) of Figure 2.*

$$\widehat{P}(x_V) = \frac{[\{1,2\}][\{1,3\}][\{2,3,4,5\}][\{2,5,6\}]}{n \cdot [\{1\}][\{2,3\}][\{2,5\}]}.$$

Node 4 is contained only in the f-clique $\{2,3,4,5\}$. Furthermore, if $pa(4) = \{2,3\}$ is made into a complete subgraph in the graph, node 4 is contained in the unique clique $\{2,3,4,5\}$ in graph

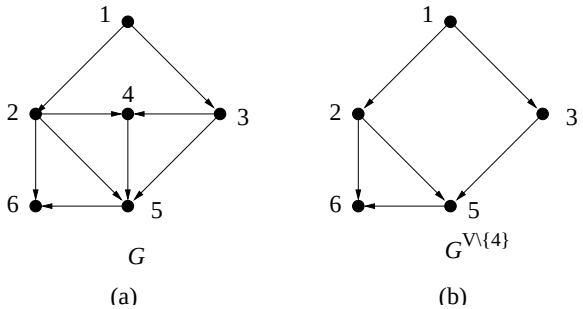

Figure 2: A recursive model $\mathcal{G}$ is of six nodes where nodes 4 and 6 are removable and DAG $\mathcal{G}_{(4)}$ is of five nodes where node 4 is removed from $\mathcal{G}$.

$\mathcal{G}$. *The induced subgraph* $\mathcal{G}_{(4)}$ *of* $\mathcal{G}$ *is in panel* (b). *Thus we have at* $x_{(4)}$

$$\begin{aligned}\widehat{P}(x_{(4)}) &= \frac{[\{1,2\}][\{1,3\}]\Big(\sum_{x_{\{4\}}}[\{2,3,4,5\}](x_{\{2,3,4,5\}})\Big)[\{2,5,6\}]}{n\cdot[\{1\}][\{2,3\}][\{2,5\}]}\\ &= \frac{[\{1,2\}][\{1,3\}][\{2,3,5\}][\{2,5,6\}]}{n\cdot[\{1\}][\{2,3\}][\{2,5\}]}\\ &= \widehat{P}_{\mathcal{G}_{(4)}}(x_{(4)}),\end{aligned}$$

where the last equality is immediate from $\mathcal{G}_{(4)}$. *We see that the three conditions* $(i),(ii),(iii)$ *of Theorem 3.2 are satisfied with graph* $\mathcal{G}$ *in panel* (a). *Node* 4 *is thus removable.* □

4 COLLAPSIBILITY IN MORAL GRAPH AND SEQUENTIAL NODE-REMOVAL

In this section, we will compare collapsibility between a recursive model and its moralized graph and show that if a set of nodes are removable, they are removable in a sequential manner. We will denote the moral graph of a recursive model $\mathcal{G}=(V,E)$ by $\mathcal{G}^m=(V,E^m)$.

Definition 4.1. *Consider an undirected graph* $\mathcal{G}^u=(V,E^u)$. *We say that* $V=(V_1,V_2)$ *is* graphically collapsible *over* V_2 *in* $\mathcal{G}^u$ *if and only if the boundary (of each connectivity component) of* V_2 *is complete in* $\mathcal{G}^u$.

Note that if V is graphically collapsible over V_2 in $\mathcal{G}^u$, then, for a distribution $\mathcal{P}$ which is Markov with respect to $\mathcal{G}^u$, it holds, with respect to $\mathcal{G}^u$, that

$$\widehat{P}(x_{V_1})=\widehat{P}_{\mathcal{G}^u_{V_1}}(x_{V_1})$$

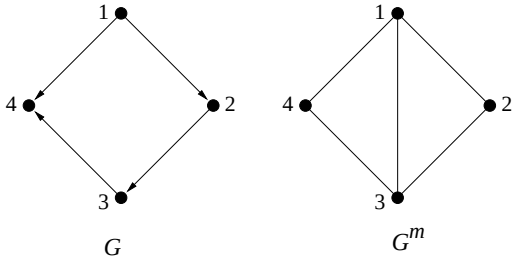

Figure 3: A recursive model and its moral graph. Nodes 2 and 3 are not sequentially removable

(Theorem 5.4 of Frydenberg (1990a)).

Theorem 4.2. *Let the moral graph of a recursive model $\mathcal{G} = (V, E)$ be $\mathcal{G}^m = (V, E^m)$. If a node v is removable from $\mathcal{G}$, then V is graphically collapsible over v in $\mathcal{G}^m$.*

Proof: If node v is removable, then, by Theorem 3.2, $clan(v)$ is the only clique containing node v in $\mathcal{G}^m$. So, the boundary of a node v is complete in the moral graph, making $\mathcal{G}^m$ graphically collapsible over v. □

The converse of this theorem does not hold. For example, Figure 3 displays a recursive model ($\mathcal{G}$) and its moral graph ($\mathcal{G}^m$). $\mathcal{G}^m$ is graphically collapsible over node 2, but the node is not removable from $\mathcal{G}$.

Consider a recursive model $\mathcal{G} = (V, E)$. If all the nodes in $A \subseteq V$ can be removed from $\mathcal{G}$ one after another in some order, we will say that A is *sequentially removable* from $\mathcal{G}$. Of course, $\emptyset$ and V are sequentially removable. However, every subset of a set of sequentially removable nodes is not necessarily sequentially removable as we see in the example below. To represent a set of nodes that are sequentially removable, we will use the symbol $\{\cdot\}^{\prec}$. For example, that $\{5, 2, 3\}^{\prec}$ is sequentially removable means that the nodes $5, 2, 3$ are sequentially removable in that order.

Example 4.3. *Consider a recursive model $\mathcal{G} = (V, E)$ with $V = \{1, 2, 3, 4\}$ and $E = \{(1, 2), (1, 3), (1, 4), (2, 3), (2, 4)\}$ as depicted in Figure 4. Let $A_1 = \{3, 2\}^{\prec}$ and $A_2 = \{4, 2\}^{\prec}$. Then A_1 and A_2 are sequentially removable, but $A_1 \cap A_2 = \{2\}$ is not.* □

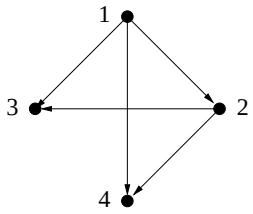

Figure 4: A recursive model of four variables

This example illustrates that the intersection of two sets of removable nodes in a recursive model $\mathcal{G}$ is not necessarily removable from $\mathcal{G}$. But, we can easily show that, if $\mathcal{G}$ is model-collapsible onto A and B, then $\mathcal{G}$ is model-collapsible onto $A \cap B$. A similar result is proved in Madigan and Mosurski (1990, Lemma 3.4) with regard to graphical log-linear models. By definition, for any set of removable nodes in a recursive model $\mathcal{G}$, there must exist a sequence according to which the nodes are removable. We summarize this as a theorem and its proof is given for interested readers.

Theorem 4.4. *For a recursive model $\mathcal{G} = (V, E)$, $A \subseteq V$ is removable from $\mathcal{G}$ if and only if A is sequentially removable from $\mathcal{G}$.*

Proof: See Appendix B.

It is important to note that, as illustrated in Figure 3, model collapsibility with regard to a recursive model cannot be determined by using the moralized graph of the recursive model. Since collapsibility is examined one node after another, we can apply Theorem 3.2 for sequential checking of the collapsibility.

5 CONCLUDING REMARKS

There is a burgeoning literature on the subject of collapsibility which goes back to Simpson (1951). This notion became a common word since parametric collapsibility was introduced in the text by Bishop et al (1975, p. 47). Asmussen and Edwards (1983) derived necessary and sufficient conditions of model collapsibility with regard to a hierarchical log-linear model for a multidimensional contingency table. Madigan and Mosurski (1990) combined these necessary and sufficient conditions with graph-theoretic algorithms to provide useful classes of sub-tables which are collapsible onto. They extended the results of Asmussen and Edwards to graphical log-linear models and their interaction graphs and showed that there is a close relationship

between model collapsibility and graphical collapsibility (see Lemma 3.3 of Madigan and Mosurski). Under the positivity assumption of the probability distribution, Frydenberg (1990a) derived several conditions, with regard to a collection of graphical interaction models and their corresponding interaction graphs, that are equivalent to model collapsibility.

We have seen in this paper, that the model collapsibility of a recursive model can not be evaluated with the moralized graph of the recursive model. The two conditions that are equivalent to model collapsibility with regard to a recursive model are however easy to apply. A node v is removable in a recursive model $\mathcal{G}$ if and only if $clan(v) = fa(ch(v) \cup \{v\})$ forms a clique in $\mathcal{G}$ when the nodes of $pa(v)$ become adjacent to each other. Asmussen and Edwards (1983) showed that there exists a sequence of nodes in a recursive model that are removable if the sequence starts from a terminal node. Theorem 3.2 enables us to check if a non-terminal node in a recursive model $\mathcal{G}$ is removable.

APPENDIX A: Markov equivalence among DAGs

In this appendix, we will describe the notion of Markov equivalence among directed acyclic graphs (DAG's) and present a theorem which is useful in proving Theorem 3.2.

Let $\mathcal{P}$ be a probability distribution defined on a product probability space $\mathcal{X} = \prod_{v \in V} \mathcal{X}_v$, where each $\mathcal{X}_v$ is sufficiently regular to ensure the existence of regular conditional probabilities. Let $\mathbf{X}$ be a random vector with its probability distribution $\mathcal{P}$. For three disjoint subsets A, B and C of V, we write $A \perp B|C\ [\mathcal{P}]$ to mean that X_A and X_B are conditionally independent given X_C under $\mathcal{P}$. For a DAG $\mathcal{G} = (V, E)$ and a subset $A \subseteq V$, if $bd(v) \subseteq A$ for all $v \in A$, we say that A is an ancestral set. Thus, the intersection of a collection of ancestral sets in $\mathcal{G}$ is again an ancestral set in $\mathcal{G}$. This implies that, for any subset A of V, there is a smallest ancestral set containing A, which we will denote by $An(A)$. If, for a DAG $\mathcal{G}$, $\mathcal{G}_{\{u,v,w\}} = (\{u, v, w\},\ \{(u, v), (w, v)\})$, we call the triple (u, v, w) an *immorality* of $\mathcal{G}$. We say that DAG's $\mathcal{G}_1$ and $\mathcal{G}_2$ are *graphically equivalent*, and write $\mathcal{G}_1 \sim \mathcal{G}_2$, if $\mathcal{G}_1^u = \mathcal{G}_2^u$ and they have the same immoralities. Chickering (1995) proposes an algorithm which is useful for checking graphical equivalency.

Given a DAG $\mathcal{G}$, we say that a probability measure $\mathcal{P}$ on a product space $\mathcal{X}_V$ is *global $\mathcal{G}$-Markovian* if $A \perp B|S\ [\mathcal{P}]$ whenever S separates A and B in $\left(\mathcal{G}_{An(A \cup B \cup S)}\right)^m$. See Proposition 3.25 of Lauritzen (1996) for an equivalent condition of the global $\mathcal{G}$-Markov condition. Two DAG's $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent on $\mathcal{X}_V$ if the classes of global $\mathcal{G}_1$-Markovian and global

$\mathcal{G}_2$-Markovian probability measures on $\mathcal{X}_V$ coincide. If $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent on every such product space $\mathcal{X}_V$, $\mathcal{G}_1$ and $\mathcal{G}_2$ are said to be *Markov equivalent*.

The theorem below was discovered by Verma and Pearl (1990, Theorem 1) and, independently, by Frydenberg (1990b, Theorem 5.6) for the more general class of chain graphs under the condition that the probability distribution $\mathcal{P}$ satisfies that

$$A \perp B|D \cup C \text{ and } A \perp C|D \cup B \text{ implies } A \perp B \cup C|D \tag{A.1}$$

whenever A, B, C, and D are disjoint subsets of V. Andersson et al. (1996) proved that the theorem holds even when $\mathcal{P}$ does not satisfy (A.1).

Theorem A.1. *Two DAG's $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if $\mathcal{G}_1 \sim \mathcal{G}_2$.*

APPENDIX B: Proofs of Theorems 3.2 and 4.4

Proof of Theorem 3.2

Let $clan(v^*) = \{v_1, \cdots, v_m, \cdots, v_\kappa\}$ where v_i's are ordered such that $v_m = v^*$ and $ch(v_m) = \{v_{m+1}, \cdots, v_\kappa\}$. Then we have, for $x \in \mathcal{X}_V$,

$$\begin{aligned}
\widehat{P}(x) &= \prod_{v \in V} \frac{[fa(v)]}{[pa(v)]} \\
&= \Big(\prod_{v \in (V \setminus clan(v_m))} \frac{[fa(v)]}{[pa(v)]} \Big) \Big(\prod_{v \in clan(v_m)} \frac{[fa(v)]}{[pa(v)]} \Big) \\
&= \Big(\prod_{v \in (V \setminus clan(v_m))} \frac{[fa(v)]}{[pa(v)]} \Big) \Big(\prod_{i=1}^{m-1} \frac{[fa(v_i)]}{[pa(v_i)]} \Big) \Big(\frac{[fa(v_m)]}{[pa(v_m)]} \Big) \Big(\prod_{i=m+1}^{\kappa} \frac{[fa(v_i)]}{[pa(v_i)]} \Big)
\end{aligned} \tag{B.1}$$

Since v_m is contained only in $fa(v_i)$, $m \le i \le \kappa$, and $pa(v_i)$, $m+1 \le i \le \kappa$, we consider only these $fa(v_i)$, $m \le i \le \kappa$, and $pa(v_i)$, $m+1 \le i \le \kappa$. For convenience's sake, let

$$F = \frac{[fa(v_m)]}{[pa(v_m)]} \prod_{i=m+1}^{\kappa} \frac{[fa(v_i)]}{[pa(v_i)]} \tag{B.2}$$

and

$$R = \Big(\prod_{v \in (V \setminus clan(v_m))} \frac{[fa(v)]}{[pa(v)]} \Big) \Big(\prod_{i=1}^{m-1} \frac{[fa(v_i)]}{[pa(v_i)]} \Big). \tag{B.3}$$

First of all, we will prove that conditions (i), (ii), and (iii) are satisfied when v_m is a terminal node. When $ch(v_m) = \emptyset$,

$$F = \frac{[fa(v_m)]}{[pa(v_m)]}.$$

This means that v_m is contained in one and only one f-clique $fa(v_m)$ in expression (B.1), which satisfies condition (i) of the theorem. Since v_m is terminal in $\mathcal{G}$, $clan(v_m)$ becomes a clique if $pa(v_m)$ is made complete. This satisfies condition (ii) of the theorem. Furthermore, $clan(v_m) = fa(v_m)$ and

$$\begin{aligned}\widehat{P}(x) &= R \cdot F \\ &= \left(\left(\prod_{v \in (V \setminus clan(v_m))} \frac{[fa(v)]}{[pa(v)]} \right) \left(\prod_{i=1}^{m-1} \frac{[fa(v_i)]}{[pa(v_i)]} \right) \right) \left(\frac{[fa(v_m)]}{[pa(v_m)]} \right) \\ &= \left(\widehat{P}_{\mathcal{G}_{(v_m)}}(x_{(v_m)}) \right) \left(\frac{[fa(v_m)]}{[pa(v_m)]} \right).\end{aligned}$$

Thus, from $fa(v_m) = pa(v_m) \cup \{v_m\}$, we have at $x_{(v_m)}$

$$\widehat{P}(x_{(v_m)}) = \sum_{x_{\{v_m\}}} \widehat{P}(x) = \left(\widehat{P}_{\mathcal{G}_{(v_m)}}(x_{(v_m)}) \right) \left(\sum_{x_{\{v_m\}}} \frac{[fa(v_m)]}{[pa(v_m)]} \right) = \widehat{P}_{\mathcal{G}_{(v_m)}}(x_{(v_m)}).$$

In other words, v_m is removable from $\mathcal{G}$. Therefore, nodes which are terminal always satisfy this theorem.

From now on, we will consider only v_m such that $ch(v_m) \neq \emptyset$. We prove the theorem by showing, first, equivalence of condition (i) with condition (ii), and then equivalence of condition (ii) with (iii). We assume condition (ii). From condition (ii) follows that $fa(v_i) = pa(v_{i+1})$, $m \leq i \leq \kappa - 1$. So, expression (B.2) becomes

$$F = \frac{[fa(v_\kappa)]}{[pa(v_m)]}$$

and in expression (B.1), v_m is involved in the term $[fa(v_k)]$ only. In other words, v_m is contained in the f-clique, $fa(v_\kappa)$, only.

To prove the sufficiency of condition (i) for condition (ii), we assume that v_m is contained in one f-clique only, say C_{v_m}. Then

$$F = \frac{[C_{v_m}]}{[pa(v_m)]} \tag{B.4}$$

since $v_m \notin pa(v_m)$. Equation (B.4), when multiplied by $[pa(v_m)]$, becomes

$$[fa(v_m)] \prod_{i=m+1}^{\kappa} \frac{[fa(v_i)]}{[pa(v_i)]} = [C_{v_m}]. \tag{B.5}$$

The left-hand side of (B.5) is a function of the X variables indexed in a set which contains $pa(v_m) \cup \{v_m, \cdots, v_\kappa\}$. According to the structure of $clan(v_m)$, $v_{m+1}, \cdots, v_\kappa$ are dependent

upon v_m at the very least. This means that $[C_{v_m}]$ must be a function of the X variables indexed in a set which contains $\{v_m, \cdots, v_\kappa\}$. In other words,

$$\{v_m, \cdots, v_\kappa\} \subseteq C_{v_m}. \tag{B.6}$$

However, $v_\kappa \notin \cup_{i=m+1}^{\kappa-1} fa(v_i)$ and $v_\kappa \in fa(v_\kappa)$. From (4) and (B.6), it follows that

$$C_{v_m} \in \{fa(v_i) \mid m \leq i \leq \kappa\}.$$

This implies that $fa(v_\kappa) = C_{v_m}$ since $v_\kappa \in C_{v_m}$. Thus we have $[fa(v_\kappa)] = [C_{v_m}]$. That is,

$$[fa(v_m)] \prod_{i=m+1}^{\kappa} \frac{[fa(v_i)]}{[pa(v_i)]} = [fa(v_\kappa)]. \tag{B.7}$$

By the definition of f-clique and (B.7), we can see that for each i, $m+1 \leq i \leq \kappa$, there exist j, $m \leq j \leq \kappa - 1$, such that

$$pa(v_i) = fa(v_j).$$

Thus, for $m \leq i \leq \kappa$,

$$pa(v_i) \in \{fa(v_j) \mid m \leq j \leq \kappa - 1\}. \tag{B.8}$$

Since $pa(v_\kappa) \subseteq fa(v_\kappa) = C_{v_m}$, we have, from (B.6) and the fact that $pa(v_\kappa) \cup \{v_\kappa\} = fa(v_\kappa)$,

$$\{v_m, \cdots, v_{\kappa-1}\} \subseteq pa(v_\kappa). \tag{B.9}$$

However, $v_{\kappa-1} \notin \cup_{j=m}^{\kappa-2} fa(v_j)$, and $v_{\kappa-1} \in fa(v_{\kappa-1})$. Thus, by (B.8) and (B.9),

$$fa(v_{\kappa-1}) = pa(v_\kappa), \tag{B.10}$$

which means $[fa(v_{\kappa-1})] = [pa(v_\kappa)]$. Proceeding to $v_{\kappa-1}$, we have from (B.8) and (B.10) that

$$pa(v_{\kappa-1}) \in \{fa(v_j) \mid m \leq j \leq \kappa - 2\}. \tag{B.11}$$

Since $pa(v_{\kappa-1}) \subseteq fa(v_{\kappa-1}) = pa(v_\kappa)$, we have, from (B.9) and the fact that $pa(v_{\kappa-1}) \cup \{v_{\kappa-1}\} = fa(v_{\kappa-1})$,

$$\{v_m, \cdots, v_{\kappa-2}\} \subseteq pa(v_{\kappa-1}).$$

However, $v_{\kappa-2} \notin \cup_{j=m}^{\kappa-3} fa(v_j)$, and $v_{\kappa-2} \in fa(v_{\kappa-2})$. Thus, by (B.11),

$$fa(v_{\kappa-2}) = pa(v_{\kappa-1}) \tag{B.12}$$

since $v_{\kappa-2} \in pa(v_{\kappa-1})$. Hence,

$$[fa(v_{\kappa-2})] = [pa(v_{\kappa-1})].$$

As for $v_{\kappa-2}$, we have from (B.8), (B.10), and (B.12),

$$pa(v_{\kappa-2}) \in \{fa(v_i) \mid m \leq i \leq \kappa - 3\}.$$

By applying the same argument as for (B.12), we can have $fa(v_j) = pa(v_{j+1})$ for $m \leq j \leq \kappa-3$. Since $fa(v_j) = pa(v_{j+1})$ for $m \leq j \leq \kappa - 1$, all the nodes in $clan(v_m)$ becomes a clique when $pa(v_m)$ is made complete and v_m is contained in $clan(v_m)$ only since v_m is surrounded by $pa(v_m)$ and $ch(v_m)$. This completes the proof for the sufficiency of condition (i) for condition (ii).

We will next prove the sufficiency of (ii) for (iii). Assuming condition (ii), we can construct a DAG $\mathcal{G}^* = (V, E^*)$ which is the same as $\mathcal{G} = (V, E)$ except that (v, v_m) is in E^* whenever either (v, v_m) or (v_m, v) is in E. We will show that $\mathcal{G}^* \sim \mathcal{G}$. Suppose that there is an edge (v_m, v) in $\mathcal{G}$ which cannot be replaced with (v, v_m) without creating or destroying an immorality in $\mathcal{G}$. If the replacement creates an immorality, it means that $v \notin pa(v_m)$ which is impossible by the definition of $clan(v_m)$. If the replacement destroys an immorality, it means that v_m is not adjacent to a parent node of v, which is also impossible by the definition of $clan(v_m)$. Hence, we have $\mathcal{G}^* \sim \mathcal{G}$.

By Theorem A.1, $\mathcal{G}^*$ and $\mathcal{G}$ are Markov equivalent. Since v_m is a terminal node in $\mathcal{G}^*$, v_m is removable from $\mathcal{G}^*$, and so is it from $\mathcal{G}$.

Finally, for the proof of the necessity of (ii) regarding (iii), we begin by supposing that condition (ii) does not hold. Then, without loss of generality, we can think of three possible situations which are displayed in $(a), (b)$, and (c) in Figure B.1. In the figure, the three situations are featured by three types of elementary violations of condition (ii). The violations are no edge between a parent and a child of v_m as depicted in panel (a), no edge between a pair of children of v_m as in panel (b), and no edge between v_m and a parent of a child of v_m as depicted in panel (c). A general form of violation may be given in a mixture of three elementary violations.

In other words, a violation of condition (ii) implies existence of at least one of the three elementary violations in $\mathcal{G}_{clan(v_m)}$. Let there be the elementary violation as in panel (a) of Figure B.1 and let $A = \{u, v_m, w\}$. It is always possible to find a data set $\{\delta(x),\ x \in \mathcal{X}_V\}$ for which it holds that

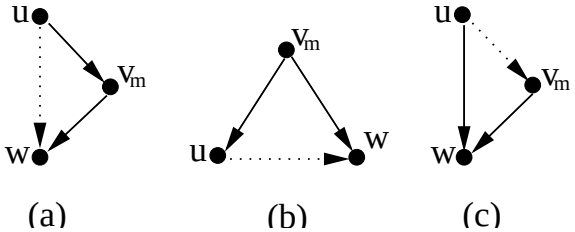

Figure B.1: A display of the three elementary violations of condition (ii) is given in panels $(a), (b)$, and (c). The dotted arrows are needed to satisfy condition (ii).

$$\frac{\delta(x)}{\delta(x_A)} = \frac{1}{|\mathcal{X}_{V\setminus A}|} \tag{B.13}$$

and

$$\sum_{x_{v_m}} \frac{\delta(x_u, x_{v_m})\delta(x_{v_m}, x_w)}{\delta(x_{v_m})} \neq \frac{\delta(x_u)\delta(x_w)}{|\delta|}, \tag{B.14}$$

where $|\delta| = \sum_{x\in\mathcal{X}_V} \delta(x)$.

Given the data set, we have

$$\begin{aligned}\widehat{P}(x) &= \prod_{v\in V} \widehat{P}_{v|pa(v)}(x_v|x_{pa(v)}) = \prod_{v\in V\setminus A} \widehat{P}_{v|pa(v)}(x_v|x_{pa(v)}) \prod_{v\in A} \widehat{P}_{v|pa(v)}(x_v|x_{pa(v)}) \\ &= \frac{1}{|\mathcal{X}_{V\setminus A}|} \prod_{v\in A} \widehat{P}_{v|pa(v)}(x_v|x_{pa(v)}) = \frac{1}{|\mathcal{X}_{V\setminus A}|} \frac{\delta(x_u, x_{v_m})\delta(x_{v_m}, x_w)}{\delta(x_{v_m})}.\end{aligned}$$

As for the MLE of the marginal $P_{\mathcal{G}_{(v_m)}}$ of P onto $\mathcal{G}_{(v_m)}$, we have

$$\begin{aligned}\widehat{P}_{\mathcal{G}_{(v_m)}}(x_{(v_m)}) &= \prod_{v\in V\setminus A} \widehat{P}^{\mathcal{G}_{(v_m)}}_{v|pa(v)}(x_v|x_{pa(v)}) \prod_{v\in A\setminus\{v_m\}} \widehat{P}^{\mathcal{G}_{(v_m)}}_{v|pa(v)}(x_v|x_{pa(v)}) \\ &= \frac{1}{|\mathcal{X}_{V\setminus A}|} \prod_{v\in A\setminus\{v_m\}} \widehat{P}^{\mathcal{G}_{(v_m)}}_{v|pa(v)}(x_v|x_{pa(v)}) = \frac{1}{|\mathcal{X}_{V\setminus A}|} \frac{\delta(x_u)\delta(x_w)}{\delta}\end{aligned}$$

where $\widehat{P}^{\mathcal{G}_{(v_m)}}_{v|pa(v)}$ denotes the MLE of $P_{v|pa(v)}$ under the model structure $\mathcal{G}_{(v_m)}$, since node u is not adjacent to node w in $\mathcal{G}_{(v_m)}$. Thus, (B.14) means that

$$\widehat{P}(x_{(v_m)}) \neq \widehat{P}_{\mathcal{G}_{(v_m)}}(x_{(v_m)}). \tag{B.15}$$

We can apply the same argument as above for the elementary violation in panel (b). As for the elementary violation in panel (c), we can find a data set $\{\delta'(x),\ x \in \mathcal{X}_V\}$ for which (B.13) holds with δ therein replaced with δ' as well as

$$\delta'(x_u) \sum_{x_{v_m}} \delta'(x_{v_m}) \frac{\delta'(x_A)}{\delta'(x_u, x_{v_m})} \neq \delta'(x_u, x_w).$$

Thus, given the data $\delta'(x)$, we end up with (B.15). In a nutshell, it is always possible to find a data set for which (B.15) holds if condition (ii) is violated. □

Proof of Theorem 4.4:

The "if" part is trivial. To prove the "only if" part, we assume the collapsibility condition given by

$$\widehat{P}(x_{(A)}) = \widehat{P}_{\mathcal{G}_{(A)}}(x_{(A)}), \tag{B.16}$$

and we will show that there exists a sequence of node-removals from $\mathcal{G}$.

By the collapsibility condition, there must exist at least one removable node, say v_1. If not, every node in A must be contained in two or more f-cliques in the expression of $\widehat{P}(x)$ in (3) by Theorem 3.2, making equation (B.16) not guaranteed in general. For a removable node, say v_1, we have at $x_{(v_1)}$,

$$\widehat{P}(x_{(v_1)}) = \widehat{P}_{\mathcal{G}_{(v_1)}}(x_{(v_1)}). \tag{B.17}$$

From equation (B.17), we have at $x_{(A)}$

$$\widehat{P}(x_{(A)}) = \sum_{x_{A\setminus\{v_1\}}} \widehat{P}(x_{(v_1)}) = \sum_{x_{A\setminus\{v_1\}}} \widehat{P}_{\mathcal{G}_{(v_1)}}(x_{(v_1)}).$$

So, from (B.16) follows that

$$\widehat{P}_{\mathcal{G}_{(A)}}(x_{(A)}) = \sum_{x_{A\setminus\{v_1\}}} \widehat{P}_{\mathcal{G}_{(v_1)}}(x_{(v_1)}). \tag{B.18}$$

Equation (B.18) means that $A \setminus \{v_1\}$ is removable from $\mathcal{G}_{(v_1)}$. So, there must exist at least one removable node, say v_2 in $A \setminus \{v_1\}$. In other words, if we let $\mathcal{G}_{(v_1)} = \mathcal{G}_1$, $\mathcal{G}_{(\{v_1,v_2\})} = \mathcal{G}_2$, and $B_2 = \{v_1, v_2\}$, we have

$$\widehat{P}_{\mathcal{G}_1}(x_{(B_2)}) = \widehat{P}_{\mathcal{G}_2}(x_{(B_2)}). \tag{B.19}$$

Applying the same argument as is applied to obtain (B.19) to the set of nodes $A \setminus B_2$ yields a sequence of nodes of A. Let the sequence be $v_1, v_2, \cdots, v_{|A|}$, and let $B_i = \{v_1, v_2, \cdots, v_i\}$. Then, we have a generalized form of (B.19) as

$$\widehat{P}_{\mathcal{G}_{(B_{i-1})}}(x_{(B_i)}) = \widehat{P}_{\mathcal{G}_{(B_i)}}(x_{(B_2)}), \quad i = 2, \cdots, |A|.$$

The corresponding sequence of induced subgraphs is given by $\mathcal{G}_{(B_1)}, \mathcal{G}_{(B_2)}, \cdots, \mathcal{G}_{(B_{|A|})}$. □

REFERENCES

Andersson, S. A., Madigan, D., and Perlman, M. D. (1997). On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs, *Scan. J. Statist.* **24**, 81-102.

Andersson, S. A., Madigan, D., and Perlman, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.* **25**(2), 505-541.

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

Bishop, Y. M. M. (1971). Effects of collapsing multidimensional contingency tables. *Biometrics* **27**, 545-562.

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

Chickering, D.M. (1995). A transformational charaterization of equivalent Bayesian network structures. In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence* (P. Besnard and S. Hanks, eds.) 87-98. Morgan Kaufmann, San Mateo, CA.

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

Didelez, V. and Edwards, D. (2004). Collapsibility of graphical CG-regression models. *Scan. J. Statist.*. To appear.

Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data* (2nd ed.). Cambridge, MA: MIT Press.

Frydenberg, M. (1990a). Marginalization and collapsibility in graphical interaction models. *Ann. Statist.* **18**, 790-805.

Frydenberg, M. (1990b). The chain graph Markov property. *Scan. J. Statist.* **17**, 333-353.

Geng, Z. (1992). Collapsibility of relative risk in contingency tables with a response variable. *J. Roy. Statist. Soc. Ser. B* **54**, 583-593.

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

Madigan, D. and Mosurski, K. (1990). An extension of the results of Asmussen and Edwards on collapsibility in contingency tables. *Biometrika* **77**, 315-319.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc. Ser. B* **13**, 238-241.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Uncertainty in Artificial Intelligence: Proceedings of the Sixth Conference* (M. Henrion, R. Shachter, L. Kanal, and J. Lemmer, eds.) 220-227. Morgan Kaufman, San Francisco.

Wermuth, N. (1987). Parametric collapsibility and the lack of moderating effects in contingency tables with a dichotomous response variable. *J. Roy. Statist. Soc. Ser. B* **49**, 353-364.

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

Whittemore, A. S. (1978). Collapsibility of multidimensional contingency tables. *J. Roy. Statist. Soc. Ser. B* **40**, 328-340.