MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

Fibers of sample size two of hierarchical models and Markov bases of decomposable models for contingency tables

Hisayuki HARA, Satoshi AOKI and Akimichi TAKEMURA

METR 2006–66 December 2006

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.i.u-tokyo.ac.jp/mi/mi-e.htm

Fibers of sample size two of hierarchical models and Markov bases of decomposable models for contingency tables

Hisayuki Hara
Department of Geosystem Engineering
University of Tokyo
Satoshi Aoki
Department of Mathematics and Computer Science
Kagoshima University
and
Akimichi Takemura
Graduate School of Information Science and Technology
University of Tokyo

December 2006

Abstract

We study Markov bases of hierarchical models in general and those of decomposable models in particular for multiway contingency tables by determining the structure of fibers of sample size two. We prove that the number of elements of fibers of sample size two are powers of two and we characterizes the primitive moves of Markov bases in terms of connected components of a certain graph defined from the generating class of a hierarchical model. This allows us to derive a complete description of minimal Markov bases and minimal invariant Markov bases for decomposable models in view of the fact that they posses Markov bases consisting of primitive moves, i.e. square-free moves of degree two.

1 Introduction

Hierarchical models are of basic importance for statistical analysis of multiway contingency tables (e.g. [13]). Decomposable models defined in terms of chordal graphs are particularly useful submodels of hierarchical models. Chordal graphs find applications in many fields as essential components of explicit solutions for some classes of nonlinear

problems. For estimation of parameters, the maximum likelihood estimators of decomposable models are explicitly written as rational functions of marginal frequencies of the model. It is also possible to give explicit improvements of maximum likelihood estimators under the Poisson sampling scheme in a decision theoretic framework ([10],[9]).

For testing goodness of fit of hierarchical models, Markov chain Monte Carlo approach based on Markov bases is very useful ([5]). However the structure of Markov bases for general hierarchical model is very difficult as illustrated in [2]. Again decomposable models are particularly simple in this respect because they possess Markov bases consisting of primitive moves, i.e. square-free moves of degree two. ([6],[11],[7]). These Markov bases are explicitly constructed based on a clique tree of the chordal graph defining the decomposable model and one Markov basis suffices for performing goodness of fit tests in applications. However from theoretical viewpoint it is important to study and clarify the properties of Markov bases for decomposable models, because they give insights into Markov bases for more general and difficult cases.

The present authors have been studying Markov basis from the viewpoint of minimality ([2], [16]) and invariance ([1], [3]). In this paper we clarify structures of primitive moves and prove that the sizes of fibers of minimal Markov bases are powers of two. All elements of fibers of sample size two are indispensable monomials ([4]). This result enables us to explicitly describe minimal Markov bases and minimal invariant Markov bases for decomposable models. We also give a necessary and sufficient condition for the uniqueness of the minimal Markov basis for decomposable models.

The organization of the paper is as follows. In Section 2 we setup notations for this paper and summarize preliminary results. In Section 3 we clarify structures of fibers of sample size two. Using this characterization in Section 4 we give a complete description of minimal Markov bases and minimal invariant Markov bases for decomposable models. In Section 5 we briefly discuss reduced Gröbner bases for decomposable models and we end the paper with some concluding remarks in Section 6.

2 Preliminaries

2.1 Preliminaries on contingency tables and Markov bases

In this section we setup appropriate notations of multiway contingency tables and summarize some preliminary results on decomposable models and Markov bases needed for this paper.

Concerning the notation for multiway contingency tables we mostly follow [13], [11] and [6]. Let $\Delta = \{1, \ldots, m\}$ denote the set of variables of an m-way contingency table. Let I_δ, $\delta \in \Delta$, denote the number of levels of the variable δ. For convenience we take the set of levels of the variable δ as $\{0, 1, \ldots, I_\delta - 1\}$ starting from 0 as in [11]. The cells of the contingency table are indexed by

$$i = (i_1, \ldots, i_m) \in \mathcal{I} = \{0, 1, \ldots, I_1 - 1\} \times \cdots \times \{0, 1, \ldots, I_m - 1\}.$$

$n(i)$ denotes the frequency of the cell i and $\boldsymbol{n} = \{n(i)\}_{i \in \mathcal{I}}$ denotes an m-way contingency table. The set of positive cells $\mathrm{supp}(\boldsymbol{n}) = \{i \in \mathcal{I} \mid n(i) > 0\}$ is the *support* of $\boldsymbol{n}$.

For a subset $D \subset \Delta$ of the variables, the D-marginal $\boldsymbol{n}_D$ of $\boldsymbol{n}$ is the contingency table with marginal cells $i_D \in \mathcal{I}_D = \prod_{\delta \in D} \{0, 1, \ldots, I_\delta - 1\}$ and entries given by $n_D(i_D) = \sum_{i_{D^C} \in \mathcal{I}_{D^C}} n(i_D, i_{D^C})$. Here we are denoting $i = (i_D, i_{D^C})$ by ignoring the order of the indices.

Next we summarize some terminology and facts on hierarchical models and decomposable models from [13]. Let $\mathcal{D} = \{D_1, \ldots, D_r\}$ be a set of subsets of Δ, such that there is no inclusion relation among D_i's and $\Delta = \cup_{i=1}^r D_i$. $\mathcal{D}$ is called the generating class of a hierarchical model. A hierarchical model with the generating class $\mathcal{D}$ is called graphical if $\mathcal{D} = \{D_1, \ldots, D_r\}$ is the set of (maximal) cliques of an undirected graph $\mathcal{G}$ with the set of vertices Δ. In this paper by a clique we mean the set of vertices of a maximal complete induced subgraph. A graphical model is called *decomposable* if $\mathcal{G}$ is chordal (decomposable, triangulated), i.e. every cycle of $\mathcal{G}$ with length greater than three has a chord. A *clique tree* (or a junction tree) $\mathcal{T}$ of a chordal graph $\mathcal{G}$ is a tree, such that the vertices of $\mathcal{T}$ are cliques of $\mathcal{G}$ and it satisfies the following property:

$$D_s \cap D_t \subset D_u \quad \text{for all } D_u \text{ on the path between } D_s \text{ and } D_t \text{ in } \mathcal{T}.$$

An intersection S of neighboring cliques in a clique tree is called a separator. S separates $\mathcal{T}$ into two subtrees and let A and B denote the unions of cliques of two subtrees, respectively. Then S decomposes $\mathcal{G}$ as $\{A \setminus S, B \setminus S, S\}$, where S is a minimal vertex separator between any vertex in $A \setminus S$ and any vertex in $B \setminus S$. In the following $\mathcal{S}$ denotes the set of minimal vertex separators of a chordal graph. In this paper, when $\mathcal{G}$ is not connected, we regard the empty set $\emptyset$ as a minimal vertex separator of $\mathcal{G}$.

For a clique $D \in \mathcal{D}$, let $\mathrm{Simp}(D)$ denote the set of simplicial vertices in D and let $\mathrm{Sep}(D)$ denote the set of non-simplicial vertices in D. Then $D = \mathrm{Simp}(D) \cup \mathrm{Sep}(D)$ is a partition (disjoint union) of D ([8]). If $\mathrm{Simp}(D) \neq \emptyset$, D is called a simplicial clique. A simplicial clique D is called a *boundary clique* if there exists another clique $D' \in \mathcal{D}$ such that $\mathrm{Sep}(D) = D \cap D'$ ([14]). Simplicial vertices in boundary cliques are called simply separated vertices ([8]). Hara and Takemura[8] showed that a clique D is boundary clique if and only if there exists a clique tree $\mathcal{T}$ such that D is its endpoint. Hence there exists at least two boundary clique in any chordal graph.

Finally we summarize some relevant facts on fibers and Markov bases ([16], [17]). Given the generating class $\mathcal{D} = \{D_1, \ldots, D_r\}$ of a hierarchical model, we denote the set of marginal frequencies as

$$\boldsymbol{b} = \{n_{D_j}(i_{D_j}), i_{D_j} \in \mathcal{I}_{D_j}, j = 1, \ldots, r\}.$$

We consider $\boldsymbol{b}$ as a column vector with dimension $d = \sum_{j=1}^r \prod_{\delta \in D_j} I_\delta$, where the elements are ordered according to an appropriate lexicographical order. We also order the elements of $\boldsymbol{n}$ appropriately and consider $\boldsymbol{n}$ as a column vector. Then the relation between the joint frequencies $\boldsymbol{n}$ and the marginal frequencies $\boldsymbol{b}$ is written simply as

$$\boldsymbol{b} = A\boldsymbol{n},$$

where A is a $d \times (\prod_{\delta \in \Delta} I_\delta)$ matrix consisting of 0's and 1's. A is the "incidence matrix" of cells and marginals with 1 indicating that the corresponding cell (column) is included in the corresponding marginal (row).

The marginal tables $\boldsymbol{n}_{D_1}, \ldots, \boldsymbol{n}_{D_r}$ are *consistent* ([6]) if, for any r_1, r_2, the $(D_{r_1} \cap D_{r_2})$-marginal of $\boldsymbol{n}_{D_{r_1}}$ is equal to the $(D_{r_1} \cap D_{r_2})$-marginal of $\boldsymbol{n}_{D_{r_2}}$.

Given $\boldsymbol{b}$, the set

$$\mathcal{F}_{\boldsymbol{b}} = \{\boldsymbol{n} \geq 0 \mid \boldsymbol{b} = A\boldsymbol{n}\}$$

of contingency tables sharing the same marginal frequencies $\boldsymbol{b}$ is called a *fiber* or $\boldsymbol{b}$-fiber. All contingency tables $\boldsymbol{n}$ in the same fiber $\mathcal{F}_{\boldsymbol{b}}$ has the same total frequency $n = \sum_{i \in \mathcal{I}} n(i)$. We call this common total frequency *sample size* or *degree* of $\boldsymbol{b}$ and denote it by $\deg \boldsymbol{b}$. Therefore "fibers of sample size two" in the title of this paper means $\mathcal{F}_{\boldsymbol{b}}$ with $\deg \boldsymbol{b} = 2$. For brevity in the following we use the term "degree two fibers".

An integer array $\boldsymbol{z}$ of the same dimension as $\boldsymbol{n}$ is called a *move* if $A\boldsymbol{z} = \boldsymbol{0}$, i.e., all the marginal sums of $\boldsymbol{z}$ are zeros. Moves are used for steps of Markov Chain Monte Carlo simulation within each fiber. If we add a move or subtract a move $\boldsymbol{z}$ to $\boldsymbol{n} \in \mathcal{F}_{\boldsymbol{b}}$, then $\boldsymbol{n} \pm \boldsymbol{z} \in \mathcal{F}_{\boldsymbol{b}}$ and we can move from $\boldsymbol{n}$ to another state $\boldsymbol{n} + \boldsymbol{z}$ (or $\boldsymbol{n} - \boldsymbol{z}$) in the same fiber $\mathcal{F}_{\boldsymbol{b}}$, as long as there is no negative element in $\boldsymbol{n} \pm \boldsymbol{z}$. A finite set $\mathcal{M}$ of moves is called a *Markov basis* if for every fiber the states become mutually accessible by the moves from $\mathcal{M}$. A Markov basis $\mathcal{M}$ is *minimal* if every proper subset of $\mathcal{M}$ is no longer a Markov basis. When we separate positive elements and negative elements of a move, then each move $\boldsymbol{z}$ is written as difference of its positive part and negative part as $\boldsymbol{z} = \boldsymbol{z}^+ - \boldsymbol{z}^-$. Then $A\boldsymbol{z}^+ = A\boldsymbol{z}^-$. Therefore the positive part and the negative of a move belong to the same fiber. In this case we simply say that a move $\boldsymbol{z}$ belongs to the fiber $\mathcal{F}_{A\boldsymbol{z}^+}$. Minimal Markov bases may not be unique, but the fibers of the moves of all minimal Markov bases are common. See the definition of the minimum fiber Markov basis in [17]. In this paper we refer to the set of fibers common to all minimal Markov bases as the fibers of the minimum fiber Markov basis.

Note that there exists no move of degree 1, since we are considering m-way contingency tables with $m \geq 2$. Suppose that a degree two fiber $\mathcal{F}_{\boldsymbol{b}}$ contains more than one element ($|\mathcal{F}_{\boldsymbol{b}}| \geq 2$). Then no two elements $\boldsymbol{n}, \boldsymbol{n}'$ of the fiber share a support:

$$\deg \boldsymbol{b} = 2,\ |\mathcal{F}_{\boldsymbol{b}}| \geq 2,\ \boldsymbol{n} \neq \boldsymbol{n}' \in \mathcal{F}_{\boldsymbol{b}} \quad \Rightarrow \quad \mathrm{supp}(\boldsymbol{n}) \cap \mathrm{supp}(\boldsymbol{n}') = \emptyset.$$

It follows that each element of a degree two fiber with more than one element is an indispensable monomial ([4]), i.e., each contingency table of sample size two is isolated and has to be connected to some other tables in the same fiber by a degree two move of a Markov basis. Hence each degree two fiber with more than one element has to be a fiber of the minimum fiber Markov basis. This fact holds for any hierarchical model. Note however that for some hierarchical models, such as no-three factor interaction models ([2]), every degree two fiber has only one element.

On the other hand for decomposable models, Dobra[6] and Hoşten and Sullivant[11] have shown that there exists a Markov basis consisting of primitive moves, i.e. square-free moves of degree two. It implies that for decomposable models it suffices to study

degree two fibers. In particular the fibers of the minimum fiber Markov bases are exactly the degree two fibers with more than one element. Furthermore by the characterization of the uniqueness of minimal Markov bases in Takemura and Aoki [16], it follows that minimal Markov basis for a decomposable model is unique if and only if all degree two fibers contain at most two elements. Based on this result we will give a necessary and sufficient condition for the uniqueness of minimal Markov bases for decomposable models (Theorem 3 below) in terms of the properties of their chordal graphs.

3 Structure of degree two fibers

In this section we study the structure of degree two fibers. Let $\mathcal{D} = \{D_1, \ldots, D_r\}$ be the generating class of a hierarchical model. Let $\boldsymbol{b}$ be the set of marginal frequencies of a contingency table with sample size two. We are interested in the structure of the degree two fiber $\mathcal{F}_{\boldsymbol{b}}$. Because the sample size is two, for each $D \in \mathcal{D}$, there exists at most two marginal cells i_D with positive marginal frequency $n_D(i_D) > 0$. The same reasoning holds for each variable $\delta \in \Delta$, namely in the one-dimensional marginal table $\{n_{\{\delta\}}(i_\delta), i_\delta \in \{0, 1, \ldots, I_\delta - 1\}\}$ there exist at most two levels i_δ such that $n_{\{\delta\}}(i_\delta) > 0$. For a given $\boldsymbol{b}$ we say that the variable δ is *degenerate* if there exists a unique level i_δ such that $n_{\{\delta\}}(i_\delta) = 2$. Otherwise, if there exist two levels $i_\delta \neq i'_\delta$ such that $n_{\{\delta\}}(i_\delta) = n_{\{\delta\}}(i'_\delta) = 1$, then we say that the variable δ is *nondegenerate*.

If a variable δ is degenerate, then the level of the variable δ is uniquely determined from the marginal and it is common for all contingency tables $\boldsymbol{n} \in \mathcal{F}_{\boldsymbol{b}}$. In particular if all the variables $\delta \in \Delta$ are degenerate, then $\mathcal{F}_{\boldsymbol{b}} = \{\boldsymbol{n}\}$ is a one-element fiber with frequency $n(i) = 2$ at a particular cell i. Since this case is trivial, below we consider the case that at least one variable is nondegenerate.

From the fact that there exist at most two levels with positive one-dimensional marginals for each variable, it follows that we only need to consider $2 \times \cdots \times 2$ tables for studying degree two fibers. Therefore for our purposes we let $I_1 = \cdots = I_m = 2$, $\mathcal{I} = \{0, 1\}^k$, without loss of generality.

For a given $\boldsymbol{b}$ of degree two let $\bar{\Delta}_{\boldsymbol{b}}$ denote the set of nondegenerate variables. As noted above we assume that $\bar{\Delta}_{\boldsymbol{b}} \neq \emptyset$. Each $\boldsymbol{n} \in \mathcal{F}_{\boldsymbol{b}}$ has frequency one in two different cells $i = (i_1, \ldots, i_m) \neq i' = (i'_1, \ldots, i'_m)$, $1 = n(i) = n(i')$. Furthermore for nondegenerate $\delta \in \bar{\Delta}_{\boldsymbol{b}}$ the levels of the variable δ in i and i' are different:

$$\{i_\delta, i'_\delta\} = \{0, 1\}, \quad \forall \delta \in \bar{\Delta}_{\boldsymbol{b}},$$

or equivalently $i'_\delta = 1 - i_\delta$, $\forall \delta \in \bar{\Delta}_{\boldsymbol{b}}$. In the following we use the notation $i^*_\delta = 1 - i_\delta$. More generally for a subset $D = \{\delta_1, \ldots, \delta_k\}$ of the variables and a marginal cell $i_D = (i_{\delta_1}, \ldots, i_{\delta_k})$ we write

$$i^*_D \equiv (i^*_{\delta_1}, \ldots, i^*_{\delta_k}) = (1 - i_{\delta_1}, \ldots, 1 - i_{\delta_k}).$$

Let us identify $\boldsymbol{n} \in \mathcal{F}_{\boldsymbol{b}}$ with the set $\{i, i'\}$ of its two cells of frequency one. Then we see that the number of elements $|\mathcal{F}_{\boldsymbol{b}}|$ of the fiber is at most $2^{|\bar{\Delta}_{\boldsymbol{b}}|-1}$. However some choice

of $\{i, i'\}$ with

$$i_\delta, i^*_\delta \in \{0, 1\}, \quad \forall \delta \in \bar{\Delta}_{\boldsymbol{b}},$$

may not be in the fiber $\mathcal{F}_{\boldsymbol{b}}$. This is because if δ and δ' belong to a common $D \in \mathcal{D}$, then the values of i_δ and $i_{\delta'}$ are tied together. For example let $D = \{1, 2\} \in \mathcal{D}$ and consider the $\{1, 2\}$-marginal specified as

$$n_{\{1,2\}}(0, 0) = n_{\{1,2\}}(1, 1) = 1, \quad n_{\{1,2\}}(0, 1) = n_{\{1,2\}}(1, 0) = 0.$$

Then if we choose $i_1 = 0$, then we have to choose $i_2 = 0$. In [18] we considered a very similar problem in the framework of swapping of observations among two records in a microdata set for the purpose of statistical disclosure control. As in [18] we make the following definition.

Let $\mathcal{G}(\bar{\Delta}_{\boldsymbol{b}})$ be a graph with the set of vertices $\bar{\Delta}_{\boldsymbol{b}}$ and an edge between $\delta \in \bar{\Delta}_{\boldsymbol{b}}$ and $\delta' \in \bar{\Delta}_{\boldsymbol{b}}$ if and only if there exists some $D \in \mathcal{D}$ such that $\delta, \delta' \in \mathcal{D}$. Namely there exists an edge between two nondegenerate variables if and only if these two variables appear together in some marginal tables of $\mathcal{D}$. As discussed above in this case the values of i_δ and $i_{\delta'}$ are tied together and once the value of i_δ is chosen, e.g. $i_\delta = 0$, then the value of $i_{\delta'}$ becomes fixed, depending on the specifications of the marginals n_D.

Another way of defining $\mathcal{G}(\bar{\Delta}_{\boldsymbol{b}})$ is as follows. Given a generating class $\mathcal{D}$, we define a graph $\mathcal{G}^{\mathcal{D}}$ *generated* by $\mathcal{D}$ with the vertex set Δ and an edge between $\delta, \delta' \in \Delta$ if and only if there exists $D \in \mathcal{D}$ such that $\delta, \delta' \in D$. Note that the graphical model associated with $\mathcal{G}^{\mathcal{D}}$ is the smallest graphical model containing the hierarchical model with the generating class $\mathcal{D}$. Then $\mathcal{G}(\bar{\Delta}_{\boldsymbol{b}})$ is the induced subgraph of $\mathcal{G}^{\mathcal{D}}$ with the vertices restricted to $\bar{\Delta}_{\boldsymbol{b}}$.

We summarize the above argument in the following lemma.

Lemma 1. *Suppose that $\boldsymbol{b}$ is a set of consistent marginal frequencies of a contingency table with sample size two. Let Γ be any subset of a connected component in $\mathcal{G}(\bar{\Delta}_{\boldsymbol{b}})$. Then the marginal table $\boldsymbol{n}_\Gamma = \{n_\Gamma(i_\Gamma) \mid i_\Gamma \in \mathcal{I}_\Gamma\}$ is uniquely defined.*

Proof. Let $r(\Gamma)$ be the number of generating sets $D \in \mathcal{D}$ satisfying $\Gamma \cap D \neq \emptyset$. We prove this lemma by induction on $r(\Gamma)$. When $r(\Gamma) = 1$, the lemma holds from the consistency of $\boldsymbol{b}$. Suppose that the lemma holds for all $r(\Gamma) < r$ and we now assume that $r(\Gamma) = r$. Let $\Gamma_1 \subset \Gamma$ and $\Gamma_2 \subset \Gamma$ satisfy

$$\Gamma_1 \cup \Gamma_2 = \Gamma, \quad \Gamma_1 \cap \Gamma_2 \neq \emptyset, \quad r(\Gamma_1) < r, \quad r(\Gamma_2) < r.$$

Since $r(\Gamma_1) < r$ and $r(\Gamma_2) < r$ both $\boldsymbol{n}_{\Gamma_1}$ and $\boldsymbol{n}_{\Gamma_2}$ are uniquely defined. Suppose that

$$n_{\Gamma_1}(i_{\Gamma_1 \setminus \Gamma_2}, i_{\Gamma_1 \cap \Gamma_2}) = 1, \quad n_{\Gamma_1}(i^*_{\Gamma_1 \setminus \Gamma_2}, i^*_{\Gamma_1 \cap \Gamma_2}) = 1. \tag{1}$$

Then there uniquely exists $i_{\Gamma_2 \setminus \Gamma_1} \in \mathcal{I}_{\Gamma_2 \setminus \Gamma_1}$ such that

$$n_{\Gamma_2}(i_{\Gamma_2 \setminus \Gamma_1}, i_{\Gamma_1 \cap \Gamma_2}) = 1, \quad n_{\Gamma_2}(i^*_{\Gamma_2 \setminus \Gamma_1}, i^*_{\Gamma_1 \cap \Gamma_2}) = 1. \tag{2}$$

Hence the table $\boldsymbol{n}_\Gamma = \{n(j_\Gamma) \mid j_\Gamma \in \mathcal{I}_\Gamma\}$ such that

$$n(j_\Gamma) = \begin{cases} 1, & \text{if} \quad j_\Gamma = (i_{\Gamma_1 \setminus \Gamma_2}, i_{\Gamma_1 \cap \Gamma_2}, i_{\Gamma_2 \setminus \Gamma_1}) \quad \text{or} \quad j_\Gamma = (i^*_{\Gamma_1 \setminus \Gamma_2}, i^*_{\Gamma_1 \cap \Gamma_2}, i^*_{\Gamma_2 \setminus \Gamma_1}), \\ 0, & \text{otherwise,} \end{cases}$$

is consistent with the marginal $\boldsymbol{b}$.

Suppose that there exists another marginal table $\boldsymbol{n}'_\Gamma$ which is consistent with $\boldsymbol{b}$ such that $n_\Gamma(j_\Gamma) = n_\Gamma(j^*_\Gamma) = 1$ and $j_\Gamma \neq (i_{\Gamma_1 \setminus \Gamma_2}, i_{\Gamma_1 \cap \Gamma_2}, i_{\Gamma_2 \setminus \Gamma_1})$. Then we have at least

$$n_\Gamma(i_{\Gamma_1}) = 0 \quad \text{or} \quad n_\Gamma(i_{\Gamma_2}) = 0.$$

This contradicts (1) and (2). □

By using the result of Lemma 1, we provide the following main theorem.

Theorem 1. *Let $\mathcal{F}_{\boldsymbol{b}}$ be a degree two fiber such that $\bar{\Delta}_{\boldsymbol{b}} \neq \emptyset$ and let c be the number of connected components of $\mathcal{G}(\bar{\Delta}_{\boldsymbol{b}})$. Then*

$$|\mathcal{F}_{\boldsymbol{b}}| = 2^{c-1}.$$

Proof. Denote by $\Gamma_1, \ldots, \Gamma_c$ the connected components of $\mathcal{G}(\bar{\Delta}_{\boldsymbol{b}})$. Define Γ_{c+1} by $\Gamma_{c+1} = \Delta \setminus \bar{\Delta}_{\boldsymbol{b}}$. Then there exists $i_{\Gamma_{c+1}}$ such that

$$i_{\Gamma_{c+1}} = \{i_\delta \mid \delta \in \Gamma_{c+1},\ n_{\{\delta\}}(i_\delta) = 2\}.$$

From Lemma 1 the marginal cells i_{Γ_k} such that $n_{\Gamma_k}(i_{\Gamma_k}) = n_{\Gamma_k}(i^*_{\Gamma_k}) = 1$ uniquely exists for $k = 1, \ldots, c$. Now define $\mathcal{I}_{\boldsymbol{b}}$ by

$$\mathcal{I}_{\boldsymbol{b}} = \{i_{\Gamma_1}, i^*_{\Gamma_1}\} \times \{i_{\Gamma_2}, i^*_{\Gamma_2}\} \times \cdots \times \{i_{\Gamma_c}, i^*_{\Gamma_c}\} \times \{i_{\Gamma_{c+1}}\},$$

where $\times$ denotes the direct product of sets. Suppose that $j \in \mathcal{I}_{\boldsymbol{b}}$. Define $\boldsymbol{n}_j = \{n_j(i) \mid i \in \mathcal{I}\}$ by

$$n_j(i) = \begin{cases} 1, & \text{if} \quad i = j \ \text{ or } \ i = j^* \\ 0, & \text{otherwise.} \end{cases}$$

Then we have $\mathcal{F}(\mathcal{I}_{\boldsymbol{b}}) = \{\boldsymbol{n}_j \mid j \in \mathcal{I}_{\boldsymbol{b}}\} \subseteq \mathcal{F}_{\boldsymbol{b}}$ and $|\mathcal{F}(\mathcal{I}_{\boldsymbol{b}})| = 2^{c-1}$.

If there exists $\boldsymbol{n}' = \{n'(i) \mid i \in \mathcal{I}\}$ such that $\boldsymbol{n}' \in \mathcal{F}_{\boldsymbol{b}}$ and $\boldsymbol{n}' \notin \mathcal{F}(\mathcal{I}_{\boldsymbol{b}})$, then there exists a cell $j \in \mathcal{I}$ and $1 \leq k \leq c+1$ such that $n(j) = 1$ and $j_{\Gamma_k} \neq i_{\Gamma_k}$. This implies that there exists $D_l \in \mathcal{D}$ such that $n'(i_{D_l}) \neq n(i_{D_l})$. Hence we have $|\mathcal{F}(\boldsymbol{b})| = 2^{c-1}$. □

In general for a consistent $\boldsymbol{b}$ such that $\deg \boldsymbol{b} > 2$, it is known that $\mathcal{F}_{\boldsymbol{b}}$ is not necessarily non-empty (e.g. [12]). However Theorem 1 shows that in the case of $\deg \boldsymbol{b} = 2$ if a consistent $\boldsymbol{b}$ such that $\bar{\Delta}_{\boldsymbol{b}} \neq \emptyset$ is given, then $\mathcal{F}_{\boldsymbol{b}} \neq \emptyset$ for any hierarchical model.

It is helpful to consider permuting the labels $0 \leftrightarrow 1$ for each variable and state Theorem 1 in a canonical form. This amounts to considering invariance of hierarchical models with respect to permutation of levels of each variable as studied in [1]. Although we have reduced our consideration to 2^m tables in treating degree two fibers, we are really considering general hierarchical models of $I_1 \times \cdots \times I_m$ tables. Note that hierarchical models posses the symmetry with respect to relabeling the levels of each variable, i.e. it is invariant under the action of the direct product of symmetric group $S_{I_1} \times \cdots \times S_{I_m}$ acting on the set of cells. If we again restrict our consideration to degree two fibers, we

only need to consider the action of $S_2^m = S_2 \times \cdots \times S_2$. It is clear that structures of degree two fibers are invariant under the action of S_2^m.

In particular as a "representative fiber", we can consider $\boldsymbol{b}$ such that the levels of all degenerate variables are determined as 0. Also for such a $\boldsymbol{b}$, let $\Gamma \subset \bar{\Delta}_{\boldsymbol{b}}$ be the set of vertices of a connected component of $\mathcal{G}(\bar{\Delta}_{\boldsymbol{b}})$. Then we can without loss of generality assume that two Γ-marginal cells of frequency 1 is specified as

$$1 = n_\Gamma(0, 0, \ldots, 0) = n_\Gamma(1, 1, \ldots, 1). \tag{3}$$

This can be achieved by interchanging the levels of each variable in $\bar{\Delta}_{\boldsymbol{b}}$. Under this standardization the proof of Theorem 1 is easier to understand, because for each connected component of $\mathcal{G}(\bar{\Delta}_{\boldsymbol{b}})$ we either choose all 0's or all 1's for the component.

This standardization is also useful in determining the setwise stabilizer of $\mathcal{F}_{\boldsymbol{b}}$ in S_2^m (Section 3.1 of [3]). If we standardize the levels as (3), then the setwise stabilizer of $\mathcal{F}_{\boldsymbol{b}}$ is isomorphic to c-fold direct product of S_2's:

$$S_2^c = S_2 \times \cdots \times S_2.$$

In the next section we use this fact in determining minimal invariant Markov bases for decomposable models.

Finally we prove the following theorem on a sufficient condition for non-uniqueness of minimal Markov bases.

Theorem 2. *Let $\mathcal{D} = \{D_1, \ldots, D_r\}$ be the generating class of a hierarchical model. Suppose that $m \geq 3$ and there exist three variables k_1, k_2, k_3 which are not connected to each other in $\mathcal{G}^{\mathcal{D}}$. Then minimal Markov bases for the hierarchical model with the generating class $\mathcal{D}$ are not unique.*

Proof. It suffices to find a degree two fiber with more than two elements. From the condition of the theorem $\mathcal{G}^{\mathcal{D}}$ has at least three connected components. Therefore $|\mathcal{F}_{\boldsymbol{b}}| \geq 4$. This completes the proof. □

4 Markov bases for decomposable models

4.1 Minimal and unique minimal Markov bases

In this section we investigate Markov bases of decomposable models in detail based on Theorem 1. We also give a necessary and sufficient condition for the uniqueness of minimal Markov bases.

As already discussed at the end of Section 1, for decomposable models there exists a Markov basis with primitive moves and the set of fibers of the minimum fiber Markov bases coincides with the set of degree two fibers with more than one element. Combined with Theorem 1 of the previous section, this gives a complete description of minimal Markov bases of decomposable models.

Let $\deg \boldsymbol{b} = 2$. As mentioned in the previous section, $\boldsymbol{b}$ is in a one-to-one correspondence to $\bar{\Delta}_{\boldsymbol{b}}$. Let $\mathcal{T}_{\boldsymbol{b}}$ be any tree whose set of nodes is $\mathcal{F}_{\boldsymbol{b}}$. Denote the set of edges in $\mathcal{T}_{\boldsymbol{b}}$ by $\mathcal{M}_{\boldsymbol{b}}$. We note that we can identify each edge $(\boldsymbol{n}, \boldsymbol{n}') \in \mathcal{M}_{\boldsymbol{b}}$ with a move $\boldsymbol{z} = \boldsymbol{n} - \boldsymbol{n}'$. So we identity $\mathcal{M}_{\boldsymbol{b}}$ with the set of moves for $\mathcal{F}_{\boldsymbol{b}}$. In considering Markov bases, we ignore the sign of $\boldsymbol{z}$ and identify $\boldsymbol{z} = \boldsymbol{n} - \boldsymbol{n}'$ with $-\boldsymbol{z} = \boldsymbol{n}' - \boldsymbol{n}$ and consider the edges in $\mathcal{T}_{\boldsymbol{b}}$ as undirected. In contrast when we consider Gröbner basis, we distinguish $\boldsymbol{z}$ from $-\boldsymbol{z}$ and correspondingly consider directed edges.

Let B denote the set of non-degenerate $\boldsymbol{b}$. Then we define $\mathcal{M}$ as follows,

$$\mathcal{M} = \bigcup_{\boldsymbol{b} \in B} \mathcal{M}_{\boldsymbol{b}}. \tag{4}$$

By following Dobra[6] and Takemura and Aoki[16], we easily obtain the following theorem.

Theorem 3. *$\mathcal{M}$ is a minimal Markov basis and (4) is a disjoint union.*

From Theorem 1 and 3, we can derive a necessary and sufficient condition on decomposable models to have the unique minimal Markov bases.

Lemma 2. *There exists the unique minimal Markov basis for a decomposable model if and only if the number of connected components in any induced subgraphs of $\mathcal{G}^{\mathcal{D}}$ is less than three.*

Proof. Suppose that $\mathcal{G}(\bar{\Delta}_{\boldsymbol{b}})$ has more than two connected components. Then since $|\mathcal{F}_{\boldsymbol{b}}| \geq 4$ from Theorem 1, $\mathcal{T}_{\boldsymbol{b}}$ is not uniquely defined. If there exists another tree $\mathcal{T}_{\boldsymbol{b}} = (\mathcal{F}_{\boldsymbol{b}}, \mathcal{M}'_{\boldsymbol{b}})$, $\mathcal{M}_{\boldsymbol{b}} \neq \mathcal{M}'_{\boldsymbol{b}}$. Hence the minimal Markov base is not unique either.

Conversely assume that the number of connected components of $\mathcal{G}(\bar{\Delta}_{\boldsymbol{b}})$ for all $\boldsymbol{b} \in \mathcal{B}$ is two. Then $\mathcal{T}_{\boldsymbol{b}}$ for all $\boldsymbol{b} \in \mathcal{B}$ is uniquely defined. Hence the minimal Markov basis is unique. □

For decomposable models $\mathcal{G}^{\mathcal{D}}$ is chordal. From the graph theoretical viewpoint the above lemma can be rewritten as follows.

Theorem 4. *There exists the unique minimal Markov basis for a decomposable model if and only if $\mathcal{G}^{\mathcal{D}}$ has only two boundary cliques D and D' and they satisfy $D'' \subset D \cup D'$ for all $D'' \in \mathcal{D}$.*

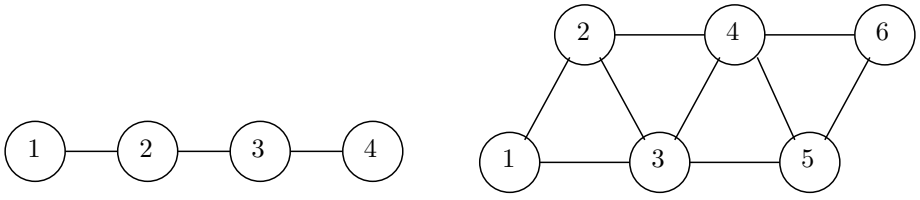
Proof. Suppose that $\mathcal{G}^{\mathcal{D}}$ has two boundary cliques D and D' such that $D'' \subset D \cup D'$ for all $D'' \in \mathcal{D}$. Then any vertex in D'' is adjacent to D or D'. Hence the number of connected components for any induced subgraphs of $\mathcal{G}^{\mathcal{D}}$ is two.

Conversely suppose that there exists $D'' \in \mathcal{D}$ such that $D'' \nsubseteq D \cup D'$. Then the subgraph induced by the union of $D'' \setminus (D \cup D')$, $\mathrm{Simp}(D)$ and $\mathrm{Simp}(D')$ has three connected components. □

The graphs with $r = 2$ always satisfy the conditions of the corollary. For $r \geq 3$ the graph with

$$\mathcal{D} = \{\{1, \ldots, r-1\}, \{2, \ldots, r\}, \ldots, \{r, \ldots, 2r-2\}\} \tag{5}$$

satisfies the conditions of the corollary. Figure 1 presents the graphs satisfying (5) for $r = 3, 4$. We can easily see that any induced subgraphs of the graphs in the figure has at most two connected components.



$r = 3$ $\qquad$ $r = 4$

Figure 1: Examples of the graphs satisfying the condition of Theorem 4

Let $\mathcal{T} = (\mathcal{D}, \mathcal{E})$ be a clique tree for $\mathcal{G}^{\mathcal{D}}$. Denote by $\mathcal{T} = (\mathcal{D}_e, \mathcal{E}_e)$ and $\mathcal{T}' = (\mathcal{D}'_e, \mathcal{E}'_e)$ the two induced subtrees of $\mathcal{T}$ obtained by removing the edge $e \in \mathcal{D}$. Define V_e and V'_e by

$$V_e = \bigcup_{D \in \mathcal{D}_e} D, \quad V'_e = \bigcup_{D \in \mathcal{D}'_e} D.$$

Let $\mathcal{M}^{\mathcal{T}}(V_e, V'_e)$ be the set of all primitive moves for the decomposable model determined by the chordal graph whose set of cliques is $\{V_e, V'_e\}$. Dobra[6] showed that

$$\mathcal{M}^{\mathcal{T}} = \bigcup_{e \in \mathcal{E}} \mathcal{M}^{\mathcal{T}}(V_e, V'_e) \tag{6}$$

is a Markov basis. We call $\mathcal{M}^{\mathcal{T}}$ a Dobra's Markov basis. From a viewpoint of the minimality of Markov bases, we have the following theorem.

Theorem 5. *Dobra's Markov basis $\mathcal{M}^{\mathcal{T}}$ is a minimal Markov basis if and only if the decomposable model has the unique minimal Markov basis.*

Proof. $\mathcal{M}^{\mathcal{T}}$ is a Markov basis. Hence if the decomposable model satisfy the condition of Lemma 2, $\mathcal{M}^{\mathcal{T}}$ is minimal.

Next we suppose that there exist three vertices in $\mathcal{G}$ which are not adjacent to each other. In the same way as the proof of Theorem 2, let $1, 2$ and 3 be such three vertices and assume that $l \in D_l$, $D_l \in \mathcal{D}$, for $l = 1, 2, 3$. Define $\{1, 2, 3\}^{\mathrm{c}} = \Delta \setminus \{1, 2, 3\}$. Consider a degree two fiber $\mathcal{F}_{\boldsymbol{b}}$ such that $\bar{\Delta}_{\boldsymbol{b}} = \{1, 2, 3\}$ and $n_{\{1,2,3\}^{\mathrm{c}}}(i_{\{1,2,3\}^{\mathrm{c}}}) = 2$. Then $|\mathcal{F}_{\boldsymbol{b}}| = 4$ from Theorem 1 and we can denote these four elements by

$$\begin{aligned}
\boldsymbol{n}_1 &= (000\ i_{\{1,2,3\}^{\mathrm{c}}})(111\ i_{\{1,2,3\}^{\mathrm{c}}}), \\
\boldsymbol{n}_2 &= (001\ i_{\{1,2,3\}^{\mathrm{c}}})(110\ i_{\{1,2,3\}^{\mathrm{c}}}), \\
\boldsymbol{n}_3 &= (010\ i_{\{1,2,3\}^{\mathrm{c}}})(101\ i_{\{1,2,3\}^{\mathrm{c}}}), \\
\boldsymbol{n}_4 &= (011\ i_{\{1,2,3\}^{\mathrm{c}}})(100\ i_{\{1,2,3\}^{\mathrm{c}}}),
\end{aligned} \tag{7}$$

where $\boldsymbol{n} = (i)(j)$ denotes a contingency table of sample size 2 having frequency 1 at the cell i and j. Let $\mathcal{T} = (\mathcal{D}, \mathcal{E})$ be a clique tree for $\mathcal{G}^{\mathcal{D}}$ and $\mathcal{T}' = (\mathcal{D}', \mathcal{E}')$ be the smallest subtree of $\mathcal{T}$ satisfying $D_l \in \mathcal{D}'$ for $l = 1, 2$ and 3. Then we can assume that $\mathcal{T}'$ satisfies either of the following two conditions,

(i) D_2 is an interior point and D_1 and D_3 are endpoints on the path ;

(ii) all of D_1, D_2 and D_3 are endpoints of $\mathcal{T}'$.

In both cases there exists $e \in \mathcal{E}$ such that $D_1, D_2 \subset V_e$ and $D_3 \subset V'_e$. Then $\mathcal{M}^{\mathcal{T}}(V_e, V'_e)$ includes the following two moves,

$$\boldsymbol{z}_1 = \boldsymbol{n}_1 - \boldsymbol{n}_2, \quad \boldsymbol{z}_2 = \boldsymbol{n}_3 - \boldsymbol{n}_4.$$

On the other hand there also exists $e' \in \mathcal{E}$ such that $D_1 \subset V_{e'}$ and $D_2, D_3 \subset V_{e'}$. In this case $\mathcal{M}^{\mathcal{T}}(V_{e'}, V'_{e'})$ includes the following two moves,

$$\boldsymbol{z}_3 = \boldsymbol{n}_1 - \boldsymbol{n}_4, \quad \boldsymbol{z}_4 = \boldsymbol{n}_2 - \boldsymbol{n}_3.$$

Thus $\mathcal{M}^{\mathcal{T}}$ includes at least four moves for the fiber $\mathcal{F}_{\boldsymbol{b}}$, which implies that $\mathcal{M}^D$ is not minimal for the model which does not have the unique minimal Markov basis. □

4.2 Minimal invariant Markov bases

We here consider Markov bases from a viewpoint of the invariance under the action of the symmetric group $G = S_{I_1} \times \cdots \times S_{I_m}$ on the labels of variables.

According to [1], we first give a brief review on the invariance of the set of moves. $\mathcal{B}$ is called *G-invariant* if

$$\forall g \in G, \ \forall \boldsymbol{z} \in \mathcal{B} \quad \Rightarrow \quad g\boldsymbol{z} \in \mathcal{B} \ \text{ or } \ -g\boldsymbol{z} \in \mathcal{B}.$$

$\mathcal{B}$ is called a *G-invariant Markov basis* for $\mathcal{D}$ if it is a Markov basis and also G-invariant. An invariant Markov basis is minimal invariant if no proper G-invariant subset of $\mathcal{B}$ is a Markov basis.

Let $\mathcal{F}_{\boldsymbol{b}}$ be a representative fiber, i.e. $\boldsymbol{n}_0 = (0 \cdots 0)(1 \cdots 1) \in \mathcal{F}_{\boldsymbol{b}}$. Define $B_0 = \{\boldsymbol{b}' \mid \boldsymbol{n}_0 \in \mathcal{F}_{\boldsymbol{b}}\}$. Denote by $\mathcal{B}_{\boldsymbol{b}}$ a set of moves in $\mathcal{F}_{\boldsymbol{b}}$. Then any $\boldsymbol{n} \in \mathcal{F}_{\boldsymbol{b}}$ is expressed as follows,

$$\boldsymbol{n} = (\overbrace{0 \cdots 0}^{|\Gamma_1|} \ i_{\Gamma_2} \cdots i_{\Gamma_c} \ \overbrace{0 \cdots 0}^{|\Delta \setminus \bar{\Delta}_{\boldsymbol{b}}|})(\overbrace{1 \cdots 1}^{|\Gamma_1|} \ i^*_{\Gamma_2} \cdots i^*_{\Gamma_c} \ \overbrace{0 \cdots 0}^{|\Delta \setminus \bar{\Delta}_{\boldsymbol{b}}|})$$

Let $G_{\boldsymbol{b}}$ be the group which acts on the set of cells such that

$$g \in G_{\boldsymbol{b}}, \quad g(\boldsymbol{n}) = (\overbrace{0 \cdots 0}^{|\Gamma_1|}, g^*_2(i_{\Gamma_2}) \cdots g^*_c(i_{\Gamma_c}) \ \overbrace{0 \cdots 0}^{|\Delta \setminus \bar{\Delta}_{\boldsymbol{b}}|})(\overbrace{1 \cdots 1}^{|\Gamma_1|}, g^*_2(i^*_{\Gamma_2}) \cdots g^*_c(i^*_{\Gamma_c}) \ \overbrace{0 \cdots 0}^{|\Delta \setminus \bar{\Delta}_{\boldsymbol{b}}|}),$$

$$g^*_l(i_{\Gamma_l}) = \{g_l(i_\delta) \mid g_l \in S_{I_l}, \delta \in \Gamma_l\}, \quad l = 2, \ldots, c.$$

Then we note that for any $\boldsymbol{n} \in \mathcal{F}_{\boldsymbol{b}}$, there exists $g \in G_{\boldsymbol{b}}$ such that $\boldsymbol{n} = G(\boldsymbol{n}_0)$. Suppose that $\mathcal{B}_{\boldsymbol{b}}$ is $G_{\boldsymbol{b}}$-invariant. Then as representative moves in $G_{\boldsymbol{b}}$ - orbits in $\mathcal{B}_{\boldsymbol{b}}$ we can consider $\boldsymbol{z}^{\boldsymbol{b}} = \boldsymbol{n}_0 - \boldsymbol{n} \in \mathcal{B}_{\boldsymbol{b}}$. Let $\kappa(\boldsymbol{b})$ be the minimum number of orbits required to connect $\mathcal{F}_{\boldsymbol{b}}$. Suppose that $\mathcal{B}_{\boldsymbol{b}}$ contain $\kappa(\boldsymbol{b})$ representative moves and connects $\mathcal{F}_{\boldsymbol{b}}$ for all $\boldsymbol{b} \in B_0$. Denote the set of representative moves in $\mathcal{B}_{\boldsymbol{b}}$ by $\mathcal{M}^0_{\boldsymbol{b}} = \{\boldsymbol{z}^{\boldsymbol{b}}_1, \ldots, \boldsymbol{z}^{\boldsymbol{b}}_{\kappa(\boldsymbol{b})}\}$. Then

$$\mathcal{M} = \bigcup_{\boldsymbol{b} \in B_0} \bigcup_{k=1}^{\kappa(\boldsymbol{b})} G(\boldsymbol{z}^{\boldsymbol{b}}_k) \tag{8}$$

is a minimal G-invariant Markov basis. Hence in order to clarify the structure of the minimal G-invariant Markov basis, it suffices to investigate $\kappa(\boldsymbol{b})$ and $\mathcal{M}^0_{\boldsymbol{b}}$ for each $\mathcal{F}_{\boldsymbol{b}}$. Since we consider the case where $\deg \boldsymbol{b} = 2$, we can restrict our consideration to $2 \times \cdots \times 2$ tables.

By following the argument in Lemma 1, the structure of $\mathcal{F}_{\boldsymbol{b}}$ is equivalent to the one of the fiber with $\bar{\Delta}_{\boldsymbol{b}} = \Delta = \{1, \ldots, c\}$. We first consider the structure of such fibers. Then $\mathcal{F}_{\boldsymbol{b}}$ is

$$\mathcal{F}_{\boldsymbol{b}} = \{(0\, i_2 \cdots i_c)(1\, i^*_2 \cdots i^*_c) \mid (i_2 \cdots i_c) = i_{\Delta \setminus \{1\}} \in \mathcal{I}_{\Delta \setminus \{1\}}\}. \tag{9}$$

Hence a representative move is expressed by

$$\boldsymbol{z}^{\boldsymbol{b}} = (0 \cdots 0)(1 \cdots 1) - (0\, i_{\Delta \setminus \{1\}})(1\, i^*_{\Delta \setminus \{1\}}), \quad i_k \in \mathcal{I}_k, \quad k = 1, \ldots, c.$$

Then we note that we can identify $G_{\boldsymbol{b}}$ with S_2^{c-1}. We first consider to derive $\kappa(\boldsymbol{b})$ and $\mathcal{M}^0_{\boldsymbol{b}}$ for this fiber. Let $\mathcal{V}^{c-1} = \{0,1\}^{c-1}$ denote the $(c-1)$-dimensional vector space over the finite field GF(2), where the addition of two vectors is defined to be the exclusive OR of the elements. Let $\circ$ denote the operator of composition defined on S_2^{c-1}. Then we obtain the following lemma.

Lemma 3. *S_2^{c-1} is isomorphic to $\mathcal{V}^{c-1}$;*

Proof. Consider the map $\phi : S_2^{c-1} \to \mathcal{V}^{c-1}$ such that $\phi(g) = \boldsymbol{v} = (v_2, \ldots, v_c)$, $g = \langle g_2, \ldots, g_c \rangle \in S_2^{c-1}$, $g_l \in S_2$, $\boldsymbol{v} \in \mathcal{V}^{c-1}$, where

$$v_l = \begin{cases} 0, & \text{if } g_l(i_l) = i_l, \\ 1, & \text{if } g_l(i_l) = i^*_l, \end{cases}$$

for $l = 2, \ldots, c$ and $\{i_l, i^*_l\} = \{0,1\}$. For $g' = \langle g'_2, \ldots, g'_c \rangle \in S_2^{c-1}$, $g'_l \in S_2$, and $\boldsymbol{v}' \in \mathcal{V}^{c-1}$, define $\phi(g') = \boldsymbol{v}' = (v'_2, \ldots, v'_c)$, Then we have $\phi(g \circ g') = \tilde{\boldsymbol{v}} = (\tilde{v}_2, \ldots, \tilde{v}_c)$, $\tilde{\boldsymbol{v}} \in \mathcal{V}^{c-1}$, where

$$\tilde{v}_l = \begin{cases} 0, & \text{if } g_l \circ g'_l(i_l) = i_l, \\ 1, & \text{if } g_l \circ g'_l(i_l) = i'_l, \end{cases}$$

for $l = 2, \ldots, c$. Hence we have

$$\tilde{v}_l = v_l \oplus v'_l, \quad k = 1, \ldots, c$$

and therefore ϕ is homomorphism. It is obvious that ϕ is a bijection. Therefore S_2^{c-1} is isomorphic to $\mathcal{V}^{c-1}$. □

Based on this lemma, we can obtain the following theorem.

Theorem 6. *Let $\mathcal{V}^0 = \{\boldsymbol{v}_k = (v_{k2}, \ldots, v_{kc}),\ k = 2, \ldots, c\}$ be any basis of $\mathcal{V}^{c-1}$. Define $\boldsymbol{n}_0, \boldsymbol{n}_{\boldsymbol{v}_k} \in \mathcal{F}_{\boldsymbol{b}}$ by*

$$\boldsymbol{n}_0 = (00\cdots0)(11\cdots1), \quad \boldsymbol{n}_{\boldsymbol{v}_k} = (0\ v_{k2}\cdots v_{kc})(1\ v_{k2}^*\cdots v_{kc}^*),$$

where $v_{kl}^* = 1 \oplus v_{kl}$ and $\oplus$ denotes the XOR operation. Then $\kappa(\boldsymbol{b}) = c-1$ and the represetntative moves in the orbits are expressed by $\boldsymbol{z}_l^{\boldsymbol{b}} = \boldsymbol{n}_0 - \boldsymbol{n}_{\boldsymbol{v}_l}$, $l = 2, \ldots, c$.

Proof. Suppose that $\mathcal{B}_{\boldsymbol{b}}$ is minimal S_2^{c-1}-invariant set of moves which connects $\mathcal{F}_{\boldsymbol{b}}$ and that $\mathcal{B}_{\boldsymbol{b}}$ includes $\kappa(\boldsymbol{b})$ orbits as $S_2^{c-1}(\boldsymbol{z}_1), \ldots, S_2^{c-1}(\boldsymbol{z}_{\kappa(\boldsymbol{b})})$, where

$$\boldsymbol{z}_k = (0\cdots0)(1\cdots1) - (0\ i_{k2}\cdots i_{kc})(1\ i_{k2}^*\cdots i_{kc}^*),$$

for some $i_{kl} \in \mathcal{I}_l$, $k = 1, \ldots, \kappa(\boldsymbol{b})$, $l = 2, \ldots, c$. Denote $\boldsymbol{n}_k = (0\ i_{k2}\cdots i_{kc})(0\ i_{k2}^*\cdots i_{kc}^*)$. Let $g^k \in S_2^c$ satisfy $g^k(\boldsymbol{n}_0) = \boldsymbol{n}_k$ for $k = 1, \ldots, \kappa(\boldsymbol{b})$. Define $G = \{g^1, \ldots, g^{\kappa(\boldsymbol{b})}\} \subseteq S_2^{c-1}$. As mentioned above, $\mathcal{F}_{\boldsymbol{b}}$ can be expressed as in (9). Hence for any $\boldsymbol{n}, \boldsymbol{n}' \in \mathcal{F}_{\boldsymbol{b}}$ there exists $g \in S_2^{c-1}$ satisfying $\boldsymbol{n}' = g(\boldsymbol{n})$. Then G satisfies

$$\forall g \in S_2^{c-1}, \quad \exists p \le \kappa(\boldsymbol{b}), \quad \exists g_*^1 \in G, \ldots, \exists g_*^p \in G \quad \text{s.t.} \quad g = g_*^p \circ \cdots \circ g_*^1. \tag{10}$$

and no proper subset of G satisfies (10). Denote $\mathcal{V}' = \phi(G) \subseteq \mathcal{V}$. Then the minimality of $\mathcal{B}_{\boldsymbol{b}}$ is equivalent to

$$\forall \boldsymbol{v} \in \mathcal{V}, \quad \exists \boldsymbol{v}^1 \in \mathcal{V}', \ldots, \exists \boldsymbol{v}^p \in \mathcal{V}' \quad \text{s.t.} \quad \boldsymbol{v} = \boldsymbol{v}^1 \oplus \cdots \oplus \boldsymbol{v}^p \tag{11}$$

and no proper subset of $\mathcal{V}'$ satisfies (11). This implies that $\mathcal{V}'$ is a basis of $\mathcal{V}$ and hence $\kappa(\boldsymbol{b}) = c-1$. Let g^k be $g^k = \langle g_{k2}, \ldots, g_{kc}\rangle = \phi^{-1}(\boldsymbol{v}_k)$, where $g_{kl}(0) = v_{kl}$. Then $g^k(\boldsymbol{n}_0) = \boldsymbol{n}_k$. Hence the representative moves of the orbits are $\boldsymbol{z}_k^0$, $k = 2, \ldots, c$.

□

For example we can set $V = \{\boldsymbol{v}_2, \ldots, \boldsymbol{v}_c\}$ as

$$\boldsymbol{v}_2 = (11\cdots11), \quad \boldsymbol{v}_3 = (01\cdots11), \quad \ldots \quad \boldsymbol{v}_{c-1} = (00\cdots011), \quad \boldsymbol{v}_c = (00\cdots01),$$

and the representative moves in a minimal G-invariant Markov basis is

$$\begin{aligned}
\boldsymbol{z}_2^0 &= (00\cdots0)(11\cdots1) - (011\cdots11)(100\cdots00)\\
\boldsymbol{z}_3^0 &= (00\cdots0)(11\cdots1) - (001\cdots11)(110\cdots00)\\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots\\
\boldsymbol{z}_c^0 &= (00\cdots0)(11\cdots1) - (000\cdots01)(111\cdots10).
\end{aligned} \tag{12}$$

So far we focus on $\mathcal{F}_{\boldsymbol{b}}$ such that $\bar{\Delta}_{\boldsymbol{b}} = \{1, \ldots, c\}$. Now we consider the fiber for a general $\boldsymbol{b}$. Let $g^k \in S_2^{c-1}$ be $g^k = \langle g_{k2}, \ldots, g_{kc}\rangle = \phi^{-1}(\boldsymbol{v}^k)$ for $k = 2, \ldots, c$ and define $G_{\boldsymbol{b}} = \{\tilde{g}^1, \ldots, \tilde{g}^{\kappa(\boldsymbol{b})}\}$ by

$$\tilde{g}^k(\boldsymbol{n}) = (\overbrace{0\cdots0}^{|\Gamma_1|}, g_{k2}^*(i_{\Gamma_2})\cdots g_{kc}^*(i_{\Gamma_c})\ \overbrace{0\cdots0}^{|\Delta\setminus\bar{\Delta}_{\boldsymbol{b}}|})(\overbrace{1\cdots1}^{|\Gamma_1|}, g_{k2}^*(i_{\Gamma_2}^*)\cdots g_{kc}^*(i_{\Gamma_c}^*)\ \overbrace{0\cdots0}^{|\Delta\setminus\bar{\Delta}_{\boldsymbol{b}}|}),$$

$$g_{kl}^*(i_{\Gamma_l}) = \{g_{kl}(i_\delta) \mid g_{kl} \in S_2, \delta \in \Gamma_l\}, \quad l = 2, \ldots, c,$$

Denote $\boldsymbol{n}_{\boldsymbol{v}_k} = \tilde{g}^k(\boldsymbol{n}_0)$ and $\boldsymbol{z}_l^{\boldsymbol{b}} = \boldsymbol{n}_0 - \boldsymbol{n}_{\boldsymbol{v}_k}$. Based on (8) and the results of Theorem 6, we can easily obtain the following result.

Theorem 7. *For any $\boldsymbol{b} \in B_0$, $\kappa(\boldsymbol{b}) = c - 1$ and the represetntative moves in the orbits are expressed by $\boldsymbol{z}_k^{\boldsymbol{b}} = \boldsymbol{n}_0 - \boldsymbol{n}_{\boldsymbol{v}_l}$, $k = 2, \ldots, c$. Then*

$$\mathcal{M} = \bigcup_{\boldsymbol{b} \in B_0} \bigcup_{k=2}^{c} G(\boldsymbol{z}_k^{\boldsymbol{b}})$$

is a minimal G-invariant Markov basis.

Next we consider Dobra's Markov basis $\mathcal{M}^{\mathcal{T}}$ from a viewpoint of invariance. Since $\mathcal{M}^{\mathcal{T}}$ does not depend on labels, $\mathcal{M}^{\mathcal{T}}$ is S_2^m-invariant. Based on the result of Theorem 6, we can show that $\mathcal{M}^{\mathcal{T}}$ is not always a minimal invariant Markov bases.

Theorem 8. *$\mathcal{M}^{\mathcal{T}}$ is minimal invariant if and only if $\mathcal{T}$ has only two endpoints.*

Proof. Suppose that $\mathcal{T} = (\mathcal{D}, \mathcal{E})$ has more than two endpoints. Let D_1, D_2 and D_3 be three of them. Then they are boundary cliques. Suppose $1, 2, 3 \in \Delta$ are simply separated vertices in D_1, D_2 and D_3, respectively. In the same way as the argument in the proof of Theorem 5, there exist $e, e', e'' \in \mathcal{E}$ such that

$$\begin{array}{ll} D_1, D_2 \in V_e, & D_3 \in V_e', \\ D_2, D_3 \in V_{e'}, & D_1 \in V_{e'}', \\ D_3, D_1 \in V_{e''}, & D_2 \in V_{e''}'. \end{array}$$

Consider the moves for the fiber $\mathcal{F}_{\boldsymbol{b}}$ for $\boldsymbol{b}$ such that $\bar{\Delta}_{\boldsymbol{b}} = \{1, 2, 3\}$. Define $\boldsymbol{z}_5$ and $\boldsymbol{z}_6$ by

$$\boldsymbol{z}_5 = \boldsymbol{n}_1 - \boldsymbol{n}_3, \quad \boldsymbol{z}_6 = \boldsymbol{n}_2 - \boldsymbol{n}_4,$$

where $\boldsymbol{n}_1, \ldots, \boldsymbol{n}_4$ are defined in (7). Then we have

$$\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{M}^T(V_e, V_e'), \quad \boldsymbol{z}_3, \boldsymbol{z}_4 \in \mathcal{M}^T(V_{e'}, V_{e'}'), \quad \boldsymbol{z}_5, \boldsymbol{z}_6 \in \mathcal{M}^T(V_{e''}, V_{e''}').$$

We note that $\{\boldsymbol{z}_1, \boldsymbol{z}_2\}$, $\{\boldsymbol{z}_3, \boldsymbol{z}_4\}$ and $\{\boldsymbol{z}_5, \boldsymbol{z}_6\}$ are S_2^{c-1}-orbits in the moves for $\mathcal{F}_{\boldsymbol{b}}$. Since $c = 3$, $\mathcal{M}^{\mathcal{T}}$ is not minimal invariant.

Suppose that $\mathcal{T}$ has only two endpoints. Then $\mathcal{T}$ is expressed as in Figure 2. Let $\Gamma_1(\boldsymbol{b}), \ldots, \Gamma_c(\boldsymbol{b})$ be c connected components of $\mathcal{G}(\bar{\Delta}_{\boldsymbol{b}})$. Suppose that $v_i \in \Gamma_i(\boldsymbol{b})$. Then the structure of $\mathcal{F}_{\boldsymbol{b}}$ is equivalent to the one of $\mathcal{F}_{\boldsymbol{b}'}$ such that $\bar{\Delta}_{\boldsymbol{b}'} = \{v_1, \ldots, v_{c-1}\}$. Hence we consider the moves in $\mathcal{F}_{\boldsymbol{b}'}$. Let $\mathcal{B}_{\boldsymbol{b}'}$ denote the set of all moves in $\mathcal{F}_{\boldsymbol{b}'}$. Without loss of generality we can assume that $v_i \in D_{\pi(i)}$, where $\pi(1) < \cdots < \pi(c)$. Define $e_i = (D_{i-1}, D_i) \in \mathcal{E}$ for $i = 2, \ldots, c$, $S_i = D_{i-1} \cap D_i$, $V_i = V_{e_i} \setminus S_i$ and $V_i' = V_{e_i}' \setminus S_i$. Then the moves in $\mathcal{M}^{\mathcal{T}}(V_i, V_i')$ are expressed by

$$\boldsymbol{z} = (i_{V_i}, i_{V_i'}, i_{S_i})(j_{V_i}, j_{V_i'}, i_{S_i}) - (i_{V_i}, j_{V_i'}, i_{S_i})(j_{V_i}, i_{V_i'}, i_{S_i}), \tag{13}$$

$$i_{V_i}, j_{V_i} \in \mathcal{I}_{V_i}, \quad i_{V_i'}, j_{V_i'} \in \mathcal{I}_{V_i'}, \quad i_{S_i} \in \mathcal{I}_{S_i}.$$

If $V_{e_i} \cap \Delta_{\boldsymbol{b}} = \emptyset$ or $V'_{e_i} \cap \Delta_{\boldsymbol{b}} = \emptyset$, then we have $\mathcal{M}^{\mathcal{T}}(V_{e_i}, V'_{e_i}) \cap \mathcal{B}_{\boldsymbol{b}} = \emptyset$. If $V_{e_i} \cap \Delta_{\boldsymbol{b}} \neq \emptyset$ and $V'_{e_i} \cap \Delta_{\boldsymbol{b}} \neq \emptyset$, there exists $2 \leq k(e_i) \leq c$ satisfying $v_k \in V_i$ for all $k < k(e_i)$ and $v_k \in V_i'$ for all $k \geq k(e_i)$. Then

$$\mathcal{M}^{\mathcal{T}}(V_{e_i}, V'_{e_i}) \cap \mathcal{B}_{\boldsymbol{b}} = S_2^{c-1}(\boldsymbol{z}_{k(e_i)}),$$

where $\boldsymbol{z}_{k(e_i)}$ is defined as in (12). Hence we have

$$\begin{aligned} \mathcal{M}^{\mathcal{T}} \cap \mathcal{B}_{\boldsymbol{b}} &= \bigcup_{e_i \in \mathcal{E}} \mathcal{M}^{\mathcal{T}}(V_{e_i}, V'_{e_i}) \cap \mathcal{B}_{\boldsymbol{b}} \\ &= \bigcup_{i=1}^{c-1} S_2^{c-1}(\boldsymbol{z}_k). \end{aligned} \tag{14}$$

Hence $\mathcal{M}^{\mathcal{T}} \cap \mathcal{B}_{\boldsymbol{b}}$ is minimal invariant. □

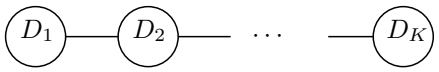


Figure 2: The clique tree with two endpoints

As an example consider the 4-way independence model $\mathcal{D} = \{D_i = \{i\}, i = 1, \ldots, 4\}$. Then both of $\mathcal{T}_1$ and $\mathcal{T}_2$ in Figure 3 are clique trees for $\mathcal{D}$. From Theorem 8, $\mathcal{M}^{T_1}$ is a minimal S_2^{c-1}-invariant Markov bases and $\mathcal{M}^{T_2}$ is not a minimal S_2^{c-1}-invariant Markov bases. Hence in general the minimality of $\mathcal{M}^T$ depends on clique trees $\mathcal{T}$.

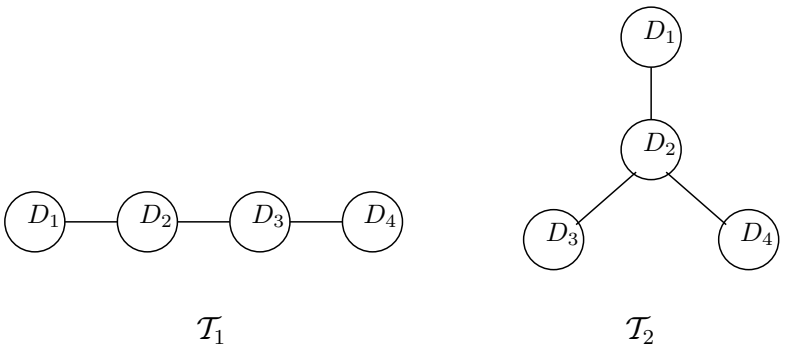


Figure 3: The clique trees for the 4-way independence model

5 Gröbner bases for decomposable models

So far we have been discussing Markov bases. In this section we briefly discuss Gröbner basis. For decomposable models, Theorem 4.17 of Hoşten and Sullivant[11] gives a recursive method for determining the term order and the corresponding Gröbner basis consisting of primitive moves only. It gives a Gröbner basis version of Dobra's Markov basis in (6). In Theorem 5 we saw that Dobra's construction gives a minimal Markov basis only in a special case. The same phenomenon can be observed with respect to the *reducedness* of Gröbner basis if we simply apply Theorem 4.17 of Hoşten and Sullivant recursively, i.e., the operation of Theorem 4.17 of Hoşten and Sullivant does not preserve reducedness in general. Here we are interested in explicit description of appropriate term order and the reduced Gröbner basis for decomposable models. We prove that for decomposable models, there exists a term order such that the reduced Gröbner basis is explicitly described and furthermore it is minimal as a Markov basis.

In obtaining a nice Gröbner basis, the term order has to be carefully chosen. For example consider the simple case of 3×3 two-way contingency tables with fixed row sums and columns sums. Proposition 5.4 of Sturmfels [15] shows that the set of 9 primitive moves of the form

$$\pm \begin{array}{|cc|} \hline +1 & -1 \\ -1 & +1 \\ \hline \end{array}$$

form a reduced Gröbner basis when the cells are lexicographically ordered and the term order is chosen to be the reverse lexicographic term order. However if we order the 9 cells as

1	8	6
4	2	9
7	5	3

and use the lexicographic order, then the reduced Gröbner contains the following degree 3 move

0	-1	1
1	0	-1
-1	1	0

in addition to 9 primitive moves. This example shows that the existence of a reduced Gröbner basis consisting of primitive moves depends on the choice of term order.

We need several steps in constructing a nice term order for a decomposable model of an m-way contingency table. First, we order m variables. Choose a boundary clique of the chordal graph corresponding to the decomposable model and order the variables in the boundary cliques as lowest variables. Then remove the boundary clique from the chordal graph, choose a boundary clique from the smaller graph and order the variables from the boundary clique as the next lowest variables. By recursively removing boundary cliques we obtain an ordering of variables. The resulting order is a perfect elimination scheme but has a stronger property. Second, we order the cells of an m-way contingency table lexicographically. Finally, as the term order $\succ$ we use the reverse lexicographic term order.

As in Section 4.1 let $\mathcal{B}$ denote the set of non-degenerate $\boldsymbol{b}$. In each fiber $\mathcal{F}_{\boldsymbol{b}}$ there exists the lowest element $\boldsymbol{n}^*_{\boldsymbol{b}}$ with respect to the above term order $\succ$. Define

$$\mathcal{M}^{GB} = \bigcup_{\boldsymbol{b}\in\mathcal{B}} \bigcup_{\substack{\boldsymbol{n}\in\mathcal{F}_{\boldsymbol{b}} \\ \boldsymbol{n}\neq\boldsymbol{n}^*_{\boldsymbol{b}}}} \{\boldsymbol{n} - \boldsymbol{n}^*_{\boldsymbol{b}}\}$$

Then we have the following theorem.

Theorem 9. *$\mathcal{M}^{GB}$ is a reduced Gröbner basis and it is minimal as a Markov basis.*

We omit the details of the proof. By generalizing the proof of Proposition 5.4 of Sturmfels [15] we can show that $\mathcal{M}^{GB}$ is indeed a Gröbner basis. Reducedness is obvious. Minimality is also obvious from Theorem 3.

6 Concluding remarks

In this paper we investigated the structure of degree two fibers of a general hierarchical model and clarified the structure of minimal Markov bases and minimal invariant Markov bases for decomposable models. We have also shown that decomposable models posses Gröbner basis which is at the same time a minimal Markov basis.

For future research it is important to investigate structures of degree three fibers, degree four fibers etc. In Takemura and Aoki [16] we gave a characterization of minimal Markov bases. It shows that minimal Markov bases can be constructed "from below", i.e., combining moves from fibers of degree 1,2,3,.... Although at the moment the construction can not be implemented as an algorithm, it shows the importance of studying fibers of low degrees. We see that the study of degree two fibers in this paper led to some interesting results.

References

[1] Satoshi Aoki and Akimichi Takemura. Invariant minimal Markov basis for sampling contingency tables with fixed marginals. Technical Report METR 2003-25, Univertisy of Tokyo, 2003. To appear in *Ann. Inst. Statist. Math.*

[2] Satoshi Aoki and Akimichi Takemura. Minimal basis for a connected Markov chain over $3 \times 3 \times K$ contingency tables with fixed two-dimensional marginals. *Aust. N. Z. J. Stat.*, 45(2):229–249, 2003.

[3] Satoshi Aoki and Akimichi Takemura. The largest group of invariance for Markov bases and toric ideals. Technical Report METR 2005-14, Univertisy of Tokyo, 2005.

[4] Satoshi Aoki, Akimichi Takemura, and Ruriko Yoshida. Indispensable monomials of toric ideals and Markov bases. METR 2005-34, 2005, arXiv:math.ST/0511290.

[5] Persi Diaconis and Bernd Sturmfels. Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.*, 26(1):363–397, 1998.

[6] Adrian Dobra. Markov bases for decomposable graphical models. *Bernoulli*, 9(6):1093–1108, 2003.

[7] Dan Geiger, Chris Meek, and Bernd Sturmfels. On the toric algebra of graphical models. *Ann. Statist.*, 34(3):1463–1492, 2006.

[8] Hisayuki Hara and Akimichi Takemura. Boundary cliques, clique trees and perfect sequences of maximal cliques of a chordal graph, 2006, arXiv:cs.DM/0607055.

[9] Hisayuki Hara and Akimichi Takemura. Simultaneous estimation of the mans in some Poisson log linear models. *Journal of the Japan Statistical Society*, 36(1):17–36, 2006.

[10] Hisayuki Hara and Akimichi Takemura. Improving on the maximum likelihood estimators of the means in Poisson decomposable graphical models. *Journal of Multivariate Analysis*, 98:410–434, 2007.

[11] Serkan Hoşten and Seth Sullivant. Gröbner bases and polyhedral geometry of reducible and cyclic models. *J. Combin. Theory Ser. A*, 100(2):277–301, 2002.

[12] R. W. Irving and M. R. Jerrum. Three-dimensional statistical data security problems. *SIAM J. Comput.*, 23(1):170–184, 1994.

[13] Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, Oxford, 1996.

[14] Y. Shibata. On the tree representation of chordal graphs. *J. Graph Theory*, 12(12):421–428, 1988.

[15] Bernd Sturmfels. *Gröbner Bases and Convex Polytopes*, volume 8 of *University Lecture Series*. American Mathematical Society, Providence, RI, 1996.

[16] Akimichi Takemura and Satoshi Aoki. Some characterizations of minimal Markov basis for sampling from discrete conditional distributions. *Ann. Inst. Statist. Math.*, 56(1):1–17, 2004.

[17] Akimichi Takemura and Satoshi Aoki. Distance reducing Markov bases for sampling from a discrete sample space. *Bernoulli*, 11(5):793–813, 2005.

[18] Akimichi Takemura and Hisayuki Hara. Conditions for swappability of records in a microdata set when some marginals are fixed, 2006, arXiv:math.ST/0603603.