MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

Hierarchical Subspace Models for Contingency Tables

Hisayuki HARA, Tomonari SEI and Akimichi TAKEMURA

METR 2009–44 September 2009

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.keisu.t.u-tokyo.ac.jp/research/techrep/index.html

Hierarchical subspace models for contingency tables

Hisayuki Hara*, Tomonari Sei† and Akimichi Takemura‡§

September 2009

Abstract

For statistical analysis of multiway contingency tables we propose modeling interaction terms in each maximal compact component of a hierarchical model. By this approach we can search for parsimonious models with smaller degrees of freedom than the usual hierarchical model, while preserving conditional independence structures in the hierarchical model. We discuss estimation and exacts tests of the proposed model and illustrate the advantage of the proposed modeling with some data sets.

Keywords : context specific interaction model, divider, Markov bases, split model, uniform association model.

1 Introduction

Modeling of the interaction term is an important topic for two-way contingency tables, because there is a large gap between the independence model and the saturated model. This problem is clearly of importance for contingency tables with three or more factors. However modeling strategies of higher order interaction terms have not been fully discussed in literature. In this paper we propose modeling interaction terms of multiway contingency tables by considering each maximal compact component of a hierarchical model.

For two-way contingency tables the uniform association model (Goodman [1979, 1985]) and the RC association model (Goodman [1979, 1985], Kuriki [2005]) are often used for modeling interaction terms. In the analysis of agreement among raters, where data are summarized as square contingency tables with the same categories, many models with interaction in diagonal elements and their extension to multiway tables have been considered (e.g. Tanner and Young [1985], Tomizawa [2009]). Hirotsu [1997] proposed

*Department of Technology Management for Innovation, University of Tokyo
†Graduate School of Information Science and Technology, University of Tokyo
‡Graduate School of Information Science and Technology, University of Tokyo
§CREST, JST

a two-way change point model and Hara et al. [2009] generalized it to a subtable sum model. For multiway contingency tables Højsgaard [2003] considered the split model as a generalization of graphical models. The context specific interaction model defined by Højsgaard [2004] is a more general model than the split model.

We give a unified treatment of these models as submodels of hierarchical models. Malvestuto and Moscarini [2000] showed that a hierarchical model possesses a compaction, i.e. the variables are grouped into maximal compact components separated by dividers. Variables of different maximal compact components separated by a divider are conditionally independent given the variables of the divider. Furthermore the likelihood function factors as a rational function of marginal likelihoods, where the numerator corresponds to maximal compact components and the denominator corresponds to dividers. By this factorization, statistical inference on a hierarchical model can be localized to each maximal compact component. In the case of decomposable model, maximal compact components and dividers reduce to maximal cliques and minimal vertex separators of a chordal graph, respectively, and the above factorization is well known (e.g. Section 4.4 of Lauritzen [1996]).

In a usual hierarchical model each maximal interaction effect is saturated, i.e. there is no restriction on the parameters for maximal interaction effects. However, as in the two-way tables, we can consider submodels for interaction effects. In the modeling process, it is advantageous to treat each maximal compact component of a hierarchical model separately and to keep the compaction of the hierarchical model. By respecting the compaction of the hierarchical model, conditional independence property and the localization property of the hierarchical model are preserved. We call a resulting model a hierarchical subspace model. We prove some properties of our proposed model and illustrate its advantage with some data sets.

The organization of the paper is as follows. For the rest of this section, as a motivating example, we consider a submodel of the conditional independence model for three-way contingency tables. In Section 2 we define the hierarchical subspace model and discuss the localization of inference through the decomposition of the model into maximal extended compact components. We also discuss maximum likelihood estimation of the proposed model. In Section 3 we study the split model in the framework of this paper. In Section 4 we present construction of Markov bases for conditional tests of the model. Fitting of the proposed model to several real data sets is presented in Section 5. Some concluding remarks are given in Section 6.

1.1 A motivating example: subspace conditional independence model for three-way tables

As an illustration of hierarchical subspace models we discuss a submodel of conditional independence model for three-way tables. Denote the sample size by $n = x_{+++}$. Consider an $I \times J \times K$ contingency table and let p_{ijk} denote the probability of the cell. Marginal probabilities are denoted by p_{i++}, p_{ij+}, etc. Similar notation is used for the frequencies $\boldsymbol{x} = \{x_{ijk}\}$ of the contingency table. Consider the conditional independence model $i \perp\!\!\!\perp$

$k \mid j$, which is expressed by

$$\log p_{ijk} = a_{ij} + b_{jk}. \tag{1}$$

In the usual conditional independence model, a_{ij}'s and b_{jk}'s are free parameters. Now suppose that we have known functions ϕ_{ij} depending on i and j and ψ_{jk} depending on j and k. Separating main effects, consider the following submodel of the conditional independence model

$$\log p_{ijk} = \alpha_i + \beta_j + \gamma_k + \delta\phi_{ij} + \delta'\psi_{jk}. \tag{2}$$

The parameters of this model are $\{\alpha_i\}_{i=1}^I, \{\beta_j\}_{j=1}^J, \{\gamma_k\}_{k=1}^K$ and δ, δ'. The uniform association model is specified by $\phi_{ij} = ij$. The two-way change point model in Hirotsu [1997] is specified by

$$\phi_{ij} = \begin{cases} 1, & \text{if } i \le I_1 \text{ and } j \le J_1, \\ 0, & \text{otherwise,} \end{cases}$$

where $1 \le I_1 < I$, $1 \le J_1 < J$. Similarly we can specify ψ_{jk} according to many well known models.

As a submodel of the conditional independence model, $i \perp\!\!\!\perp k \mid j$ holds for (2) and we can write

$$p_{ijk} = \frac{p_{ij+}p_{+jk}}{p_{+j+}}.$$

Moreover, since $\{\beta_j\}_{j=1}^J$ are free parameters, the model is saturated for the main effect of the second factor. This implies that the maximum likelihood estimate (MLE) of the model is obtained as

$$\hat{p}_{ijk} = \frac{\hat{p}_{ij+}\hat{p}_{+jk}}{x_{+j+}/n}. \tag{3}$$

Here $\hat{p}_{ij+}$ is the MLE of the following model for the marginal probability

$$\log p_{ij+} = \alpha_i + \beta_j + \delta\phi_{ij} \tag{4}$$

and $\hat{p}_{ij+}$ only depends on the marginal frequencies $\{x_{ij+}\}$. Similarly $\hat{p}_{+jk}$ is the MLE for (j,k)-marginal probabilities:

$$\log p_{+jk} = \beta_j + \gamma_k + \delta'\psi_{jk}. \tag{5}$$

In this way the maximum likelihood estimation of the model (2) is localized to estimations of two marginal models.

In our model (2) it is important to note that δ and δ' are free parameters. Consider an additional constraint $H: \delta = \delta'$ to (2):

$$\log p_{ijk} = \alpha_i + \beta_j + \gamma_k + \delta(\phi_{ij} + \psi_{jk}). \tag{6}$$

This model is still log-affine and contained in the conditional independence model. However, because both the (i,j)-marginals and the (j,k)-marginals are relevant for the estimation of the common value of δ, we can not localize estimation of the parameters to two marginal tables. This consideration shows that it is convenient to set up a log-affine model, which is "conformal" to the conditional independence model. We will appropriately define the notion of conformality in Section 2.

In our model we can also allow certain patterns of structural zeros. Consider the data on song sequence of a wood pewee in Section 7.5.2 of Bishop et al. [1975]. The data is shown in Table 1. The wood pewee has a repertoire of four distinctive phrases. The observed data consists of 198 triplets of consecutive phrases $(i, j, k) \in \{1, 2, 3, 4\}^3$. It is a $4 \times 4 \times 4$ contingency table with the cells of the form (i, i, k) and (i, j, j) being structural zeros.

Table 1: Triples of phrases in a song sequence of a wood pewee, with repeats deleted.

		Third place			
First place	Second place	A	B	C	D
A	A	—	—	—	—
	B	19	—	2	2
	C	2	26	—	0
	D	12	5	0	—
B	A	—	9	6	12
	B	—	—	—	—
	C	24	1	—	1
	D	1	2	0	—
C	A	—	4	22	0
	B	3	—	22	0
	C	—	—	—	—
	D	1	0	0	—
D	A	—	11	0	4
	B	5	—	1	1
	C	0	0	—	0
	D	—	—	—	—

Source: Craig [1943]

The model considered in Bishop et al. [1975] is written as

$$p_{ijk} = 1_{\{i \neq j\}} e^{a_{ij}} 1_{\{j \neq k\}} e^{b_{jk}}, \tag{7}$$

where a_{ij} and b_{jk} are free parameters. With some abuse of notation (7) can be written as

$$\log p_{ijk} \;=\; a_{ij} 1_{\{i \neq j\}} + (-\infty) 1_{\{i=j\}} + b_{jk} 1_{\{j \neq k\}} + (-\infty) 1_{\{j=k\}}.$$

As (2), this model is a conditional independence model and also it is saturated for the main effect of the second factor. Therefore the MLE for this model is again expressed as (3). An appropriate handling of $(-\infty)$ and further analysis are given in Section 5.

In Section 3, we also consider the split model defined by Højsgaard [2003] as an important example of the hierarchical subspace model. Here we give an example of the split model for the three-way table $\{x_{ijk}\}$ by

$$\log p_{ijk} \;=\; a_i + b_j + c_k + (d_{ij} + e_{jk})1_{\{j \in \mathcal{J}_1\}},$$

where $\mathcal{J}_1$ is a subset of $\{1, \dots, J\}$. This model means that the interaction between i and j (resp. j and k) exists only if $j \in \mathcal{J}_1$. The general definition of split models is given in Section 3 and a numerical example of it is given in Section 5.

We now consider the conditional goodness of fit test of the model based on the Markov basis methodology (Diaconis and Sturmfels [1998]). Assume that ϕ_{ij}'s and ψ_{jk}'s in (2) are integers. Furthermore suppose that Markov bases $\mathcal{B}_{12}$ and $\mathcal{B}_{23}$ are already obtained for two marginal models (4) and (5). Following the notation in Section 2 of Dobra and Sullivant [2004], write a particular move $\boldsymbol{z} = \{z(i,j)\}$ of degree d in the Markov basis $\mathcal{B}_{12}$ for (4) as

$$\boldsymbol{z} = [\{(i_1, j_1), \dots, (i_d, j_d)\} \| \{(i'_1, j'_1), \dots, (i'_d, j'_d)\}], \tag{8}$$

where $(i_1, j_1), \dots, (i_d, j_d)$ are cells (with replication) of positive elements of $\boldsymbol{z}$ and (i'_1, j'_1), $\dots, (i'_d, j'_d)$ are cells of negative elements of $\boldsymbol{z}$. By replication we mean that the same cell (i,j) is repeated $|z(i,j)|$ times in (8). Extend the move $\boldsymbol{z}$ to three-way table as

$$\boldsymbol{z}^{k_1,\dots,k_d} = [\{(i_1, j_1, k_1), \dots, (i_d, j_d, k_d)\} \| \{(i'_1, j'_1, k_1), \dots, (i'_d, j'_d, k_d)\}],$$

where $k_1, \dots, k_d \in \{1, \dots, K\}$ are arbitrary. Similarly for a move

$$\tilde{\boldsymbol{z}} = [\{(j_1, k_1), \dots, (j_d, k_d)\} \| \{(j'_1, k'_1), \dots, (j'_d, k'_d)\}] \in \mathcal{B}_{23}$$

let

$$\tilde{\boldsymbol{z}}^{i_1,\dots,i_d} = [\{(i_1, j_1, k_1), \dots, (i_d, j_d, k_d)\} \| \{(i_1, j'_1, k'_1), \dots, (i_d, j'_d, k'_d)\}].$$

Let $\mathcal{B}_{\{1,2\},\{2,3\}}$ denote a Markov basis for conditional independence model (1). Following the argument in Dobra and Sullivant [2004] we easily see that

$$\begin{aligned}\mathcal{B} = \mathcal{B}_{\{1,2\},\{2,3\}} &\cup \{\boldsymbol{z}^{k_1,\dots,k_{\deg \boldsymbol{z}}}, \boldsymbol{z} \in \mathcal{B}_{12}, 1 \le k_1, \dots, k_{\deg \boldsymbol{z}} \le K\} \\ &\cup \{\tilde{\boldsymbol{z}}^{i_1,\dots,i_{\deg \tilde{\boldsymbol{z}}}}, \tilde{\boldsymbol{z}} \in \mathcal{B}_{23}, 1 \le i_1, \dots, i_{\deg \tilde{\boldsymbol{z}}} \le I\}\end{aligned}$$

is a Markov basis for (2). Therefore the problem of conditional test of the model (2) is also localized to two marginal models.

For the rest of this paper, we generalize the above results to a log-affine model.

2 Hierarchical subspace models and their decompositions

In this section we give a definition of hierarchical subspace models for $I_1 \times \cdots \times I_m$ contingency tables and discuss its decomposition by partial edge separators. In Section 2.1 we define a hierarchical subspace model. In Section 2.2 we prove that for any log-affine model there exists a natural smallest decomposable model, such that the the log-affine model is a hierarchical subspace model of the decomposable model. In Section 2.3 we discuss properties of hierarchical models containing a given log-affine model.

2.1 Definition of a hierarchical subspace model

We give a definition of a hierarchical subspace model and also discuss the localization of maximum likelihood estimation of the proposed model. We follow notation of Lauritzen [1996] and Malvestuto and Moscarini [2000].

Let $V = \mathbb{R}^{I_1\times\cdots\times I_m}$ denote the set of $I_1 \times \cdots \times I_m$ tables with real entries, where $I_j \geq 2$ for all j. V is considered as an $I_1 \times \cdots \times I_m$-dimensional real vector space of functions (tables) from $\mathcal{I} = [I_1] \times \cdots \times [I_m]$ to $\mathbb{R}$, where $[J]$ denotes $\{1, \ldots, J\}$. A probability distribution over $\mathcal{I}$ is denoted by $\{p(\boldsymbol{i}), \boldsymbol{i} \in \mathcal{I}\}$. Let L be a linear subspace of V containing the constant function $1 \in L$. A *log-affine model* specified by L is given by $\log p(\cdot) \in L$, where $\log p(\cdot)$ denotes $\{\log p(\boldsymbol{i}), \boldsymbol{i} \in \mathcal{I}\}$. In the following we only consider linear subspaces of V containing the constant function 1.

Let D be a subset of $[m] = \{1, \ldots, m\}$. $\boldsymbol{i}_D = \{i_j, j \in D\}$ is a D-marginal cell. $p(\boldsymbol{i}_D)$ denotes the marginal probability of a probability distribution $p(\cdot)$. Similarly $x(\boldsymbol{i}_D)$ denotes the marginal frequency of the contingency table $\boldsymbol{x} = \{x(\boldsymbol{i}), \boldsymbol{i} \in \mathcal{I}\}$. As in Sei et al. [2008] let

$$L_D = \{\psi \in V \mid \psi(i_1, \ldots, i_m) = \psi(i'_1, \ldots, i'_m) \text{ if } i_h = i'_h, \forall h \in D\}$$

denote the set of functions depending only on $\boldsymbol{i}_D$. L_D is considered as $\mathbb{R}^{I_D}$, where $I_D = \prod_{h\in D} I_h$. Hence we have $L_{[m]} = V$. For a subspace L and $D \subset [m]$, we say that D is *saturated* in L if $L_D \subset L$. D is saturated in L if and only if the sufficient statistic for L fixes all the D-marginals of the contingency table. Note that if D is saturated in L, then every $E \subset D$ is saturated in L.

Let Δ denote a simplicial complex on $[m]$ and let $\operatorname{red}\Delta$ denote the set of maximal elements, i.e. facets, of Δ. Then the hierarchical model associated with Δ is defined as

$$\log p(\cdot) \in L_\Delta \stackrel{\text{def}}{=} \sum_{D\in\operatorname{red}\Delta} L_D,$$

where the right-hand side is the summation of vector spaces. We note that $\operatorname{red}\Delta$ is considered as a hypergraph. Here we summarize some notions on hypergraphs according to Malvestuto and Moscarini [2000]. A subset of a hyperedge of $\operatorname{red}\Delta$ is called a *partial edge*. A partial edge S is a separator of $\operatorname{red}\Delta$ if the subhypergraph of $\operatorname{red}\Delta$ induced by $[m] \setminus S$ is disconnected. A partial edge separator S of $\operatorname{red}\Delta$ is called a *divider* if there exist two vertices $u, v \in [m]$ that are separated by S but by no proper subset of S. If two vertices $u, v \in [m]$ are not separated by any partial edge, u and v are called tightly connected. A subset $C \subset [m]$ is called a *compact component* if any two variables in C are tightly connected. Denote the set of maximal compact components of $\operatorname{red}\Delta$ by $\mathcal{C}$. Then there exists a sequence of maximal compact components $C_1, \ldots, C_{|\mathcal{C}|}$ such that

$$(C_1 \cup \cdots \cup C_{k-1}) \cap C_k = S_k$$

and S_k, $k = 2, \ldots, |\mathcal{C}|$ are dividers of $\operatorname{red}\Delta$. We denote $\mathcal{S} = \{S_2, \ldots, S_{|\mathcal{C}|}\}$. $\mathcal{S}$ is a multiset in general.

Let $W_1, \ldots, W_K$ be linear subspaces of V and let $W = W_1 + \cdots + W_K$. We say that a subspace L is *conformal* to $\{W_j\}_{j=1}^K$ if

$$L = (L \cap W_1) + \cdots + (L \cap W_K).$$

Any L conformal to $\{W_j\}_{j=1}^K$ is clearly a subspace of W. Note that the relation $L = L \cap W \supset (L \cap W_1) + \cdots + (L \cap W_K)$ always holds but the inclusion is strict in general. Consider the models (2) and (6) in Section 1.1. Let $K = 2$ and let $W_1 := L_{\{1,2\}}$ and $W_2 := L_{\{2,3\}}$. In the case of the model (2),

$$L \cap W_1 = \{\alpha_i + \beta_j + \delta\phi_{ij}\}, \quad L \cap W_2 = \{\beta_j + \gamma_k + \delta'\psi_{jk}\},$$

where $\alpha_i, \beta_j, \gamma_k, \delta, \delta' \in \mathbb{R}$ are free parameters. Hence $L = (L \cap W_1) + (L \cap W_2)$ and (2) is conformal to two marginal spaces $\{L_{\{1,2\}}, L_{\{2,3\}}\}$. In the case of the model (6), however,

$$L \cap W_1 = \{\alpha_i + \beta_j\}, \quad L \cap W_2 = \{\beta_j + \gamma_k\}.$$

Hence $(L \cap W_1) + (L \cap W_2) = \{\alpha_i + \beta_j + \gamma_k\}$ and the model (6) is not conformal to $\{L_{\{1,2\}}, L_{\{2,3\}}\}$.

Given a subspace L consider a hierarchical model $L_\Delta \supset L$. We present the following definition of a hierarchical subspace model.

Definition 1. *L is a hierarchical subspace model (HSM) of L_Δ if the following conditions hold:*

1. *$L \subset L_\Delta$.*
2. *Each divider $S \in \mathcal{S}$ of* $\mathrm{red}\,\Delta$ *is saturated in L, i.e. $L_S \subset L$.*
3. *L is conformal to the set of subspaces $\{L_C, C \in \mathcal{C}\}$.*

Condition 2 guarantees that the MLE $\hat{p}(\boldsymbol{i})$ satisfies

$$\hat{p}(\boldsymbol{i}) = \frac{\prod_{C\in\mathcal{C}} \hat{p}(\boldsymbol{i}_C)}{\prod_{S\in\mathcal{S}} \hat{p}(\boldsymbol{i}_S)} = \frac{\prod_{C\in\mathcal{C}} \hat{p}(\boldsymbol{i}_C)}{\prod_{S\in\mathcal{S}} x(\boldsymbol{i}_S)/n}. \tag{9}$$

By Condition 3 the marginal probability $\hat{p}(\boldsymbol{i}_C)$ in the numerator of (9) coincides with the MLE of the model $L \cap L_C$, which is computed only on the marginal table $x(\boldsymbol{i}_C)$. We confirm this fact. By induction, it is sufficient to consider the case $\mathcal{C} = \{C, R\}$ with $S = C \cap R$. The MLE of the model L is the maximizer of $\sum_{\boldsymbol{i}} x(\boldsymbol{i}) \log p(\boldsymbol{i})$ subject to $\log p(\cdot) \in L$ and $\sum_{\boldsymbol{i}} p(\boldsymbol{i}) = 1$. By Condition 3 we write $\log p(\cdot) = \theta_C + \theta_R$ with $\theta_C \in L \cap L_C$ and $\theta_R \in L \cap L_R$. Since L_S is saturated both in $L \cap L_C$ and $L \cap L_R$, we can assume $\sum_{\boldsymbol{i}_{C\setminus S}} e^{\theta_C(\boldsymbol{i}_C)} = 1$ for each $\boldsymbol{i}_S$ without loss of generality. Hence the problem is decomposed into two parts: maximization of $\sum_{\boldsymbol{i}_C} x(\boldsymbol{i}_C)\theta_C(\boldsymbol{i}_C)$ subject to $\theta_C \in L \cap L_C$ and $\sum_{\boldsymbol{i}_{C\setminus S}} e^{\theta_C(\boldsymbol{i}_C)} = 1$, and maximization of $\sum_{\boldsymbol{i}_R} x(\boldsymbol{i}_R)\theta_R(\boldsymbol{i}_R)$ subject to $\theta_R \in L \cap L_R$ and $\sum_{\boldsymbol{i}_R} e^{\theta_R(\boldsymbol{i}_R)} = 1$. Since the maximizer $\hat{\theta}_C$ does not depend on R, it is computed from the case $R = S$. We have $\hat{\theta}_C(\boldsymbol{i}) = \log\{\check{p}(\boldsymbol{i}_C)/(x(\boldsymbol{i}_S)/n)\}$, where $\check{p}(\boldsymbol{i}_C)$ is the MLE of the model $L \cap L_C$. Hence the computation of the MLE of an HSM of L_Δ is localized to each $C \in \mathcal{C}$.

2.2 Ambient decomposable model of a log-affine model

In Definition 1, L is an HSM of a particular L_Δ. Note that by definition every log-affine model L is an HSM of the saturated model $L_{[m]}$. Therefore every log-affine model L has a hierarchical model for which L is an HSM and a natural question is to look for a small simplicial complex Δ such that L is an HSM of L_Δ. As the main theoretical result of this paper we show in Theorem 1 below that for each log-affine model L there exists a natural smallest decomposable model $L_\mathcal{H}$ such that L is an HSM of $L_\mathcal{H}$. We call such $L_\mathcal{H}$ the *ambient decomposable model* of L.

We want to define the ambient decomposable model in such a way that the conditional independence model $i \perp\!\!\!\perp k \mid j$ is the ambient decomposable model for (2) and the saturated model $L_{[3]}$ is the ambient decomposable model for (6).

In order to define the ambient decomposable model, we first define connectedness of L and a partial edge separator of L. L is called disconnected if there exists a non-empty proper subset A of $[m]$ such that L is conformal to $\{L_A, L_{A^C}\}$, where A^C denotes the complement of A. This means that the variables in A and the variables in A^C are independent and independently modeled in L. We call L *connected* if L is not disconnected. It is obvious that under this definition L can be decomposed into its connected components and each connected component can be investigated separately. Therefore from now on we assume that L is connected.

Definition 2. *A non-empty subset S of $[m]$ is called a* partial edge separator *of L if $[m]$ is partitioned into three non-empty and disjoint subsets $\{A_1, A_2, S\}$ such that*

1. *S is saturated in L.*
2. *L is conformal to $\{L_{A_1\cup S}, L_{A_2\cup S}\}$.*

Then we call the triple (A_1, A_2, S) a decomposition *of L. When the model L has a partial edge separator, we call L* reducible. *A pair of vertices i and j are called* tightly connected *in L if there does not exist a decomposition (A_1, A_2, S) of L such that $i \in A_1$ and $j \in A_2$. When L is not reducible, we call L* prime.

A set of vertices such that any two of them are tightly connected in L is called *extended compact component* of L. The set of maximal extended compact components of L is considered as a hypergraph and is denoted by $\mathcal{H}$. Denote by $L_\mathcal{H}$ the hierarchical model induced by $\mathcal{H}$. Then we have the following theorem.

Theorem 1. *$L_\mathcal{H}$ is the smallest decomposable model with respect to inclusion relation such that L is an HSM of $L_\mathcal{H}$.*

The following corollary is obvious from (9).

Corollary 1. *The MLE $\hat{p}(\boldsymbol{i})$ satisfies*

$$\hat{p}(\boldsymbol{i}) = \frac{\prod_{C\in\mathcal{H}} \hat{p}(\boldsymbol{i}_C)}{\prod_{S\in\mathcal{S}} x(\boldsymbol{i}_S)/n},$$

where $\mathcal{S}$ is the set of dividers of $\mathcal{H}$ and $\hat{p}(\boldsymbol{i}_C)$ depends only on the marginal table $x(\boldsymbol{i}_C)$.

The rest of this subsection is devoted to a proof of Theorem 1. Before we give the proof, we present some lemmas required to prove the theorem.

Lemma 1. *If S is a partial edge separator of L, S is also a partial edge separator of $\mathcal{H}$.*

Proof. Since S is saturated in L, S is an extended compact component. Hence S is a partial edge of $\mathcal{H}$. Denote by $\mathcal{H}([m] \setminus S)$ the subhypergraph of $\mathcal{H}$ induced by $[m] \setminus S$. Assume that S is not a separator of $\mathcal{H}$. Then $\mathcal{H}([m] \setminus S)$ is connected.

Since S is a separator of L, there exists a decomposition (A, B, S) of L by definition. Define $\tilde{\mathcal{H}}(A)$ and $\tilde{\mathcal{H}}(B)$ by

$$\tilde{\mathcal{H}}(A) := \{C \in \mathcal{H} \mid A \cap C \neq \emptyset\}, \quad \tilde{\mathcal{H}}(B) := \{C \in \mathcal{H} \mid B \cap C \neq \emptyset\}.$$

Then we have $\tilde{\mathcal{H}}(A) \cap \tilde{\mathcal{H}}(B) = \emptyset$ which contradicts the fact that $\mathcal{H}([m] \setminus S)$ is connected. $\square$

By using Lemma 1, we can prove the following lemma in the same way as Theorem 5 in Malvestuto and Moscarini [2000].

Lemma 2. *$\mathcal{H}$ is acyclic.*

Denote by $\mathcal{S}$ the set of dividers of $\mathcal{H}$.

Lemma 3. *Suppose $S \in \mathcal{S}$ is a divider of $\mathcal{H}$ with a decomposition (A, B, S). Then S is a partial edge separator of L with a decomposition (A, B, S).*

Proof. Since S is a divider, there exists a pair of vertices $\{u, v\}$ such that S is the unique minimal partial edge separating u and v. Then there exists a decomposition (A, B, S) such that $u \in A$ and $v \in B$. Any vertices in A and any vertices in B are not tightly connected in L. This implies that there exists a partial edge separator $S' \subset S$ of L and a decomposition (A', B', S') of L satisfying $A' \supset A$ and $B' \supset B$. From Lemma 1, S' is also a partial edge separator of $\mathcal{H}$. Noting that S is the unique minimal partial edge of $\mathcal{H}$ separating u and v, we have $S' = S$. Then (A, B, S) is a decomposition of L. $\square$

Now we provide a proof of Theorem 1.

Proof of Theorem 1. It is obvious that $L \subset L_{\mathcal{H}}$. From Lemma 3, every divider $S \in \mathcal{S}$ of $\mathcal{H}$ is a partial edge separator of L and hence saturated in L. From Lemma 2, $\mathcal{H}$ is considered as the set of maximal cliques of a chordal graph $\mathcal{G}^{\mathcal{H}}$. Let C_k, $k = 1, \ldots, K$, be a perfect sequence of maximal cliques in $\mathcal{G}^{\mathcal{H}}$ (see e.g. Section 2.1.3 of Lauritzen [1996]). Let

$$B_k := C_1 \cup C_2 \cup \cdots \cup C_k, \quad R_k := (C_K \cup C_{K-1} \cup \cdots \cup C_k) \setminus S_k, \quad S_k := B_{k-1} \cap C_k.$$

It is known that S_K is a divider of $\mathcal{H}$ with a decomposition (B_{K-1}, R_K, S_K). From Lemma 3, S_K is a partial edge separator in L with the same decomposition. Hence L is conformal to $\{L_{B_{K-1}}, L_{C_K}\}$, i.e.

$$L = (L \cap L_{B_{K-1}}) + (L \cap L_{C_K}).$$

In the same way S_{K-1} is a partial edge separator in L with a decomposition $(B_{K-2}, R_{K-1}, S_{K-1})$ and hence L is conformal to $\{L_{B_{K-2}}, L_{C_K \cup C_{K-1}}\}$, i.e.

$$\begin{aligned} L &= (L \cap L_{B_{K-2}}) + (L \cap L_{C_K \cup C_{K-1}}) \\ &= \big[\big((L \cap L_{B_{K-1}}) + (L \cap L_{C_K})\big) \cap L_{B_{K-2}}\big] \\ &\quad + \big[\big((L \cap L_{B_{K-1}}) + (L \cap L_{C_K})\big) \cap L_{C_{K-1} \cup C_K}\big] \\ &= (L \cap L_{B_{K-2}}) + (L \cap L_{C_{K-1}}) + (L \cap L_{C_K}). \end{aligned}$$

By iterating this procedure, we can obtain $L = (L \cap L_{C_1}) + \cdots + (L \cap L_{C_K})$. Hence L is conformal to $\{L_C, C \in \mathcal{H}\}$. Therefore L is an HSM of $L_{\mathcal{H}}$.

Suppose that there exists a smaller decomposable model $L_{\mathcal{H}'} \subset L_{\mathcal{H}}$ for which L is an HSM. Then there exist $C \in \mathcal{H}$ and a divider S' of $\mathcal{H}'$ such that $S' \subset C$. This contradicts the fact that any vertices in C are tightly connected in L. □

2.3 Hierarchical models containing a log-affine model

In Theorem 1 we have shown the existence of the smallest decomposable model containing a log-affine model. Then a natural question is to ask whether there exists the smallest hierarchical model containing a log-affine model. In general this does not hold and we here discuss properties of hierarchical models containing a log-affine model.

As an example consider the model (6). Although (6) is a submodel of the conditional independence model $i \perp\!\!\!\perp k \mid j$, (6) is not an HSM of the conditional independence model. The difficulty lies in the fact that a hierarchical model containing L may have a partial edge separator which is not a partial edge separator of L.

Given a log-affine model L, consider the set of hierarchical models L_Δ containing L: $\{L_\Delta \mid L_\Delta \supset L\}$. Because of the relation $L_\Delta \cap L_{\Delta'} = L_{\Delta \cap \Delta'}$ it follows that there exists the smallest hierarchical model in $\{L_\Delta \mid L_\Delta \supset L\}$. We call the smallest hierarchical model containing L as *hierarchical closure* of L and denote the corresponding simplicial complex and the hierarchical model by $\bar{\Delta}(L)$ and $L_{\bar{\Delta}(L)}$, respectively. Note that for both (2) and (6), the hierarchical closure is the three-way conditional independence model (1).

We call a log-affine model L a tight hierarchical subspace model if L is an HSM of $L_{\bar{\Delta}(L)}$. If L is a tight HSM, obviously $\bar{\Delta}(L)$ is the smallest simplicial complex such that L is its HSM.

We now present an example of a log-affine model L of a 5-way contingency table, which has two minimal hierarchical models Δ_1, Δ_2, such that L is an HSM of both L_{Δ_1} and L_{Δ_2}. Consider the following model L of 5-way contingency tables:

$$\begin{aligned} \log p(i_1, \ldots, i_5) = \sum_{j=1}^{5} \alpha_{\{j\}}(i_j) + \theta\big(&\psi_{\{1,2\}}(i_1, i_2) + \psi_{\{1,3\}}(i_1, i_3) + \psi_{\{2,3\}}(i_2, i_3) \\ &+ \psi_{\{2,4\}}(i_2, i_4) + \psi_{\{3,5\}}(i_3, i_5) + \psi_{\{4,5\}}(i_4, i_5)\big), \end{aligned}$$

where the main effects $\alpha_{\{j\}}$'s and θ are parameters and $\psi_{\{j,j'\}}$'s are fixed functions. The hierarchical closure of this model is given by

$$\operatorname{red} \bar{\Delta} = \{\{1,2\}, \{1,3\}, \{2,3\}, \{2,4\}, \{3,5\}, \{4,5\}\}$$

which has a divider $\{2,3\}$. Hence L is not tight. Note that L is an HSM of any L_Δ, such that L_Δ does not possess a partial edge separator and $L \subset L_\Delta$. As in Figure 1 define

$$\text{red}\,\Delta_1 = \text{red}\,\bar{\Delta} \cup \{\{1,4\}\}, \qquad \text{red}\,\Delta_2 = \text{red}\,\bar{\Delta} \cup \{\{1,5\}\}.$$

Then L is an HSM of both L_{Δ_1} and L_{Δ_2}.

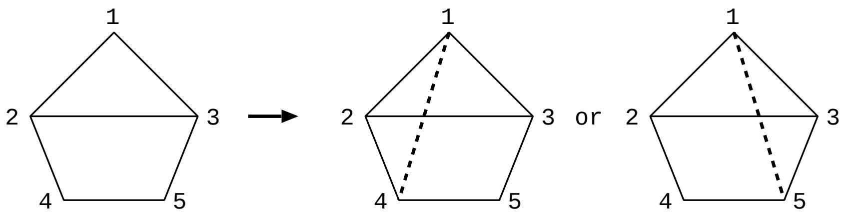


Figure 1: Two ways to cross a divider of the hierarchical closure

3 Split model as a hierarchical subspace model

We consider the split model by Højsgaard [2003]. An example of the split model for three-way tables is $L = L_{\{1\}}^{i_2=1} + L_{\{3\}}^{i_2=1} + L_{\{1,3\}}^{i_2=2}$, which represents that there exists a conditional interaction of i_1 and i_3 given i_2 only if $i_2 = 2$. A precise definition is given below.

We first define the context specific interaction (CSI) model (Højsgaard [2004]). The split model is a particular case of the CSI model. Recall that $V = \mathbb{R}^{|\mathcal{I}|}$ is the set of all tables. For any subset B of $[m]$ and $\boldsymbol{j}_B \in \mathcal{I}_B$, we consider a subspace $L^{\boldsymbol{j}_B}$ of V in that only the $\boldsymbol{j}_B$-slice has nonzero components, that is,

$$\begin{aligned} L^{\boldsymbol{j}_B} &= \{\psi \in V \mid \psi(\boldsymbol{i}) = 0 \text{ if } \boldsymbol{i}_B \neq \boldsymbol{j}_B\}. \\ &= \left\{\psi \in V \mid \psi(\boldsymbol{i}) = f(\boldsymbol{i}_{[m]\setminus B}) 1_{\{\boldsymbol{i}_B = \boldsymbol{j}_B\}},\ f : \mathcal{I}_{[m]\setminus B} \to \mathbb{R}\right\}. \end{aligned}$$

If B is empty, we define $L^{\boldsymbol{j}_\emptyset} = V$ with a dummy symbol $\boldsymbol{j}_\emptyset$. For any subsets B and D of $[m]$ and any level $\boldsymbol{j}_B \in \mathcal{I}_B$, we define a subspace

$$L_D^{\boldsymbol{j}_B} = L_{D\cup B} \cap L^{\boldsymbol{j}_B} = \left\{\psi \in V \mid \psi(\boldsymbol{i}) = f(\boldsymbol{i}_{D\setminus B}) 1_{\{\boldsymbol{i}_B = \boldsymbol{j}_B\}},\ f : \mathcal{I}_{D\setminus B} \to \mathbb{R}\right\}.$$

The subspace $L_D^{\boldsymbol{j}_B}$ represents *a context specific interaction*, that is, an interaction over $\boldsymbol{i}_D$ exists only if $\boldsymbol{i}_B = \boldsymbol{j}_B$. The following relation is easily proved:

$$L_{D\cup B} = \sum_{\boldsymbol{j}_B \in \mathcal{I}_B} L_D^{\boldsymbol{j}_B}. \tag{10}$$

A context specific interaction (CSI) model is a direct sum of subspaces $L_D^{\boldsymbol{j}_B}$ for a set of $(\boldsymbol{j}_B, D)$'s. It is easily shown that any hierarchical model is a CSI model.

Next we define split models. In order to clarify the definition, we consider a more general model, the split subspace model. The split model is a particular case of the split subspace models. Although Højsgaard [2003] defined the split model on the basis of a graphical model, we let the graphical model be a decomposable model for simplicity.

Consider a decomposable model L_Δ with the set of maximal cliques $\mathcal{C}$. For each $C \in \mathcal{C}$ choose a subset $Z(C) \subset C$. We admit the case where $Z(C)$ is empty. For each $\boldsymbol{j}_{Z(C)} \in \mathcal{I}_{Z(C)}$, choose a subspace $N_C^{\boldsymbol{j}_{Z(C)}} \subset L_C^{\boldsymbol{j}_{Z(C)}}$ such that

$$\forall C' \in \mathcal{C} \setminus \{C\}, \quad L_{C \cap C'}^{\boldsymbol{j}_{Z(C)}} \subset N_C^{\boldsymbol{j}_{Z(C)}} \subset L_C^{\boldsymbol{j}_{Z(C)}}. \tag{11}$$

Then a log-affine model L is defined by

$$L \;=\; \sum_{C \in \mathcal{C}} N_C, \quad N_C \;=\; \sum_{\boldsymbol{j}_{Z(C)} \in \mathcal{I}_{Z(C)}} N_C^{\boldsymbol{j}_{Z(C)}}. \tag{12}$$

We call L *a split subspace model with root* $\mathcal{C}$ if it satisfies (11) and (12). The following proposition holds.

Proposition 1. *Let L_Δ be a decomposable model with the cliques $\mathcal{C}$. Then any split subspace model L with root $\mathcal{C}$ is an HSM of L_Δ.*

Proof. We first check $L \subset \sum_{C \in \mathcal{C}} L_C$. Since $L = \sum_{C \in \mathcal{C}} N_C$, it is sufficient to show $N_C \subset L_C$ for each $C \in \mathcal{C}$. However, this is clear because $N_C^{\boldsymbol{j}_{Z(C)}} \subset L_C^{\boldsymbol{j}_{Z(C)}} \subset L_C$ for any $\boldsymbol{j}_{Z(C)}$. Next we prove that $L_S \subset L$ for any divider S. From the definition of dividers of decomposable models, there exist two cliques C and C' ($C \neq C'$) such that $S = C' \cap C$. By the relations (10) and (11), we have

$$L_S \;\subset\; L_{(C' \cap C) \cup Z(C)} \;=\; \sum_{\boldsymbol{j}_{Z(C)} \in \mathcal{I}_{Z(C)}} L_{C' \cap C}^{\boldsymbol{j}_{Z(C)}} \;\subset\; \sum_{\boldsymbol{j}_{Z(C)} \in \mathcal{I}_{Z(C)}} N_C^{\boldsymbol{j}_{Z(C)}} \;=\; N_C.$$

Therefore $L_S \subset L$. Lastly, we prove that L is conformal to $\{L_C \mid C \in \mathcal{C}\}$. We have already proved $N_C \subset L_C$. Since N_C is also a subspace of L, we obtain $N_C \subset L \cap L_C$ and therefore $L = \sum_{C \in \mathcal{C}} N_C \subset \sum_{C \in \mathcal{C}} (L \cap L_C)$. The opposite inclusion is obvious. □

Now we define a split model as a special case of split subspace models. We say that any decomposable model is a split model of degree zero. Then *a split model of degree one* is defined as the decomposition (12) with

$$N_C^{\boldsymbol{j}_{Z(C)}} \;=\; \sum_{D \in \mathcal{C}_C^{\boldsymbol{j}_{Z(C)}}} L_D^{\boldsymbol{j}_{Z(C)}},$$

where $\mathcal{C}_C^{\boldsymbol{j}_{Z(C)}}$ is a decomposable model with the vertex set $C \setminus Z(C)$. Here we assume

$$\forall C' \in \mathcal{C} \setminus \{C\}, \; \exists D \in \mathcal{C}_C^{\boldsymbol{j}_{Z(C)}} \text{ s.t. } (C \cap C') \setminus Z(C) \subset D \tag{13}$$

to assure the condition (11). Split models of degree greater than one are defined recursively. See Højsgaard [2003] for details.

In Section 5, we will consider an example of the split model (of degree one). The following elementary lemma is useful to obtain the MLE of split models.

Lemma 4. *Let $\mathcal{I} = \bigcup_\lambda \mathcal{J}_\lambda$ be a partition of $\mathcal{I}$ and consider subspaces $N_\lambda \subset V$ such that*

$$N_\lambda \subset \{\psi \in V \mid \psi(\boldsymbol{i}) = 0 \text{ if } \boldsymbol{i} \notin \mathcal{J}_\lambda\}.$$

Then the MLE of the model $\sum_\lambda N_\lambda$ is given by $\hat{p}(\boldsymbol{i}) = \sum_\lambda (n_\lambda/n)\hat{p}_\lambda(\boldsymbol{i}) 1_{\{\boldsymbol{i}\in\mathcal{J}_\lambda\}}$, where $\hat{p}_\lambda(\boldsymbol{i})$ is the MLE of the model N_λ with the total frequency $n_\lambda = \sum_{\boldsymbol{i}\in\mathcal{I}_\lambda} x(\boldsymbol{i})$.

4 Conditional tests of hierarchical subspace models via Markov bases

In this section we discuss conditional tests of our proposed model via Markov bases technique. In Section 1.1, we have discussed that the divide-and-conquer approach of Dobra and Sullivant [2004] still works for the model (2). In this section we generalize the argument to an HSM L.

Let $\boldsymbol{x} = \{x(\boldsymbol{i})\}_{i\in\mathcal{I}}$ denote an m-way contingency table, where $x(\boldsymbol{i})$ denotes the frequency of the cell $\boldsymbol{i} \in \mathcal{I}$. Let $\boldsymbol{b}$ be the set of sufficient statistics for L. We assume that the elements of $\boldsymbol{b}$ are integer combinations of the frequencies $x(\boldsymbol{i})$. For a hierarchical model L_Δ, $\boldsymbol{b}$ is written by

$$\boldsymbol{b} = \{x_D(\boldsymbol{i}_D), \boldsymbol{i}_D \in \mathcal{I}_D, D \in \operatorname{red}\Delta\},$$

where $x_D(\boldsymbol{i}_D) = \sum_{\boldsymbol{i}_{D^C}\in\mathcal{I}_{D^C}} x(\boldsymbol{i}_D, \boldsymbol{i}_{D^C})$. We consider $\boldsymbol{b}$ as a column vector with dimension ν. We order the elements of $\boldsymbol{x}$ appropriately and consider $\boldsymbol{x}$ as a column vector. Then the relation between the joint frequencies $\boldsymbol{x}$ and the marginal frequencies $\boldsymbol{b}$ is written simply as

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

where A is a $\nu \times |\mathcal{I}|$ integer matrix. A is called the configuration for L.

For a subset $D \subset [m]$, denote $L(D) := L \cap L_D$. Let (A_1, A_2, S) be a decomposition of L and define $V_1 := A_1 \cup S$ and $V_2 := A_2 \cup S$. Since L is conformal to $\{L_{V_1}, L_{V_2}\}$, we note that $L(V_1)$ and $L(V_2)$ are marginal models corresponding to V_1 and V_2, respectively. Denote by $A_{V_1} = \{\boldsymbol{a}_{V_1}(\boldsymbol{i}_{V_1})\}_{\boldsymbol{i}_{V_1}\in\mathcal{I}_{V_1}}$ and $A_{V_2} = \{\boldsymbol{a}_{V_2}(\boldsymbol{i}_{V_2})\}_{\boldsymbol{i}_{V_2}\in\mathcal{I}_{V_2}}$ the configurations for the marginal models $L(V_1)$ and $L(V_2)$, where $\boldsymbol{a}_{V_1}(\boldsymbol{i}_{V_1})$ and $\boldsymbol{a}_{V_2}(\boldsymbol{i}_{V_2})$ denote column vectors of A_{V_1} and A_{V_2}, respectively. Noting that $\boldsymbol{i}_{V_1} = (\boldsymbol{i}_{A_1}\boldsymbol{i}_S)$ and $\boldsymbol{i}_{V_2} = (\boldsymbol{i}_S\boldsymbol{i}_{A_2})$, the configuration A for L is written by

$$A = A_{V_1} \oplus_S A_{V_2} = \{\boldsymbol{a}_{V_1}(\boldsymbol{i}_{A_1}\boldsymbol{i}_S) \oplus \boldsymbol{a}_{V_2}(\boldsymbol{i}_S\boldsymbol{i}_{A_2})\}_{\boldsymbol{i}_{A_1}\in\mathcal{I}_{A_1}, \boldsymbol{i}_S\in\mathcal{I}_S, \boldsymbol{i}_{A_2}\in\mathcal{I}_{A_2}},$$

where

$$\boldsymbol{a}_{V_1}(\boldsymbol{i}_{A_1}\boldsymbol{i}_S) \oplus \boldsymbol{a}_{V_2}(\boldsymbol{i}_S\boldsymbol{i}_{A_2}) = \begin{pmatrix} \boldsymbol{a}_{V_1}(\boldsymbol{i}_{A_1}\boldsymbol{i}_S) \\ \boldsymbol{a}_{V_2}(\boldsymbol{i}_S\boldsymbol{i}_{A_2}) \end{pmatrix}.$$

Given $\boldsymbol{b}$, the set

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

of contingency tables sharing the same $\boldsymbol{b}$ is called a *fiber*. An integer array $\boldsymbol{z} = \{z(\boldsymbol{i})\}_{\boldsymbol{i}\in\mathcal{I}}$ of the same dimension as $\boldsymbol{x}$ is called a *move* if $A\boldsymbol{z} = \boldsymbol{0}$. As in (8), we denote $\boldsymbol{z}$ with degree d as

$$\boldsymbol{z} = [\{\boldsymbol{i}_1, \ldots, \boldsymbol{i}_d\} \| \{\boldsymbol{i}'_1, \ldots, \boldsymbol{i}'_d\}],$$

where $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_d \in \mathcal{I}$ are cells (with replication) of positive elements of $\boldsymbol{z}$ and $\boldsymbol{i}'_1, \ldots, \boldsymbol{i}'_d \in \mathcal{I}$ are cells of negative elements of $\boldsymbol{z}$. Moves are used for steps of Markov chain Monte Carlo simulation within each fiber. If we add or subtract a move $\boldsymbol{z}$ to $\boldsymbol{x} \in \mathcal{F}_{\boldsymbol{b}}$, then $\boldsymbol{x} \pm \boldsymbol{z} \in \mathcal{F}_{\boldsymbol{b}}$ and we can move from $\boldsymbol{x}$ to another state $\boldsymbol{x} + \boldsymbol{z}$ (or $\boldsymbol{x} - \boldsymbol{z}$) in the same fiber $\mathcal{F}_{\boldsymbol{b}}$, as long as there is no negative element in $\boldsymbol{x} + \boldsymbol{z}$ (or $\boldsymbol{x} - \boldsymbol{z}$). 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 $\mathcal{M}$.

Assume that $\mathcal{B}(V_1)$ and $\mathcal{B}(V_2)$ are Markov bases for $L(V_1)$ and $L(V_2)$, respectively. Let $\boldsymbol{z}_1 = \{z_1(\boldsymbol{i}_{V_1})\}_{\boldsymbol{i}_{V_1}\in\mathcal{I}_{V_1}} \in \mathcal{B}(V_1)$ and $\boldsymbol{z}_2 = \{z_2(\boldsymbol{i}_{V_2})\}_{\boldsymbol{i}_{V_2}\in\mathcal{I}_{V_2}} \in \mathcal{B}(V_2)$. Since S is saturated, we have

$$\sum_{\boldsymbol{i}_{V_1\setminus S}\in\mathcal{I}_{V_1\setminus S}} z_1(\boldsymbol{i}_{V_1}) = 0, \qquad \sum_{\boldsymbol{i}_{V_2\setminus S}\in\mathcal{I}_{V_2\setminus S}} z_2(\boldsymbol{i}_{V_2}) = 0.$$

Hence $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ can be written as

$$\boldsymbol{z}_1 = [\{(\boldsymbol{i}_1, \boldsymbol{j}_1), \ldots, (\boldsymbol{i}_d, \boldsymbol{j}_d)\} \| \{(\boldsymbol{i}'_1, \boldsymbol{j}_1), \ldots, (\boldsymbol{i}'_d, \boldsymbol{j}_d)\}], \tag{14}$$

$$\boldsymbol{z}_2 = [\{(\boldsymbol{j}_1, \boldsymbol{k}_1), \ldots, (\boldsymbol{j}_d, \boldsymbol{k}_d)\} \| \{(\boldsymbol{j}_1, \boldsymbol{k}'_1), \ldots, (\boldsymbol{j}_d, \boldsymbol{k}'_d)\}],$$

respectively, where $\boldsymbol{i}_k, \boldsymbol{i}'_k \in \mathcal{I}_{A_1}$, $\boldsymbol{j}_k \in \mathcal{I}_S$ and $\boldsymbol{k}_k, \boldsymbol{k}'_k \in \mathcal{I}_{A_2}$ for $k = 1, \ldots, d$.

Definition 3 (Dobra and Sullivant [2004]). *Define $\boldsymbol{z}_1 \in \mathcal{B}(V_1)$ as in (14). Let $\boldsymbol{k} := \{\boldsymbol{k}_1, \ldots, \boldsymbol{k}_d\} \in \mathcal{I}_{A_2} \times \cdots \times \mathcal{I}_{A_2}$. Define $\boldsymbol{z}_1^{\boldsymbol{k}}$ by*

$$\boldsymbol{z}_1^{\boldsymbol{k}} := [\{(\boldsymbol{i}_1, \boldsymbol{j}_1, \boldsymbol{k}_1), \ldots, (\boldsymbol{i}_d, \boldsymbol{j}_d, \boldsymbol{k}_d)\} \| \{(\boldsymbol{i}'_1, \boldsymbol{j}_1, \boldsymbol{k}_1), \ldots, (\boldsymbol{i}'_d, \boldsymbol{j}_d, \boldsymbol{k}_d)\}].$$

Then we define $\mathrm{Ext}(\mathcal{B}(V_1) \to L)$ *by*

$$\mathrm{Ext}(\mathcal{B}(V_1) \to L) := \{\boldsymbol{z}_1^{\boldsymbol{k}} \mid \boldsymbol{k} \in \mathcal{I}_{A_2} \times \cdots \times \mathcal{I}_{A_2}\}.$$

In the same way as Lemma 5.4 in Dobra and Sullivant [2004] we can obtain the following lemma.

Lemma 5. *Suppose that $z_1 \in \mathcal{B}(V_1)$ as in (14). Then $\mathrm{Ext}(\mathcal{B}(V_1) \to L)$ is the set of moves for L.*

Proof. Let $\boldsymbol{z} \in \mathrm{Ext}(\mathcal{B}(V_1) \to L)$. Then we have

$$A\boldsymbol{z} = \begin{pmatrix} \sum_{\boldsymbol{i}_{V_1}\in\mathcal{I}_{V_1}} \boldsymbol{a}_{V_1}(\boldsymbol{i}_{V_1}) z_{V_1}(\boldsymbol{i}_{V_1}) \\ \sum_{\boldsymbol{i}_{V_2}\in\mathcal{I}_{V_2}} \boldsymbol{a}_{V_2}(\boldsymbol{i}_{V_2}) z_{V_2}(\boldsymbol{i}_{V_2}) \end{pmatrix},$$

where

$$z_{V_1}(\boldsymbol{i}_{V_1}) = \sum_{\boldsymbol{i}_{V_1^C} \in \mathcal{I}_{V_1^C}} z(\boldsymbol{i}), \quad z_{V_2}(\boldsymbol{i}_{V_2}) = \sum_{\boldsymbol{i}_{V_2^C} \in \mathcal{I}_{V_2^C}} z(\boldsymbol{i}).$$

Since $z_{V_1}(\boldsymbol{i}_{V_1}) = z_1(\boldsymbol{i}_{V_1})$ and $z_1 \in \mathcal{B}(V_1)$, $\sum_{\boldsymbol{i}_{V_1} \in \mathcal{I}_{V_1}} \boldsymbol{a}_{V_1}(\boldsymbol{i}_{V_1}) z_{V_1}(\boldsymbol{i}_{V_1}) = 0$. From Definition 3, $z_{V_2}(\boldsymbol{i}_{V_2}) = 0$ for all $\boldsymbol{i}_{V_2} \in \mathcal{I}_{V_2}$. Hence $A\boldsymbol{z} = 0$. □

Theorem 2. *Let $\mathcal{B}(V_1)$ and $\mathcal{B}(V_2)$ be Markov bases for $L(V_1)$ and $L(V_2)$, respectively. Let $\mathcal{B}_{V_1,V_2}$ is a Markov basis for the hierarchical model with two cliques V_1 and V_2. Then*

$$\mathcal{B} := \mathrm{Ext}(\mathcal{B}(V_1) \to L) \cup \mathrm{Ext}(\mathcal{B}(V_2) \to L) \cup \mathcal{B}_{V_1,V_2}. \tag{15}$$

is a Markov basis for L.

We can prove the theorem in the same way as Theorem 5.6 in Dobra and Sullivant [2004]. Suppose that L is an HSM of $L_{\mathcal{H}}$. Then Theorem 2 implies that a Markov basis for L is obtained from $\mathcal{B}(C)$, $C \in \mathcal{H}$, by recursively using (15). This shows that the computation of a Markov basis can be localized according to the maximal extended compact components of L.

Concerning Markov bases of the split model of Section 3 we state the following lemma.

Lemma 6. *With the same notation as in Lemma 4, a Markov basis of the model $\sum_\lambda N_\lambda$ is given by union of Markov bases of N_λ.*

5 Examples

In this section we give several applications of HSMs. In Section 5.1 we analyze the data on song sequence of a wood pewee, which we already discussed in Section 1.1. In Section 5.2 we consider an example of a split model.

5.1 Sequences of unrepeated events

Consider the data on song sequence of a wood pewee in Table 1. As mentioned in Section 1.1, it is a $4 \times 4 \times 4$ contingency table with the cells of the form (i,i,k) and (i,j,j) being structural zeros. The probability function $\{p_{ijk}\}$ satisfies the condition $p_{iik} = 0$ and $p_{ijj} = 0$, or equivalently, $\log p_{iik} = -\infty$ and $\log p_{ijj} = -\infty$. Hence $\{\log p_{ijk}\}$ is not an element of $V = \mathbb{R}^{4\times4\times4}$. However we can replace V by $R^{|\bar{\mathcal{I}}|}$, where

$$\bar{\mathcal{I}} = \mathcal{I} \setminus \big(\{(i,i,j), i,j \in [4]\} \cup \{(i,j,j), i,j \in [4]\}\big),$$

and consider log-affine models of $R^{|\bar{\mathcal{I}}|}$. Formally it is more convenient to proceed with $V = \mathbb{R}^{4\times4\times4}$ allowing $\log p_{iik} = \log p_{ijj} = -\infty$.

We first consider the conditional independence model

$$L_{\mathrm{Model1}} \;=\; L_{\{1,2\}} + L_{\{2,3\}},$$

which corresponds to (7). The MLE of this model is explicitly given by

$$\hat{p}_{ijk} = \frac{x_{ij+}x_{+jk}}{nx_{+j+}} = \frac{x_{ij+}1_{\{i\neq j\}}x_{+jk}1_{\{j\neq k\}}}{nx_{+j+}}.$$

A Markov basis of the model is $\mathcal{B}_{\text{Model1}} = \mathcal{B}_{\{1,2\},\{2,3\}}$ (see Theorem 2 for the notation). An experimental result that compares the saturated model and Model 1 is given in Figure 2. Both the asymptotic and experimental estimates of the p-value are almost zero.

Although Model 1 does not fit the data, we proceed to consider a submodel of Model 1 for theoretical interest. Let

$$L_{\text{model2}} = \left\{\alpha_i + \beta_j + \gamma_k + \phi_i 1_{\{i=j\}} + \psi_j 1_{\{j=k\}}\right\}.$$

This model is an HSM of $L_{\{1,2\}} + L_{\{2,3\}}$. It represents a quasi-independence model for the three-way table. The MLE of the model is

$$\hat{p}_{ijk} = \frac{\hat{p}^{(1)}_{ij}\hat{p}^{(2)}_{jk}}{x_{+j+}/n},$$

where $\hat{p}^{(1)}_{ij}$ and $\hat{p}^{(2)}_{jk}$ are the MLE of the 2-way quasi-independence models with the diagonal structural zeros, that is,

$$\begin{aligned}
\hat{p}^{(1)}_{ii} &= e^{\hat{\alpha}_i}e^{\hat{\beta}_j}1_{\{i\neq j\}}, \quad \hat{p}^{(1)}_{i+} = x_{i++}/n, \quad \hat{p}^{(1)}_{+j} = x_{+j+}/n,\\
\hat{p}^{(2)}_{jj} &= e^{\hat{\beta}_j}e^{\hat{\gamma}_k}1_{\{j\neq k\}}, \quad \hat{p}^{(2)}_{j+} = x_{+j+}/n, \quad \hat{p}^{(2)}_{+k} = x_{++k}/n.
\end{aligned}$$

They are computed by the iterative proportional fitting method. By Theorem 2, a Markov basis is given by

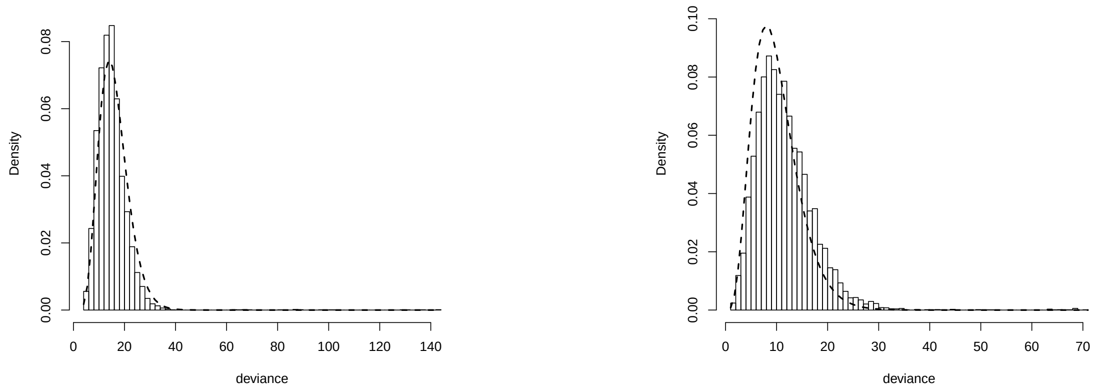
$$\mathcal{B}_{\text{Model2}} = \mathcal{B}_{\{1,2\},\{2,3\}} \cup \text{Ext}(\mathcal{B}(\{1,2\}) \to V) \cup \text{Ext}(\mathcal{B}(\{2,3\}) \to V)$$

where $\mathcal{B}(\{1,2\})$ and $\mathcal{B}(\{2,3\})$ are the Markov bases of the 2-way quasi-independence model with structural zeros obtained by Aoki and Takemura [2005]. An experimental result that compares the Model 1 and Model 2 is given in Figure 2.

5.2 WAM data

Here we deal with a real data called *women and mathematics (wam) data* used in Højsgaard [2003]. The data is shown in Table 2. The data consists of the following six factors: (1) Attendance in math lectures (attended=1, not=2), (2) Sex (female=1, male=2), (3) School type (suburban=1, urban=2), (4) Agree in statement "I'll need mathematics in my future work" (agree=1, disagree=2), (5) Subject preference (math-science=1, liberal arts=2) and (6) Future plans (college=1, job=2). We consider two models Højsgaard [2003] treated. The first model is a decomposable model

$$L_{\text{Model1}} = L_{\{1,2,3,5\}} + L_{\{2,3,4,5\}} + L_{\{3,4,5,6\}}.$$



(a) Deviance of Model 1 ($G^2 = 142.4$). (b) Deviance of Model 2 from Model 1 ($G^2 = 66.9$).

Figure 2: The empirical distribution and asymptotic distribution of deviance G^2 for the wood pewee data. The degree of freedom is 16 and 10, respectively. The number of steps in the MCMC procedure is 10^5.

By Theorem 2, a Markov basis of this model is given by

$$\mathcal{B}_{\mathrm{Model1}} \;=\; \mathcal{B}_{\{1,2,3,5\},\{2,3,4,5,6\}} \cup \mathcal{B}_{\{1,2,3,4,5\},\{3,4,5,6\}}.$$

The second model is a split model

$$L_{\mathrm{Model2}} \;=\; L_{\{1,2,3,5\}} + L^{j_3=1}_{\{2,5\}} + L^{j_3=1}_{\{4,5\}} + L^{j_3=2}_{\{2,4,5\}} + L_{\{3,4,5,6\}}.$$

This model is indeed a split model (of degree one) with

$$\begin{aligned}
\mathcal{C} &= \{\{1,2,3,5\},\{2,3,4,5\},\{3,4,5,6\}\},\\
Z(\{1,2,3,5\}) &= \emptyset, \quad \mathcal{C}^{\boldsymbol{j}_\emptyset}_{\{1,2,3,5\}} \;=\; \{\{1,2,3,5\}\},\\
Z(\{2,3,4,5\}) &= \{3\}, \quad \mathcal{C}^{j_3=1}_{\{2,3,4,5\}} \;=\; \{\{2,5\},\{4,5\}\}, \quad \mathcal{C}^{j_3=2}_{\{2,3,4,5\}} \;=\; \{\{2,4,5\}\},\\
Z(\{3,4,5,6\}) &= \emptyset, \quad \mathcal{C}^{\boldsymbol{j}_\emptyset}_{\{3,4,5,6\}} \;=\; \{\{3,4,5,6\}\}.
\end{aligned}$$

The condition (13) is easily checked. The MLE is calculated if one decomposes the table into those for $j_3 = 1$ and $j_3 = 2$ and then calculates the MLE separately (Lemma 4). By Theorem 2 and Lemma 6, a Markov basis of this model is

$$\mathcal{B}_{\mathrm{Model2}} \;=\; \mathcal{B}^{i_3=1}_{\{1,2,5\},\{4,5,6\}} \cup \mathcal{B}_{\{1,2,3,5\},\{2,3,4,5,6\}} \cup \mathcal{B}_{\{1,2,3,4,5\},\{3,4,5,6\}},$$

where we put $\mathcal{B}^{i_3=1}_{\{1,2,5\},\{4,5,6\}} = \mathcal{B}_{\{1,2,5\},\{4,5,6\}} \cap L^{i_3=1}$.

We calculate the p-value of the deviance of Model 2 from Model 1 by the MCMC method. The number of steps in the MCMC procedure is 10^5. The result is as follows.

Table 2: Survey data concerning the attitudes of high-school students in New Jersey towards mathematics.

School		Suburban school				Urban school			
Sex		Female		Male		Female		Male	
Plans	Preference	Attend	Not	Attend	Not	Attend	Not	Attend	Not
College	Math-sciences								
	Agree	37	27	51	48	51	55	109	86
	Disagree	16	11	10	19	24	28	21	25
	Liberal arts								
	Agree	16	15	7	6	32	34	30	31
	Disagree	12	24	13	7	55	39	26	19
Job	Math-sciences								
	Agree	10	8	12	15	2	1	9	5
	Disagree	9	4	8	9	8	9	4	5
	Liberal arts								
	Agree	7	10	7	3	5	2	1	3
	Disagree	8	4	6	4	10	9	3	6

Source: Fowlkes et al. [1988]

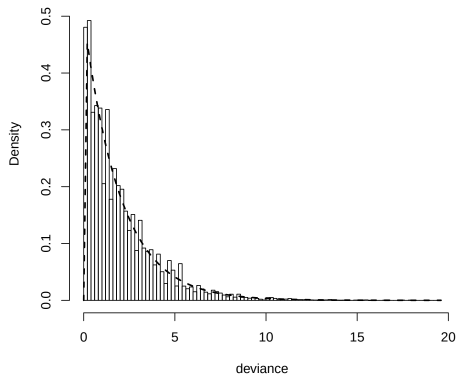


Figure 3: The empirical and asymptotic distributions of the deviance of Model 2 from Model 1.

Deviance	df	p-value (asymptotic)	p-value (MCMC)
1.851	2	0.396	0.399±0.012

The confidence interval of the p-value is computed on the basis of the batch-means method. The empirical distribution and asymptotic distribution of the deviance are given in Figure 3. Since the sample size of the data is large, the results of the asymptotic method and MCMC method are almost the same.

6 Concluding remarks

We proposed a hierarchical subspace model, by defining the notion of conformality of linear subspaces to a given hierarchical model. The notion of an HSM gives a modeling strategy of multiway tables and unifies various models of interaction effects in the literature. We illustrated practical advantage of our modeling strategy with some data sets.

In this paper we only considered log-affine model. Note that there are some nonlinear models of interaction terms for two-way tables, such as the RC association model. It seems clear that we can separately fit a nonlinear model to each maximal compact component of a hierarchical model, as long as the models for dividers are saturated. However conformality of a general nonlinear model with respect to a given hierarchical model has to be carefully defined and this is left to our future study.

The separation by dividers are closely related to the notion of collapsibility (e.g. Asmussen and Edwards [1983]) of hierarchical models. Localization of statistical inference to the marginal table of a maximal compact component seems to correspond to the collapsibility to the component. Also our results for Markov bases for HSMs are closely related to those of Sullivant [2007]. Sullivant [2007] is more concerned with Markov bases for models with latent variables and marginalization of latent variables. Collapsibility and marginalization properties of HSM require further investigation.

In the computation of the MLE for the hierarchical models, it is known that the algorithm can be localized into the marginal tables of maximal cliques for chordal extension of the simplicial complex associated with the model, which is smaller than maximal compact component (e.g. Badsberg and Malvestuto [2001]). By using the notion of ambient hierarchical model discussed in Section 2.3, it may be possible to localize the inference to smaller units than maximal extended compact component also in the HSMs.

Another important question on hierarchical subspace model is the necessity of saturation of the model for dividers. Saturation of the model for dividers is a sufficient condition for localization of statistical inference, but it may not be a necessary condition. There may exist some important models, for which statistical inferences can be localized to extended compact components without the requirement of saturation of dividers. This question also needs a careful investigation.

References

Satoshi Aoki and Akimichi Takemura. Markov chain Monte Carlo exact tests for incomplete two-way contingency table. *Journal of Statistical Computation and Simulation*, 75(10):787–812, 2005.

Søren Asmussen and David Edwards. Collapsibility and response variables in contingency tables. *Biometrika*, 70(3):567–578, 1983. ISSN 0006-3444.

J. H. Badsberg and F. M. Malvestuto. An implementaition of the iterative proportional fitting procecure by propagation trees. *Comput. Statist. Data. Anal.*, 37:297–322, 2001.

Yvonne M. M. Bishop, Stephen E. Fienberg, and Paul W. Holland. *Discrete multivariate analysis: theory and practice*. The MIT Press, Cambridge, Mass.-London, 1975. With the collaboration of Richard J. Light and Frederick Mosteller.

W. Craig. The song of the wood pewee. *Bull. N. Y. State Museum*, 334:1–186, 1943.

Persi Diaconis and Bernd Sturmfels. Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.*, 26(1):363–397, 1998. ISSN 0090-5364.

Adrian Dobra and Seth Sullivant. A divide-and-conquer algorithm for generating Markov bases of multi-way tables. *Comput. Statist.*, 19(3):347–366, 2004. ISSN 0943-4062.

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

Leo A. Goodman. Simple models for the analysis of association in cross-classifications having ordered categories. *J. Amer. Statist. Assoc.*, 74(367):537–552, 1979. ISSN 0003-1291.

Leo A. Goodman. The analysis of cross-classified data having ordered and/or unordered categories: association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Ann. Statist.*, 13(1):10–69, 1985. ISSN 0090-5364.

Hisayuki Hara, Akimichi Takemura, and Ruriko Yoshida. Markov bases for two-way subtable sum problems. *J. Pure Appl. Algebra*, 213(8):1507–1529, 2009. doi:10.1016/j.jpaa.2008.11.019.

Chihiro Hirotsu. Two-way change-point model and its application. *Australian Journal of Statistics*, 39(2):205–218, 1997.

Søren Højsgaard. Split models for contingency tables. *Comput. Statist. Data. Anal.*, 42: 621–645, 2003.

Søren Højsgaard. Statistical inference in context specific interaction models for contingency tables. *Scand. J. Statist.*, 31:143–158, 2004.

Satoshi Kuriki. Asymptotic distribution of inequality-restricted canonical correlation with application to tests for independence in ordered contingency tables. *J. Multivariate Anal.*, 94(2):420–449, 2005. ISSN 0047-259X.

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

F. M. Malvestuto and M. Moscarini. Decomposition of a hypergraph by partial-edge separators. *Theoret. Comput. Sci.*, 237:57–79, 2000.

Tomonari Sei, Satoshi Aoki, and Akimichi Takemura. Perturbation method for determining group of invariance of hierarchical models. arXiv:0808.2725v1, 2008. *Advances in Applied Mathematics*, doi:10.1016/j.aam.2009.02.005.

Seth Sullivant. Toric fiber products. *J. Algebra*, 316(2):560–577, 2007. ISSN 0021-8693.

Martin A. Tanner and Michael A. Young. Modeling agreement among raters. *J. Amer. Statist. Assoc.*, 80:175–180, 1985.

Sadao Tomizawa. Analysis of square contingency tables in statistics. In *Selected Papers on Probability and Statistics*, volume 227 of *Translations, Series 2*, pages 147–174. American Mathematical Society, Providence, Rhode Island, 2009.