Journal of Multivariate Analysis 103 (2012) 19–34

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Hierarchical subspace models for contingency tables

Hisayuki Hara [a,*], Tomonari Sei [b], Akimichi Takemura [c,d]

[a] Faculty of Economics, Niigata University, 8050 Ikarashi 2-no-cho, Nishi-ku, Niigata 950-2042, Japan
[b] Department of Mathematics, Faculty of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan
[c] Graduate School of Information Science and Technology, University of Tokyo, 7-3-1 Hongo Bunkyo-ku, Tokyo 113-8656, Japan
[d] CREST, JST, Japan

ARTICLE INFO

Article history:
Received 13 March 2010
Available online 16 June 2011

AMS 2000 subject classifications:
62H17
62H05

Keywords:
Context specific interaction model
Divider
Markov bases
Split model
Uniform association model

ABSTRACT

For the 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 the localization property of the inference in the hierarchical model. This approach also enables us to evaluate the localization property of a given log-affine model. We discuss estimation and exact tests of the proposed model and illustrate the advantage of the proposed modeling with some data sets.

1. Introduction

Modeling of the interaction term is an important topic for two-way contingency tables, because there is a large gap between the complete 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 the literature. In this paper, we establish a general mathematical framework for 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 [14,15] and the RC association model [14,15,22] 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. [27,28]). Hirotsu [19] proposed a two-way change point model and Hara et al. [17] generalized it to a subtable sum model. For multiway contingency tables, [20] considered the split model as a generalization of graphical models. The context specific interaction model defined by Højsgaard [21] is a more general model than the split model. In this article, we give a unified treatment of these models as submodels of hierarchical models and consider their extension to the models for higher-dimensional tables from the viewpoints of decomposition and conditional independence structure of the models.

Conditional independence structure of a log-affine model is described by a graph. Such a graph is called an independence graph. In a usual hierarchical model, the likelihood is factorized to submodels induced by each compact component [24] of the simplicial complex determining the model. By this factorization, statistical inference on a hierarchical model can be

* Corresponding author.
E-mail addresses: hara@tmi.t.u-tokyo.ac.jp, hara@econ.niigata-u.ac.jp (H. Hara), sei@stat.t.u-tokyo.ac.jp (T. Sei), takemura@stat.t.u-tokyo.ac.jp (A. Takemura).

0047-259X/$ – see front matter
doi:10.1016/j.jmva.2011.06.003

localized through the decomposition of the simplicial complex for the model. The possibility of localizing the inference of a given hierarchical model has been well studied by many authors (e.g. [16,13,24,3,23]).

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, we can consider the modeling for interaction effects of a given hierarchical model. In the modeling process, it is sometimes advantageous to preserve the conditional independence structure and localization property of the hierarchical model and to treat each marginal model corresponding to each compact component of the hierarchical model separately. The resulting model is a submodel of the hierarchical model. Throughout this paper we assume that the model is log-affine. When a log-affine model is a submodel of a given hierarchical model, the log-affine model has the same conditional independence structure as the hierarchical model. As we will discuss in Section 3, however, the log-affine model does not necessarily have the same localization property as the hierarchical model. Therefore the localization property of a given log-affine model is not trivial in general.

In this article we define a hierarchical subspace model by a log-affine model possessing the same localization property as a given hierarchical model and discuss the localization property of the log-affine model. As pointed out by referees, ideas similar to our hierarchical subspace model have been discussed in many contexts. Sociologists have been employing marginal modeling, where a few important marginals are first modeled and they are combined into a joint model. Dobra and Fienberg [10] presented maximum likelihood estimation and bounds for cell entries for reducible models and discuss generalizations to nongraphical log-affine models. By our formulation of the hierarchical subspace model we can discuss these models in a unified framework.

The organization of the paper is as follows. In Section 2 we give a brief review on log-affine models and we summarize some basic facts on graphs and hypergraphs. In Section 3 we define the hierarchical subspace model and discuss the localization of inference through the decomposition of the model. We show that for a given log-affine model there exists the smallest decomposable model possessing the same localization property of the inference. In Section 4 we study the split model in the framework of this paper. In Section 5 we present construction of Markov bases for conditional tests of our model based on the argument in [11] for the hierarchical model. In Section 6 we show some real data examples. Some concluding remarks are given in Section 7.

2. Definitions and notations

2.1. Log-affine model and hierarchical model for contingency tables

In this section we summarize basic definitions and notations of log-affine model and hierarchical model. We follow definitions and notations of [7,23].

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. A *log-affine model* $\mathcal{M}(L)$ specified by L is given by the class of probability functions satisfying $\log p(\cdot) \in L$, where $\log p(\cdot)$ denotes the vector $\{\log p(\boldsymbol{i}), \boldsymbol{i} \in \mathcal{I}\}$ (Chapter 4 of [23]). In the following we only consider linear subspaces of V containing the constant function 1.

Let D be a subset of $[m]$. $\boldsymbol{i}_D = \{i_j, j \in D\}$ is a D-marginal cell. $\mathcal{I}_D = \prod_{j \in D} [I_j]$ denotes the set of D-marginal cells. $p(\boldsymbol{i}_D)$ and $x(\boldsymbol{i}_D)$ denote the marginal probability of a probability distribution $p(\cdot)$ and the marginal frequency of a contingency table $\boldsymbol{x} = \{x(\boldsymbol{i}), \boldsymbol{i} \in \mathcal{I}\}$, respectively, that is,

$$p(\boldsymbol{i}_D) := \sum_{i_{[m]\setminus D} \in \mathcal{I}_{[m]\setminus D}} p(\boldsymbol{i}), \qquad x(\boldsymbol{i}_D) := \sum_{i_{[m]\setminus D} \in \mathcal{I}_{[m]\setminus D}} x(\boldsymbol{i}).$$

Define $n := \sum_{\boldsymbol{i} \in \mathcal{I}} x(\boldsymbol{i})$, which is the total frequency. Denote by $\hat{p}(\boldsymbol{i})$ and $\hat{p}(\boldsymbol{i}_D)$ the maximum likelihood estimator (MLE) of $p(\boldsymbol{i})$ and $p(\boldsymbol{i}_D)$, respectively. As in [7] or [23], let

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

denote the set of functions depending only on $\boldsymbol{i}_D$. F_D can be identified with $\mathbb{R}^{I_D}$, where $I_D = \prod_{h \in D} I_h$, and especially we note that $F_{[m]} = V$. For a subspace L of V and $D \subset [m]$, we say that D is *saturated* in L if $F_D \subset L$. Then we note the following proposition.

Proposition 1. *D is saturated in L if and only if the sufficient statistic for $\mathcal{M}(L)$ fixes all the D-marginals of the contingency table.*

Proof. The sufficient statistic for $\mathcal{M}(L)$ is usually described by taking a basis of L. Let $d = \dim L$ and take a basis $\phi_1, \ldots, \phi_d$ of L. Then a sufficient statistic for $\mathcal{M}(L)$ is given as $\{\sum_{\boldsymbol{i} \in \mathcal{I}} \phi_j(\boldsymbol{i}) x(\boldsymbol{i}), j = 1, \ldots, d\}$. However if we allow redundancy, we can define the sufficient statistic of L just by $\{\sum_{\boldsymbol{i} \in \mathcal{I}} \phi(\boldsymbol{i}) x(\boldsymbol{i}), \ \forall \phi(\cdot) \in L\}$. On the other hand the sufficient statistic for F_D is given by the set of D-marginal frequencies $\{x(\boldsymbol{i}_D), \boldsymbol{i}_D \in \mathcal{I}_D\}$, or equivalently by $\{\sum_{\boldsymbol{i} \in \mathcal{I}} \phi(\boldsymbol{i}) x(\boldsymbol{i}), \ \forall \phi(\cdot) \in F_D\}$ if we allow redundancy. Hence the sufficient statistic of L fixes all $x(\boldsymbol{i}_D)$ if and only if $F_D \subset L$. $\square$

Note that if D is saturated in L, then every $E \subset D$ is saturated in L because $F_E \subset F_D$.

Let Δ denote a simplicial complex on $[m]$ and let $\operatorname{red} \Delta$ denote the set of maximal elements, i.e. facets, of Δ (Chapter 2 of [23]). For a subset D, define the subcomplex $\Delta(D) := \{D \cap E \mid E \in \Delta\}$. The hierarchical model $\mathcal{M}(H_\Delta)$ associated with Δ

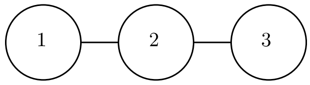


Fig. 1. Three-way conditional independence model.

is defined as

$$\log p(\cdot) \in H_\Delta := \sum_{D \in \mathrm{red}\,\Delta} F_D,$$

where the right-hand side is the summation of vector spaces. Noting that

$$H_\Delta = \left\{ \sum_{D \in \mathrm{red}\,\Delta} \phi_D(\cdot) \mid \phi_D(\cdot) \in F_D, D \in \mathrm{red}\,\Delta \right\},$$

we have $H_{\Delta \cap \Delta'} = H_\Delta \cap H_{\Delta'}$.

Let G_Δ be a graph with the vertex set $[m]$ and an edge between $v, v' \in [m]$ if and only if there exists $D \in \Delta$ such that $v, v' \in D$. Then G_Δ is called an *independence graph* of Δ [11]. G_Δ shows an conditional independence structure of $\mathcal{M}(H_\Delta)$, i.e., if two vertices v and v' are not adjacent each other, the corresponding variables are conditionally independent given the rest of variables. If red Δ is the set of maximal cliques of G_Δ, $\mathcal{M}(H_\Delta)$ is called a graphical model. When G_Δ is chordal, a graphical model H_Δ is called a decomposable model.

2.2. Basic facts on hypergraphs

We note that red Δ is considered as a hypergraph. Here we summarize some notions on hypergraphs according to [23,24].

A hypergraph is reduced if its edges are pairwise inclusion-incomparable sets. Hence red Δ is reduced. A subset of a hyperedge is called a *partial edge*. A subhypergraph of red Δ is a hypergraph whose edges are all partial edges of red Δ. A subhypergraph of red Δ induced by a non-empty subset E of $[m]$ is red $\Delta(E)$. We note that red $\Delta(E)$ is a reduced hypergraph whose edges are the maximal edges of the hypergraph $\{D \cap E \mid D \in \mathrm{red}\,\Delta\}$.

Two vertices v and v' are called adjacent in red Δ when they are also adjacent in G_Δ. Two vertices v and v' are connected if they are connected in G_Δ. A hypergraph is connected if every pair of two vertices is connected. A hypergraph is called disconnected if it is not connected.

A partial edge S is a separator of red Δ if the subhypergraph of red Δ induced by $[m] \setminus S$ is disconnected. For every partial edge separator, there exist three non-empty and disjoint subsets $\{A, B, S\}$, $A \cup B \cup S = [m]$ satisfying that red $\Delta(A)$ and red $\Delta(B)$ are disconnected. Then $\{A, B, S\}$ is called a decomposition of red Δ. For two vertices u and v, if there is a decomposition $\{A, B, S\}$ such that $u \in A$ and $v \in B$, we say S separates u and v. A partial edge separator S of red Δ 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 edges, u and v are called tightly connected. A subset $C \subset [m]$ is called a *compact component* if any two vertices in C are tightly connected. Denote the set of maximal compact components of red Δ 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 red Δ. We denote $\mathcal{S} = \{S_2, \ldots, S_{|\mathcal{C}|}\}$. $\mathcal{S}$ is a multiset in general. $\mathcal{C}$ is obtained by decomposing red Δ recursively by dividers.

By definition it is clear that v and v' are adjacent to each other in red Δ if and only if they are adjacent in G_Δ. Therefore red Δ also gives the conditional independence structure of the hierarchical model $\mathcal{M}(H_\Delta)$. The cell probability $p(\boldsymbol{i})$ of hierarchical model $\mathcal{M}(H_\Delta)$ is factorized as

$$p(\boldsymbol{i}) = \frac{\prod_{C \in \mathcal{C}} p(\boldsymbol{i}_C)}{\prod_{S \in \mathcal{S}} p(\boldsymbol{i}_S)}, \tag{1}$$

where the marginal models $p(\boldsymbol{i}_C)$ and $p(\boldsymbol{i}_S)$ are hierarchical models $\mathcal{M}(H_{\Delta(C)})$ and $\mathcal{M}(H_{\Delta(S)})$, respectively. Then the MLE is written as

$$\hat{p}(\boldsymbol{i}) = \frac{\prod_{C \in \mathcal{C}} \hat{p}(\boldsymbol{i}_C)}{\prod_{S \in \mathcal{S}} \hat{p}(\boldsymbol{i}_S)} = \frac{\prod_{C \in \mathcal{C}} \hat{p}(\boldsymbol{i}_C)}{\prod_{S \in \mathcal{S}} x(\boldsymbol{i}_S)/n}, \tag{2}$$

and the computation of the MLE is localized to the marginal model corresponding to each compact component and the localization corresponds to the decomposition of red Δ.

Example 1. Consider the decomposable graphical model for three-way contingency tables corresponding to the graph in Fig. 1. The model is described as

$$\log p(\boldsymbol{i}) = a(i_1, i_2) + b(i_2, i_3). \tag{3}$$

In this model $\Delta = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{2,3\}\}$ and red $\Delta = \{\{1,2\}, \{2,3\}\}$, respectively, and the corresponding linear subspace is $H_\Delta = F_{\{1,2\}} + F_{\{2,3\}}$. We note that $a(i_1, i_2)$'s and $b(i_2, i_3)$'s are free parameters. Since the model satisfies $i_1 \perp\!\!\!\perp i_3 \mid i_2$, $p(\boldsymbol{i})$ is written by

$$p(\boldsymbol{i}) = \frac{p(\boldsymbol{i}_{\{1,2\}})p(\boldsymbol{i}_{\{2,3\}})}{p(i_2)}. \tag{4}$$

The marginal models $p(\boldsymbol{i}_{\{1,2\}})$, $p(\boldsymbol{i}_{\{2,3\}})$ and $p(i_2)$ are saturated models corresponding to $F_{\{1,2\}}$, $F_{\{2,3\}}$ and $F_{\{2\}}$, respectively. Then the MLE of $p(\boldsymbol{i})$ is obtained by

$$\hat{p}(\boldsymbol{i}) = \frac{\hat{p}(\boldsymbol{i}_{\{1,2\}})\hat{p}(\boldsymbol{i}_{\{1,2\}})}{\hat{p}(i_2)} = \frac{x(\boldsymbol{i}_{\{1,2\}})x(\boldsymbol{i}_{\{2,3\}})}{nx(i_2)}, \tag{5}$$

where $\hat{p}(\boldsymbol{i}_{\{1,2\}})$, $\hat{p}(\boldsymbol{i}_{\{2,3\}})$ and $\hat{p}(i_2)$ are the MLE of $p(\boldsymbol{i}_{\{1,2\}})$, $p(\boldsymbol{i}_{\{2,3\}})$ and $p(i_2)$, respectively.

Now consider modeling of two-way interaction terms. Suppose that we have known functions $\phi(\boldsymbol{i}_{\{1,2\}})$ depending only on $\boldsymbol{i}_{\{1,2\}} = (i_1, i_2)$ and $\psi(\boldsymbol{i}_{\{2,3\}})$ depending only on $\boldsymbol{i}_{\{2,3\}} = (i_2, i_3)$. Separating main effects, consider the following submodel of (3),

$$\log p(\boldsymbol{i}) = \alpha(i_1) + \beta(i_2) + \gamma(i_3) + \delta\phi(\boldsymbol{i}_{\{1,2\}}) + \delta'\psi(\boldsymbol{i}_{\{2,3\}}). \tag{6}$$

The model (3) is still log-affine. Let L be the linear subspace corresponding to this model. Then L is a linear subspace of F_Δ.

The parameters of this model are $\{\alpha(i_1)\}_{i_1=1}^{I_1}$, $\{\beta(i_2)\}_{i_2=1}^{I_2}$, $\{\gamma(i_3)\}_{i_3=1}^{I_3}$ and δ, δ'. The uniform association model is specified by $\phi(\boldsymbol{i}_{\{1,2\}}) = i_1 i_2$. The change point model in [19] is specified by

$$\phi(\boldsymbol{i}_{\{1,2\}}) = \begin{cases} 1, & \text{if } i_1 \le I_1' \text{ and } i_2 \le I_2', \\ 0, & \text{otherwise,} \end{cases}$$

where $1 \le I_1' < I_1$, $1 \le I_2' < I_2$. Similarly we can specify $\psi(\boldsymbol{i}_{\{2,3\}})$ according to many well known models.

Since the model (6) is a submodel of the model (3), $i_1 \perp\!\!\!\perp i_3 \mid i_2$ still holds for (6) and $p(\boldsymbol{i})$ is written as (4), where we note that the marginal models $p(\boldsymbol{i}_{\{1,2\}})$ and $p(\boldsymbol{i}_{\{2,3\}})$ are written by

$$\log p(\boldsymbol{i}_{\{1,2\}}) = \alpha(i_1) + \beta(i_2) + \delta\phi(\boldsymbol{i}_{\{1,2\}}) \tag{7}$$

and

$$\log p(\boldsymbol{i}_{\{2,3\}}) = \beta(i_2) + \gamma(i_3) + \delta'\psi(\boldsymbol{i}_{\{2,3\}}), \tag{8}$$

respectively. Moreover, since $\{\beta(i_2)\}_{i_2=1}^{I_2}$ in (6) are free parameters, $F_2 \to F_{\{2\}}$ is saturated in L. Therefore the MLE of $p(\boldsymbol{i})$ is written by

$$\hat{p}(\boldsymbol{i}) = \frac{\hat{p}(\boldsymbol{i}_{\{1,2\}})\hat{p}(\boldsymbol{i}_{\{1,2\}})}{\hat{p}(i_2)} = \frac{\hat{p}(\boldsymbol{i}_{\{1,2\}})\hat{p}(\boldsymbol{i}_{\{2,3\}})}{x(i_2)/n}. \tag{9}$$

Therefore the maximum likelihood estimation of the model (6) is also localized to estimations of two marginal models in the same way as the hierarchical model (5).

Note that although we use the same notation for $\beta(i_2)$ in (6)–(8) for simplicity, they are different parameters (as functions of cell probabilities). If we distinguish them by $\beta(i_2)^{(123)}$, $\beta(i_2)^{(12)}$, $\beta(i_2)^{(23)}$ in (6), (7), (8), respectively, then they are connected as $\beta(i_2)^{(123)} = \beta(i_2)^{(12)} + \beta(i_2)^{(23)} - \log p(i_2)$. Accordingly, in view of (9), the maximum likelihood estimates are connected as $\hat{\beta}(i_2)^{(123)} = \hat{\beta}(i_2)^{(12)} + \hat{\beta}(i_2)^{(23)} - \log(x(i_2)/n)$.

When a log-affine model has the same localization property as a given hierarchical model as seen in this example, we call the model a hierarchical subspace model of the hierarchical model. Actually the model (6) is a hierarchical subspace model of (3). In the next section we give a precise definition of the hierarchical subspace model.

3. Hierarchical subspace models and their decompositions

3.1. Conformality of log-affine model

For defining our hierarchical subspace model, we introduce the notion of conformality of a hierarchical model. As an illustrating example, we again consider the three-way conditional independence model in Example 1. In (6) it is important to note that δ and δ' are free parameters. Now consider the following model imposing an additional constraint $H : \delta = \delta'$ on (6):

$$\log p(\boldsymbol{i}) = \alpha(i_1) + \beta(i_2) + \gamma(i_3) + \delta(\phi(\boldsymbol{i}_{\{1,2\}}) + \psi(\boldsymbol{i}_{\{2,3\}})). \tag{10}$$

This model is still log-affine and the conditional independence $i_1 \perp\!\!\!\perp i_3 \mid i_2$ holds. However, since δ is shared by two interaction terms for $\boldsymbol{i}_{\{1,2\}}$ and $\boldsymbol{i}_{\{2,3\}}$, both $x(\boldsymbol{i}_{\{1,2\}})$ and $x(\boldsymbol{i}_{\{2,3\}})$ are relevant for the estimation of the common value of δ. Therefore we cannot localize estimation of the parameters to two marginal tables. We now formulate the above notion of no restriction on parameters across maximal compact components by defining the notion of conformality of linear subspaces.

Definition 1. Let $W_1, \ldots, W_K$ be linear subspaces of V. 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 $W = W_1 + \cdots + W_K$. Note that if L is a subspace of W then 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. We note that H_Δ satisfies

$$H_\Delta = \sum_{C \in \mathcal{C}} H_\Delta \cap F_C \tag{11}$$

and therefore H_Δ is conformal to $\{F_C, C \in \mathcal{C}\}$.

Example 2. Consider the models (6) and (10) again. Let L and L' denote the corresponding subspaces of the models (6) and (10), respectively. Let $K = 2$ and let $W_1 := F_{\{1,2\}}$ and $W_2 := F_{\{2,3\}}$. In the case of the model (6),

$$L \cap W_1 = \{\alpha(i_1) + \beta(i_2) + \delta\phi(\boldsymbol{i}_{12})\}, \qquad L \cap W_2 = \{\beta(i_2) + \gamma(i_3) + \delta'\psi(\boldsymbol{i}_{23})\}.$$

Hence $L = (L \cap W_1) + (L \cap W_2)$ is conformal to two marginal spaces $\{F_{\{1,2\}}, F_{\{2,3\}}\}$. In the case of the model (10), however,

$$L' \cap W_1 = \{\alpha(i_1) + \beta(i_2)\}, \qquad L' \cap W_2 = \{\beta(i_2) + \gamma(i_3)\}.$$

Hence $(L' \cap W_1) + (L' \cap W_2) = \{\alpha(i_1) + \beta(i_2) + \gamma(i_3)\}$ and L' is not conformal to $\{F_{\{1,2\}}, F_{\{2,3\}}\}$.

3.2. Hierarchical subspace model

We now present the following definition of a hierarchical subspace model.

Definition 2. Let Δ be a simplicial complex and H_Δ be the subspace of the corresponding hierarchical model. Then the log-affine model $\mathcal{M}(L)$ for a subspace L is a hierarchical subspace model (HSM) of H_Δ if the following conditions hold:

1. Each divider $S \in \mathcal{S}$ of red Δ is saturated in L, i.e. $F_S \cap L = F_S$.
2. L is conformal to the set of subspaces $\{F_C, C \in \mathcal{C}\}$.

By condition 1 of HSM the conditional independence structure of H_Δ is preserved in L. Condition 2 together with condition 1 guarantees that the statistical inference is localized to each C.

On the computation of the MLE we can generalize (2) to HSM as follows.

Theorem 1. *The MLE $\hat{p}(\boldsymbol{i})$ of cell probabilities for HSM of H_Δ satisfies*

$$\hat{p}(\boldsymbol{i}) = \frac{\prod_{C \in \mathcal{C}} \hat{p}(\boldsymbol{i}_C)}{\prod_{S \in \mathcal{S}} \hat{p}(\boldsymbol{i}_S)} = \frac{\prod_{C \in \mathcal{C}} \hat{p}(\boldsymbol{i}_C)}{\prod_{S \in \mathcal{S}} x(\boldsymbol{i}_S)/n}, \tag{12}$$

where $\hat{p}(\boldsymbol{i}_C)$ coincides with the MLE of the model associated with the linear space $L \cap F_C$, which is computed only on the marginal table $x(\boldsymbol{i}_C)$.

Proof. By induction on the number of compact components $|\mathcal{C}|$ of red Δ, it is sufficient to consider the case $\mathcal{C} = \{C_1, C_2\}$ with $S = C_1 \cap C_2$. The MLE of the model $\mathcal{M}(L)$ 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 2 we write $\log p(\cdot) = \theta_{C_1} + \theta_{C_2}$ with $\theta_{C_1} \in L \cap F_{C_1}$ and $\theta_{C_2} \in L \cap F_{C_2}$. Since F_S is saturated both in $L \cap F_{C_1}$ and $L \cap F_{C_2}$, we can assume $\sum_{\boldsymbol{i}_{C_1 \setminus S}} e^{\theta_{C_1}(\boldsymbol{i}_{C_1})} = 1$ for each $\boldsymbol{i}_S$ without loss of generality. Hence the problem is decomposed into two parts: maximization of $\sum_{\boldsymbol{i}_{C_1}} x(\boldsymbol{i}_{C_1})\theta_{C_1}(\boldsymbol{i}_{C_1})$ subject to $\theta_{C_1} \in L \cap F_{C_1}$ and $\sum_{\boldsymbol{i}_{C_1 \setminus S}} e^{\theta_{C_1}(\boldsymbol{i}_{C_1})} = 1$, and maximization of $\sum_{\boldsymbol{i}_{C_2}} x(\boldsymbol{i}_{C_2})\theta_{C_2}(\boldsymbol{i}_{C_2})$ subject to $\theta_{C_2} \in L \cap F_{C_2}$ and $\sum_{\boldsymbol{i}_{C_2}} e^{\theta_{C_2}(\boldsymbol{i}_{C_2})} = 1$. Since the maximizer $\hat{\theta}_{C_1}$ does not depend on C_2, it is computed from the case $C_2 = S$. We have $\hat{\theta}_{C_1}(\boldsymbol{i}_{C_1}) = \log\{\hat{p}(\boldsymbol{i}_{C_1})/(x(\boldsymbol{i}_S)/n)\}$, where $\hat{p}(\boldsymbol{i}_{C_1})$ is the MLE of the model $\mathcal{M}(L \cap F_{C_1})$. □

This theorem shows that the computation of the MLE of an HSM of H_Δ is localized to each $C \in \mathcal{C}$. We note that Theorem 1 depends on Condition 1. Even if Condition 1 is not satisfied, the conditional independence structure of $\mathcal{M}(H_\Delta)$ is preserved. But $\hat{p}(\boldsymbol{i}_C)$ is not necessarily the MLE for the marginal model $\mathcal{M}(L \cap F_C)$.

Example 3. By following the argument in Example 2, we can easily show that the model (6) is an HSM of (3). On the other hand, since the model (10) is not conformal to $F_{\{1,2\}}$ and $F_{\{2,3\}}$, the model (10) is not an HSM of (3). Although the model (10) has the same conditional independence structure $i_1 \perp\!\!\!\perp i_3 \mid i_2$ depicted in the graph in Fig. 1, the inference is not localized in the same way as the decomposition of the graph.

As seen in this example, we note that even if a given log-affine model $\mathcal{M}(L)$ is a subset of a hierarchical model $\mathcal{M}(H_\Delta)$, the localization property of $\mathcal{M}(H_\Delta)$ is not necessarily preserved in L.

However we note that the model (10) is an HSM of the three-way saturated model. In the saturated model, red $\Delta = \mathcal{C} = [m]$ and there is no divider in red Δ. Therefore every log-affine model is an HSM of the saturated model. This also means that every log-affine model $\mathcal{M}(L)$ has a hierarchical model for which $\mathcal{M}(L)$ is an HSM.

3.3. Ambient decomposable model of a log-affine model

Suppose that a conditional independence structure of the model is given by a hypergraph red Δ. By following Definition 2, we can formulate an HSM of H_Δ by modeling interaction terms $L \cap F_D$, $D \in$ red Δ, under the conditions of conformality (11) and $F_S \subset L$, $S \in \mathcal{S}$. Then the resulting model preserves the same localization property as H_Δ.

Since every log-affine model $\mathcal{M}(L)$ has a hierarchical model for which $\mathcal{M}(L)$ is an HSM, a next natural question is to look for a small simplicial complex Δ such that $\mathcal{M}(L)$ is an HSM of H_Δ. As mentioned in Example 3, even if $L \subset H_\Delta$, the localization property of $\mathcal{M}(L)$ does not necessarily correspond to the decomposition of red Δ. Therefore the question is not trivial. We will show in Theorem 2 below that for each log-affine model $\mathcal{M}(L)$ there exists a natural smallest decomposable model $\mathcal{M}(H_{\mathcal{H}})$ with respect to inclusion relation, such that $\mathcal{M}(L)$ is an HSM of $H_{\mathcal{H}}$. Here $\mathcal{H}$ is the hypergraph corresponding to the decomposable model. We call such $\mathcal{M}(H_{\mathcal{H}})$ the *ambient decomposable model* of $\mathcal{M}(L)$. The notion of ambient decomposable model is also interpreted as a classification of log-affine models in terms of decomposition of the models.

In order to define the ambient decomposable model, we first introduce the notion of connectedness and decomposition of a subspace L separately from those of hypergraphs. L is called disconnected if there exists a non-empty proper subset A of $[m]$ such that L is conformal to $\{F_A, F_{A^C}\}$, where A^C denotes the complement of A in $[m]$. We call L *connected* if L is not disconnected. Now we note the following proposition.

Proposition 2. *When L is disconnected, the variables in A and the variables in A^C are independent.*

Proof. $L = (L \cap F_A) + (L \cap F_{A^C})$ means that $\mathcal{M}(L)$ is described as $\log p(\boldsymbol{i}) = \phi(\boldsymbol{i}_A) + \psi(\boldsymbol{i}_{A^C})$, where $\phi(\cdot) \in F_A$ and $\psi(\cdot) \in F_{A^C}$. Therefore A and A^C are independent. $\square$

Under this definition L can be decomposed into its connected components. By the above proposition, variables in different connected components are independent. Therefore they can be independently modeled in L and can be investigated separately. Therefore from now on we assume that L is connected.

We need to generalize the notion of partial edge separator of a hypergraph to our setting.

Definition 3. For a subspace L, a non-empty subset S of $[m]$ is called an L-separator 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 $\{F_{A_1 \cup S}, F_{A_2 \cup S}\}$.

Then we call the triple (A_1, A_2, S) a *decomposition* of L. When the subspace L has a L-separator, we call L *reducible*. A pair of vertices v and v' are called *tightly connected* in L if there does not exist a decomposition (A_1, A_2, S) of L such that $v \in A_1$ and $v' \in A_2$. 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 an *extended compact component* of L. We note that the notions of L-separator, tight connectivity in L and extended compact component for a hierarchical model $\mathcal{M}(H_\Delta)$ are exactly the same as the notions of partial edge separator, tight connectivity and compact component of the hypergraph red Δ.

The set of maximal extended compact components of L is also considered as a hypergraph and we denote it by $\mathcal{H}$. Denote by $H_{\mathcal{H}}$ the subspace of the hierarchical model induced by $\mathcal{H}$. Then we have the following theorem.

Theorem 2. *$\mathcal{M}(H_{\mathcal{H}})$ is the smallest decomposable model with respect to inclusion relation such that $\mathcal{M}(L)$ is an HSM of $H_{\mathcal{H}}$.*

The following corollary is obvious from (12).

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 2. Before we give the proof, we present some lemmas required to prove the theorem.

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

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

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

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. □

When there exists a chordal graph whose set of maximal clique is $\mathcal{H}$, $\mathcal{H}$ is called acyclic. By using Lemma 1, we can prove the following lemma in the same way as Theorem 5 in [24].

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 an L-separator 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 an L-separator $S' \subset S$ 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. □

Now we provide a proof of Theorem 2.

Proof of Theorem 2. It is obvious that $L \subset H_{\mathcal{H}}$. From Lemma 3, every divider $S \in \mathcal{S}$ of $\mathcal{H}$ is an L-separator 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 [23]). Let

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

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 an L-separator with the same decomposition. Hence L is conformal to $\{F_{B_{K-1}}, F_{C_K}\}$, i.e.

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

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

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

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

Suppose that there exists a smaller decomposable model associated with a subspace $F_{\mathcal{H}'} \subset H_{\mathcal{H}}$ for which $\mathcal{M}(L)$ 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. □

3.4. Hierarchical models containing a log-affine model

In Theorem 2 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 a smallest hierarchical model with respect to inclusion relation containing a log-affine model as an HSM. 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 (10) again. As seen in Example 3, (10) is a submodel of (3) but is not an HSM of (3). The difficulty lies in the fact that a hierarchical model containing L may have a partial edge separator which is not an L-separator.

Given a subspace L consider the subspace of hierarchical models H_Δ containing L : $\{H_\Delta \mid H_\Delta \supset L\}$. As mentioned in Section 2.1, $H_\Delta \cap H_{\Delta'} = H_{\Delta \cap \Delta'}$. It follows that there exists the smallest hierarchical model in $\{\mathcal{M}(H_\Delta) \mid H_\Delta \supset L\}$. We call the smallest hierarchical model containing L as *hierarchical closure* of L and denote the corresponding simplicial complex and the subspace by $\bar{\Delta}(L)$ and $H_{\bar{\Delta}(L)}$, respectively. Note that for both (6) and (10), the hierarchical closure is the three-way

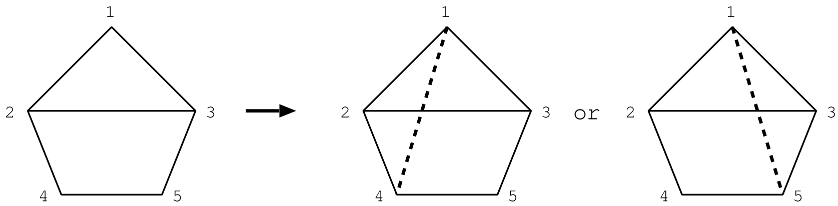


Fig. 2. Two ways to cross a divider of the hierarchical closure.

conditional independence model (3). We note that L does not necessarily satisfy the conformality with respect to the linear subspaces for red $\bar{\Delta}(L)$. We call $\mathcal{M}(L)$ a tight hierarchical subspace model if $\mathcal{M}(L)$ is an HSM of $H_{\bar{\Delta}(L)}$. If $\mathcal{M}(L)$ is a tight HSM, obviously $\bar{\Delta}(L)$ is the smallest simplicial complex such that $\mathcal{M}(L)$ is its HSM of $H_{\bar{\Delta}(L)}$.

We now present an example of a log-affine model L of a five-way contingency table, which has two minimal hierarchical models $\mathcal{M}(H_{\Delta_1})$, $\mathcal{M}(H_{\Delta_2})$, such that $\mathcal{M}(L)$ is an HSM of both of them. Consider the following model $\mathcal{M}(L)$ of five-way contingency tables:

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

where the main effects $\alpha_{\{j\}}$'s and θ are parameters and $\psi_{\{j,j'\}}$'s are fixed functions. The set of facets of $\bar{\Delta}(L)$ is given by

$$\mathrm{red}\,\bar{\Delta}(L) = \{\{1,2\},\{1,3\},\{2,3\},\{2,4\},\{3,5\},\{4,5\}\},$$

which has a divider $\{2,3\}$. On the other hand, since $\psi_{\{2,3\}}(\cdot)$ is a fixed function, $L\cap F_{\{2,3\}}$ is not saturated in L and hence $\{2,3\}$ is not an L-separator. Therefore $\mathcal{M}(L)$ is not an HSM of $H_{\bar{\Delta}(L)}$ and is not tight. Note that $\mathcal{M}(L)$ is an HSM of any H_{Δ}, such that H_{Δ} does not possess a partial edge separator and $L\subset H_{\Delta}$. As in Fig. 2 define

$$\mathrm{red}\,\Delta_1 = \mathrm{red}\,\bar{\Delta}(L)\cup\{\{1,4\}\},\qquad \mathrm{red}\,\Delta_2 = \mathrm{red}\,\bar{\Delta}(L)\cup\{\{1,5\}\}.$$

Then $\mathcal{M}(L)$ is an HSM of both H_{Δ_1} and H_{Δ_2}.

4. Split model as a hierarchical subspace model

In this section we give a brief review on the split model by Højsgaard [20]. We first define the context specific interaction (CSI) model [21]. 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 $F^{\boldsymbol{j}_B}$ of V in which only the $\boldsymbol{j}_B$-slice has nonzero components, that is,

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

If B is empty, we define $F^{\boldsymbol{j}_\emptyset}=V$ 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

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

The subspace $F_D^{\boldsymbol{j}_B}$ 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:

$$F_{D\cup B} = \sum_{\boldsymbol{j}_B\in\mathcal{I}_B} F_D^{\boldsymbol{j}_B}. \tag{13}$$

A context specific interaction (CSI) model is a direct sum of subspaces $F_D^{\boldsymbol{j}_B}$ 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 [20] 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 $\mathcal{M}(H_\Delta)$ 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 $N_C^{\boldsymbol{j}_{Z(C)}}\subset F_C^{\boldsymbol{j}_{Z(C)}}$ such that

$$\forall C'\in\mathcal{C}\setminus\{C\},\quad F_{C\cap C'}^{\boldsymbol{j}_{Z(C)}}\subset N_C^{\boldsymbol{j}_{Z(C)}}\subset F_C^{\boldsymbol{j}_{Z(C)}}. \tag{14}$$

Then a log-affine model $\mathcal{M}(L)$ is defined by

$$L = \sum_{C \in \mathcal{C}} N_C, \qquad N_C = \sum_{\boldsymbol{j}_{Z(C)} \in \mathcal{I}_{Z(C)}} N_C^{\boldsymbol{j}_{Z(C)}}. \tag{15}$$

We call $\mathcal{M}(L)$ *a split subspace model with root* $\mathcal{C}$ if L satisfies (14) and (15). The following proposition holds.

Proposition 3. *Let $\mathcal{M}(H_\Delta)$ be a decomposable model with the cliques $\mathcal{C}$. Then any split subspace model $\mathcal{M}(L)$ with root $\mathcal{C}$ is an HSM of H_Δ.*

Proof. First we prove that $F_S \subset L$ 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 (13) and (14), we have

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

Therefore $F_S \subset L$. Next, we prove that L is conformal to $\{F_C \mid C \in \mathcal{C}\}$. Note that $N_C^{\boldsymbol{j}_{Z(C)}} \subset F_C^{\boldsymbol{j}_{Z(C)}} \subset F_C$ for any $\boldsymbol{j}_{Z(C)}$ and we have $N_C \subset F_C$ for each $C \in \mathcal{C}$. Since N_C is also a subspace of L, we obtain $N_C \subset L \cap F_C$ and therefore $L = \sum_{C \in \mathcal{C}} N_C \subset \sum_{C \in \mathcal{C}} (L \cap F_C)$. 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 (15) with

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

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

$$\forall C' \in \mathcal{C} \setminus \{C\}, \quad \exists D \in \mathcal{C}_C^{\boldsymbol{j}_{Z(C)}} \text{ s.t. } (C \cap C') \setminus Z(C) \subset D \tag{16}$$

to assure the condition (14). Split models of degree greater than one are defined recursively. See [20] for details.

In Section 6, 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 associated with the subspace $\sum_\lambda N_\lambda$ 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 $\mathcal{M}(N_\lambda)$ with the total frequency $n_\lambda = \sum_{\boldsymbol{i} \in \mathcal{I}_\lambda} x(\boldsymbol{i})$.

5. Conditional tests of hierarchical subspace models via Markov bases

So far we have discussed the localization of the computation of the MLE for the log-affine model. In the hierarchical model, Dobra and Sullivant [11] showed that the computation of Markov bases is also localized to the computation of the Markov bases of the marginal model corresponding to each maximal compact component. In this section we generalize the argument to an HSM.

In this section we first give a brief review on Markov bases and conditional tests based on Markov basis methodology [8]. Next we generalize the argument of [11] to the HSM.

5.1. Markov basis and conditional test

Let $\boldsymbol{b}$ be the set of sufficient statistics for $\mathcal{M}(L)$. We assume that the elements of $\boldsymbol{b}$ are integer combinations of the frequencies $x(\boldsymbol{i})$. For a hierarchical model $\mathcal{M}(H_\Delta)$, $\boldsymbol{b}$ is written by

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

We consider $\boldsymbol{b}$ as a column vector with dimension ν.

We order the elements of a contingency table $\boldsymbol{x}$ lexicographically 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 $\mathcal{M}(L)$.

The conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{b}$ is exactly a hypergeometric distribution. Usually the goodness of fit of the model is assessed by large sample approximation. However when the sample size is not large, it is desirable to use conditional tests based on the exact distribution of test statistics. 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*. If we can enumerate all the elements of the fiber which $\boldsymbol{x}$ belongs to, we can evaluate the null distribution of a test statistic exactly based on the conditional hypergeometric distribution of $\boldsymbol{x}$. However since the number of elements of fibers is too large in general, it is difficult to evaluate the null distribution of a test statistic by the enumeration of elements of 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 $A\boldsymbol{z} = 0$. A move is expressed as a difference of its positive part and negative part $\boldsymbol{z} = \boldsymbol{z}^+ - \boldsymbol{z}^-$, where $\boldsymbol{z}^+$ and $\boldsymbol{z}^-$ are two contingency tables in the same fiber. We denote a move $\boldsymbol{z}$

$$\boldsymbol{z} = [\{\boldsymbol{i}_1, \ldots, \boldsymbol{i}_d\} \parallel \{\boldsymbol{i}'_1, \ldots, \boldsymbol{i}'_d\}], \tag{17}$$

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 positive elements of $\boldsymbol{z}^-$. d is the sample size of $\boldsymbol{z}^+$ (or $\boldsymbol{z}^-$) and is called a degree of $\boldsymbol{z}$.

Example 4. Consider a 3×3 common diagonal effect model discussed in [18],

$$\log p(\boldsymbol{i}) = \alpha(i_1) + \beta(i_2) + \delta\phi(\boldsymbol{i}), \tag{18}$$

where

$$\phi(\boldsymbol{i}) = \begin{cases} 1 & i_1 = i_2, \\ 0 & \text{otherwise}. \end{cases} \tag{19}$$

The sufficient statistic $\boldsymbol{b}$ of this model is the set of row sums, column sums and the diagonal sum,

$$\boldsymbol{b} = \left\{ x(i_1), i_1 \in \{1, 2, 3\},\ x(i_2), i_2 \in \{1, 2, 3\},\ \sum_{i_1=1}^{3} x(i_1 i_1) \right\}.$$

Then an integer array

$$\boldsymbol{z} := \begin{array}{cc} & i_2 \\ i_1 & \begin{array}{|c|c|c|}\hline 0 & 1 & -1 \\ \hline -1 & 0 & 1 \\ \hline 1 & -1 & 0 \\ \hline \end{array} \end{array} = \begin{array}{|c|c|c|}\hline 0 & 1 & 0 \\ \hline 0 & 0 & 1 \\ \hline 1 & 0 & 0 \\ \hline \end{array} - \begin{array}{|c|c|c|}\hline 0 & 0 & 1 \\ \hline 1 & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline \end{array} \tag{20}$$

is a degree three move of the model (18). Actually we easily see that row sums, column sums and diagonal sums of $\boldsymbol{z}$ are all zeros. By following the notation in (17), $\boldsymbol{z}$ is written as

$$\boldsymbol{z} = [\{(1, 2), (2, 3), (3, 1)\} \parallel \{(3, 2), (1, 3), (2, 1)\}]. \tag{21}$$

For this model only one move $\boldsymbol{z}$ forms a Markov basis [18].

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 $\mathcal{M}$. If we have a Markov basis, we can generate a Markov chain of contingency tables from any fiber whose stationary distribution is the conditional hypergeometric distribution [8]. In this way Markov basis methodology enables us to evaluate a test statistics based on the exact distribution.

Dobra [9] showed that the decomposable model has a Markov basis consisting of only degree two moves. Markov bases for some other log-affine model have been discussed in [17,18,26] etc. In general, however it is not easy to obtain an exact list of Markov basis for the log-affine model, even for the hierarchical model. In hierarchical model [11] developed an algorithm to compute a Markov basis recursively from Markov bases of the maximal prime submodels corresponding to maximal compact components. In the next section we generalize the result to the HSM.

5.2. Local computation of Markov basis of HSM

Most of the arguments and the notations in this section follow those in [11]. For a subset $D \subset [m]$, denote $L(D) := L \cap F_D$. 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 $\{F_{V_1}, F_{V_2}\}$, we note that $\mathcal{M}(L(V_1))$ and $\mathcal{M}(L(V_2))$ 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 $\mathcal{M}(L(V_1))$ and $\mathcal{M}(L(V_2))$, 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 $\mathcal{M}(L)$ 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

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

Assume that $\mathcal{B}(V_1)$ and $\mathcal{B}(V_2)$ are Markov bases for $\mathcal{M}(L(V_1))$ and $\mathcal{M}(L(V_2))$, 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, the sufficient statistic $\boldsymbol{b}$ fixes $x(\boldsymbol{i}_S)$. Hence we have

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

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

$$\begin{aligned}\boldsymbol{z}_1 &= [\{(\boldsymbol{i}_{A_1}^1, \boldsymbol{i}_S^1), \dots, (\boldsymbol{i}_{A_1}^d, \boldsymbol{i}_S^d)\} \parallel \{(\boldsymbol{j}_{A_1}^1, \boldsymbol{j}_S^1), \dots, (\boldsymbol{j}_{A_1}^d, \boldsymbol{j}_S^d)\}], \\ \boldsymbol{z}_2 &= [\{(\boldsymbol{i}_S^1, \boldsymbol{i}_{A_2}^1), \dots, (\boldsymbol{i}_S^d, \boldsymbol{i}_{A_2}^d)\} \parallel \{(\boldsymbol{j}_S^1, \boldsymbol{j}_{A_2}^1), \dots, (\boldsymbol{j}_S^d, \boldsymbol{j}_{A_2}^d)\}],\end{aligned} \tag{22}$$

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

Definition 4 (*Dobra and Sullivant [11]*). Define $\boldsymbol{z}_1 \in \mathcal{B}(V_1)$ as in (22). Let $\eta := \{\boldsymbol{i}_{A_2}^1, \dots, \boldsymbol{i}_{A_2}^d\} \in \mathcal{I}_{A_2} \times \cdots \times \mathcal{I}_{A_2}$. Define $\boldsymbol{z}_1^{\boldsymbol{k}}$ by

$$\boldsymbol{z}_1^\eta := [\{(\boldsymbol{i}_{A_1}^1, \boldsymbol{i}_S^1, \boldsymbol{i}_{A_2}^1), \dots, (\boldsymbol{i}_{A_1}^d, \boldsymbol{i}_S^d, \boldsymbol{i}_{A_2}^d)\} \parallel \{(\boldsymbol{j}_{A_1}^1, \boldsymbol{j}_S^1, \boldsymbol{i}_{A_2}^1), \dots, (\boldsymbol{j}_{A_1}^d, \boldsymbol{j}_S^d, \boldsymbol{i}_{A_2}^d)\}].$$

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

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

In the same way as Lemma 5.4 in [11] we can obtain the following lemma.

Lemma 5. *Suppose that $z_1 \in \mathcal{B}(V_1)$ as in (22). 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

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

where

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

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 4, $z_{V_2}(\boldsymbol{i}_{V_2}) = 0$ for all $\boldsymbol{i}_{V_2} \in \mathcal{I}_{V_2}$. Hence $A\boldsymbol{z} = 0$. □

Example 5. Consider a $3 \times 3 \times 3$ model in the class (6),

$$\log p(\boldsymbol{i}) = \alpha(i_1) + \beta(i_2) + \gamma(i_3) + \delta\phi(\boldsymbol{i}_{\{1,2\}}) + \delta'\phi(\boldsymbol{i}_{\{2,3\}}), \tag{23}$$

where $\phi(\cdot)$ is defined as in (19). The sufficient statistic for this model is the set of one-dimensional marginals $x(i_k)$, $i_k \in \mathcal{I}_k$, $k = 1, 2, 3$ and two-dimensional diagonal sums $\sum_{\boldsymbol{i}:i_1=i_2} x(\boldsymbol{i})$, $\sum_{\boldsymbol{i}:i_2=i_3} x(\boldsymbol{i})$.

As discussed in Example 3, this model is an HSM of (3). Hence we can set $V_1 = \{1, 2\}$ and $V_2 = \{2, 3\}$ and $L(V_i) = L \cap F_{V_i}$, $i = 1, 2$, are both 3×3 common diagonal effect models (18).

Let $\boldsymbol{z}_1 := \boldsymbol{z}$ in (21). As mentioned in Example 4, $\boldsymbol{z}_1$ forms a Markov basis for the model (18), that is, $\mathcal{B}(V_1) = \{\boldsymbol{z}_1\}$. We see that $\boldsymbol{z}_1$ is written in the form (22). Let $\eta := (i_3, i_3', i_3'')$. Then $\boldsymbol{z}_1^\eta$ is written by

$$\boldsymbol{z}_1^\eta = [\{(1, 2, i_3), (2, 3, i_3'), (3, 1, i_3'')\} \parallel \{(3, 2, i_3), (1, 3, i_3'), (2, 1, i_3'')\}].$$

When $\eta = (1, 2, 3)$, $\boldsymbol{z}_1^\eta$ is written in array expression as in (20) by

$$\boldsymbol{z}_1^\eta = \begin{array}{c}i_3 = 1\\ \begin{array}{|c|c|c|}\hline 0 & 1 & 0\\ \hline 0 & 0 & 0\\ \hline 0 & -1 & 0\\ \hline\end{array}\end{array}, \quad \begin{array}{c}i_3 = 2\\ \begin{array}{|c|c|c|}\hline 0 & 0 & -1\\ \hline 0 & 0 & 1\\ \hline 0 & 0 & 0\\ \hline\end{array}\end{array}, \quad \begin{array}{c}i_3 = 3\\ \begin{array}{|c|c|c|}\hline 0 & 0 & 0\\ \hline -1 & 0 & 0\\ \hline 1 & 0 & 0\\ \hline\end{array}\end{array}.$$

Table 1
Triples of phrases in a song sequence of a wood pewee, with repeats deleted.
Source: Craig [6].

First place	Second place	Third 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	–	–	–	–

We easily see that one-dimensional marginals and two-dimensional diagonal sums of $\boldsymbol{z}_1^\eta$ are all zeros and hence that $\boldsymbol{z}_1^\eta$ is a move for (23). $\mathrm{Ext}(\mathcal{B}(V_1) \to L)$ is

$$\mathrm{Ext}(\mathcal{B}(V_1) \to L) = \{\boldsymbol{z}_1^\eta \mid i_3, i_3', i_3'' \in \{1, 2, 3\}\}.$$

Consider a decomposable model $\mathcal{M}(H_\Delta)$ such that $\mathrm{red}\,\Delta = \{V_1, V_2\}$. Dobra [9] showed that the set of all degree two moves

$$\boldsymbol{z}_{V_1,V_2} = [\{(\boldsymbol{i}_{A_1}^1, \boldsymbol{i}_S^1, \boldsymbol{i}_{A_2}^1), (\boldsymbol{i}_{A_1}^2, \boldsymbol{i}_S^2, \boldsymbol{i}_{A_2}^2)\} \parallel \{(\boldsymbol{i}_{A_1}^1, \boldsymbol{i}_S^1, \boldsymbol{i}_{A_2}^2), (\boldsymbol{i}_{A_1}^2, \boldsymbol{i}_S^2, \boldsymbol{i}_{A_2}^1)\}],$$

where $\boldsymbol{i}_{A_1}^k \in \mathcal{I}_{A_1}$, $\boldsymbol{i}_S^d \in \mathcal{I}_S$ and $\boldsymbol{i}_{A_2}^k \in \mathcal{I}_{A_2}$ for $k = 1, 2$, forms a Markov basis and denote it by $\mathcal{B}_{V_1,V_2}$.

Theorem 3. *Let $\mathcal{B}(V_1)$ and $\mathcal{B}(V_2)$ be Markov bases for $\mathcal{M}(L(V_1))$ and $\mathcal{M}(L(V_2))$, respectively. Then*

$$\mathcal{B} := \mathrm{Ext}(\mathcal{B}(V_1) \to L) \cup \mathrm{Ext}(\mathcal{B}(V_2) \to L) \cup \mathcal{B}_{V_1,V_2} \tag{24}$$

is a Markov basis for $\mathcal{M}(L)$.

We can prove the theorem in the same way as Theorem 5.6 in [11]. Suppose that $\mathcal{M}(L)$ is an HSM of $H_{\mathcal{H}}$. Then Theorem 3 implies that a Markov basis for L is obtained from $\mathcal{B}(C)$, $C \in \mathcal{H}$, by recursively using (24). This shows that the computation of a Markov basis can be localized according to submodels corresponding to maximal extended compact components of L.

Concerning Markov bases of the split model of Section 4 we state the following lemma.

Lemma 6. *With the same notation as in Lemma 4, a Markov basis of the model associated with the subspace $\sum_\lambda N_\lambda$ is given by union of Markov bases of $\mathcal{M}(N_\lambda)$.*

6. Examples

In this section we give several applications of conditional tests of HSMs by using Markov bases. In Section 6.1 we discuss conditional tests for models of multiway tables with structural zeros. In Section 6.2 we present an example of a split model. The models in this section are relatively small and intended to illustrate the notions of this paper, rather than being examples of large scale data analyses.

6.1. *Conditional tests for models with structural zeros*

Table 1 is the data on song sequence of a wood pewee in Section 7.5.2 of [4]. 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. As discussed in [5], we consider this sequence as a Markov chain. The main interest is the order of the chain. As an example of conditional tests for the model with structural zeros, we consider the goodness-of-fit test of two Markov chain models of first order for this data. Aoki and Takemura [1] provided a complete description of Markov basis for the quasi-independence model for two-way tables and proposed conditional test by using the Markov basis. However its extension to the model for multiway tables has not yet been studied.

First we consider the model discussed by Bishop et al. [4] for this data,

$$p_{ijk} = 1_{\{i\neq j\}}\mathrm{e}^{a_{ij}}1_{\{j\neq k\}}\mathrm{e}^{b_{jk}}, \tag{25}$$

where a_{ij} and b_{jk} are free parameters. With some abuse of notation (25) can be written as

$$\log p_{ijk} = a_{ij}1_{\{i\neq j\}} + (-\infty)1_{\{i=j\}} + b_{jk}1_{\{j\neq k\}} + (-\infty)1_{\{j=k\}}. \tag{26}$$

We note that this model is also in the class (6). 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 $V = \mathbb{R}^{4\times 4\times 4}$. However we can replace V by $R^{|\bar{\mathcal{I}}|}$, where

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

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 $\mathcal{M}(F_{\mathrm{Model1}})$, where

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

which corresponds to (25). 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 $\mathcal{B}_{\mathrm{Model1}} = \mathcal{B}_{\{1,2\},\{2,3\}}$ (see Theorem 3 for the notation). An experimental result that compares the saturated model and Model 1 is given in Fig. 3. 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

$$F_{\mathrm{model2}} = \{\alpha_i + \beta_j + \gamma_k + \phi_i 1_{\{i=j\}} + \psi_j 1_{\{j=k\}}\}.$$

$\mathcal{M}(F_{\mathrm{model2}})$ is an HSM of $F_{\{1,2\}} + F_{\{2,3\}}$. 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 two-way quasi-independence models with the diagonal structural zeros, that is,

$$\hat{p}^{(1)}_{ij} = \mathrm{e}^{\hat{\alpha}_i}\mathrm{e}^{\hat{\beta}_j}1_{\{i\neq j\}}, \qquad \hat{p}^{(1)}_{i+} = x_{i++}/n, \qquad \hat{p}^{(1)}_{+j} = x_{+j+}/n,$$
$$\hat{p}^{(2)}_{jk} = \mathrm{e}^{\hat{\beta}'_j}\mathrm{e}^{\hat{\gamma}_k}1_{\{j\neq k\}}, \qquad \hat{p}^{(2)}_{j+} = x_{+j+}/n, \qquad \hat{p}^{(2)}_{+k} = x_{++k}/n,$$

where $\hat{\beta}_j$ and $\hat{\beta}'_j$ are different in general as discussed in Example 1. They are computed by the iterative proportional fitting method. By Theorem 3, a Markov basis is given by

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

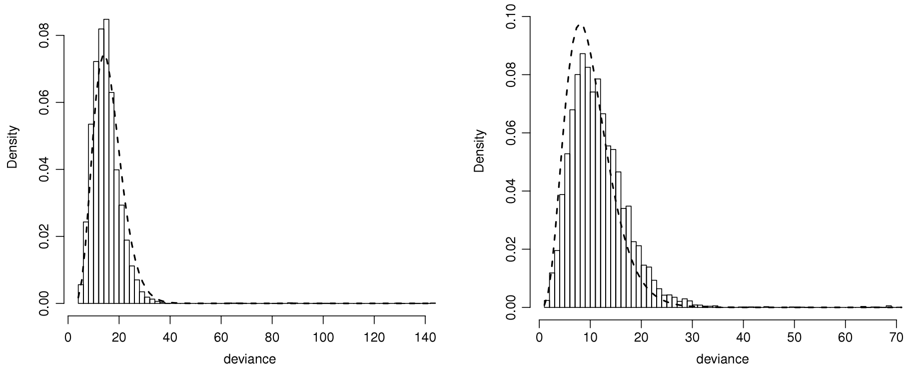
where $\mathcal{B}(\{1,2\})$ and $\mathcal{B}(\{2,3\})$ are the Markov bases of the two-way quasi-independence model with structural zeros obtained by Aoki and Takemura [1]. An experimental result that compares the Model 1 and Model 2 is given in Fig. 3. These results show that we can conclude the chain is at least of second order.

In this way we can perform conditional test for the models of multiway tables with structural zeros.

6.2. Conditional test for the split model

In this section we give an example of conditional test of the split model. Here we deal with a real data called *women and mathematics (wam) data* used in [20]. 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 [20] treated. The first model is a decomposable model $\mathcal{M}(F_{\mathrm{model1}})$

$$F_{\mathrm{Model1}} = F_{\{1,2,3,5\}} + F_{\{2,3,4,5\}} + F_{\{3,4,5,6\}}.$$



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

Fig. 3. 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.

Table 2
Survey data concerning the attitudes of high-school students in New Jersey toward mathematics.
Source: Fowlkes et al. [12].

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

By Theorem 3, a Markov basis of this model is given by

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

The second model is a split model $\mathcal{M}(F_{\text{model2}})$

$$F_{\text{Model2}} = F_{\{1,2,3,5\}} + F_{\{2,5\}}^{j_3=1} + F_{\{4,5\}}^{j_3=1} + F_{\{2,4,5\}}^{j_3=2} + F_{\{3,4,5,6\}}.$$

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

$$\mathcal{C} = \{\{1,2,3,5\},\{2,3,4,5\},\{3,4,5,6\}\},$$
$$Z(\{1,2,3,5\}) = \emptyset, \qquad \mathcal{C}_{\{1,2,3,5\}}^{\boldsymbol{j}_\emptyset} = \{\{1,2,3,5\}\},$$
$$Z(\{2,3,4,5\}) = \{3\}, \qquad \mathcal{C}_{\{2,3,4,5\}}^{j_3=1} = \{\{2,5\},\{4,5\}\}, \qquad \mathcal{C}_{\{2,3,4,5\}}^{j_3=2} = \{\{2,4,5\}\},$$
$$Z(\{3,4,5,6\}) = \emptyset, \qquad \mathcal{C}_{\{3,4,5,6\}}^{\boldsymbol{j}_\emptyset} = \{\{3,4,5,6\}\}.$$

The condition (16) 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 3 and Lemma 6, a Markov basis of this model is

$$\mathcal{B}_{\text{Model2}} = \mathcal{B}_{\{1,2,5\},\{4,5,6\}}^{j_3=1} \cup \mathcal{B}_{\{1,2,3,5\},\{2,3,4,5,6\}} \cup \mathcal{B}_{\{1,2,3,4,5\},\{3,4,5,6\}},$$

where we put $\mathcal{B}_{\{1,2,5\},\{4,5,6\}}^{j_3=1} = \mathcal{B}_{\{1,2,5\},\{4,5,6\}} \cap F^{j_3=1}$.

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.

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

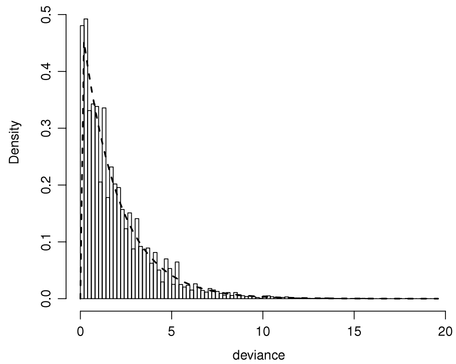


Fig. 4. The empirical and asymptotic distributions of the deviance of Model 2 from Model 1.

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 Fig. 4. In this way we can perform conditional test for the split model.

7. 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 our modeling strategy with some data sets. As a referee pointed out, our approach is novel in the sense that the localization properties are described not only by means of graph-theoretical criteria but also using the properties of the linear subspaces encoding these models.

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. [2]) of hierarchical models. Localization of statistical inference to the marginal table of a maximal extended compact component seems to correspond to the collapsibility to the component. Also Theorem 1 suggests the effectiveness of using mixed parameterization for contingency tables, i.e., we fit log-linear models for maximal extended compact components and connect them by marginal probabilities as in (12). Furthermore our results for Markov bases for HSMs are closely related to those of [25]. Sullivant [25] 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. [3]). By using the notion of ambient hierarchical model discussed in Section 3.4, 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.

Acknowledgments

The authors are grateful to the three anonymous referees for constructive and detailed comments.

References

[1] S. Aoki, A. Takemura, Markov Chain Monte Carlo exact tests for incomplete two-way contingency table, J. Stat. Comput. Simul. 75 (10) (2005) 787–812.
[2] S. Asmussen, D. Edwards, Collapsibility and response variables in contingency tables, Biometrika (ISSN: 0006-3444) 70 (3) (1983) 567–578.
[3] J.H. Badsberg, F.M. Malvestuto, An implementaition of the iterative proportional fitting procecure by propagation trees, Comput. Statist. Data. Anal. 37 (2001) 297–322.
[4] Y.M.M. Bishop, S.E. Fienberg, P.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.
[5] C. Chatfield, R.E. Lemon, Analysing sequences of behavioral events, J. Theoret. Biol. 29 (1970) 427–445.
[6] W. Craig, The song of the wood pewee, Bull. NY State Museum 334 (1943) 1–186.

[7] J.N. Darroch, T.P. Speed, Additive and multiplicative models and interactions, Ann. Statist. 11 (2009) 724–738.
[8] P. Diaconis, B. Sturmfels, Algebraic algorithms for sampling from conditional distributions, Ann. Statist. (ISSN: 0090-5364) 26 (1) (1998) 363–397.
[9] A. Dobra, Markov bases for decomposable graphical models, Bernoulli (ISSN: 1350-7265) 9 (6) (2003) 1093–1108.
[10] A. Dobra, S.E. Fienberg, Bounds for cell entries in contingency tables given marginal totals and decomposable graphs, Proc. Natl. Acad. Sci. USA (ISSN: 1091-6490) 97 (22) (2000) 11885–11892. doi:10.1073/pnas.97.22.11885.
[11] A. Dobra, S. Sullivant, A divide-and-conquer algorithm for generating Markov bases of multi-way tables, Comput. Statist. (ISSN: 0943-4062) 19 (3) (2004) 347–366.
[12] E.B. Fowlkes, A.E. Freeny, J.M. Landwehr, Evaluating logistic models for large contingency tables, J. Amer. Statist. Assoc. 83 (1988) 611–622.
[13] Z. Geng, Decomposability and collapsibility for log-linear models, Appl. Statist. 38 (1989) 189–197.
[14] L.A. Goodman, Simple models for the analysis of association in cross-classifications having ordered categories, J. Amer. Statist. Assoc. (ISSN: 0003-1291) 74 (367) (1979) 537–552.
[15] L.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. (ISSN: 0090-5364) 13 (1) (1985) 10–69.
[16] S. Haberman, The Analysis of Frequency Data, University of Chicago Press, Chicago, Illinoi, 1974.
[17] H. Hara, A. Takemura, R. Yoshida, Markov bases for two-way subtable sum problems, J. Pure Appl. Algebra 213 (8) (2009) 1507–1529. doi:10.1016/j.jpaa.2008.11.019.
[18] H. Hara, A. Takemura, R. Yoshida, A Markov basis for conditional test of common diagonal effect in quasi-independence model for square contingency tables, Comput. Stat. Data Anal. 53 (2009) 1006–1014.
[19] C. Hirotsu, Two-way change-point model and its application, Austr. J. Stat. 39 (2) (1997) 205–218.
[20] S. Højsgaard, Split models for contingency tables, Comput. Statist. Data. Anal. 42 (2003) 621–645.
[21] S. Højsgaard, Statistical inference in context specific interaction models for contingency tables, Scand. J. Statist. 31 (2004) 143–158.
[22] S. Kuriki, Asymptotic distribution of inequality-restricted canonical correlation with application to tests for independence in ordered contingency tables, J. Multivariate Anal. (ISSN: 0047-259X) 94 (2) (2005) 420–449.
[23] S.L. Lauritzen, Graphical Models, Oxford University Press, Oxford, 1996.
[24] F.M. Malvestuto, M. Moscarini, Decomposition of a hypergraph by partial-edge separators, Theoret. Comput. Sci. 237 (2000) 57–79.
[25] S. Sullivant, Toric fiber products, J. Algebra (ISSN: 0021-8693) 316 (2) (2007) 560–577.
[26] A. Takemura, H. Hara, Markov chain Monte Carlo test of toric homogeneous Markov chains, 2010. arXiv:1004.3599v1.
[27] M.A. Tanner, M.A. Young, Modeling agreement among raters, J. Amer. Statist. Assoc. 80 (1985) 175–180.
[28] S. Tomizawa, Analysis of square contingency tables in statistics, in: Selected Papers on Probability and Statistics, in: Translations, Series 2, vol. 227, American Mathematical Society, Providence, Rhode Island, 2009, pp. 147–174.