

分割表のグラフィカルモデルとその周辺

竹村 彰通

項目

1. 多元分割表解析の問題点
2. 多元分割表の記法
3. 対数線形モデルの階層モデルと部分モデル
4. グラフィカルモデル
5. 分解可能モデル

多元分割表解析の問題点

- 本稿の目的: 多元分割表の階層モデルについて基本的事項を整理する。
 - 現状で Lauritzen の教科書を除いてあまり一般的に書いていない。
- 多元分割表: 元数が大きくなると急速に難しくなる。
2元, 3元, ..., 8元, ..., 20元, ..., 300元, ...
 - 総セル数が指数的に増大
 - 可能なモデル数がさらに速く増大

- 例: 2 岐選択問題 20 問に回答すると 2^{20} 型の 20 元表 .
セル数は 100 万 . ほとんどのセルは観測ゼロとなる .
一方可能なモデルの総数は数えられないくらい多い .
- 連続な観測値については “ $p > n$ ” 問題なども考えられている . しかし分割表で考えると , 次元 p は元数に対応するので , $p > n$ のようなケースは扱いにくい .

- 以下では対数線形モデル (≡ 階層モデル) について述べる
- ただし対数線形モデルにも問題点がある
 - ⇒ 周辺分布をとることに閉じていない
 - 通常の周辺分布 (cf. シンプソンのパラドックス)
 - カテゴリーの併合
- もし対数線形モデルの混合をうまく扱えればかなり一般性がある。

(過分散の観点や, ベイズ推測の観点からの議論はかなりなされている.)

多元分割表の記法

- $\Delta = \{1, \dots, m\}$: 変数の集合
- $\delta \in \Delta$: 個々の変数
- $\mathcal{I}_\delta = \{1, \dots, I_\delta\}$: δ の水準の集合
- セルの集合

$$\mathcal{I} = \prod_{\delta \in \Delta} \mathcal{I}_\delta \quad (\text{直積})$$

- $i = (i_1, \dots, i_m)$: 個々のセル

- $a, b, \dots \subset \Delta$: 変数の部分集合
- a -周辺セル

$$i_a = (i_\delta)_{\delta \in a} \in \mathcal{I}_a = \prod_{\delta \in a} \mathcal{I}_\delta.$$

\mathcal{I}_a は a -周辺セルの集合 .

- $x(i)$ あるいは $n(i)$: セル i の頻度
- $p(i)$: セル i の生起確率
- $x(i_a)$: 周辺頻度 , $p(i_a)$: 周辺確率

- “ a -周辺のみに依存する関数”
 - 各周辺セル $i_a \in \mathcal{I}_a$ に実数を対応させる関数 $\psi : \mathcal{I}_a \mapsto \mathbb{R}$ を (a を明示して) ψ_a と書く .
 - さらに 引数を i に拡張して $\psi_a(i) \stackrel{\text{def}}{=} \psi_a(i_a)$ と書く . (i の関数だが i_a のみに依存すると言っても同じ .)
 - 例: 2 元分割表の独立モデル

$$\log p(i, j) = \mu_{\emptyset}(i, j) + \alpha_{\{1\}}(i, j) + \beta_{\{2\}}(i, j)$$

- 「 a -周辺のみに依存する関数」の集合は線形空間となっていることに注意
- $b \subset a$ とする時, b -周辺のみに依存する関数は a -周辺のみに依存する関数の特殊な場合である
- すなわち a -周辺のみに依存する関数の集合は, b -周辺のみに依存する関数の集合をふくむ.

対数線形モデルの階層モデルと部分モデル

- 階層モデルの定義
- \mathcal{A} : Δ の部分集合の族
 - 例: 3変数交互作用のない3元分割表のモデル:

$$\mathcal{A} = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$$

- \mathcal{A} に対する階層モデル:

$$\log p(i) = \sum_{a \in \mathcal{A}} \mu_a(i) \quad (1)$$

- 指数型分布族で, $\{\mu_a\}_{a \in \mathcal{A}}$ が自然母数であり, 自然母数について線形な部分モデル
- 階層モデルの (完備) 十分統計量 T は, \mathcal{A} に属する周辺和の全体

$$T = \{x(i_a) \mid i_a \in \mathcal{I}_a, a \in \mathcal{A}\}$$

こちらが期待値母数に対応

- $b \subset a \in \mathcal{A}$ とすると, (1) 式の右辺には $\mu_b(i)$ の項が自動的に含まれていると考える
- そこで, \mathcal{A} には次の性質を要求することとする.

$$b \subset a, a \in \mathcal{A} \Rightarrow b \in \mathcal{A} \quad (2)$$

\Rightarrow 「抽象的単体的複体」(abstract simplicial complex)

[各 $\delta \in \Delta$ について $\{\delta\} \in \mathcal{A}$ を要求することもある。「主効果は必ず含む」ことに対応.]

- 階層モデルの研究は数学的には抽象的単体的複体の研究と (水準数の考察等を除いて) 同等
- \mathcal{A} の中で包含関係の意味で極大なもののみを残して考えてもよい
- 記法: $\text{red } \mathcal{A}$
- $\text{red } \mathcal{A}$ の要素間には包含関係がない . **Antichain**, clutter, Sperner system.
- 階層モデルの文脈では , $\text{red } \mathcal{A}$ を生成集合 (族) とよぶことが多い (generating class) .

補集合で考える

- \mathcal{A} には含まれない相互作用項 $a \subset \Delta$, $a \notin \mathcal{A}$ の全体を考える .
- すなわち , 2^Δ における \mathcal{A} の補集合 \mathcal{A}^C . (2^Δ は Δ の部分集合の全体 . 巾集合 .)
- \mathcal{A}^C : セル確率に何の仮定もおかない飽和モデルから排除する相互作用項を指定
- (2) 式は次と同値

$$b \supset a, a \in \mathcal{A}^C \Rightarrow b \in \mathcal{A}^C \quad (3)$$

- \mathcal{A}^C の包含関係の意味での極小な集合の族，すなわち階層モデルに含まれない極小な相互作用の族，が重要な意味を持つ．これも $\text{red } \mathcal{A}^C$ と書く．

階層モデルの数 \doteq antichain の数 = デデキント数

(主効果のいくつかが無いモデルも含めた数)

Table 1: デデキント数

2	3	4	5	6	7	8
4	18	166	7579	7828352	2414682040996	56130437228687557907786

- $m = 9$ の正確な Dedekind 数は神のみぞ知る .
- デデキント数の漸近的評価は $2^{\binom{m}{\lfloor m/2 \rfloor}}$ とされる .
- 階層モデルの部分モデルを考えることが重要:

分解可能モデル \subset グラフィカルモデル \subset 階層モデル

Table 2: グラフィカルモデルと分解可能モデルの個数

m	グラフィカル	分解可能 (同型判定後)
2	2	2 (2)
3	8	8 (4)
4	64	61 (10)
5	1024	820 (27)
6	32768 ($= 2^{15}$)	18154 (96)
7	2097152 ($= 2^{21}$)	617675 (469)
8	268435456 ($= 2^{28}$)	30888596 (3734)

グラフィカルモデル

- 階層モデルにおいて生成集合 $\text{red } \mathcal{A}$ があるグラフ G の極大クリークの族となっているモデル
- クリーク: 互いに辺 (あるいは枝) によって結ばれた頂点の集合
- 統計のグラフィカルモデルでは単にクリークと言うと極大クリークをさすことが多い.
- \mathcal{A}^C に属する極小な集合による特徴づけ

補題 1 \mathcal{A} がグラフィカルモデルであるための必要十分条件は $\text{red } \mathcal{A}^C$ が Δ の 2 要素集合のみからなることである .

[証明は予稿にあり . やさしいがいい練習問題 .]

- 例: 無 3 因子交互作用モデル

$$\begin{aligned}\text{red } \mathcal{A} &= \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}, \\ \text{red } \mathcal{A}^C &= \{\{1, 2, 3\}\}\end{aligned}$$

独立グラフ: 必ずしもグラフィカルとは限らないモデルに関して考える .

- $\{p(i)\}_{i \in \mathcal{I}}$: 確率分布
- $\{p(i)\}_{i \in \mathcal{I}}$ の「独立グラフ」 G
 - δ, δ' 間に辺が無い \Leftrightarrow 「 δ, δ' 以外のすべての変数の値を所与とした時に δ, δ' が条件つき独立になる」
- 一般の階層モデル \mathcal{A} に対しては , その独立グラフ $G = G(\mathcal{A})$ において δ と δ' の間に辺があることと , ある $a \in \text{red } \mathcal{A}$ が存在して $\{\delta, \delta'\} \subset a$ となることが同値 .

- $G(\mathcal{A})$ に対応するグラフィカルモデルは、階層モデル \mathcal{A} を含むグラフィカルモデルのうち最小のモデルである。
- $\mathcal{A} \mapsto G(\mathcal{A})$ は多対1写像
 - 例: \mathcal{A} : 3元表の無3因子交互作用モデルの時、 $G(\mathcal{A})$ は飽和モデル。
 - 各グラフィカルモデル G には、それを制約した階層モデルの集合が張りついでいて、ファイバー構造をなしている。

- 単体的複体の用語を用いれば，各ファイバーは 1-skeleton を共有する単体的複体の族．
(1-skeleton とは 2 要素集合の集合「骨格」)
- Hammersley-Clifford の定理はファイバーに関するもの？

[Hammersley-Clifford の定理: グラフの分解 (あるいは分離) の観点からグラフィカルモデルの条件つき独立性が完全に記述される (必要十分) .]

分解可能モデル

- 分解可能モデルは，グラフィカルモデルの部分モデルであり，グラフ G がコーダルグラフの場合
- G がコーダルとは，長さ 4 以上の閉路には途中の頂点間を結ぶ「弦」が必ず存在することを言う．
“triangulated” とも言う．
- コーダルグラフは性質の良いグラフであり，統計のみならずさまざまな分野に現れる．
- ここでは階層モデルの分解という観点から分解可能モデルを考える (原尚幸．研究会資料．2007 年 6 月) ．

- 分解可能モデルは最近ではグラフィカルモデルの部分モデルととらえることが多いが，歴史的には分解可能モデルの概念のほうが先に定義された．

定義 1 (Haberman の本) 階層モデル \mathcal{A} が分解可能であるとは， $\text{red } \mathcal{A}$ が一つの集合からなるか，あるいは二つの分解可能モデル $\mathcal{A}_1, \mathcal{A}_2$ が存在して， $\text{red } \mathcal{A} = \text{red } \mathcal{A}_1 \cup \text{red } \mathcal{A}_2$, $\text{red } \mathcal{A}_1 \cap \text{red } \mathcal{A}_2 = \emptyset$ ，と分割され，かつ $a \in \text{red } \mathcal{A}_1, b \in \text{red } \mathcal{A}_2$ が存在して，

$$\left[\bigcup_{a' \in \mathcal{A}_1} a' \right] \cap \left[\bigcup_{b' \in \mathcal{A}_2} b' \right] = a \cap b$$

となることである．

- 定義中の $a \cap b$ は単体的複体を「左右に分離」する感じになっている。
- コーダルグラフに関しては，定義中の $a \cap b$ は minimal vertex separator とよばれるものとなる。
 - Minimal vertex separator とは，二つの頂点を分離するような頂点の集合 (関所の集合) の中で，包含の意味で極小な集合を言う。
 - グラフがコーダルグラフであるための必要十分条件として，任意の minimal vertex separator S が complete (すなわち $S \in \mathcal{A}$) であることが古典的な事実として知られている。

- また $\text{red } \mathcal{A}$ の要素はコーダルグラフ G の極大クリークの族である .
- コーダルグラフの構造は , 極大クリークの集合 $\mathcal{C} = \mathcal{A}$ と , “minimal vertex separator” の集合 \mathcal{S} によって完全に指定される .

- ただし S の各要素には重複度 (正整数) が付随している．ラフに言えば，重複度とは「 G を何個に分解するか」に対応している．
- そこで S を “multiset” とし，各要素が重複度の回数だけ含まれるものと定義する．
- コーダルグラフの極大クリークの集合を \mathcal{A} とする時， $\text{red } \mathcal{A}^C$ の特徴づけが気になる．統計的に有意な特徴づけはないように思われるが，代数学の観点からは Froberg による特徴づけが知られている．

- 統計的推測の観点からは，分解可能モデルは様々な良い性質を持っている．

- 最尤推定量の明示的表現:

$$\hat{p}^{\text{ML}}(i) = \begin{cases} \frac{1}{n} \frac{\prod_{c \in \mathcal{C}} x(i_c)}{\prod_{s \in \mathcal{S}} x(i_s)}, & \text{if } x(i_c) > 0, \forall c \in \mathcal{C}, \\ 0, & \text{otherwise,} \end{cases}$$

- 十分統計量を与えた時の分割表の条件つき分布

$$\begin{aligned} & p(\{x(i)\}_{i \in \mathcal{I}} \mid \{x(i_a)\}_{i_a \in \mathcal{I}_a}, a \in \mathcal{A}) \\ &= \frac{\prod_{c \in \mathcal{C}} x(i_c)!}{n! \prod_{i \in \mathcal{I}} x(i)! \times \prod_{s \in \mathcal{S}} x(i_s)!} \end{aligned}$$

- 自明な不偏推定量:

$$\tilde{p}(i) = \frac{x(i)}{n}$$

- Rao-Blackwell の定理の適用

$$\hat{p}^{\text{ML}}(i) = E[\tilde{p}(i) \mid \{x(i_a)\}_{i_a \in \mathcal{I}_a, a \in \mathcal{A}}]$$

となることが見える .

- すなわち分解可能モデルにおいては MLE は UMVU でもある (十分統計量の完備性より) .

- 分解可能モデル以外での最尤推定値は Lauritzen の教科書によれば，母数空間の境界を含めば常に一意に存在し，比例反復法 (IPS, IPF) によって数値的に求めることができる．(IPS とは周辺和を順次あわせるように比例的にセル頻度を調整していく方法．)
- 一方で，上の形の Rao-Blackwell の定理の適用は分解可能モデル以外でも可能である．十分統計量の完備性も成り立つ．MLE と UMVU の関係は？
- 推定値が境界に落ちてしまうようなデータの特徴づけは代数的な問題で，やさしくないようである．

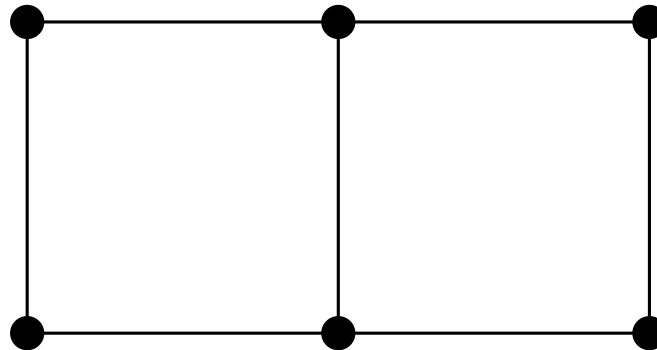
- 定義 1 の分解が最後まで進んで最終的に極大クリークまで分解されるのが分解可能モデル .
- しかし , 最終的に極大クリークまで分解されなくても , 分解自体は統計的推測にとって基本的な重要性を持つ .
- そこで以下の定義を与える .

定義 2 階層モデル \mathcal{A} が $s \in \mathcal{A}$ により分解されるとは、二つの階層モデル $\mathcal{A}_1, \mathcal{A}_2$ が存在して、 $\text{red } \mathcal{A} = \text{red } \mathcal{A}_1 \cup \text{red } \mathcal{A}_2, \text{red } \mathcal{A}_1 \cap \text{red } \mathcal{A}_2 = \emptyset,$ と分割され、かつ $a \in \text{red } \mathcal{A}_1, b \in \text{red } \mathcal{A}_2$ が存在して

$$s = a \cap b, \quad \left[\bigcup_{a' \in \mathcal{A}_1} a' \right] \cap \left[\bigcup_{b' \in \mathcal{A}_2} b' \right] = s$$

を満たすことである。

- 定義 2 を満たす s を “divider” と呼ぶ (cf. Malvestuto and Moscarini).

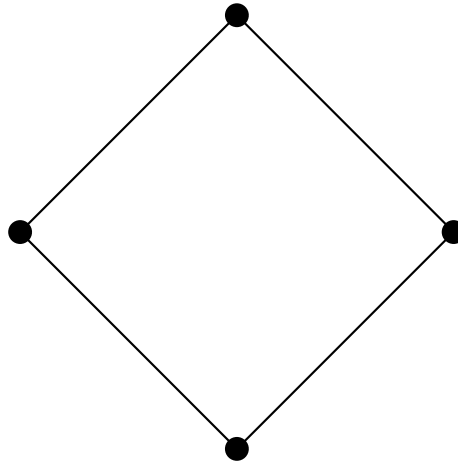


- \mathcal{A} 自体が分解可能モデルである場合には, divider の定義は minimal vertex separator の定義と同等
- 一般に, divider を持たない \mathcal{A} を “compact” とよぶ. (あまりいい用語とは思えない.)

- 統計的には, s が divider であれば, (s 以外の) \mathcal{A}_1 に属する変数と \mathcal{A}_2 に属する変数は条件つき独立になる.
- ただし divider としては s が \mathcal{A} に属することを要求していることに注意.
- 例:4 cycle model

$$\mathcal{A} = \{\{1, 2\}, \{2, 3\}, \{3, 4\}, \{4, 1\}\}$$

においては, $\{2, 4\}$ を与えた時に 1 と 3 は条件つき独立であるが, $\{2, 4\} \notin \mathcal{A}$ であるから $\{2, 4\}$ は divider ではない.



- グラフの場合

- A がグラフ G に対応する場合には, divider であることと, クリークをなす minimal vertex separator であることが同値.
- compact は prime graph とよばれ, 極大部分 compact は maximal prime subgraph とよばれる.

- Divider の基本的な重要性

- 定義 2 を再帰的に適用して \mathcal{A} を分解していくと、適用の順序にかかわらず分解は一意に定まる。
- 分解の結果は \mathcal{A} の極大な部分 compact の族となる。
- この分解の操作を “compaction” とよぶ。
- 極大部分 compact 間の関係は、コーダルグラフにおける極大クリーク間関係と全く同様である。

- すなわち 極大部分 compact の perfect sequence や, 極大部分 compact 間を結ぶ junction tree などが, コーダルグラフの場合と全く同様に定義される.
- 統計的観点からは 極大部分 compact ごとに推定や検定の手続きを分解することができる.
 - 最尤推定においては各極大部分 compact ごとの最尤推定を, 分解可能モデルの MLE に対応する形で組み合わせることによって, モデル全体の最尤推定値が得られる.

- モデルの適合度検定においても，尤度比が compaction に対応する形で分解される．
- また正確検定をおこなうためのマルコフ基底やグレブナー基底に関しても，各極大部分 compact ごとのマルコフ基底やグレブナー基底を組合せて，モデル全体のマルコフ基底やグレブナー基底を構成することができる．
- このように compaction は階層モデルの推測に基本的な重要性を持つが， compaction 自体がまだあまり知られていないために，階層モデルの推測のどの段階で compaction を考えるべきについてはあまり議論がなされていない．

時間があまった時のための余談

- H.Wynn 氏も最近「グラフでなく単体的複体で考えるべきだ」と語っていた．
- 本稿を書くうちに，自分のその感じを強くした．その意味では，事後的にタイトルには不満が残った．
- 一方で，単体的複体まで考えなくても， $\text{red } A$ の要素の積集合全体からなる intersection poset の構造のみから定まる部分も多いのではないかという感じがする．
- 例えば，自由度の計算などは，包除原理を用いておこなうが，包除原理の適用は本質的には intersection

poset のメビウス関数を扱っていることにあたる．

- 分解可能モデルは intersection poset の構造が非常に特殊であるように思われる．例えば分解可能モデルの自由度の計算は，クリークの自由度の和から，minimal vertex separator の自由度の和を引くだけで求まってしまい，包除原理の観点からすると2項目までである．