

分割表のモデルと計算代数統計*

竹村 彰通 (東大情報理工)

*本講演は解説であり，研究内容や文献等のサーベイはあまり含みません．

項目

1. 分割表とは
2. 分割表の確率モデル
(2元独立モデル, 3元条件つき独立モデルとシンプソンのパラドックス等)
3. 有限標本空間の指数型分布族と toric model
4. 多元分割表の問題点, 記法
5. 対数線形モデルの階層モデルと部分モデル
6. グラフィカルモデル
7. 分解可能モデル

分割表とは

Table 1: あるクラスの数学演習の成績

幾何 \ 統計	5	4	3	2	1	計
5	2	1	1	0	0	4
4	8	3	3	0	0	14
3	0	2	1	1	1	5
2	0	0	0	1	1	2
1	0	0	0	0	1	1
計	10	6	5	2	3	26

- 分割表: 有限個の値のみをとる複数の確率変数の頻度を表にしたもの .
- 前ページの例は「2元の 5×5 の分割表」と言う .
- 成績の各々の組み合わせ (i, j) を「セル」とよぶ .
- 幾何学単独あるいは統計学単独の成績の分布は , 行和 (行計) あるいは列和 (列計) として示されている . これを周辺頻度とよぶ . x_{i+} などと書く .
- 3科目以上であれば「多元配列」となる . 以下では多元の場合を一般に考察したい .
- セルの集合が直積集合となっていることが特徴 .

分割表の確率モデル

- 基本的な例: 2元分割表の独立モデル ($I \times J$)

$$p_{ij} = p_{i+} \times p_{+j} \quad (\text{周辺確率の積と見る})$$
$$= \alpha_i \times \beta_j \quad (\text{単に積の形に書けていると見る})$$

$$i = 1, \dots, I, \quad j = 1, \dots, J.$$

- 伝統的には対数をとって $\log p_{ij} = \log \alpha_i + \log \beta_j$ の形に書き「対数線形モデル」と呼ぶ。
- 計算代数統計では $p_{ij} = \alpha_i \times \beta_j$ をそのまま「単項式」と見る (“toric model”)

単なる視点の違いだが、違いは結構大きい。

次の例: 3元分割表の条件つき独立モデル

- p_{ijk} : 同時確率
- j 所与のもとでの i の条件つき確率

$$p_{i|j} = \frac{p_{ij+}}{p_{+j+}}$$

- j 所与のもとでの (i, k) の条件つき確率

$$p_{ik|j} = \frac{p_{ijk}}{p_{+j+}}$$

- 条件つき独立モデル “ $i \perp k \mid j$ ”

$$p_{ik|j} = p_{i|j} \times p_{k|j}$$

- これを同値変形すると

$$\begin{aligned} p_{ijk} &= \frac{p_{ij+} p_{+jk}}{p_{+j+}} \\ &= \alpha_{ij} \times \beta_{jk} \end{aligned}$$

- 条件つき独立モデルとシンプソンのパラドックス

例: 二つの学部別, 男女別の入試の合格者 (架空の例)

学部 A

	合格	不	計
男	54	36	90
女	6	4	10
計	60	40	100

学部 B

	合格	不	計
男	3	7	10
女	27	63	90
計	30	70	100

学部の区別をなくして, 二つの表の数字を足すと

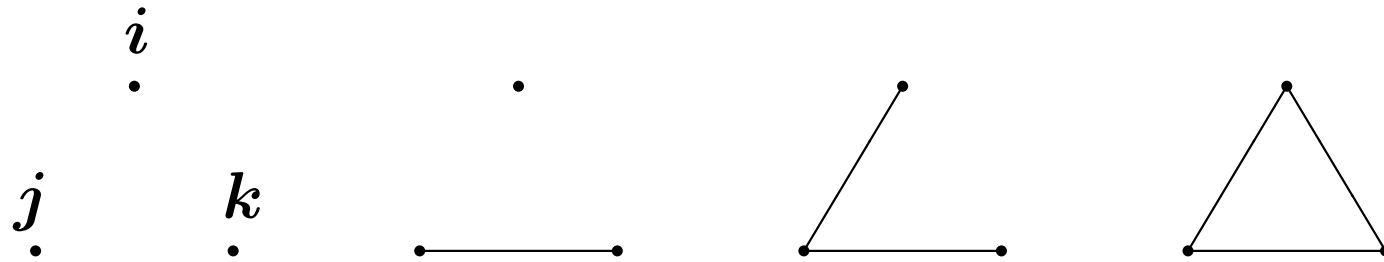
2 学部計

	合格	不	計
男	57	43	100
女	33	67	100
計	90	110	200

- 学部ごとには男女の合格率は全く同じなのに，学部の区別を無くすと男子の合格率が高くなっている．
- 理由：男子がやさしい学部を多く受けた
- 3 元表で条件つき独立モデルが成り立っても，2 元表に周辺化すると独立でなくなることもある．

3元分割表のその他のモデル

Figure 1: グラフとの対応で考える



左から

$$p_{ijk} = \alpha_i \beta_j \gamma_k, \quad p_{ijk} = \alpha_i \beta_{jk}, \quad p_{ijk} = \alpha_{ij} \beta_{jk}$$

- ただし一番右の三角形には二つの場合が考えられる .

$$p_{ijk} : \text{制限なし} \quad \text{or} \quad p_{ijk} = \alpha_{ij}\beta_{jk}\gamma_{ik}$$

- simplicial complex として中身がつまっているかが ,
グラフの表示だけではわからない .
- “graphical model” と呼ぶ時は中身はすべて詰める .
- $p_{ijk} = \alpha_{ij}\beta_{jk}\gamma_{ik}$ は「無三因子交互作用モデル」とよばれ , toric ideal の観点からは非常に興味深い

有限標本空間の指数型分布族と toric model

- 分割表では, セルの集合が直積集合となっているが, ここでは単なる有限集合 $\Omega = \{\omega_1, \dots, \omega_N\}$ とする.
- $p_j, j = 1, \dots, N,$ がそれぞれの点の確率とする.
- $p = (p_1, \dots, p_N)$ は R^N の単体 $S = \{p \mid p_j \geq 0, \sum_j p_j = 1\}$ 上にのっている.

- 不定元の集合 β_1, \dots, β_L によって各 p_j が monomial

$$p_j = \beta_1^{a_{j1}} \dots \beta_L^{a_{jL}}$$

と表されるようなモデルを toric model という。
(a_{jl} は所与の非負整数)。

- 伝統的には対数をとって

$$\log p_j = a_{j1}\theta_1 + \dots + a_{jL}\theta_L, \quad \theta_l = \log \beta_l$$

あるいは

$$p_j = \exp(a_{j1}\theta_1 + \dots + a_{jL}\theta_L)$$

の形に表し, 対数線形モデルという。

- より一般には「指数型分布族」と言う。

- 多項式環の準同型

$$\pi : k[p_1, \dots, p_N] \rightarrow k[\beta_1, \dots, \beta_L]$$

$$\pi : p_j \mapsto \beta_1^{a_{j1}} \dots \beta_L^{a_{jL}}$$

の kernel が toric ideal .

- Toric ideal の生成系は「マルコフ基底」(Diaconis and Sturmfels) とよばれ , toric model の検定に本質的な役割を果たす .

- 指数型分布族に慣れた人には $a_{jl} = T_l(j)$, さらには $j \rightarrow x$ と記法を変えて

$$p(x) = \exp (T_1(x)\theta_1 + \cdots + T_L(x)\theta_L)$$

と書けば見やすい . 十分統計量 $(T_1(x), \dots, T_L(x))$ が整数ベクトルの場合が toric model.

多元分割表解析の問題点，記法

以下では，一般の多元分割表のモデルについて考える^a．ここでの目的は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ことにある．

- 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

^aここからは6月の応用統計学会での講演の再利用になるので，話が速くなります．

- 可能なモデル数がさらに速く増大 (階層モデルであれば二重指数的)

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 $i_a = (i_\delta)_{\delta \in a} \in \mathcal{I}_a = \prod_{\delta \in a} \mathcal{I}_\delta$.
- $x(i)$ 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 $i_a \in \mathcal{I}_a$ に実数に対応させる関数 $\psi : \mathcal{I}_a \mapsto \mathbb{R}$ を (a を明示して) ψ_a と書く .
 - 引数を i に拡張して $\psi_a(i) \stackrel{\text{def}}{=} \psi_a(i_a)$ と書く .

- 例: 2元分割表の独立モデル $\log p_{ij} = \alpha_i + \beta_j$ を

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

と書く .

- 「 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: 無 3 因子交互作用モデル:

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

- \mathcal{A} に対する階層モデル:

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

- $b \subset a \in \mathcal{A}$ 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) .

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 2: 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 数は困難 .
- 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 3: 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。

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 .

- $\mathcal{A} \mapsto G(\mathcal{A})$ 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」)

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。
- 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。

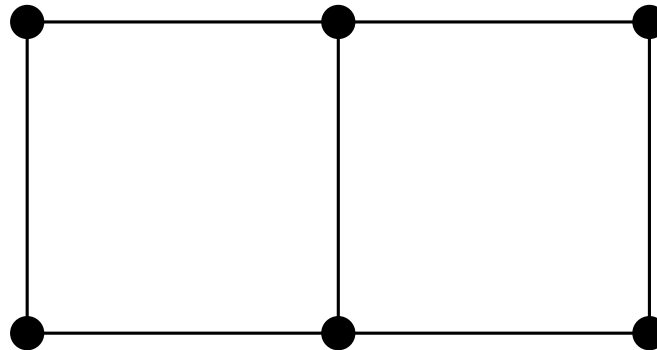
- そこで以下の定義を与える .

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

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

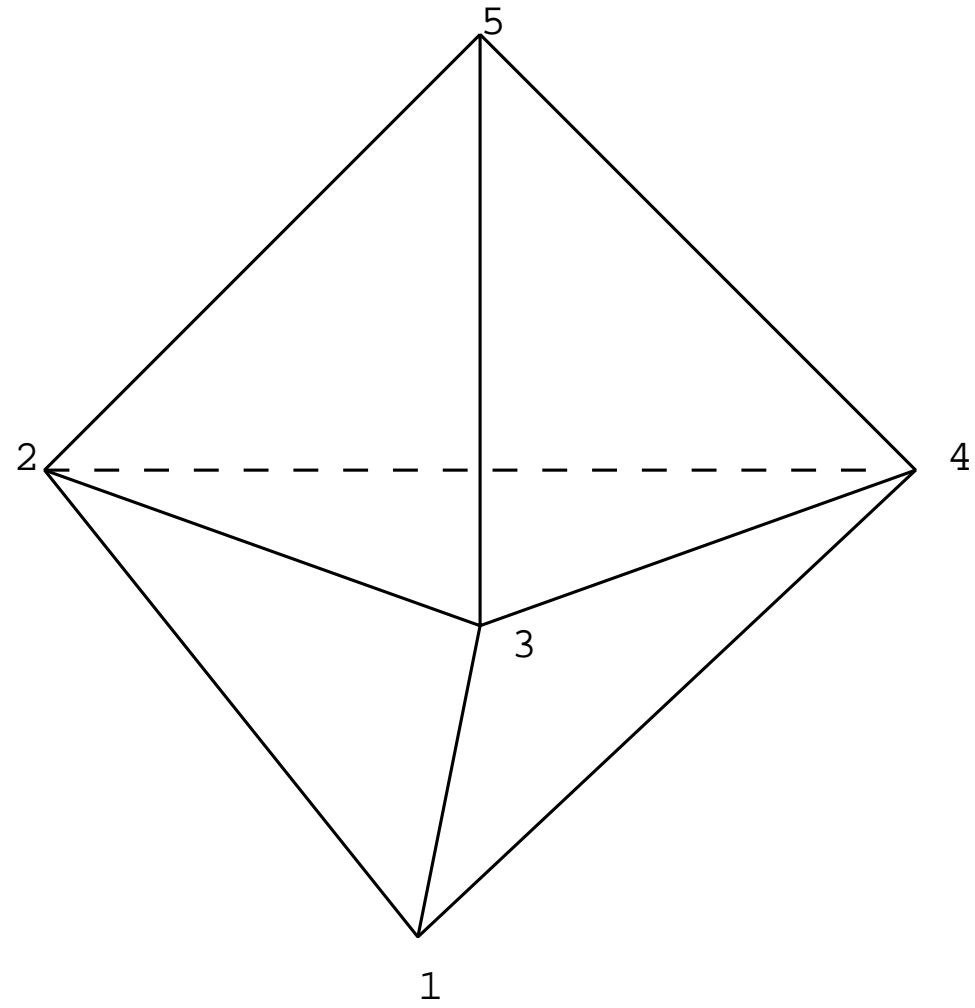
を満たすことである .

- 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.)

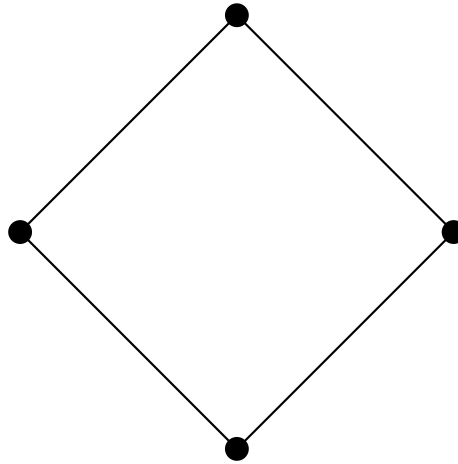
6面体の例



- 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の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．

まとめ 以下の事項について説明した．

- 分割表の基本的事項，条件つき独立性．
- toric model (指数型分布族)．
- 一般の多元分割表の階層モデルが simplicial complex と同値であること．
- 階層モデルの部分モデルとしてのグラフィカルモデル，分解可能モデル．
- 階層モデルの観点から重要となる simplicial complex の諸概念 (特に分離の概念)．

余談及び補足

- compaction によるモデルの分類と, 1-skeleton によるモデルの分類の関係が自明でない. 6面体の例.
- 単体的複体まで考えなくても, $\text{red } \mathcal{A}$ 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．

- 有向グラフについても今後考えたい．DAG から moralization によって得られる simplicial complex は，必ずしもグラフには対応しないはず．