数理解析研究所講究録
1138巻 2000年 172-179

Gröbner basis による分割表の数え上げ

倉敷芸術科学大学　中川　重和(Sigekazu Nakagawa) *

1　はじめに

行和 $\boldsymbol{r}=(r_1,r_2,\cdots,r_I)$ と列和 $\boldsymbol{c}=(c_1,c_2,\cdots,c_J)$ が与えられた $I\times J$ 分割表全体の集合を $\Omega(\boldsymbol{r},\boldsymbol{c})$ とする $(N=\sum_{i=1}^{I}r_i=\sum_{j=1}^{J}c_j)$. つまり, $\boldsymbol{u}=(u_{ij})\in\Omega(\boldsymbol{r},\boldsymbol{c})$ は $u_{ij}\in\boldsymbol{N}=\{0,1,2,\cdots\}$, $\sum_{j=1}^{J}u_{ij}=r_i$, $\sum_{i=1}^{I}u_{ij}=c_j$ を満足する. 分割表の数え上げとは, すべての $\boldsymbol{u}=(u_{ij})\in\Omega(\boldsymbol{r},\boldsymbol{c})$ を列挙することで, (統計学における) Fisher の正確確率法による p 値計算において必要である.

分割表の数え上げ問題は, 古くから知られた基本的な問題であり, Fisher の正確確率法以来, 考えられている. 計算機を意識した算法として, 行の置換によるクラス分けに基づく Fortran プログラム [13] やネットワークアルゴリズム [10] [11] などが 1980 年代に研究されている.

1990 年代に入って, Markov Chain Monte Carlo(MCMC) 法が統計学の有効な手法として大きな話題を呼び, この流れは分割表解析にも及んだ. 統計学の標準的な教科書では, 分割表の独立性の検定において, (経験則として) すべてのセルの度数 (u_{ij}) が 5 以上のとき, 検定統計量が χ^2 分布に漸近的に従うとして検定せよ, またそうでないときは, Fisher の正確確率法により p 値計算をせよ, とある. しかし, いずれにも当てはまらない例がある. 例えば, [7], pp.364, Table 1. この例では, いくつかのセルの度数が 5 以下 (χ^2 分布に漸近的に従わない) であるにもかかわらず, $\#\Omega(\boldsymbol{r},\boldsymbol{c})$ が多き過ぎて, 事実上すべての分割表を列挙できない.

このような問題にも頑健な手法として, 分割表解析における MCMC 法の研究が進んでいる. 中でも, Markov 連鎖の推移パターン (Markov basis という) の構成に Gröbner basis を用いたアルゴリズムを提案した [7] の貢献は大きい.

MCMC 法を分割表解析に適用する際, 問題となるのが収束性の問題と精度の問題である. 本稿では精度の問題に注目するが, その場合 (計算できる) exact な値との比較が不可欠となる. [14] に基づき, Gröbner basis と 有向グラフの backward search による数え上

*nakagawa@soft.kusa.ac.jp

げを実現し，exact p 値を計算する．そして，既存の MCMC での結果 ([1]) との数値比較を行う．

2 節では，記号の定義とおさらいの意味を込めて，Gröbner basis と分割表の関連について述べる．3 節では，Backward search とその実装について述べる．3.1 節では，$I \times J$ 分割表において，Gröbner basis と 有向グラフの backward search による方法で，どの程度の問題までが計算できるかを示す．3.2 節では，本稿での方法が多元分割表にも適用可能であることを示す．

なお，本報告に関連する話題として分割表の総数数え上げがある．総数の近似として，正規分布による近似 [8] がある．また対称式の内積計算 [6]，分割統治法 [12]，などが総数の完全な数え上げとして研究されている．

2 分割表の数え上げと Gröbner basis

定義 1 $\boldsymbol{Z}^I$, $\boldsymbol{Z}^J$ の 標準基底をそれぞれ $\{\boldsymbol{e}_i\}$, $\{\boldsymbol{e}'_j\}$ として，線形写像

$$\begin{array}{rccc} \pi: & \boldsymbol{N}^{IJ} & \to & \boldsymbol{Z}^I \oplus \boldsymbol{Z}^J \\ & u_{ij} & \mapsto & \sum_{i,j} u_{ij} \boldsymbol{e}_i \oplus \boldsymbol{e}'_j \end{array} \tag{1}$$

を定めるとき，$\Omega(\boldsymbol{r},\boldsymbol{c}) = \{\boldsymbol{u} \in \boldsymbol{N}^{IJ} \mid \pi(\boldsymbol{u}) = \begin{bmatrix} \boldsymbol{r} & \boldsymbol{c} \end{bmatrix}'\}$ である．π の定義域を $\boldsymbol{Z}^{IJ}$ に拡張して $ker(\pi) \subset \boldsymbol{Z}^{IJ}$ を考えるとき (今後，$ker(\pi) \subset \boldsymbol{Z}^{IJ}$ とする)，$ker(\pi)$ の任意の元は周辺和を不変にする．したがって，$ker(\pi)$ の基底が $\Omega(\boldsymbol{r},\boldsymbol{c})$ 上のひとつの *Markov basis* を与える．ここで，$\{\boldsymbol{m}_1, \cdots, \boldsymbol{m}_L\} \subset ker(\pi)$ が *Markov basis* であるとは，すべての $\boldsymbol{u}, \boldsymbol{u}' \in \Omega(\boldsymbol{r},\boldsymbol{c})$ に対して，$(\epsilon_1, \boldsymbol{m}_{i_1}), \ldots, (\epsilon_A, \boldsymbol{m}_{i_A})$ が存在して (ただし，$\epsilon_i = \pm 1$)，

- $\boldsymbol{u}' = \boldsymbol{u} + \sum_{j=1}^{A} \epsilon_j \boldsymbol{m}_{i_j}$ および
- $\boldsymbol{u} + \sum_{j=1}^{a} \epsilon_j \boldsymbol{m}_{i_j} \geq 0 \ (1 \leq a \leq A)$

をみたすことである．

我々の問題へ Gröbner basis の理論を持ち込むために，(1) で与えた線形写像を多項式写像へと持ち上げる：

$$\begin{array}{rccc} \hat{\pi}: & k[\boldsymbol{x}] & \to & k[\boldsymbol{t}] \\ & x_{ij} & \mapsto & \boldsymbol{t}^{\boldsymbol{e}_i \oplus \boldsymbol{e}'_j}. \end{array} \tag{2}$$

ただし，$\boldsymbol{x} = (x_{11}, x_{12}, \cdots, x_{IJ})$ であり，$\boldsymbol{t} = (t_1, t_2, \cdots, t_{I+J})$ である．このとき，

$$\langle x_{ij} - \boldsymbol{t}^{\boldsymbol{e}_i \oplus \boldsymbol{e}'_j} \rangle \tag{3}$$

は $IJ + I + J$ 個の変数 $\boldsymbol{x}, \boldsymbol{t}$ からなる多項式環 $k[\boldsymbol{x}, \boldsymbol{t}]$ のイデアルであり，$I := \ker(\hat{\pi})$ は $k[\boldsymbol{x}]$ のイデアルである．I と Markov basis の関係は次の命題で与えられる：

命題 2 ([7]) $\boldsymbol{N}^{IJ}$ 上の任意の *term order* $\succ$ に対し，有限集合 $M \subset ker(\pi)$ で

$$\left\{\boldsymbol{x}^{\boldsymbol{m}^+} - \boldsymbol{x}^{\boldsymbol{m}^-} \mid \boldsymbol{m} \in M\right\}$$

が I の *Gröbner basis* となる M が存在する．そしてこの M が $\Omega(\boldsymbol{r}, \boldsymbol{c})$ 上の *Markov basis* を与える．

命題 2 から，Markov basis を求めることは (2) で定まる多項式写像の核 $\ker(\hat{\pi})$ の生成元を求める問題に帰着される．この問題の解法は Gröbner basis 理論では既知の事実 ([5]) であり，以下のように Markov basis を求めるアルゴリズムが構成できる．ただし，3.2 節にある多元分割表の場合には少々の工夫が必要である ([9], [2])．

Markov basis を求めるアルゴリズム ([5])

- $\langle x_{ij} - \boldsymbol{t}^{\boldsymbol{e}_i \oplus \boldsymbol{e}'_j} \rangle$ $(\subset k[\boldsymbol{x}, \boldsymbol{t}])$ の reduced Gröbner basis (with elimination order $\boldsymbol{t} \succ \boldsymbol{x}$) G を求めよ．
- $G' := G \cap k[\boldsymbol{x}]$ が I の ($k[\boldsymbol{x}]$ における) Gröbner basis i.e. $I = \langle G' \rangle$. G' が求める Markov basis である．

$\ker(\hat{\pi})$ の reduced Gröbner basis (drl with $x_{11} \prec x_{12} \prec \cdots \prec x_{IJ}$) は，

$$\{x_{i\ell}x_{kj} - x_{ij}x_{k\ell} \mid 1 \leq i < k \leq I, 1 \leq j < \ell \leq J\}$$

であり，対応する $\boldsymbol{m} = (m_{ab}) \in M$ は

$$m_{ab} = \begin{cases} -1 & (a = i, b = j) \\ 1 & (a = i, b = l) \\ 1 & (a = k, b = j) \\ -1 & (a = k, b = l) \\ 0 & (\text{その他}) \end{cases} \tag{4}$$

である．

定義 3 $M \subset ker(\pi)$ に対し，$\Omega(\boldsymbol{r}, \boldsymbol{c})$ に付随する無向グラフ $\mathcal{G}_M$ を以下のように定義する．$\Omega(\boldsymbol{r}, \boldsymbol{c})$ を $\mathcal{G}_M$ の頂点集合とする．$\boldsymbol{u}, \boldsymbol{u}'$ が辺で結ばれるのは，ある $\boldsymbol{m} \in M$ によって，$\boldsymbol{u} = \boldsymbol{u}' + \epsilon \cdot \boldsymbol{m} (\epsilon = \pm 1)$ のときである．

このとき，命題は次のように読み換えることができる．

命題 4 ([5]) $\mathcal{G}_M$ が連結グラフである $\Longleftrightarrow \left\{\boldsymbol{x}^{\boldsymbol{m}^+} - \boldsymbol{x}^{\boldsymbol{m}^-} \mid \boldsymbol{m} \in M\right\}$ がイデアル I を生成する．

定義 5 $\boldsymbol{N}^{IJ}$ 上の任意の *term order* に対し，$\Omega(\boldsymbol{r},\boldsymbol{c})$ に付随する有向グラフ $\mathcal{G}_{M,\succ}$ を以下のように定義する．無向グラフ $\mathcal{G}_M$ において，$\boldsymbol{u}$ から $\boldsymbol{u}'$ への有向辺を $\boldsymbol{u}' \prec \boldsymbol{u}$ のときに定義する．

このとき，以下が成立する：

命題 6 ([5]) $\mathcal{G}_{M,\succ}$ が *unique sink* をもつ $\Longleftrightarrow$ $\{\boldsymbol{x}^{\boldsymbol{m}^+} - \boldsymbol{x}^{\boldsymbol{m}^-} \mid \boldsymbol{m} \in M\}$ が I の $\succ$ に関する *Gröbner basis* である．

3 Backward search

前節の議論より，分割表の数え上げは有向グラフの backward search により，実行できる ([14])．

Input: 周辺和 $\boldsymbol{r},\boldsymbol{c}$
Output: $\Omega(\boldsymbol{r},\boldsymbol{c})$ のすべての点の列挙

1. M : $\ker(\hat{\pi})$ の Gröbner basis G の計算
2. $\boldsymbol{u}' \in \Omega(\boldsymbol{r},\boldsymbol{c})$ を一つ見つけよ (観測データ)
3. $\boldsymbol{x}^{\boldsymbol{u}'} \xrightarrow{G}_{+} \boldsymbol{x}^{\boldsymbol{u}''}$ の計算 : $\boldsymbol{u}''$ は $\Omega(\boldsymbol{r},\boldsymbol{c})$ に付随する有向グラフの unique sink
4. 初期化 ; `Active` $:= \{\boldsymbol{u}''\}$, `Passive` $:= \emptyset$;
5. **while** (`Active` $\neq \emptyset$) **do**

   Choose $\boldsymbol{u} \in$ `Active`

   forall $\boldsymbol{m} = \boldsymbol{m}^+ - \boldsymbol{m}^- \in M\,(\boldsymbol{m}^+ \succ \boldsymbol{m}^-)$ **do**

   if $(\boldsymbol{u} - \boldsymbol{m}^- \geq 0)$ **and** $(\boldsymbol{u} + \boldsymbol{m} \notin$ `Passive`$)$ **then** `Active` $:=$ `Active` $\cup\{\boldsymbol{u} + \boldsymbol{m}\}$;

   `Active` $:=$ `Active`$\backslash\{\boldsymbol{u}\}$;
   `Passive` $:=$ `Passive`$\cup\{\boldsymbol{u}\}$;

ステップ 1 から 3 は数式処理システム Asir [3] で実装している．ステップ 4, 5 を C 言語で実装している．変数 `Active` と `Passive` は探索が頻繁に起こるためデータ構造として 2 分木を採用している．

3.1 $I \times J$ 分割表の数え上げ

表 11 は [4] から抜粋したふさぎ込み症候群のデータである．米国の 4 つの病院において，十二指腸潰瘍の摘出度合により，無気力，ふさぎ込みなどになる患者の頻度をまとめたものである．なお，手術法は

A　十二指腸 0% 摘出　　B　十二指腸 25% 摘出
C　十二指腸 50% 摘出　　D　十二指腸 75% 摘出

に層別している．

表 12 は 分割表それぞれの unique sink である．各病院毎の 4 つの 4×3 分割表に対し，backward search を実行した結果が表 13 である．左からふさぎ込み症候群データの総数，exact p 値 (DEC Alpha station 400MHz, 256MB での CPU Time(sec)) および MCMC(10^6 回) による p 値である．p 値とは

$$p = \sum_{\boldsymbol{v} \in \mathcal{T}} \Pr(\boldsymbol{v})$$

であり，$\mathcal{T} = \{\boldsymbol{v} \in \Omega(\boldsymbol{r}, \boldsymbol{c}) \mid \Pr(\boldsymbol{v}) \leq \Pr(\boldsymbol{u})\}$ としている．なお，病院 4 については，この実装では計算できなかった[1)]．

表 11: ふさぎ込み症候群データ

病院	手術法	ふさぎ込み症候群			計
		無	軽	重	
1	A	18	6	1	25
	B	18	6	2	26
	C	13	13	2	28
	D	9	15	2	26
		58	40	7	105
2	A	8	6	3	17
	B	12	4	4	20
	C	11	6	2	19
	D	7	7	4	18
		38	23	13	74

病院	手術法	ふさぎ込み症候群			計
		無	軽	重	
3	A	12	9	1	22
	B	15	3	2	20
	C	14	8	3	25
	D	13	6	4	23
		54	26	10	90
4	A	23	7	2	32
	B	23	10	5	38
	C	20	13	5	38
	D	24	10	6	40
		90	40	18	148

[1)] [11] を用いて計算すると，総数は 15,272,124 であり，exact p 値は 0.7677 となる

表 12: unique sink

1			
25	0	0	25
26	0	0	26
7	21	0	28
0	19	7	26
58	40	7	105

2			
17	0	0	17
20	0	0	20
1	18	0	19
0	5	13	18
38	23	13	74

3			
22	0	0	22
20	0	0	20
12	13	0	25
0	13	10	23
54	26	10	90

4			
32	0	0	32
38	0	0	38
20	18	0	38
0	22	18	40
90	40	18	148

表 13: ふさぎ込み症候群データの総数，p 値と MCMC

	$\#\Omega(\boldsymbol{r}, \boldsymbol{c})$	exact p 値 (CPU Time, sec)	MCMC による p 値
1	1,106,454	0.0610 (34,635)	0.0633
2	1,107,960	0.7849 (35,279)	0.7856
3	944,944	0.5280 (24,574)	0.5340
4	—	—	0.7698

3.2 $3 \times 3 \times 3$ 分割表の数え上げ

Backward search による数え上げの利点は，[10]，[11]，[13] などに対し，多元分割表にも適用可能なことである．ここでは，$3 \times 3 \times 3$ 分割表の数え上げを実行する．$3 \times 3 \times 3$

分割表の数え上げとは，以下の 27 個の line sums

$$u_{.jk} := \sum_{i=1}^{3} u_{ijk},\ u_{i.k} := \sum_{i=1}^{3} u_{ijk},\ u_{ij.} := \sum_{i=1}^{3} u_{ijk}$$

が与えられたもとで，これを満足するようなすべての (u_{ijk}) を列挙することである．$u_{.jk} = u_{i.k} = u_{ij.} = s$ について実行した結果が表 14 である (DEC Alpha station 400MHz, 256MB)．表 15 は $3 \times 3 \times 3$ 分割表の Markov basis である ([2])．

表 14: $u_{.jk} = u_{i.k} = u_{ij.} = s$

s	総数	CPU Time
1	847	0.3
2	43,687	307
3	619,219	47,699

表 15: $3 \times 3 \times 3$ 分割表の Markov basis

$$\begin{array}{ccc|ccc|ccc} -&+&0&+&-&0&0&0&0 \\ +&-&0&-&+&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0 \end{array} \quad 27\text{個}\quad 4\text{次}$$

$$\begin{array}{ccc|ccc|ccc} +&-&0&-&+&0&0&0&0 \\ -&0&+&+&0&-&0&0&0 \\ 0&+&-&0&-&+&0&0&0 \end{array}$$

18 個　　6 次

$$\begin{array}{ccc|ccc|ccc} +&-&0&-&0&+&0&+&- \\ -&+&0&+&0&-&0&-&+ \\ 0&0&0&0&0&0&0&0&0 \end{array}$$

36 個　　6 次

$$\begin{array}{ccc|ccc|ccc} -&+&0&+&0&-&0&-&+ \\ +&0&-&-&0&+&0&0&0 \\ 0&-&+&0&0&0&0&+&- \end{array} \quad 28\text{個}\quad 7\text{次}$$

$$
\begin{array}{cccccccccc}
+ & - & 0 & - & 0 & + & 0 & + & - & \\
- & 0 & + & 0 & 0 & 0 & + & 0 & - & \\
0 & + & - & + & 0 & - & - & - & +2 & \\
 & & & & 1\text{個} & & & 9\text{次} & &
\end{array}
$$

参 考 文 献

[1] 中川重和 (1999a). 分割表上の Markov basis と Gröbner basis. 日本計算機統計学会第13回大会論文集, 54-57.

[2] 中川重和 (1999b). 3次元分割表上の Markov chain の Gröbner 基底による構成. 第67回日本統計学会講演報告集, 57-58.

[3] 野呂正行, 下山武司 (1995). *Asir User's Manual.* `ftp://endeavor.fujitsu.co.jp`.

[4] 柳川尭 (1986). 離散多変量データの解析, 共立出版.

[5] Adams, W.W. and Loustaunau, P.(1994). *An Introduction to Gröbner bases.* AMS.

[6] Diaconis, P. and Gangolli, A(1995). Rectangular arrays with fixed margins. In *Discrete Probability and Algorithms*, Springer, New York, 15–41.

[7] Diaconis, P. and Strumfels, B.(1998). Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.*, **26**, 363–397.

[8] Gail, M. and Mantel, N.(1977). Counting the Number of $r \times c$ Contingency Tables with Fixed Margins. *J. of the American Statistical Association*, **72**, No. 360, 859–862.

[9] Hosten, S. and Strumfels, B.(1995). GRIN: An Implementation of Gröbner Bases for Integer Programming. Springer LNCS, **920**, 267–276.

[10] Metha, C. R. and Patel, N. R.(1983). A Network Algorithm for Performing Fisher's Exact Test in $r \times c$ Contingency Tables. *J. of the American Statistical Association*, **78**, No. 382, 427–434.

[11] Metha, C. R.(1986). ALGORITHM 643: FEXACT: A FORTRAN Subroutine for Fisher's Exact Test on Unordered $r \times c$ Contingency Tables. *ACM Transactions on Mathematical Software*, **12**, No. 2, 154–161.

[12] Mount, J.(1995). Application of Convex Sampling to Optimization and Contingency Table Generation/Counting, Ph.D. Dissertation, Department of Computer Science, Carnegie Mellon University.

[13] Saunders, I. W.(1984). AS 205: Enumeration of $R \times C$ Tables with Related Row Totals. *J. Royal Statistical Society*, 340–352.

[14] Strumfels, B.(1996). *Gröbner bases and convex polytopes.* AMS.