学術フロンティア講義 数理工学のすすめ

「計算代数統計」（理論編）

2018 年 4 月 26 日 (木) 配布[1]

清 智也 sei@mist.i.u-tokyo.ac.jp

1 背景

1.1 記号と用語

- 非負整数全体を $\mathbb{N}$ とおく。η と ν を正の整数とする。
- $\boldsymbol{x} \in \mathbb{N}^{\eta}$ を**度数データ**という。分割表はその特別な場合と考える（$\rightarrow$ 例 1)。
- 行列 $\boldsymbol{A} \in \mathbb{N}^{\nu \times \eta}$ で，ある $\boldsymbol{c} \in \mathbb{R}^{\nu}$ に対して $\boldsymbol{c}^{\top}\boldsymbol{A} = (1, \ldots, 1)$ となるようなものを**配置**という。($\rightarrow$ 例 1)。
- $\boldsymbol{t} \in \mathbb{N}^{\nu}$ に対し，$F_{\boldsymbol{t}} = \{\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{t}\}$ を**ファイバー**という。

例 1. サイズ 2×3 の分割表の場合，周辺度数を求める配置は

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

と表すことができる。 例えば「学食データ」

	渋谷側から	吉祥寺側から	その他
学食	4	5	1
学食以外	4	3	1

は $\boldsymbol{x} = (4, 5, 1, 4, 3, 1)^{\top}$ と表され，$\boldsymbol{t} = \boldsymbol{A}\boldsymbol{x} = (10, 8, 8, 8, 2)^{\top}$ となる。 □

演習問題 1. 上の例に対し，ファイバー $F_{\boldsymbol{t}}$ の要素数を求めよ。 □

- ファイバー上の確率分布

$$f(\boldsymbol{x}) = \frac{C}{\prod_{i=1}^{\eta} x_i!}, \quad \boldsymbol{x} \in F_{\boldsymbol{t}}, \tag{1}$$

を**超幾何分布**という。比例定数 C は $\sum_{\boldsymbol{x} \in F_{\boldsymbol{t}}} f(\boldsymbol{x}) = 1$ となるように定められる。

[1]次回も使います。

演習問題 2. $\nu = 1$ で $\boldsymbol{A} = (1, \ldots, 1)$ の場合，超幾何分布は多項分布と一致することを確かめよ。

例 2. 「学食データ」の $\boldsymbol{x}$ と $\boldsymbol{t}$ に対する超幾何分布は

$$f(\boldsymbol{x}) = \frac{C}{4!5!1!4!3!1!} = \frac{{}_8\mathrm{C}_4 \times {}_8\mathrm{C}_5 \times {}_2\mathrm{C}_1}{{}_{18}\mathrm{C}_{10}}$$

で与えられる。2 番目の等号については黒板で説明する。 □

演習問題 3 (多項分布の条件付き分布)**.** 多項分布

$$p(\boldsymbol{x}) = \frac{N!}{\prod_{i=1}^{\eta} x_i!} \prod_{i=1}^{\eta} \pi_i^{x_i}, \quad \sum_i \pi_i = 1, \quad \sum_i x_i = N$$

を考える。いま $\boldsymbol{A} = (A_{ji})$ を配置，$\theta_1, \ldots, \theta_\nu$ を実数とし，$\pi_i = e^{\sum_{j=1}^{\nu} \theta_j A_{ji}}$ とおく。ただし $\sum_i \pi_i = 1$ は満たされると仮定する。このとき $\boldsymbol{x} \in F_{\boldsymbol{t}}$ ならば

$$\frac{p(\boldsymbol{x})}{\sum_{\tilde{\boldsymbol{x}} \in F_{\boldsymbol{t}}} p(\tilde{\boldsymbol{x}})}$$

が超幾何分布となることを示せ。この式は $\boldsymbol{Ax} = \boldsymbol{t}$ を与えたもとでの $\boldsymbol{x}$ の条件付き分布を表している。 □

1.2 背景：何が問題か？

ファイバー上の確率分布 $f(\boldsymbol{x})$ と実数値関数 $h(\boldsymbol{x})$ に対し，期待値

$$\mathrm{E}[h] = \sum_{\boldsymbol{x} \in F_{\boldsymbol{t}}} h(\boldsymbol{x}) f(\boldsymbol{x}) \tag{2}$$

を計算したい場面がある。$f(\boldsymbol{x})$ は超幾何分布でなくてもよい。

例 3 (仮説検定)**.** $\boldsymbol{x}^0 \in \mathbb{N}^\eta$ をデータとする。いま「$\boldsymbol{x}^0$ は超幾何分布に従っている」という仮説を立て，この仮説が妥当かどうかを判断したいとする。このようなとき，例えば尤度比検定統計量

$$T(\boldsymbol{x}) = \min_{\boldsymbol{\theta} \in \Theta} 2 \sum_{i=1}^{\eta} x_i \log \frac{x_i}{\pi_i(\boldsymbol{\theta})}, \quad \pi_i(\boldsymbol{\theta}) = e^{\sum_j \theta_j A_{ji}}, \quad \Theta = \{\boldsymbol{\theta} \mid \textstyle\sum_i \pi_i(\theta) = 1\},$$

という関数を定義し，$T(\boldsymbol{x}^0)$ が大きければ仮説を棄却する。ここで「$T(\boldsymbol{x}^0)$ が大きい」ということを客観的に判断するため，p 値を計算する。p 値とは，式 (2) において $f(\boldsymbol{x}) =$ (超幾何分布)，$h(\boldsymbol{x}) = \mathbb{I}_{\{T(\boldsymbol{x}) \geq T(\boldsymbol{x}^0)\}}$ としたものである。p 値が非常に小さいとき，仮説を棄却する。詳しくは後述の「グレブナー道場」4 章を参照せよ。 □

ところで，ファイバー $F_{\boldsymbol{t}}$ はしばしば非常に大きな集合となり，そのような場合には式 (2) の和を計算するのが現実的に不可能となる場合がある。そこで次節で述べるマルコフ連鎖モンテカルロ法 (MCMC) が必要となる。MCMC は汎用的な手法であるが，ここでは式 (2) の近似計算に特化した形でその方法を述べる。

2　マルコフ基底とその応用

2.1　マルコフ基底

有限集合 $B \subset \ker_{\mathbb{Z}} \boldsymbol{A} = \{\boldsymbol{z} \in \mathbb{Z}^\eta \mid \boldsymbol{A}\boldsymbol{z} = \boldsymbol{0}\}$ が $\boldsymbol{A}$ に関する**マルコフ基底**であるとは，任意の $\boldsymbol{t} \in \mathbb{N}^\nu$ および任意の $\boldsymbol{x}, \boldsymbol{y} \in F_{\boldsymbol{t}}$ に対し，ある $L \in \mathbb{N}$, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_L \in B$ および $s_1, \ldots, s_L \in \{-1, 1\}$ が存在して

$$\boldsymbol{x} + \sum_{i=1}^{\ell} s_i \boldsymbol{z}_i \in F_{\boldsymbol{t}}, \quad 1 \leq \ell \leq L-1,$$

かつ

$$\boldsymbol{x} + \sum_{i=1}^{L} s_i \boldsymbol{z}_i = \boldsymbol{y}$$

が成り立つことである。

例 4. $\boldsymbol{A} = \begin{pmatrix} 1 & 1 \end{pmatrix} \in \mathbb{R}^{1\times 2}$ のとき，マルコフ基底 (の一つ) は $B = \{\binom{1}{-1}\}$ である。 □

例 5. 例 1 の配置 $\boldsymbol{A}$ の場合，マルコフ基底（の一つ）を分割表の形で表すと

$$B = \left\{ \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & -1 \\ -1 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix} \right\}$$

となる。 □

マルコフ基底の定義は，グラフ理論の用語を使うと次のように言い換えることができる。$F_{\boldsymbol{t}}$ を頂点集合とし，$\boldsymbol{x}, \boldsymbol{y} \in F_{\boldsymbol{t}}$ が $\boldsymbol{x} - \boldsymbol{y} \in B$ または $\boldsymbol{y} - \boldsymbol{x} \in B$ を満たすとき $\boldsymbol{x}$ と $\boldsymbol{y}$ の間に辺を結ぶ。この規則で作ったグラフが（各 $\boldsymbol{t}$ について）連結であるとき，B をマルコフ基底という。

2.2　マルコフ連鎖モンテカルロ法 (MCMC)

以下のようにして乱数列 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ を作り，式 (2) の $\mathrm{E}[h]$ の近似値を求める方法をマルコフ連鎖モンテカルロ法 (Markov Chain Monte Carlo = **MCMC**) という。

1. $\boldsymbol{x}_1 \in F_{\boldsymbol{t}}$ を適当に選ぶ。(仮説検定の例では $\boldsymbol{x}_1 = \boldsymbol{x}^0$ でよい。)
2. $i = 2, \ldots, N$ に対して以下を行う。
   (i) $\boldsymbol{z} \in B$ を等確率で選ぶ。
   (ii) $s \in \{-1, 1\}$ を等確率で選ぶ。

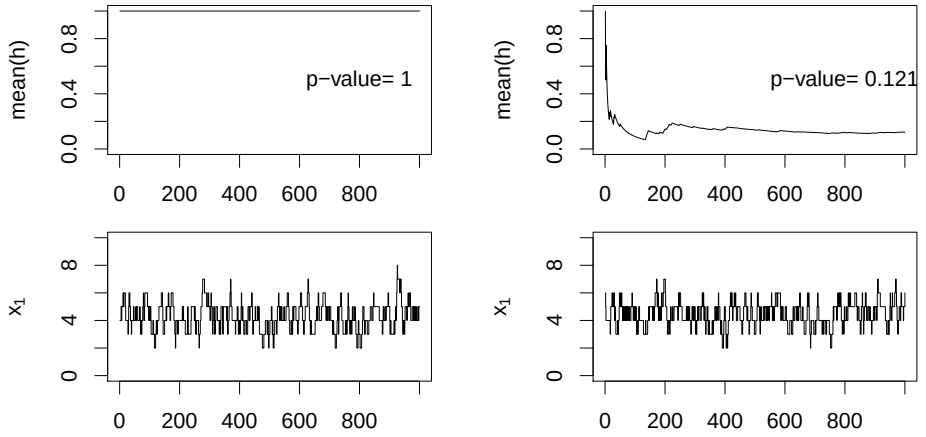
(iii) $\tilde{\boldsymbol{x}} = \boldsymbol{x}_{i-1} + s\boldsymbol{z}$ とおく。$\tilde{\boldsymbol{x}} \notin F_{\boldsymbol{t}}$ ならば $\boldsymbol{x}_i = \boldsymbol{x}_{i-1}$ とおく。$\tilde{\boldsymbol{x}} \in F_{\boldsymbol{t}}$ ならば，

$$q = \min\{f(\tilde{\boldsymbol{x}})/f(\boldsymbol{x}_{i-1}), 1\}$$

とおき，確率 q で $\boldsymbol{x}_i = \tilde{\boldsymbol{x}}$，確率 $1-q$ で $\boldsymbol{x}_i = \boldsymbol{x}_{i-1}$ とおく。

3. 標本平均 $N^{-1}\sum_{i=1}^{N} h(\boldsymbol{x}_i)$ を $\mathrm{E}[h]$ の近似値として出力する。この出力は $N \to \infty$ のとき $\mathrm{E}[h]$ に収束することが知られている。

例 6. 「学食データ」について，仮説検定の p 値を計算した例を図に示す。なお，この例では計算機によって p 値を正確に計算することができる。その値は，$\boldsymbol{x}^0 = (4,5,1,4,3,1)^\top$ の場合は 1.00，$\boldsymbol{x}^0 = (6,4,0,2,4,2)^\top$ の場合は約 0.17 となる。



(a) $\boldsymbol{x}^0 = (4,5,1,4,3,1)^\top$ の場合。　(b) $\boldsymbol{x}^0 = (6,4,0,2,4,2)^\top$ の場合。

図 1. 学食データの仮説検定。上側は h の標本平均，下側は x_1 の推移を表す。

2.3 理論的な研究課題

「色々な配置 $\boldsymbol{A}$ に対し，そのマルコフ基底を求めよ。」

マルコフ基底に関する諸結果がまとめられた一冊として

S. Aoki, H. Hara and A. Takemura (2012). *Markov Bases in Algebraic Statistics*, Springer

がある。また，マルコフ基底のデータベースが

`http://markov-bases.de/`

に公開されているので興味ある者は眺められたい。

3　グレブナー基底とマルコフ基底

グレブナー基底の理論を使うと，マルコフ基底が「有限時間で」求められる。このことを概観する。詳しくは,

JST CREST 日比チーム編 (2012). 「グレブナー道場」, 共立出版

の第 1 章と第 4 章を参照されたい。

3.1　記号と用語

- $K = \mathbb{Q}$（有理数体）とおく。以下の議論は $K = \mathbb{R}$, $K = \mathbb{C}$ としても成立する。
- $\boldsymbol{p} = (p_1, \ldots, p_\eta)$ および $\boldsymbol{q} = (q_1, \ldots, q_\nu)$ を不定元（独立変数）とする。
- K を係数とする $\boldsymbol{p}$ の多項式全体を $K[\boldsymbol{p}]$ と表す。$K[\boldsymbol{p}]$ には通常の意味で和と積の演算が定義される。このとき $K[\boldsymbol{p}]$ を多項式環という。同様に，$\boldsymbol{p}, \boldsymbol{q}$ の多項式全体を $K[\boldsymbol{p}, \boldsymbol{q}]$ と表す。たとえば

  $$p_1^3 p_3 - 3p_2^2 \in K[\boldsymbol{p}], \quad 5p_1 q_2^4 + 6p_3 + 7 \in K[\boldsymbol{p}, \boldsymbol{q}]$$

  などである。
- 多項式環 $K[\boldsymbol{p}]$ の部分集合 I が次の性質を満たすとき，これをイデアルという。

  (i) $f \in K[\boldsymbol{p}]$ かつ $g \in I$ ならば $fg \in I$。

  (ii) $f, g \in I$ ならば $f + g \in I$。
- 集合 $S \subset K[\boldsymbol{p}]$ に対し，S を含む最小のイデアルを $\langle S \rangle$ と書き，S で生成されるイデアルという。$K[\boldsymbol{p}, \boldsymbol{q}]$ についても同様である。
- 配置 $\boldsymbol{A}$ に対し,

  $$J_{\boldsymbol{A}} = \left\langle p_i - \prod_{j=1}^{\nu} q_j^{A_{ji}} \mid i = 1, \ldots, \eta \right\rangle$$

  と定義する。$\boldsymbol{x} \in F_{\boldsymbol{t}}$ のとき $\prod_i p_i^{x_i} \equiv \prod_j q_j^{t_j} \pmod{J_{\boldsymbol{A}}}$ となる。

定理 (Diaconis and Sturmfels 1998)**.** 有限集合 $B \subset \mathbb{Z}^\eta$ が配置 $\boldsymbol{A}$ に対するマルコフ基底であるための必要十分条件は,

$$J_{\boldsymbol{A}} \cap K[\boldsymbol{p}] = \langle \boldsymbol{p}^{\boldsymbol{z}^+} - \boldsymbol{p}^{\boldsymbol{z}^-} \mid \boldsymbol{z} \in B \rangle \tag{3}$$

が成り立つことである。ただし $\boldsymbol{p}^{\boldsymbol{z}^+} = \prod_{i:z_i>0} p_i^{z_i}$, $\boldsymbol{p}^{\boldsymbol{z}^-} = \prod_{i:z_i<0} p_i^{-z_i}$ とする。

例 7. $\boldsymbol{A} = \begin{pmatrix} 1 & 1 \end{pmatrix}$ の場合，(q_1 を q と略記することにして)

$$\begin{aligned} J_{\boldsymbol{A}} &= \langle p_1 - q, p_2 - q \rangle \\ &= \{(p_1 - q)f + (p_2 - q)g \mid f, g \in K[p_1, p_2, q]\} \end{aligned}$$

となる。$f = \sum_{i=0}^{m} c_i q^i$, $g = \sum_{i=0}^{n} d_i q^i$ ($c_i, d_i \in K[p_1, p_2]$, $c_m \neq 0$, $d_n \neq 0$) とおき，係数を比較すると，$J_{\boldsymbol{A}} \cap K[\boldsymbol{p}] = \langle p_1 - p_2 \rangle$ となることが示される。よって $B = \{\begin{pmatrix} 1 \\ -1 \end{pmatrix}\}$ はマルコフ基底である。 □

演習問題 4. 上の例を確かめよ。

3.2 グレブナー基底

イデアルの生成系として性質のよいものに，グレブナー基底がある。特に，Buchberger アルゴリズムと消去理論と呼ばれる方法により，式 (3) を通じてマルコフ基底を有限時間で求めることができる。

グレブナー基底の定義と Buchberger アルゴリズムは，実はそれほど難しいものではないのだが（前述の本を参照のこと），ここでは具体例だけ見ることにしよう。

例 8. $\boldsymbol{A} = \begin{pmatrix} 1 & 1 \end{pmatrix}$ の場合，$J_{\boldsymbol{A}} = \langle p_1 - q, p_2 - q \rangle$ のグレブナー基底は，以下の手続きで求まる。ただし辞書式順序 $q > p_2 > p_1$ を用いる。$f := q - p_1$, $g := q - p_2$ として，

$$\begin{aligned} S(f, g) &= f - g \quad (f \text{ と } g \text{ の先頭項が打ち消されるように決める}) \\ &= p_2 - p_1 \\ &=: h, \\ S(f, h) &= p_2 f - q h \\ &= p_1 q - p_1 p_2 \\ &\equiv p_1^2 - p_1 p_2 \pmod{f} \\ &= -p_1 p_2 + p_1^2 \\ &\equiv -p_1^2 + p_1^2 \pmod{h} \ = 0, \\ S(g, h) &= p_2 g - q h = p_1 q - p_2^2 \\ &\equiv p_1^2 - p_2^2 \pmod{f} \ = -p_2^2 + p_1^2 \\ &\equiv -p_1 p_2 + p_1^2 \pmod{h} \\ &\equiv -p_1^2 + p_1^2 \pmod{h} \ = 0. \end{aligned}$$

こうして得られる $\{f, g, h\} = \{q - p_1, q - p_2, p_2 - p_1\}$ が $J_{\boldsymbol{A}}$ のグレブナー基底であり（Buchberger アルゴリズム），また $J_{\boldsymbol{A}} \cap K[p_1, p_2]$ のグレブナー基底は $\{f, g, h\} \cap K[p_1, p_2] = \{p_2 - p_1\}$ となる（消去理論）。これより，$\boldsymbol{A}$ に対するマルコフ基底は $B = \{\begin{pmatrix} -1 \\ 1 \end{pmatrix}\}$ で与えられることが分かる。 □