最近の統計学における代数的諸問題について

東大情報理工・数理情報　　竹村彰通

2009年6月24日

項目

1. 最近の話題の個人的な概観
2. Pistone-Wynn 流の実験計画入門
3. 壷のモデルとマルコフ基底
4. Naiman-Wynn の離散チューブ (“abstract tube”)
5. 感想など

最近の話題の個人的な概観

- 代数的方法は統計学でさまざまな形で使われて来た．
  - 実験計画法における有限体の理論，有限幾何
  - 多変量分布論における連続群の表現論 (“zonal polynomial” 等)
  - 統計的決定理論における変換群論とハール測度の応用
  - 離散データ解析における有限群の表現論 (Diaconis)

- しかし，最近 10 年間の「計算代数統計」は双方向性，多分野性，複合性がある．
- 計算 (応用) と 理論が車の両輪として発展
  - 「計算できる代数」がさまざまな分野で応用される中で，統計の問題も扱われてきた．グレブナー基底をキーワードとして具体的な問題に対して具体的な答を出してくれる代数学．
  - 統計で扱われている問題や概念が，代数の理論的観点からも新たな視点を与える．

計算代数統計の二つの主な流れ

1. Diaconis and Sturmfels (1998) によるグレブナー基底の分割表解析への応用 (“マルコフ基底”)．マルコフ連鎖を用いた正確検定を可能とするもの．
2. Pistone and Wynn (1996) によるグレブナー基底の実験計画法への応用．一部実施計画における non-regular design の別名関係 (推定可能性) の解析に有用．

これらについては，以下で簡単な例を用いて説明する．

その他にも興味深いテーマはあり，本講演では Naiman-Wynn の離散チューブについても紹介する．

Pistone-Wynn 流の実験計画入門

ここでは簡単のために 2 水準の一部実施計画を考える．

背景

- 実験計画法の教科書で標準的に解説されている．
- 古い本となるが，奥野・芳賀 (1969) の 8 章には以下のような記述があり，田口玄一などのわが国の貢献が強調されている．(最近は?)

多因子実験の計画と解析は,「直交表」を用いることによって，きわめてエレガントにおこなうことができる．これは，わが国独特の手法であって，田口玄一氏等に負うものである．諸外国では，いまだに直交表を用いず，面倒な割付け方をしているために，この有用な手法の普及がたいへん遅れている．

- 一部実施要因計画の別名関係の扱いには，いかにも代数的な「におい」がする．
- Pistone-Wynn 流はいわゆる regular design の別名関係の扱いを一般化するものであるが，まだあまり広く理解されていない．

2 水準一部実施計画の例 (積型記法)

- ある製品の製造工程において，品質に影響を与える要因として，温度や圧力などのいくつかの要因を考える．
- 要因数を 6 とし，それらの要因を A, B, C, D, E, F とする．
- それぞれの要因を「低水準」と「高水準」のいずれかに設定して実験をおこなうこととする．
- 「積型」の記法では，低水準を $+1$ (あるいは簡単に $+$)，高水準を -1 (あるいは $-$) と表すこととする．
- すべての組合せを行うとすると，$64 = 2^6$ 回の実験をおこなわなければならない

- そこで，実験回数を $1/4$ の 16 回で済ませることを考える．

- $1/4$ に減らすための「定義対比」として次の二つを考える:

$$I = ABE, \quad I = ACDF \tag{1}$$

- $I = ABE$ の意味は，A, B, C のそれぞれの水準 (± 1) をかけ合わせた時に，それらの積が $+1$ となるように水準の組合せが定まっていることを意味する．つまり，$A = B = +1$ の時には $E = +1$ と「割り付ける」．

- **また，各要因の値は** ± 1 **なので次のように定義する:**

$$A^2 = B^2 = C^2 = D^2 = E^2 = F^2 = I \tag{2}$$

- **このルールのもとで** (1) **式は次のように書き換えられる**

$$E = AB, \quad F = ACD \tag{3}$$

Table 1: 2^{6-2} 計画の例

A	B	C	D	E	F
$+$	$+$	$+$	$+$	$+$	$+$
$+$	$+$	$+$	$-$	$+$	$-$
$+$	$+$	$-$	$+$	$+$	$-$
$+$	$+$	$-$	$-$	$+$	$+$
$+$	$-$	$+$	$+$	$-$	$+$
$+$	$-$	$+$	$-$	$-$	$-$
$+$	$-$	$-$	$+$	$-$	$-$
$+$	$-$	$-$	$-$	$-$	$+$
$-$	$+$	$+$	$+$	$-$	$-$
$-$	$+$	$+$	$-$	$-$	$+$
$-$	$+$	$-$	$+$	$-$	$+$
$-$	$+$	$-$	$-$	$-$	$-$
$-$	$-$	$+$	$+$	$+$	$-$
$-$	$-$	$+$	$-$	$+$	$+$
$-$	$-$	$-$	$+$	$+$	$+$
$-$	$-$	$-$	$-$	$+$	$-$

加法型への書き換えと線形符号との同値性

- 「低水準」と「高水準」を 0 および 1 で符号化する:

$$+ \ \rightarrow \ 0, \qquad - \ \rightarrow \ 1$$

- 前節の積の演算は mod 2 での “exclusive or” の加法の演算と同じであることがわかる．
- 表 1 の例で，6 個の要因の水準を 6 個の「ビット」と考え $x_1, \ldots, x_6$ と表す．まず最初の 4 ビットについては full factorial (完全実施計画) の形に書いていた．
- このことは，最初の 4 ビットはそのまま「送信する」ことを意味している．

- **5 ビット目と 6 ビット目は，誤り訂正のために一定のルールで値を定めて付加する．(3) 式を加法的に表せば**

$$\begin{aligned} x_5 &\equiv x_1 + x_2 \pmod 2 \\ x_6 &\equiv x_1 + x_3 + x_4 \pmod 2 \end{aligned}$$

となる．(parity check)

- **この時，表 1 は次のように書き換えられる．**

x_1	x_2	x_3	x_4	x_5	x_6
0	0	0	0	0	0
0	0	0	1	0	1
0	0	1	0	0	1
0	0	1	1	0	0
0	1	0	0	1	0
0	1	0	1	1	1
0	1	1	0	1	1
0	1	1	1	1	0
1	0	0	0	1	1
1	0	0	1	1	0
1	0	1	0	1	0
1	0	1	1	1	1
1	1	0	0	0	1
1	1	0	1	0	0
1	1	1	0	0	0
1	1	1	1	0	1

- 加法型に書くことのメリット: 有限体 $\mathbb{F}_2$ 上の通常の線形代数として理解できる.
- 加法型のほうが数学的に「正しい」?
- 積型の記法がなぜ加法型の記法によって駆逐されなかったのか?
- 実は積型の記法のほうがより「代数的」!

- A, B, C, D, E, F の 6 個の文字を「不定元」と考えて，これらの不定元からなる多項式環 $k[A, B, C, D, E, F]$ を考える．k は適当な体．
- $k[A, B, C, D, E, F]$ の中に，次の 8 個の多項式で生成されるイデアルを考える．

$$\begin{aligned} I = \langle & A^2 - 1, B^2 - 1, C^2 - 1, D^2 - 1, E^2 - 1, F^2 - 1, \\ & ABE - 1, ACDF - 1 \rangle \end{aligned} \tag{4}$$

- Pistone and Wynn の言うところの "design ideal"
- 積型の形式的な操作は，商環 $k[A, B, C, D, E, F]/I$ の操作と全く同じ．

- 2 個の monomial が別名関係にあるための必要十分条件は，それらの差が design ideal I に属することである．
- この観点に立てば，例えば別名関係の判定はイデアル帰属問題であるから，I のグレブナー基底を求めればよい．

- **積型だとコンパクトに記述できる non-regular design の例 (5 要因)**

$$
\begin{aligned}
f(x_1, \ldots, x_5) &= \frac{1}{2} + \frac{1}{4}x_1x_2x_4 - \frac{1}{4}x_1x_2x_3 \\
&\qquad + \frac{1}{4}x_1x_2x_4x_5 + \frac{1}{4}x_1x_2x_3x_5 \\
&= \frac{1}{2} + \frac{x_1x_2\big(x_4(x_5+1) + x_3(x_5-1)\big)}{4}
\end{aligned}
$$

の零点の集合 ($x_i^2 - 1 \equiv 0, \forall i$, のもと．実験回数は 16 なので 1/2 fraction)

1	1	1	1	1
1	1	-1	1	1
1	1	-1	1	-1
1	1	-1	-1	-1
1	-1	1	1	-1
1	-1	1	-1	1
1	-1	1	-1	-1
1	-1	-1	-1	1
-1	1	1	1	-1
-1	1	1	-1	1
-1	1	1	-1	-1
-1	1	-1	-1	1
-1	-1	1	1	1
-1	-1	-1	1	1
-1	-1	-1	1	-1
-1	-1	-1	-1	-1

- **別名関係は** $\langle f, x_1^2 - 1, \ldots, x_5^2 - 1 \rangle$ **のグレブナー基底を求めれば判定できる．**

- ただし，non-regular design を許したとして，最適な design を求める問題はほとんど手がついていないと思われる．

壷のモデルとマルコフ基底

ここでは，簡単な 2 元の分割表について,「北西隅ルール」に関する簡単な問 (クイズ) を見ながら，グレブナー基底との対応を考える.

問 1　周辺が以下の分割表の中身をうまくうめてください.

$$
\begin{array}{ccc|c}
* & * & * & 6 \\
* & * & * & 5 \\
* & * & * & 5 \\
\hline
7 & 5 & 4 & 16
\end{array}
$$

解答例 (行優先で右下から)

- まず右下を $\min(4,5)=4$ とする．

$$
\begin{array}{ccc|c}
* & * & * & \mathbf{6} \\
* & * & * & \mathbf{5} \\
* & * & 4 & \mathbf{1} \\
\hline
\mathbf{7} & \mathbf{5} & \mathbf{0} & \mathbf{12}
\end{array}
$$

- 次に **(3,2)** 要素を $\min(1,5)=1$ でうめると **(3,1)** 要素も **0** となり

$$
\begin{array}{ccc|c}
* & * & * & \mathbf{6} \\
* & * & * & \mathbf{5} \\
\mathbf{0} & \mathbf{1} & \mathbf{4} & \mathbf{0} \\
\hline
\mathbf{7} & \mathbf{4} & \mathbf{0} & \mathbf{11}
\end{array}
$$

- **さらに (2,3) 要素を** $\min(0,5)=0$ **とし，続いて他の 2 行目をうめると**

$$
\begin{array}{ccc|c}
* & * & * & \mathbf{6} \\
\mathbf{1} & \mathbf{4} & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \mathbf{1} & \mathbf{4} & \mathbf{0} \\
\hline
\mathbf{6} & \mathbf{0} & \mathbf{0} & \mathbf{6}
\end{array}
$$

- **結果:**

$$\begin{array}{ccc|c} 6 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 0 & 1 & 4 & 0 \\ \hline 0 & 0 & 0 & 0 \end{array}$$

(**対角付近に数字が集まる．北西隅ルールとよばれる．**)

- **一方最後の列から，列優先でうめる方法を考えても同じ結果に到達する．**

グレブナー基底の視点からの解釈

- 右下の要素からできるだけ大きい値を入れて行く操作は，グレブナー基底の観点からは revlex 式の term order を考えていることにあたる．
- うめた結果の表は,「余り」(standard monomial) にあたる．つまり所与の term order のもとでもっとも「小さい」表である．
- 行優先でも，列優先でも同じ答えとなることは，異なる term order が同じ最適解を与える例となっている．

Sorting で考える (大杉・日比)

- 任意の可能解から出発する:

Table 2: 中身をうめた表の例

3	**2**	**1**	**6**
2	**1**	**2**	**5**
2	**2**	**1**	**5**
7	**5**	**4**	**16**

- これを単項式に対応させる (積型記法):

$$x_{11}^3 x_{12}^2 x_{13} x_{21}^2 x_{22} x_{23}^2 x_{31}^2 x_{32}^2 x_{33}$$

- 行の添字のみ集めてくると，周辺頻度より

  1 が 6 個，2 が 5 個，3 が 5 個

  である．列の添字は同様に

  1 が 7 個，2 が 5 個，3 が 4 個

  である．

- ここで，列の添字のほうが，グループとして，行の添字より後と考える．例えば，列の添字を $+3$ して x_{ij} を $x_{i,j+3}$ と表す．この場合，例えば x_{11} を x_{14} と表すことになる．

- その上で，行の添字と列の添字を区別せずに sort すると

  1 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 4 4 4 4 4 4 4 5 5 5 5 5 6 6 6 6

  となる.

- ここで，x_{ij} の 1 桁目の i の部分に上の数字を左から順にうめていく:

$$x_{1?}x_{1?}x_{1?}x_{1?}x_{1?}x_{1?}x_{2?}x_{2?}x_{2?}x_{2?}x_{2?}x_{3?}x_{3?}x_{3?}x_{3?}x_{3?}$$

- 一巡したので，つぎに後半を 2 桁目にうめていく

$$x_{14}x_{14}x_{14}x_{14}x_{14}x_{14}x_{24}x_{25}x_{25}x_{25}x_{25}x_{35}x_{36}x_{36}x_{36}x_{36}$$

- 最後に $4, 5, 6 \to 1, 2, 3$ と添字を戻すと

$$x_{11}^6 x_{21} x_{22}^4 x_{32} x_{33}^4$$

となり，sorting を一回施すだけで最適化の解を得る!

- この sorting の操作は何か不思議な感じのもの．
- この場合については，次のように古典的な「壷のモデル」で解釈できる．
- 壷のモデル: 使い古した考え方だが，組み合わせの基本．

問 2　いま，壷の中に，1 から 16 まで番号のふった 16 枚のカードがあったとします．各カードには 2 桁の数字を書く欄があります．ここにこれから数字を書いていきます．それぞれの欄とも $1,2,3$ のいずれかの数字を書きます．そして，第 1 桁には

1 を書くカードは 6 枚,　　2 を書くカードは 5 枚

3 を書くカードは 5 枚

としてください．また第 2 桁には

1 を書くカードは 7 枚,　　2 を書くカードは 5 枚

3 を書くカードは 4 枚

としてください．このような数字の埋め方で，最も規則的で簡単な方法を考えてください．

解答例: 第 1 桁には，カードの番号順に，1 を 6 枚に書き，2 を 5 枚に書き，3 を残りの 5 枚に書く．第 2 桁も同様に書く．

さらに，グレブナー基底の観点からすると，次の問が興味深い．

問 3　以上のような 16 枚への数字のうめ方と異なるうめ方をした場合，必ず 2 枚のカードを選ぶことができ，第 2 桁の数字をとりかえることによって，規則的な埋め方に近づけることができることを示してください．表 2 に対応する例が参考資料の最後の草稿用ページにあるので，それで確認してみてください．

難しい問題にチャレンジ!

(いわゆる分解可能モデルに一般化します．分解可能モデルは chordal graph とほぼ同値なものです．)

問 4　いま，壷の中に，1 から 16 まで番号のふった 16 枚のカードがあったとします．各カードには 4 桁の数字を書く欄があります．ここにこれから数字を書いていきます．それぞれの欄とも 1,2 のいずれかの数字を書きます．そして

- 第 (1,2) 桁には
  (1,1) の組合せ 6 枚，(1,2)3 枚，(2,1)2 枚，(2,2)5 枚
- 第 (2,3) 桁にはいずれの組合せも 4 枚づつ
- 第 (3,4) 桁には

$(1,1)3$ 枚，$(1,2)5$ 枚，$(2,1)3$ 枚，$(2,2)5$ 枚

になるようにしてください．このような数字の埋め方で，最も規則的で簡単な方法を考えてください．

解答例: $(1,2)$ 桁は単に順にうめて行く．$(2,3)$ 桁は，2 桁目の数字があうカードを順にひろいながら，3 桁目を順に埋めて行く．$(3,4)$ 桁は，3 桁目の数字があうカードを順にひろいながら，4 桁目を順に埋めて行く．

問 5(難)　　それでは，以上のような数字の埋め方と異なる埋め方をした場合，必ず 2 枚のカードを選ぶことができて，書いてある数字を入れ換えることによって，必ず以上の埋め方に近づくことができることを示してください．

Naiman-Wynn の離散チューブ (abstract tube)

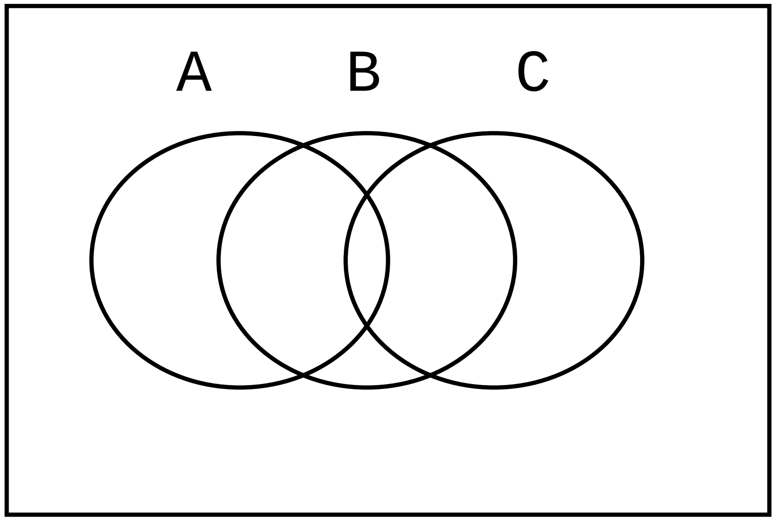
Klaus Dohmen (2003) “Improved Bonferroni Inequalities via Abstract Tubes, Inequalities and Identities of Inclusion-Exclusion Type”, Springer, Lecture Notes in Mathematics, 1826. 第 2 章の簡単な紹介．

最も簡単な例: 次のように 3 個の集合 A, B, C について

$A \cap C \subset B$ が成り立つと，包除原理は

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C)$$

と単純化される．集合の交わりに，うまい「退化」があれば，このように包除原理が単純になる．



Ω

用語と定義

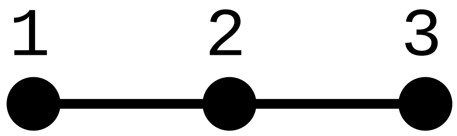
- Ω を標本空間とし，$A_1, \ldots, A_n$ を Ω の部分集合の族とする．
- $V = [n] = \{1, \ldots, n\}$ とおく．$\mathcal{S}$ を V の simplicial complex とする．
- $\mathcal{A} = \{A_1, \ldots, A_n\}$ とおく．

定義: ペア $\{\mathcal{A}, \mathcal{S}\}$ が abstract tube をなすとは，すべての $\omega \in \bigcup_{v \in V} A_v$ に対して

$$\mathcal{S}(\omega) = \{I \in \mathcal{S} \mid \omega \in \bigcap_{i \in I} A_i\}$$

が contractible (1 点可縮) であることを言う．

- 上の例では，$\mathcal{S} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{2,3\}\}$



とおけばよい．

- $\chi(B)$ を集合の定義関数とする．すなわち

$$\chi(B)(\omega) = \begin{cases} 1 & \omega \in B \\ 0 & \omega \notin B \end{cases}$$

とする．

定理 (Naiman-Wynn (1997)): $(\{A_v\}_{v\in V}, \mathcal{S})$ が abstract tube をなす時

$$\chi(\bigcup_{v\in V} A_v) \geq \sum_{\substack{I\in\mathcal{S}\\|I|\leq r}} (-1)^{|I|-1}\chi(\bigcap_{i\in I} A_i), \quad (r : \textbf{even})$$

$$\chi(\bigcup_{v\in V} A_v) \leq \sum_{\substack{I\in\mathcal{S}\\|I|\leq r}} (-1)^{|I|-1}\chi(\bigcap_{i\in I} A_i), \quad (r : \textbf{odd})$$

が成り立つ．

(つまり包除原理に現われる項を $\mathcal{S}$ に制限してしまってよい．)

- Abstract tube の定義はなかなかおもしろい
- Abstract tube を用いるには可縮性の判定をおこなう必要がある．
- Naiman-Wynn, Dohmen 等はどのような場合に可縮性が成り立つかの十分条件をいろいろと与えている．
- しかし，例えば次のような問題は未解決と思われる: 所与の $\mathcal{A}$ に関して $\{\mathcal{A}, \mathcal{S}\}$ が abstract tube となるような極小な $\mathcal{S}$ を列挙せよ．

感想など

- やさしい例で説明できる問題は話しやすい．
- 積型記法と加法型記法の違いなどを気にするのは「下手の考え休むに似たり」かも．
- 対応をつければどうせ自明に同値?
- ただし「素人」(他分野では専門性のある人) が感じる違和感には何か根拠のある場合もある．
- 日比プロジェクトもそれなりにうまく行っていると思う (宣伝)．