

分割表の分解可能モデルの個票開示問題への応用について^a

竹村 彰通

^a3 年ほど前の，遠藤祐司君との発表の内容，および原尚幸氏との発表の内容をまとめたものです．

項目

1. 多元分割表の記法
2. アメリカのセンサス個票データの例
3. 分割表のグラフィカルモデル
4. 分解可能モデル
5. センサス個票データでの計算結果
6. `swapping` による秘匿処理

1 多元分割表の記法

- m 元の分割表: $I_1 \times \cdots \times I_m$ 分割表
 - セル頻度: $f_{i_1 \dots i_m}$ あるいは $f(i_1, \dots, i_m)$
 - $\mathcal{I}_k = \{1, \dots, I_k\}$: 変数 k のカテゴリーの集合
 - セルの集合 $\mathcal{I} = \mathcal{I}_1 \times \cdots \times \mathcal{I}_m$ (直積集合)
 - 多重添字 $i = (i_1, \dots, i_m) \in \mathcal{I}$
 - セル頻度 $f(i)$
 - $|\mathcal{I}| = I_1 \times \cdots \times I_m$: 総セル数は m とともに指数的に大きくなる \Rightarrow 頻度ゼロのセルが出てくる

● 確率の導入

- $p(i_1, \dots, i_m) = p(i)$: セル i の確率 (生起確率)
- n 人の個体が互いに独立にセル i に確率 $p(i)$ で落ちてくると考える: “多項分布のモデル”
- (無条件の) 生起確率の推定値 = 相対頻度

$$\hat{p}(i) = f(i)/n$$

- 観測頻度がゼロのセルは確率が 0 か?
“ゼロカウント問題” ← 正のはず
- 観測頻度が 1 のセルの確率は $1/n$ か?
個票開示問題の関心 ← おそらくもっと小さい

- 母集団と標本: N 人の母集団から n 人の標本を観測したとする. 残りの $N - n$ 人はどこに落ちてくるか
 - 標本で観測 0 のセルは母集団でもあいかわらず 0 か?
← 観測されなかった「種」の推定問題
 - 標本で頻度 1 のセルは母集団でも頻度 1 か?
標本一意が母集団一意でもあるか?
← 母集団一意の推定問題
 - 標本一意の個体が母集団一意でもあると個体識別の可能性が生じる

2 アメリカのセンサス個票データの例

- アメリカの 1990 年の国勢調査の 1%抽出データ .
CD-ROM 販売 .
(2000 年の国勢調査のデータは web 上で入手できる .
日本の統計当局は保守的 .)
- 今回のテストデータ: ワシントン州の 1%抽出個人データから , 項目を $m = 8$ 項目, 個人を無作為に $n = 9809$ 人再抽出 .

- | | |
|-----------------|------------------|
| 1. 続柄 (14 分類) | 2. 性別 (2 分類) |
| 3. 年齢 (91 分類) | 4. 配偶関係 (5 分類) |
| 5. 出身地 (14 分類) | 6. 配偶者の有無 (7 分類) |
| 7. 実子の有無 (2 分類) | 8. 実子の年齢 (5 分類) |

- $m = 8$ 元の

$$14 \times 2 \times 91 \times 5 \times 14 \times 7 \times 2 \times 5 = 12485200$$

(約 1200 万セル) 型の分割表

- ワシントン州の人口 約 $N = 4,867,000$: 母集団サイズ
- 構造的ゼロの問題: 例 配偶者の有無

- 頻度の頻度（「サイズインデックス」）

セルサイズ	1	2	3	4	5	
頻度	2243	524	275	132	104	
	6	7	8	9	10	11以上
	60	59	34	46	19	124

- 標本中に観測されたセル総数 3620
- 2243 人の標本一意のうちで何人くらいが母集団一意でもあるか
- さらに 2243 人のうちで特に母集団一意である条件つき確率が高いのは誰か

実は変数間の相関や「構造的ゼロ」の問題が深刻

3 分割表のグラフィカルモデル

- m 個の変数の集合: $\Delta = \{1, \dots, m\}$
- 個々の変数: $\delta \in \Delta$
- δ のカテゴリーの集合: $\mathcal{I}_\delta = \{1, \dots, I_\delta\}$
- セルの集合: $\mathcal{I} = \prod_{\delta \in \Delta} \mathcal{I}_\delta$
- Δ の部分集合: $a, b, \dots,$

- a -周辺セル i_a :

$$i_a \in \mathcal{I}_a = \prod_{\delta \in a} \mathcal{I}_\delta$$

- a -周辺セル i_a の周辺頻度

$$f(i_a) = \sum_{j: j_a = i_a} f(j)$$

- $a \subset \Delta$ に対して $\mu_a : \mathcal{I} \rightarrow R$ を a に含まれる変数のみに依存する関数とする
- $G = (\Delta, E)$: Δ を頂点集合とし, 辺集合を E とする無向グラフ

- G の (極大) クリークの族を \mathcal{C} とする
- グラフィカルモデルの定義:

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

の形に表されるモデル

- グラフィカルモデルは対数線形モデルの部分モデルであり, さまざまな良い性質を持つが, 一般のグラフに関しては推定に繰り返し計算が必要となる
- 更に都合のよいサブモデルとして分解可能モデルがある

4 分解可能モデル

- 分解可能モデル

G が弦グラフ (chordal graph, triangulated graph), すなわち長さが 4 以上の閉路を持たないグラフ, であるようなグラフィカルモデル

- G を弦グラフとする時, G の minimal vertex separator S にはその重複度 $\nu(S)$ が定まる

- 弦グラフ G が連結の時, $1 + \sum_{S \in \mathcal{S}} \nu(S)$ は G の (極大) クリーク数 $|\mathcal{C}|$ に一致する
- 分解可能モデルの最尤推定値は, G が連結の場合には

$$\hat{p}(\mathbf{i}) = \frac{1}{n} \frac{\prod_{C \in \mathcal{C}} f(\mathbf{i}_C)}{\prod_{S \in \mathcal{S}} f(\mathbf{i}_S)^{\nu(S)}}$$

と有理式で明示的に書ける．ただし \mathcal{S} は G の minimal vertex separator の集合

- モデルの自由度も明示的に書ける

$$\sum_{C \in \mathcal{C}} \prod_{\delta \in C} I_{\delta} - \sum_{S \in \mathcal{S}} \nu(S) \prod_{\delta \in S} I_{\delta}$$

- モデルを選べば推定は簡単
- 対数尤度から自由度を引くことによって **AIC** も簡単

- 問題点

- $m = 8$ 程度でも分解可能モデルはたくさんある。
 m とともに弦グラフの数は急速に増える
- 多くのモデルの中でどのモデルを選ぶか、
モデル選択の基準が難しい。観測ゼロが多い状況では
AIC の理論的正当化は難しい。

Table 1: 階層モデル, グラフィカルモデル, 分解可能モデルの個数

m	階層	グラフィカル	分解可能
2	2	2	2
3	9	8	8
4	114	64	61
5	6894	1024	820
6	7785062	32768	18154
7	2414627396434	2097152	617675
8	56130437209370320359966	268435456	30888596

分解可能モデルのリストは竹村の web で公開

5 センサス個票データでの計算結果

- Ewens モデル, Pitman モデルのあてはめはサイズインデックスのみに依存する .

- これらの母集団一意数の推定値は

Ewens モデル: 6, Pitman モデル: 214

- 分解可能モデルのあてはめ: 無作為に 1 万個程度の分解可能モデルを推定して, AIC の高いもの 5 個程度を選んで推定結果を見る

- セル確率の推定値 $\hat{p}(i)$ に基づく標本一意数中の母集団一意数の推定値:

$$\sum_{i:\text{標本一意}} (1 - \hat{p}(i))^{N-n}$$

- $(1 - \hat{p}(i))^{N-n}$ は残りの $N - n$ 人の母集団から誰もセル i に落ちてこない確率

● モデル 1:

clique: {6 0 1 3}, {6 0 1 7}, {6 0 5 3}, {5 4},
{5 2}

separator: {6 0 1}, {6 0 3}, {5}, {5}

AIC/2 15073.158 対数尤度 -13009.158 自由度 2064

推定母集団一意数 54.37

標本一意セルの生起確率推定値の対数のヒストグラム

-4 ~ -5 0

-5 ~ -6 0

-6 ~ -7 1

-7 ~ -8 23

-8 ~ -9 213

-9 ~ -10 460

-10 ~ -11 498
-11 ~ -12 442
-12 ~ -13 258
-13 ~ -14 176
-14 ~ -15 87
-15 ~ -16 54
-16 ~ -17 16
-17 ~ -18 9
-18 ~ -19 5
-19 ~ -20 1
-20 ~ -21 0
-21 ~ -22 0

● モデル 2:

clique: {1 0 7 6}, {1 3 0 6}, {1 3 5 6}, {1 5 4 6},
{5 2}

separator: {1 0 6}, {1 3 6}, {1 5 6}, {5}

AIC/2 15184.494 対数尤度 -13567.494 自由度 1617

推定母集団一意数 96.35

標本一意セルの生起確率推定値の対数のヒストグラム

-4 ~ -5 0

-5 ~ -6 0

-6 ~ -7 1

-7 ~ -8 26

-8 ~ -9 202

-9 ~ -10 442

-10 ~ -11 517
-11 ~ -12 432
-12 ~ -13 238
-13 ~ -14 145
-14 ~ -15 99
-15 ~ -16 64
-16 ~ -17 37
-17 ~ -18 29
-18 ~ -19 6
-19 ~ -20 5
-20 ~ -21 0
-21 ~ -22 0

● モデル 3:

clique: {5 7 1}, {5 7 3}, {5 7 6}, {5 7 0},
{5 2 6}, {2 4}

separator: {5 7}, {5 7}, {5 7}, {5 6}, {2}

AIC/2 15433.673 対数尤度 -12291.673 自由度 3142

推定母集団一意数 41.40

標本一意セルの生起確率推定値の対数のヒストグラム

-4 ~ -5 0

-5 ~ -6 0

-6 ~ -7 0

-7 ~ -8 29

-8 ~ -9 200

-9 ~ -10 438

-10 ~ -11 555
-11 ~ -12 404
-12 ~ -13 280
-13 ~ -14 171
-14 ~ -15 99
-15 ~ -16 51
-16 ~ -17 10
-17 ~ -18 4
-18 ~ -19 2
-19 ~ -20 0
-20 ~ -21 0
-21 ~ -22 0

印象と課題

- 数値的には割合うまく動いたように思える
- それぞれのセルの生起確率が推定できるのがよい
- $m = 12$ 程度まで動かしたい．分解可能モデルの列挙は前提か．
- モデルを無作為に選ぶのではなく，当てはまりのよさそうなモデルをたどっていけないか
- そもそも構造的ゼロをまともに扱えないか

6 swapping による秘匿処理

例:

- 二つのレコード:

(男性, 年齢 55, 看護師, 東京在住)

(女性, 年齢 50, 警察官, 大阪在住)

- 「職業」を入れ換える

(男性, 年齢 55, 警察官, 東京在住)

(女性, 年齢 50, 看護師, 大阪在住)

- この場合, 職業という単一の変数のみを入れ換えており, 1次元の周辺頻度はすべて不変であるが, 職業を含

む 2次元の周辺頻度は変わってしまっている。

- 職業と年齢を同時に入れ換えると

(男性, 年齢 50, 警察官, 東京在住)

(女性, 年齢 55, 看護師, 大阪在住)

- この場合, 1次元の周辺に加えて, (年齢, 職業) の2次元の周辺も保存される。
- 目的: このような場合, 周辺をできるだけ保存したい。
あるいは, 所与の周辺分布は不変にしたい。

- $E \subset \Delta$: swap する変数集合
- 二つのレコード (あるいはセル)

$$i = (i_E, i_{E^C}), \quad j = (j_E, j_{E^C})$$

- E -swapping

$$\{(i_E, i_{E^C}), (j_E, j_{E^C})\} \rightarrow \{(i_E, j_{E^C}), (j_E, i_{E^C})\}.$$

- E -swapping によって D -周辺が不変となるための必要十分条件は以下のいずれかが成り立つこと .

$$\text{i) } D \subset E, \quad \text{ii) } D \subset E^C,$$

$$\text{iii) } i_{E \cap D} = j_{E \cap D}, \quad \text{iv) } i_{E^C \cap D} = j_{E^C \cap D}.$$

- 不変にしたい周辺が一つならばこのように簡単だが，複数の周辺を固定すると複雑になる．
- $D = \{D_1, \dots, D_r\}$: 固定したい周辺の族．階層モデルの生成集合族と概念的には同じ．
- **重要な結果:** もし D が分解可能モデルならば，二つのレコードの swapping のみを繰り返すことによって，周辺を保存する任意の他の表に変換することができる．
- この定理に基づき，前節のそれぞれのモデルでセル確率の低い50レコードについて swapping 可能かどうかを判定したところ，可能であることがわかった

- 以上では2つのレコード間での swapping を考えたが、より多くのレコード間で巡回的な swapping を考えれば、より柔軟性が大きくなる。
- 例えば4レコード間で巡回的に swapping をすれば、すべての2変数の周辺頻度を保存できる場合がある：

x_1	x_2	x_3		x_1	x_2	x_3
1	1	1		1	1	2
1	2	2	→	1	2	1
2	2	1		2	2	2
2	1	2		2	1	1

- 実は、分割表のマルコフ基底の高次の move に対応。

References

- [1] 遠藤 祐司 (2004). 「分解可能モデルの列挙アルゴリズム」 東京大学工学部計数工学科卒業論文.
- [2] Hisayuki Hara, Satoshi Aoki and Akimichi Takemura (2007). Minimal and minimal invariant Markov bases of decomposable models for contingency tables. arXiv:math/0701429v2. To appear in *Bernoulli*.
- [3] Takemura, A. and Hara, H. (2007). Conditions for swappability of records in a microdata set when some marginals are fixed. arxiv:math.ST/0603603. *Computational Statistics*, **22**, 173-185
- [4] Takemura, A. and Endo, Y. (2006). Evaluation of per-record identification risk and swappability of records in a microdata set via decomposable models. arxiv:math.ST/0603609