マルコフ基底と分割表解析への応用について

原　尚幸

東京大学・工・技術経営戦略学専攻

Nov 5 2010

於 東京大学生産技術研究所

This talk is based on some joint works with A. Takemura, S. Aoki, R. Yoshida and T. Sei

Contents

2 元分割表

- 分割表: 複数の特性値 (変数) に注目して観測値の頻度をまとめた表
- 3×3 分割表

統計 \ 解析	優	良	可	行和
優	7	5	1	13
良	10	5	6	21
可	2	6	8	16
列和	14	21	15	50

2 元分割表

- 分割表: 複数の特性値 (変数) に注目して観測値の頻度をまとめた表
- $I \times J$ 分割表 $\boldsymbol{x} = \{x_{ij}\}$

要因 1 \ 要因 2	1	$\cdots$	J	行和
1	x_{11}	$\cdots$	x_{1J}	x_{1+}
$\vdots$		$\cdots$		$\vdots$
I	x_{I1}	$\cdots$	x_{IJ}	x_{I+}
列和	x_{+1}	$\cdots$	x_{+J}	x_{++}

- x_{i+} : 行和, x_{+j} : 列和, x_{++} : 総頻度

χ^2 適合度検定

- 通常の関心：独立性の仮説

$$H_0 : p_{ij} = p_{i+} \cdot p_{+j}$$

  - 個々のセルの確率 p_{ij} が周辺確率 p_{i+}, p_{+j} の積で表される
  - このモデルを 2 元完全独立モデルと言う

- Pearson の χ^2 統計量

$$\chi^2 = \sum_{i,j} \frac{(x_{ij} - e_{ij})^2}{e_{ij}}, \quad e_{ij} := \frac{x_{i+}x_{+j}}{x_{++}}$$

  - 漸近的に自由度 $(I-1)(J-1)$ の χ^2 分布に従う
    $\Rightarrow$ これに基づいて検定

χ^2 適合度検定

- 通常の関心：独立性の仮説

$$H_0 : p_{ij} = p_{i+} \cdot p_{+j}$$

  - 個々のセルの確率 p_{ij} が周辺確率 p_{i+}, p_{+j} の積で表される
  - このモデルを 2 元完全独立モデルと言う
- Pearson の χ^2 統計量

$$\chi^2 = \sum_{i,j} \frac{(x_{ij} - e_{ij})^2}{e_{ij}}, \quad e_{ij} := \frac{x_{i+} x_{+j}}{x_{++}}$$

  - 漸近的に自由度 $(I-1)(J-1)$ の χ^2 分布に従う
    $\Rightarrow$ これに基づいて検定

正確検定

- 十分統計量

$$\boldsymbol{b} = \{x_{1+}, \ldots, x_{I+}, x_{+1}, \ldots, x_{+J}\}$$

- $\boldsymbol{b}$ を与えたときの分割表 $\boldsymbol{x}$ の条件付き分布は,
  以下の超幾何分布で明示的に与えられる

$$\Pr(\boldsymbol{x} \mid \boldsymbol{b}) = \frac{\displaystyle\prod_j \begin{pmatrix} x_{+j} \\ x_{1j} \cdots x_{Ij} \end{pmatrix}}{\begin{pmatrix} x_{++} \\ x_{+1} \cdots x_{+J} \end{pmatrix}}$$

- Fisher の正確検定
  超幾何分布に基づく検定
  - 標本数に依存しないので, 特に標本数が小さい場合に有用

正確検定の実装

- ファイバー $\mathcal{F}_{\boldsymbol{b}}$
  十分統計量 $\boldsymbol{b}$ を共有する分割表の集合
- ファイバーのすべての要素の数え上げができれば,
  検定統計量の正確分布による評価が可能
- ファイバーの要素数は一般には非常に大きいため,
  列挙による検定統計量の評価は困難

⇓

ファイバー内の状態遷移を用いて, MCMC 法により分割表をサンプリングし, それに用いて検定統計量を評価する

正確検定の実装

- ファイバー $\mathcal{F}_{\boldsymbol{b}}$
  十分統計量 $\boldsymbol{b}$ を共有する分割表の集合
- ファイバーのすべての要素の数え上げができれば,
  検定統計量の正確分布による評価が可能
- ファイバーの要素数は一般には非常に大きいため,
  列挙による検定統計量の評価は困難

$\Downarrow$

ファイバー内の状態遷移を用いて, MCMC 法により
分割表をサンプリングし, それに用いて検定統計量を
評価する

分割表の move

- 2 元表 $\boldsymbol{x} := (x_{11}, \ldots, x_{IJ}) \in \mathbb{Z}_{\geq 0}^{IJ}$
- 配置 A : 分割表 $\boldsymbol{x}$ から十分統計量 $\boldsymbol{b}$ への線形写像

$$A\boldsymbol{x} = \boldsymbol{b}$$

- move $\boldsymbol{z}$

  行和・列和がすべて 0 の整数配列 $\Leftrightarrow$ A の整数核

$$A\boldsymbol{z} = 0$$

  - move : 同一ファイバー内の 2 表の差
  - ファイバーの要素間は move で移動できる

5	3	2	10
4	2	4	10
1	5	4	10
10	10	10	30

$+$

1	-1	0	0
-1	1	0	0
0	0	0	0
0	0	0	0

$=$

6	2	2	10
3	3	4	10
1	5	4	10
10	10	10	30

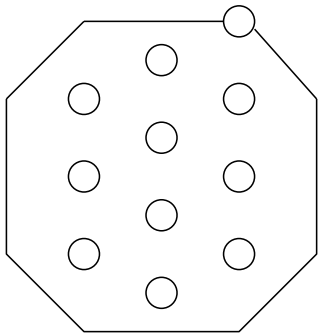
Markov 基底 $\mathcal{M}$

要素数が 2 以上のすべてのファイバーの要素を連結に結ぶ move の集合

- $\boldsymbol{x}, \boldsymbol{y}$: 同一ファイバーに属する任意の 2 表

$$\exists \boldsymbol{z}_1, \ldots, \boldsymbol{z}_K \in \mathcal{M} \text{ s.t.}$$

$$\boldsymbol{y} = \boldsymbol{x} + \sum_{k=1}^{K} \boldsymbol{z}_k, \quad \boldsymbol{x} + \sum_{k=1}^{K'} \boldsymbol{z}_k \geq 0,\ K' \leq K.$$



	i	i'
j	1	-1
j'	-1	1

  というタイプの move の集合

  $\Rightarrow$ 2 元完全独立モデルのマルコフ基底

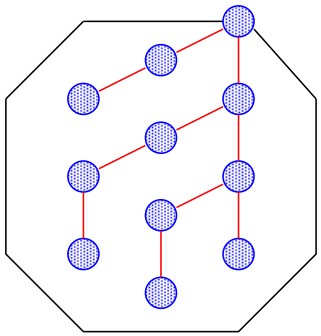
Markov 基底 $\mathcal{M}$

要素数が 2 以上のすべてのファイバーの要素を連結に結ぶ move の集合

- $\boldsymbol{x}, \boldsymbol{y}$: 同一ファイバーに属する任意の 2 表

$$\exists \boldsymbol{z}_1, \ldots, \boldsymbol{z}_K \in \mathcal{M} \text{ s.t.}$$

$$\boldsymbol{y} = \boldsymbol{x} + \sum_{k=1}^{K} \boldsymbol{z}_k, \quad \boldsymbol{x} + \sum_{k=1}^{K'} \boldsymbol{z}_k \geq 0,\ K' \leq K.$$



	i	i'
j	1	-1
j'	-1	1

  というタイプの move の集合

  $\Rightarrow$ 2 元完全独立モデルのマルコフ基底

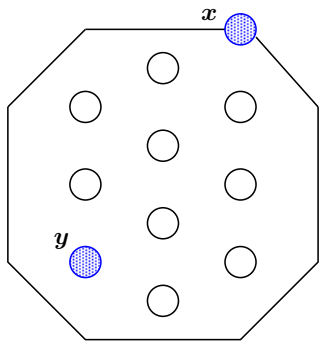
Markov 基底 $\mathcal{M}$

要素数が 2 以上のすべてのファイバーの要素を連結に結ぶ move の集合

- $\boldsymbol{x}, \boldsymbol{y}$: 同一ファイバーに属する任意の 2 表

$$\exists \boldsymbol{z}_1, \ldots, \boldsymbol{z}_K \in \mathcal{M} \text{ s.t.}$$

$$\boldsymbol{y} = \boldsymbol{x} + \sum_{k=1}^{K} \boldsymbol{z}_k, \quad \boldsymbol{x} + \sum_{k=1}^{K'} \boldsymbol{z}_k \geq 0,\ K' \leq K.$$



	i	i'
j	1	-1
j'	-1	1

  というタイプの move の集合

  $\Rightarrow$ 2 元完全独立モデルのマルコフ基底

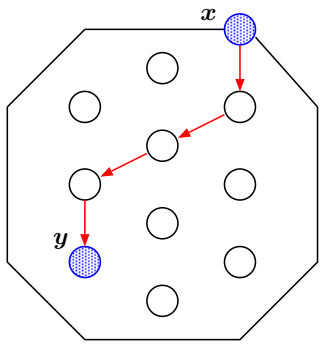
Markov 基底 $\mathcal{M}$

要素数が 2 以上のすべてのファイバーの要素を連結に結ぶ move の集合

- $\boldsymbol{x}, \boldsymbol{y}$: 同一ファイバーに属する任意の 2 表

$$\exists \boldsymbol{z}_1, \ldots, \boldsymbol{z}_K \in \mathcal{M} \ \text{ s.t.}$$

$$\boldsymbol{y} = \boldsymbol{x} + \sum_{k=1}^{K} \boldsymbol{z}_k, \quad \boldsymbol{x} + \sum_{k=1}^{K'} \boldsymbol{z}_k \geq 0, \ K' \leq K.$$



	i	i'
j	1	-1
j'	-1	1

  というタイプの move の集合

  $\Rightarrow$ 2 元完全独立モデルのマルコフ基底

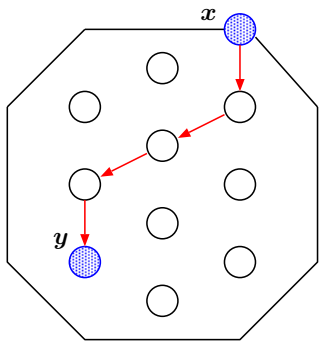
Markov 基底 $\mathcal{M}$

要素数が 2 以上のすべてのファイバーの要素を連結に結ぶ move の集合

- $\boldsymbol{x}, \boldsymbol{y}$: 同一ファイバーに属する任意の 2 表

$$\exists \boldsymbol{z}_1, \ldots, \boldsymbol{z}_K \in \mathcal{M} \text{ s.t.}$$

$$\boldsymbol{y} = \boldsymbol{x} + \sum_{k=1}^{K} \boldsymbol{z}_k, \quad \boldsymbol{x} + \sum_{k=1}^{K'} \boldsymbol{z}_k \geq 0,\ K' \leq K.$$



	i	i'
j	1	-1
j'	-1	1

  というタイプの move の集合

  $\Rightarrow$ 2 元完全独立モデルのマルコフ基底

正確検定の実装

- マルコフ基底の要素の足し引きによる状態遷移と, MCMC 法による生成確率の制御により, 任意のデータセットが属するファイバーに対して, 超幾何分布を定常分布にもつ既約なマルコフ連鎖からのサンプリングが可能になる (Diaconis and Sturmfels(1998))

[超幾何分布を定常分布に持つ分割表の発生]

Step 0. $\boldsymbol{x}$: データセット , $\Pr(\boldsymbol{x} \mid \boldsymbol{b})$: 超幾何分布

Step 1. $\boldsymbol{z} \in \mathcal{M}$ をランダムに抽出

Step 2. if $\boldsymbol{x} + \boldsymbol{z} \geq 0$

$$\boldsymbol{x} \leftarrow \boldsymbol{x} + \boldsymbol{z} \quad \text{with prob} \quad \min\left(\frac{\Pr(\boldsymbol{x} + \boldsymbol{z} \mid \boldsymbol{b})}{\Pr(\boldsymbol{x} \mid \boldsymbol{b})}, 1\right)$$

Step 1 に戻る

2 元分割表の代数的考察

- 2 元完全独立モデルの多項式表現

$$p_{ij} = p_{i+}p_{+j},\ \forall i,j \ \Leftrightarrow\ p_{ij}p_{i'j'} - p_{i'j}p_{ij'} = 0,\ \forall i \neq i', j \neq j'$$

  - モデルの空間 $\Leftrightarrow$ 2 項式の連立方程式系の解の集合

- move の 2 項式表現

	i	i'
j	1	-1
j'	-1	1

$\Leftrightarrow$

	i	i'
j	1	0
j'	0	1

$-$

	i	i'
j	0	1
j'	1	0

$$\Leftrightarrow\ \boxed{u_{ij}u_{i'j'} - u_{i'j}u_{ij'}}$$

- より一般に

$$\boldsymbol{z} = \boldsymbol{z}^+ - \boldsymbol{z}^-\ \Leftrightarrow\ \boldsymbol{u}^{\boldsymbol{z}^+} - \boldsymbol{u}^{\boldsymbol{z}^-}$$

$$\boldsymbol{u}^{\boldsymbol{z}^+} = \prod_{i,j} u_{ij}^{z_{ij}^+},\quad \boldsymbol{u}^{\boldsymbol{z}^-} = \prod_{i,j} u_{ij}^{z_{ij}^-}$$

Markov 基底の代数的定義

- 配置 $A = \{a_{kl}\}$: $d \times IJ$ 整数行列
- K : 適当な体
- $K[u_{11}, \ldots, u_{IJ}]$: 分割表の各セルに対応する多項式環
- $K[v_1, \ldots, v_d]$: 十分統計量に対応する多項式環
- 準同型写像 $\pi_A : u_{ij} \mapsto \prod_{k=1}^{d} v_k^{a_{kl}}$

定義 : トーリックイデアル

$I_A := Ker(\pi_A) =< u^{\boldsymbol{z}^+} - u^{\boldsymbol{z}^-}; A\boldsymbol{z} = 0 >$ という多項式イデアルを
モデル A に付随するトーリックイデアルと言う

Markov 基底 (Diaconis and Sturmfels(1998))

$\mathcal{M}$ がモデル A の Markov 基底 $\Leftrightarrow$ $\mathcal{M}$ が I_A の生成系

Markov 基底を用いた正確検定の課題

- Markov 基底が与えられれば正確検定は実装可能
- Markov 基底は計算代数ソフトウェアで計算が可能
- しかし計算コストが非常に高く, セル数が 100 を越えると実用時間内での計算が不可能なのが現状
- モデルが大きくなると Markov 基底の要素数も膨大になり, すべての要素をメモリに蓄積してサンプリングすることも現実的ではない

$\Downarrow$

主たる研究課題

- Markov 基底の構造分析 (極小性・不変性)
- 適応的 move 生成アルゴリズムの開発

★ Markov 基底の構造は一般には非常に複雑

Markov 基底を用いた正確検定の課題

- Markov 基底が与えられれば正確検定は実装可能
- Markov 基底は計算代数ソフトウェアで計算が可能
- しかし計算コストが非常に高く, セル数が 100 を越えると実用時間内での計算が不可能なのが現状
- モデルが大きくなると Markov 基底の要素数も膨大になり, すべての要素をメモリに蓄積してサンプリングすることも現実的ではない

$\Downarrow$

主たる研究課題

- Markov 基底の構造分析 (極小性・不変性)
- 適応的 move 生成アルゴリズムの開発

★ Markov 基底の構造は一般には非常に複雑

Markov 基底を用いた正確検定の課題

- Markov 基底が与えられれば正確検定は実装可能
- Markov 基底は計算代数ソフトウェアで計算が可能
- しかし計算コストが非常に高く, セル数が 100 を越えると実用時間内での計算が不可能なのが現状
- モデルが大きくなると Markov 基底の要素数も膨大になり, すべての要素をメモリに蓄積してサンプリングすることも現実的ではない

$\Downarrow$

主たる研究課題

- Markov 基底の構造分析 (極小性・不変性)
- 適応的 move 生成アルゴリズムの開発

★ Markov 基底の構造は一般には非常に複雑

階層モデル

- Δ：変数集合の部分集合族からなる単体的複体
- $\mathcal{D}_\Delta$：Δ の 極大要素 (facet) の集合
- 階層モデル L_Δ :

$$\log p(\boldsymbol{i}) := \sum_{D \in \mathcal{D}_\Delta} \mu_D(\boldsymbol{i}), \quad \boldsymbol{i}：\text{各セル}$$

  - グラフィカルモデル
    $\Leftrightarrow$ $\mathcal{D}_\Delta$ はグラフのクリークの集合
  - 分解可能モデル
    $\Leftrightarrow$ コーダルグラフに対するグラフィカルモデル
  - 二元完全独立モデル
    2 点のみからなるグラフに対応する分解可能モデル

階層モデルのマルコフ基底

- 分解可能モデルは, 2 次のマルコフ基底を持つことが知られている (Dobra (2003))
  - 2 次のマルコフ基底

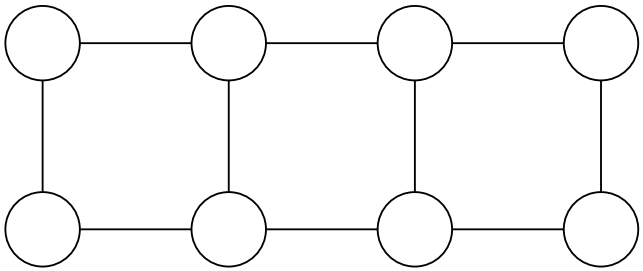
    1 が 2 つ, -1 が 2 つの move のみからなるマルコフ基底

	i	i'
j	1	-1
j'	-1	1

  - モデルに対応するトーリックイデアルが 2 次の 2 項式からなる生成系を持つ
- 分解可能モデル以外のマルコフ基底はほとんど知られていない

マルコフ基底の局所計算

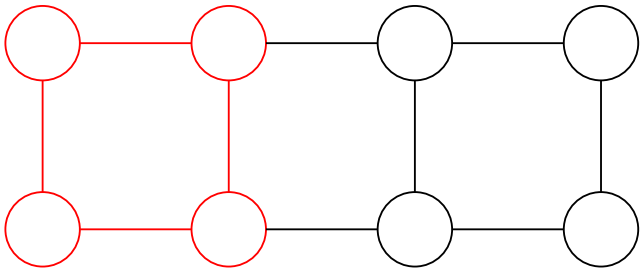
- Dobra and Sullivant (2004)
  グラフィカルモデルの場合, 極大な素部分グラフに対応する周辺モデルのマルコフ基底から, 全体のモデルのマルコフ基底が計算可能



- 素グラフに対するグラフィカルモデルのマルコフ基底が与えられれば, すべてのグラフィカルモデルのマルコフ基底が計算可能
- Hara, Sei and Takemura (2009) ではもう少し一般的なモデルの場合に拡張

マルコフ基底の局所計算

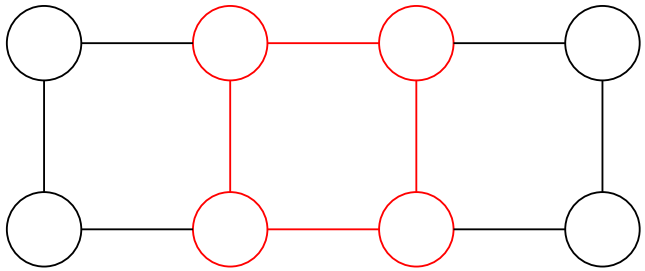
- Dobra and Sullivant (2004)
  グラフィカルモデルの場合, 極大な素部分グラフに対応する周辺モデルのマルコフ基底から, 全体のモデルのマルコフ基底が計算可能



- 素グラフに対するグラフィカルモデルのマルコフ基底が与えられれば, すべてのグラフィカルモデルのマルコフ基底が計算可能
- Hara, Sei and Takemura (2009) ではもう少し一般的なモデルの場合に拡張

マルコフ基底の局所計算

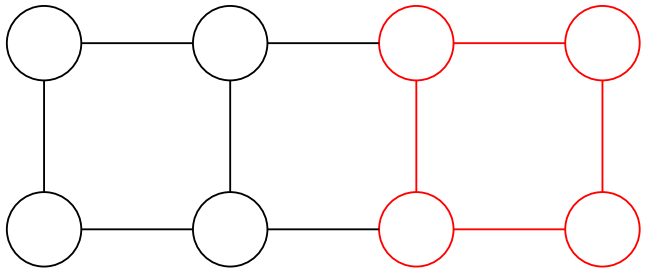
- Dobra and Sullivant (2004)
  グラフィカルモデルの場合, 極大な素部分グラフに対応する周辺モデルのマルコフ基底から, 全体のモデルのマルコフ基底が計算可能



- 素グラフに対するグラフィカルモデルのマルコフ基底が与えられれば, すべてのグラフィカルモデルのマルコフ基底が計算可能
- Hara, Sei and Takemura (2009) ではもう少し一般的なモデルの場合に拡張

マルコフ基底の局所計算

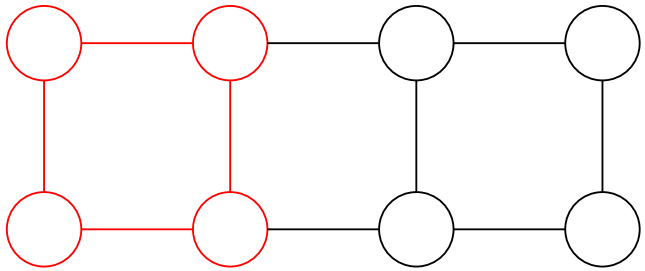
- Dobra and Sullivant (2004)
  グラフィカルモデルの場合, 極大な素部分グラフに対応する周辺モデルのマルコフ基底から, 全体のモデルのマルコフ基底が計算可能



- 素グラフに対するグラフィカルモデルのマルコフ基底が与えられれば, すべてのグラフィカルモデルのマルコフ基底が計算可能
- Hara, Sei and Takemura (2009) ではもう少し一般的なモデルの場合に拡張

マルコフ基底の局所計算

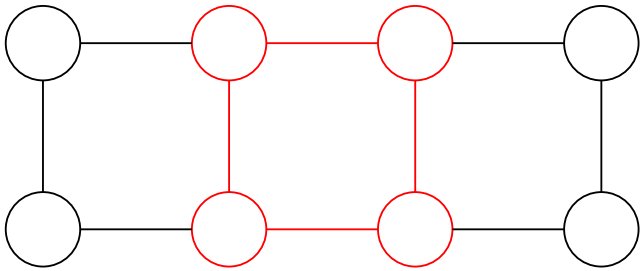
- Dobra and Sullivant (2004)
  グラフィカルモデルの場合, 極大な素部分グラフに対応する周辺モデルのマルコフ基底から, 全体のモデルのマルコフ基底が計算可能



- 素グラフに対するグラフィカルモデルのマルコフ基底が与えられれば, すべてのグラフィカルモデルのマルコフ基底が計算可能
- Hara, Sei and Takemura (2009) ではもう少し一般的なモデルの場合に拡張

マルコフ基底の局所計算

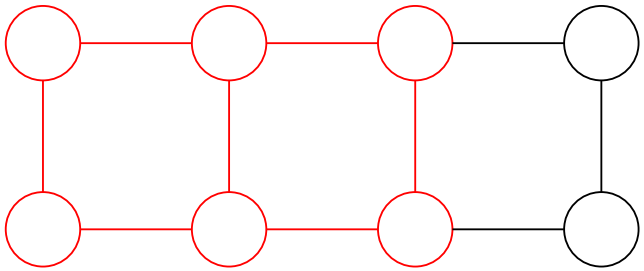
- Dobra and Sullivant (2004)
  グラフィカルモデルの場合, 極大な素部分グラフに対応する周辺モデルのマルコフ基底から, 全体のモデルのマルコフ基底が計算可能



- 素グラフに対するグラフィカルモデルのマルコフ基底が与えられれば, すべてのグラフィカルモデルのマルコフ基底が計算可能
- Hara, Sei and Takemura (2009) ではもう少し一般的なモデルの場合に拡張

マルコフ基底の局所計算

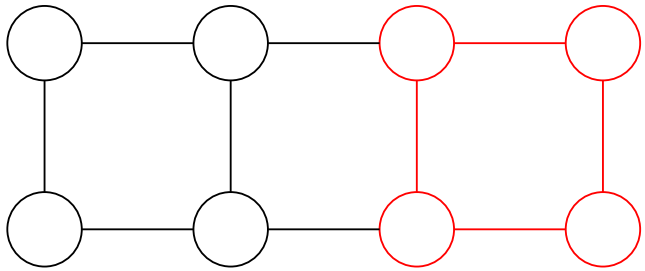
- Dobra and Sullivant (2004)
  グラフィカルモデルの場合, 極大な素部分グラフに対応する周辺モデルのマルコフ基底から, 全体のモデルのマルコフ基底が計算可能



- 素グラフに対するグラフィカルモデルのマルコフ基底が与えられれば, すべてのグラフィカルモデルのマルコフ基底が計算可能
- Hara, Sei and Takemura (2009) ではもう少し一般的なモデルの場合に拡張

マルコフ基底の局所計算

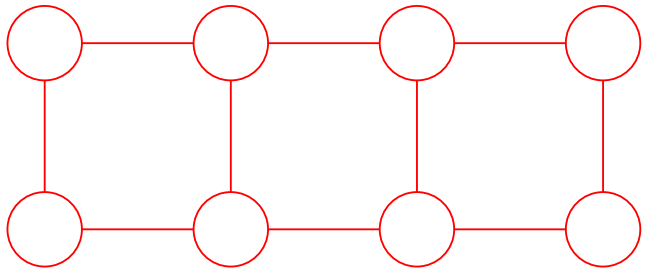
- Dobra and Sullivant (2004)
  グラフィカルモデルの場合, 極大な素部分グラフに対応する周辺モデルのマルコフ基底から, 全体のモデルのマルコフ基底が計算可能



- 素グラフに対するグラフィカルモデルのマルコフ基底が与えられれば, すべてのグラフィカルモデルのマルコフ基底が計算可能
- Hara, Sei and Takemura (2009) ではもう少し一般的なモデルの場合に拡張

マルコフ基底の局所計算

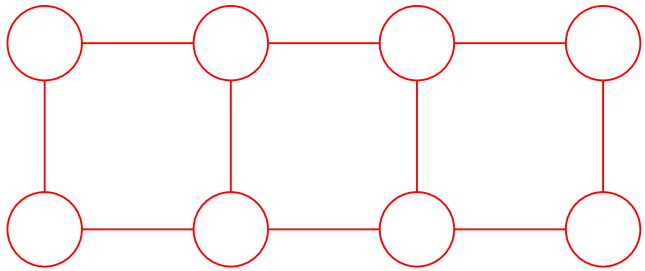
- Dobra and Sullivant (2004)
  グラフィカルモデルの場合, 極大な素部分グラフに対応する周辺モデルのマルコフ基底から, 全体のモデルのマルコフ基底が計算可能



- 素グラフに対するグラフィカルモデルのマルコフ基底が与えられれば, すべてのグラフィカルモデルのマルコフ基底が計算可能
- Hara, Sei and Takemura (2009) ではもう少し一般的なモデルの場合に拡張

マルコフ基底の局所計算

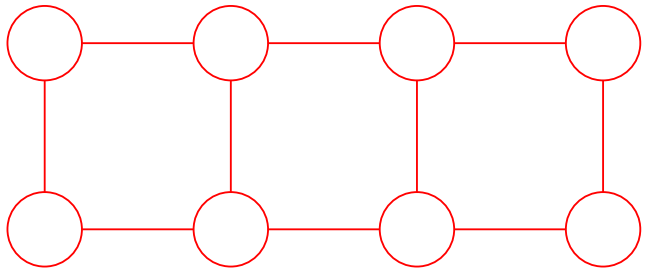
- Dobra and Sullivant (2004)
  グラフィカルモデルの場合, 極大な素部分グラフに対応する周辺モデルのマルコフ基底から, 全体のモデルのマルコフ基底が計算可能



- 素グラフに対するグラフィカルモデルのマルコフ基底が与えられれば, すべてのグラフィカルモデルのマルコフ基底が計算可能
- Hara, Sei and Takemura (2009) ではもう少し一般的なモデルの場合に拡張

マルコフ基底の局所計算

- Dobra and Sullivant (2004)
  グラフィカルモデルの場合, 極大な素部分グラフに対応する周辺モデルのマルコフ基底から, 全体のモデルのマルコフ基底が計算可能



- 素グラフに対するグラフィカルモデルのマルコフ基底が与えられれば, すべてのグラフィカルモデルのマルコフ基底が計算可能
- Hara, Sei and Takemura (2009) ではもう少し一般的なモデルの場合に拡張

我々の成果

- 比較的小さい次元のモデルの Markov 基底の構造分析
  - 分解可能モデル
    Hara, Aoki and Takemura (2010)
  - subtable sum model
    Hara, Takemura and Yoshida (2009a,b)
  - Poisson 回帰モデル, logistic 回帰モデル
    Hara, Takemura and Yoshida (2010)
  - zero-one table
    Hara and Takemura (2010a)
  - split model
    Hara, Sei and Takemura (2009)
  - 同次マルコフ連鎖モデル
    Takemura and Hara (2010)
    Hara and Takemura (2010b)
- 個票秘匿問題への応用
  Takemura and Hara (2007)

研究のアプローチ

1. 強引にマルコフ基底を導出する
2. 局所計算アルゴリズム
3. マルコフ基底の複雑さの評価
   - 最大次数, 最小次数の下限 etc
4. マルコフ基底の複雑さによるモデルの分類
   - 構造的ゼロの入り方
5. 実用的なマルコフ基底の部分集合を求める

研究のアプローチ

1. 強引にマルコフ基底を導出する
2. 局所計算アルゴリズム
3. マルコフ基底の複雑さの評価
   - 最大次数, 最小次数の下限 etc
4. マルコフ基底の複雑さによるモデルの分類
   - 構造的ゼロの入り方
5. 実用的なマルコフ基底の部分集合を求める

Markov 基底

要素数が 2 以上のすべてのファイバーにおいて,
ファイバー内の要素を連結に結ぶ move の集合

- 一般にマルコフ基底の構造は非常に複雑
- Markov 基底の定義はすべてのファイバーを連結に結ぶことを要求
- Markov 基底の move の多くは特殊なファイバーでのみ必要
- 実用上は対象とするデータが属するファイバーが連結に結ばれればよい
- 特定のファイバーの集合に着目すれば, それらが単純な構造の move のみの集合からなる Markov 基底の部分集合によって連結に結ばれることがある

  ex. logistic 回帰モデル

心臓疾患発症に関するデータ

	Blood Pressure	Serum Cholesterol (mg/100ml) 1	2	3	4	5	6	7
		< 200	200-209	210-219	220-244	245-259	260-284	> 284
1	< 117	2/53	0/21	0/15	0/20	0/14	1/22	0/11
2	117-126	0/66	2/27	1/25	8/69	0/24	5/22	1/19
3	127-136	2/59	0/34	2/21	2/83	0/33	2/26	4/28
4	137-146	1/65	0/19	0/26	6/81	3/23	2/34	4/23
5	147-156	2/37	0/16	0/6	3/29	2/19	4/16	1/16
6	157-166	1/13	0/10	0/11	1/15	0/11	2/13	4/12
7	167-186	3/21	0/5	0/11	2/27	2/5	6/16	3/14
8	> 186	1/5	0/1	3/6	1/10	1/7	1/7	1/7

Source : Cornfield(1962)

- 2/53 : 53 人中 2 人が発症
- $2 \times 8 \times 7$ 表とみなすことができる
- 説明変数は離散・等間隔
- 心臓疾患と血圧・血中コレステロール値の関係は？

ロジスティック回帰モデル

- p_{1jk}：血中コレステロール j, 血圧 k の人の発症確率
- 2 変量ロジスティック回帰モデル

$$\log\left(\frac{p_{1jk}}{1-p_{1jk}}\right) = \mu + \alpha j + \beta k,$$

- 1 変量ロジスティック回帰モデル

$$\log\left(\frac{p_{1jk}}{1-p_{1jk}}\right) = \mu + \alpha j$$

マルコフ基底を用いた適合度検定

- H_0 : 1 変量 vs H_1 : 2 変量 という検定をマルコフ基底を用いて行うことを考える
- 1 変量ロジスティック回帰モデル: $p_{ij} = p_{ijk}$

$$\log\left(\frac{p_{1j}}{1-p_{1j}}\right) = \log\left(\frac{p_{1jk}}{1-p_{1jk}}\right) = \mu + \alpha j$$

- 周辺表 $\boldsymbol{x} = \{x_{ij}\}, i = 1, 2, j = 1, \ldots, J$

	Serum Cholesterol (mg/100ml)						
	1	2	3	4	5	6	7
incidence	< 200	200-209	210-219	220-244	245-259	260-284	> 284
+	12	2	6	23	8	23	18
−	307	131	115	311	128	133	112

マルコフ基底を用いた適合度検定

- H_0 : 1 変量 vs H_1 : 2 変量 という検定をマルコフ基底を用いて行うことを考える
- 1 変量ロジスティック回帰モデル: $p_{ij} = p_{ijk}$

$$\log\left(\frac{p_{1j}}{1-p_{1j}}\right) = \log\left(\frac{p_{1jk}}{1-p_{1jk}}\right) = \mu + \alpha j$$

- 周辺表 $\boldsymbol{x} = \{x_{ij}\}$, $i = 1, 2$, $j = 1, \ldots, J$

	Serum Cholesterol (mg/100ml)						
	1	2	3	4	5	6	7
incidence	<200	200-209	210-219	220-244	245-259	260-284	>284
+	12	2	6	23	8	23	18
−	307	131	115	311	128	133	112

マルコフ基底を用いた適合度検定

- 十分統計量: $x_{1+}, \sum_{j=1}^{J} j x_{1j}$
- $x_{+1}, \ldots, x_{+J}$ も固定
- $\boldsymbol{b} := (x_{1+}, x_{+1}, \ldots, x_{+J}, \sum_{j=1}^{J} j x_{1j})$
- ファイバー：$\boldsymbol{b}$ を共有する分割表の集合
- 検定統計量：尤度比検定統計量 L_α
- $\boldsymbol{b}$ を与えたときの $\boldsymbol{x}$ の条件付分布

$$\Pr(\boldsymbol{x} \mid \boldsymbol{b}) \propto \prod_{j=1}^{J} \left(\frac{x_{+j}!}{x_{1j}! x_{2j}!} \right) \quad \Rightarrow \text{hypergeometric dist.}$$

- L_α の p-value を $\Pr(\boldsymbol{x} \mid \boldsymbol{b})$ に基づいて評価したい

move

- 1 変量ロジスティック回帰モデルの move $z = \{z_{ij}\}$ は
  $z_{1+} = 0,\ z_{+1} = 0, \ldots, z_{+J} = 0,\ \sum_{j=1}^{J} j z_{+j} = 0$
  を満たす

	12	2	6	23	8	23	18	92
	307	131	115	311	128	133	112	1237
	319	133	121	334	136	156	130	
+	1	−1	0	0	0	−1	1	0
	−1	1	0	0	0	1	−1	0
	0	0	0	0	0	0	0	
=	13	1	6	23	8	22	19	92
	306	132	115	311	128	134	111	1237
	319	133	121	334	136	156	130	

$\left(\sum_{j=1}^{7} j x_{1j} = 430\right)$

$\left(\sum_{j=1}^{7} j z_{1j} = 0\right)$ → degree 4 move

$\left(\sum_{j=1}^{7} j x_{1j} = 430\right)$

Logistic 回帰モデルのマルコフ基底

- Logistic 回帰モデルの Markov 基底の move の構造は非常に複雑で要素数も膨大

	10	11	12	13	14	15	16
最大次数	18	20	22	24	26	28	30
move の総数	1830	3916	8569	16968	34355	66066	123330

- Markov 基底のほとんどの move は $x_{+j} = 0$ という周辺セルが存在するファイバーでのみ必要
- 実験データなどでは $x_{+j} > 0$ が仮定できる場合が多い
- Hara, Takemura and Yoshida (2010)
  $x_{+j} > 0$ を満たすファイバーに限れば, 非常にシンプルな構造の move のみからなる Markov 基底の部分集合によって, すべてのファイバーを連結に結ぶことができ, MCMC による正確検定の実装が可能

Logistic 回帰モデルのマルコフ基底

- Logistic 回帰モデルの Markov 基底の move の構造は非常に複雑で要素数も膨大

	10	11	12	13	14	15	16
最大次数	18	20	22	24	26	28	30
move の総数	1830	3916	8569	16968	34355	66066	123330

- Markov 基底のほとんどの move は $x_{+j} = 0$ という周辺セルが存在するファイバーでのみ必要
- 実験データなどでは $x_{+j} > 0$ が仮定できる場合が多い
- Hara, Takemura and Yoshida (2010)
  $x_{+j} > 0$ を満たすファイバーに限れば, 非常にシンプルな構造の move のみからなる Markov 基底の部分集合によって, すべてのファイバーを連結に結ぶことができ, MCMC による正確検定の実装が可能

HTY(2010) の主定理

Theorem (HTY(2010))

$$z(j_1, j_2; k) := \begin{array}{|cccc|} \multicolumn{4}{c}{\begin{array}{cccc} j_1 & j_1+k & j_2-k & j_2 \end{array}} \\ \hline 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 \\ \hline \end{array}$$

という move 全体の集合 は $(x_{+1}, \ldots, x_{+J}) > 0$ を満たす
すべてのファイバーを連結に結ぶ

- 説明変数が 2 つの場合まで拡張が可能 (HTY(2010))
- 説明変数が dummy の場合は 3 つの場合まで拡張が可能
- 説明変数 3 つ以上, 多項ロジットへの一般化は今後の課題

心臓病発症に関するデータ

		Serum Cholesterol (mg/100ml)						
	Blood	1	2	3	4	5	6	7
	Pressure	< 200	200-209	210-219	220-244	245-259	260-284	> 284
1	< 117	2/53	0/21	0/15	0/20	0/14	1/22	0/11
2	117-126	0/66	2/27	1/25	8/69	0/24	5/22	1/19
3	127-136	2/59	0/34	2/21	2/83	0/33	2/26	4/28
4	137-146	1/65	0/19	0/26	6/81	3/23	2/34	4/23
5	147-156	2/37	0/16	0/6	3/29	2/19	4/16	1/16
6	157-166	1/13	0/10	0/11	1/15	0/11	2/13	4/12
7	167-186	3/21	0/5	0/11	2/27	2/5	6/16	3/14
8	> 186	1/5	0/1	3/6	1/10	1/7	1/7	1/7

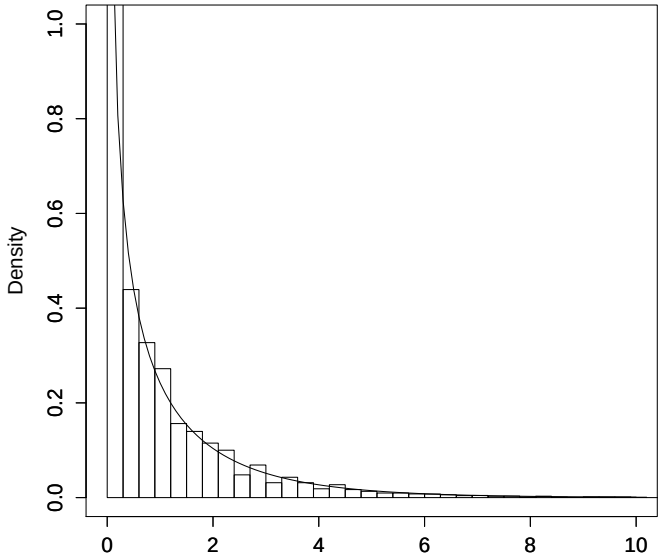
Source : Cornfield(1962)

- 2 変量ロジスティック回帰モデル

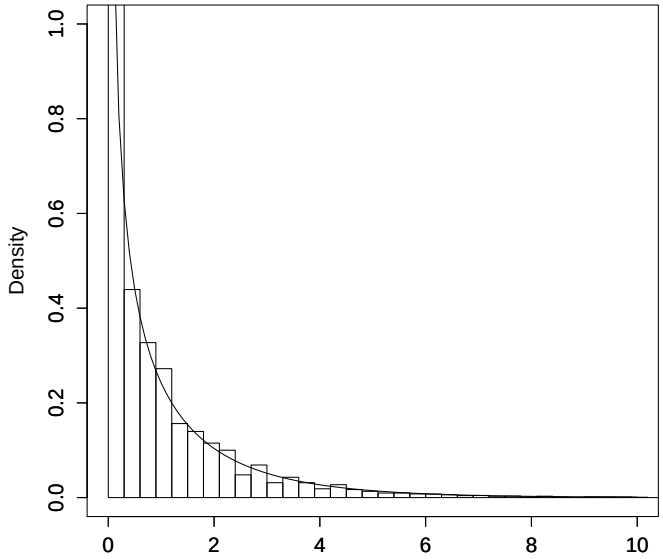
$$\log\left(\frac{p_{1jk}}{1-p_{1jk}}\right) = \mu + \alpha j + \beta k$$

- $H_\alpha : \alpha = 0$, $H_\beta : \beta = 0$ を尤度比統計量 L_α, L_β でそれぞれ検定

結果



L_α



L_β

- $L_\alpha = 18.09,\ L_\beta = 22.56$
- p-value

	χ^2_1	MCMC
L_α	2.107×10^{-5}	1.0×10^{-6}
L_β	2.037×10^{-6}	1.0×10^{-6}

結論と今後の課題

- 本講ではマルコフ基底を概説し, これまでの研究の展開を主として応用の観点から整理した
- 未解決な問題 (で, かつ解けそうなもの) はまだまだ多い
- 今後の課題としては
  - マルコフ基底の複雑さの評価
  - 格子基底を用いた正確検定アルゴリズムの開発 etc

References


A. Takemura and H. Hara (2007)
Conditions for swappability of records in a microdata set when some marginals are fixed
Computational Statistics, **22**, 173-185.


H. Hara, A. Takemura and R. Yoshida (2009a)
Markov Bases for Two-way Subtable Sum Problems
J. Pure Appl. Algebra, **213**, 1507-1521.


H. Hara, A. Takemura and R. Yoshida (2009b)
A Markov basis for conditional test of common diagonal effect in quasi-independence model for two-way contingency tables
Comput. Statist. Data Anal., **53**, 1006–1024.


Hara, H., Takemura, A. and Yoshida, R.(2010).
On connectivity of fibers with positive marginals in multiple logistic regression
J. Multivariate Anal., **101**, 909–925.


Hara, H., Aoki, S. and Takemura, A.(2010).
Minimal and minimal invariant Markov bases of decomposable models for contingency tables
Bernoulli, **16**, 208–233.

References

Hara, H. and Takemura, A.(2009).
Connecting tables with zero-one entries by a subset of a Markov basis
Algebraic Methods in Statistics and Probability (II - Urbana Volume) in AMS Contemporary Mathematics Series, to appear.

H. Hara, T. Sei and A. Takemura (2009)
Hierarchical subspace models for contingency tables
arXiv 0909.4821, submitted.

A. Takemura and H. Hara (2010)
Markov chain Monte Carlo test of toric homogeneous Markov chains
arXiv 1004.3599, submitted.

Hara, H. and Takemura, A.(2010).
A Markov basis for two-state toric homogeneous Markov chain model without initial parameters
arXiv 1005.1717, submitted.