マルコフ基底と実験計画, 問題集その 1, 2009.09.15

1. 2×3 分割表で行和、列和が固定されているものを考える。つまり

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

を考える。

   (a) Markov 基底をトーリックイデアル I_A のグレブナ基底を計算することによりもとめよ。

   (b) `4ti2` を使って Markov 基底を計算せよ。

   (c) $Ax = \beta,\ x \in \mathbf{N}_0^6$ を満たす x を $\beta = (2,3;1,2,2)$ についてすべて求めよ.

   (d) ファイバー $\mathcal{F}_{(2,3;1,2,2)}$ の連結グラフを書け。

   参考文献: [5, chap.3], [3, chap.5].

2. $3 \times 3 \times 3$ 分割表で全ての 2 次元周辺頻度 $\{x_{ij+}\}, \{x_{i+k}\}\{x_{+jk}\}$ が固定されているものを考える。

   (a) この分割表に対応する行列 A を求めよ。

   データ：`3x3x3cont.mat`

   (b) `4ti2` を使って Markov 基底を計算せよ。

   (c) $x_{ij+} = 1, x_{i+k} = 1, x_{+jk} = 1$ を満たす分割表を全て列挙せよ。

   (d) (研究課題) 与えられた Markov 基底と分割表の 1 つからファイバーのすべての元を列挙するプログラムをつくれ。またそれから連結グラフを表示するプログラムを書け。

   プログラム：`enumerate_fiber`, `connect2` を用いよ。

3. 離散確率変数 x_1, x_2, x_3 、各変数は値として $+1$ もしくは -1 をとるとする。(x_1, x_2, x_3) の発生確率を次のようにとる。

$$P(x_1, x_2, x_3) = \frac{\exp(0.2(x_1x_2 + x_2x_3 + x_1x_3))}{Z}$$

ここで Z は正規化定数とよばれるもので

$$Z = \sum_{(x_1,x_2,x_3) \in \{+1,-1\}^3} \exp(0.2(x_1x_2 + x_2x_3 + x_1x_3))$$

である。この分布に従うような (x_1, x_2, x_3) の列をマルコフ連鎖モンテカルロ法を用いてサンプリングし、$\{+1, -1\}^3$ の 8 点の出現割合がそれぞれ $P(x_1, x_2, x_3)$ に近くなることを確かめよ。

この場合のマルコフ連鎖モンテカルロ法のアルゴリズムは次の通り。

0. $\mathbf{x} = (x_1, x_2, x_3)$ に適当な初期状態を設定。
   例えば、 $\mathbf{x} \leftarrow (-1, -1, -1)$
1. x_1, x_2, x_3 のいづれかをランダムに選択。
   選んだ変数を x_i とすれば、x_i の値を $-x_i$ に置き換えたものを $\mathbf{x}'$ とする。
2. $r \leftarrow \frac{P(\mathbf{x}')}{P(\mathbf{x})}$
3. $0 \leq R < 1$ の一様乱数 R をとる
4. $r > R$ であれば、$\mathbf{x}_{\mathrm{next}} \leftarrow \mathbf{x}'$
   それ以外の場合、$\mathbf{x}_{\mathrm{next}} \leftarrow \mathbf{x}$
5. $\mathbf{x}_{\mathrm{next}}$ がサンプルとして得られる
6. $\mathbf{x} \leftarrow \mathbf{x}_{\mathrm{next}}$
   step 1 に戻る

適当なプログラミング言語で上のアルゴリズムを実装するか、もしくは、プログラム: `metropolis` を用いよ。

参考文献: [4, p.8]

4. 2×3 分割表 周辺頻度 $(5, 15; 5, 5, 10)$ のファイバー $\mathcal{F}$、すなわち

$$\mathcal{F} = \{(x_{ij}) \mid x_{ij} \in \mathbf{N}, x_{1+} = 5, x_{2+} = 15, x_{+1} = 5, x_{+2} = 5, x_{+3} = 10\}$$

から、分布 $\frac{1}{Z}\frac{1}{x_{11}!x_{12}!x_{13}!x_{21}!x_{22}!x_{23}!}$ に従って元をマルコフ連鎖モンテカルロ法を用いてサンプリングし、$2x_{11} + x_{12}$ の期待値を推定せよ。(Z は正規化定数で $\sum_{x \in \mathcal{F}} \frac{1}{x_{11}!x_{12}!x_{13}!x_{21}!x_{22}!x_{23}!}$)

プログラム: `2x3mcmc` を用いよ。

5. 2×2 分割表でデータ

疾病/喫煙	あり	なし	
あり	3	1	4
なし	2	4	6
	5	5	10

が与えられているとする。

(a) ファイバー $\mathcal{F}_{(4,6;,5,5)}$ の要素を全列挙し、
H_0 : 喫煙と疾病は無関係

が真と仮定した時のそれぞれの出現確率 $p(x) = \frac{1}{Z}\frac{1}{x_{11}!x_{12}!x_{21}!x_{22}!}$ を求めよ。ただし、Z は正規化定数である。

(b) 「$X_{11} > c$ ならば H_0 を棄却」という検定方式を考える。与えられたデータの p 値を計算せよ。

6. 5×5 2 元分割表でデータ

幾何/代数	5	4	3	2	1	計
5	2	1	1	0	0	4
4	8	3	3	0	0	14
3	0	2	1	1	1	5
2	0	0	0	1	1	2
1	0	0	0	0	1	1
計	10	6	5	2	3	26

が与えられているとする。([5], [6] より引用)

帰無仮説 (H_0)：代数と幾何の成績は独立。

とする。

(a) H_0 の下での当てはめ値を計算せよ。

(b) カイ 2 乗適合度統計量 $\chi^2(x)$ を計算せよ。

(c) 上の結果から漸近的 p 値を求めよ。

(d) マルコフ連鎖モンテカルロ法を用いてサンプリングを行い、p 値を近似的に求めよ。

プログラム：`5x5mcmc` を用いよ。

参考文献：[6], [5, chap.3]

7. 2^{7-3} 一部実施計画 (要因 A, B, C, D, E, F, G, 別名関係 $ABDE = ACDF = BCDG = I$) のデータとして次のようなものが与えられているとする。([1] より引用)

run	A	B	C	D	E	F	G	y
1	1	1	1	1	1	1	1	69
2	1	1	1	-1	-1	-1	-1	31
3	1	1	-1	1	1	-1	-1	55
4	1	1	-1	-1	-1	1	1	149
5	1	-1	1	1	-1	1	-1	46
6	1	-1	1	-1	1	-1	1	43
7	1	-1	-1	1	-1	-1	1	118
8	1	-1	-1	-1	1	1	-1	30
9	-1	1	1	1	-1	-1	1	43
10	-1	1	1	-1	1	1	-1	45
11	-1	1	-1	1	-1	1	-1	71
12	-1	1	-1	-1	1	-1	1	380
13	-1	-1	1	1	1	-1	-1	37
14	-1	-1	1	-1	-1	1	1	36
15	-1	-1	-1	1	1	1	1	212
16	-1	-1	-1	-1	-1	-1	-1	52

階層モデル H_0： $AC/BD/E/F/G$ のデータへの当てはまりを、以下の手順で検証せよ。

(a) モデル H_0 に対する共変量行列 M を求めよ。

データ：`covariate_mat1.mat`, `2_7-3.dat`

(b) モデル H_0 の下での当てはめ値 $\{m_1, \ldots, m_{16}\}$ を求めよ。

(c) 与えられたデータ $\{y_1, \ldots, y_{16}\}$ について、尤度比検定統計量

$$G(y) = 2\sum_{i=1}^{16} y_i \log \frac{y_i}{m_i}$$

を計算せよ。また、この値を、自由度 6 のカイ 2 乗分布の上側 5% 点と比較せよ。

(d) M' の Markov 基底を求めよ。

(e) データに対するモデル H_0 の当てはまりを、尤度比検定統計量にもとづき検証する。マルコフ連鎖モンテカルロ法により、このデータの p 値を計算せよ。

プログラム：`cov1_mcmc` を用いよ。

さらに、別の階層モデル $AB/AC/BD/E/F/G$ についても、上と同様の検定を行え。

データ：`covariate_mat2.mat`

参考文献：[1]

8. (研究課題, 社会階層の移動のマルコフモデル) L (level), G (generation) を自然数とする. i を $0 \leq i < L^G$ をみたす整数とするとき, この数の G 桁 L 進数への展開を $i_{G-1} \cdots i_1 i_0$ とする. 次のような式できまる affine toric variety を考える.

$$x_{i_{G-1} \cdots i_1 i_0} = p_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{G-2} i_{G-1}}$$

ここで x 変数は L^G 個, 添字 1 つの p 変数が L 個, 添字 2 つの p 変数が L^2 個. x 変数の 2 項関係式をきめる行列 A を求め (参考：`sstatus2.rr`), この A に対して `4ti2` で Markov 基底を計算せよ. どの位のサイズの L, G まで計算ができるか?

参考文献

[1] S.Aoki, A.Takemura, Markov chain Monte Carlo tests for designed experiments, arXiv:math.ST0611463v1

[2] S.Aoki, A.Takemura, Markov basis for design of experiments with three-level factors, arXiv:0709.4323v2

[3] B.Sturmfels, Gröbner Bases and Convex Polytopes, 1995, American Mathematical Society, University lecture series 8.

[4] 伊庭 幸人, 種村 正美, 大森 裕浩, 和合 肇, 佐藤 整尚, 高橋 昭彦, “統計科学のフロンティア 12, 計算統計 II マルコフ連鎖モンテカルロ法とその周辺”, 岩波書店, 2005

[5] 日比孝之編, グレブナー基底の現在, 数学書房, 2006

[6] 青木敏, マルコフ連鎖、モンテカルロ法による分割表の解析, 応用統計学会第 27 回シンポジウム発表資料, 2005