Vol.2014-MPS-100 No.13
2014/9/25

分割表のマルコフ基底における計算の一考察

大沢 泰貴[†1,a)]　大谷 康介[†1]　松田 健[†1,b)]

概要：分割表で表現されるデータ集合の妥当性は, マルコフ基底を用いて検定することが可能である. マルコフ基底はある多項式環上のトーリックイデアルにおける被約グレブナー基底に対応しており, 一意的に計算できる手法が確立されているが, 高次元分割表を対象としたとき、導出される被約グレブナー基底は増加し, 計算上のコストも増加する. 本研究では, 分割表の周辺頻度に対応する不定元を消去したとき, マルコフ基底にどのような影響を与えるかということについて, 自明なマルコフ基底のみを持つ分割表を対象に考察する.

1. はじめに

自然科学や社会科学などの分野では, 与えられたデータ集合から母集団の性質に関する仮説の妥当性を, 統計的仮説検定により検証することがしばしばある. しかしサンプル数が少ない, もしくは分割表がスパースであるようなとき, 漸近理論に基づく統計的検定では当てはまりが悪いことがある. そのような場合, Fisher の正確検定を用いることになるが, 周辺頻度を共有する分割表をすべて列挙することは一般的に困難である [1]. そこで分割表の配置行列に付随するマルコフ基底とマルコフ連鎖モンテカルロ法を組み合わせることで, 周辺頻度を共有する分割表を確率的に抽出し, 抽出された分割表の検定値を相対比較する手法が提案されている. マルコフ基底はある多項式環のトーリックイデアルの被約グレブナー基底として導出することが可能であるが, 高次元の分割表におけるマルコフ基底の導出は, 被約グレブナー基底の肥大化による計算コストの問題から, 非常に困難である [2]. 効率的な計算のために, トーリックイデアルが持つ対称性を利用したアルゴリズムが提案されているが. 一般的なマルコフ基底の構造は, 殆ど知られていない [3].

本稿では, 自明なマルコフ基底を持つ 2×2 分割表と $2\times2\times2$ 分割表に対し, 周辺頻度に対応するある不定元を 1 とした (以下, 消去するという) とき, 被約グレブナー基底にマルコフ基底が含まれるような不定元の組み合わせが存在することを, 実際に計算することで示す. さらに被約グレブナー基底にマルコフ基底が含まれないような, 消去する不定元の組み合わせについて定理を与える. ここで自明なマルコフ基底を, 自明な move に対応する多項式のみで構成されるマルコフ基底として定義する.

†1 現在，静岡理工科大学
Presently with Shizuoka Institute of Science and Technology
a) gs14003@ym.sist.ac.jp
b) tmatsuda@cs.sist.ac.jp

2. Move とマルコフ基底

2×2 分割表において, 周辺頻度を共有する分割表は以下で定義されるファイバー $F_{\mathbf{t}}$ の元である.

$$F_{\mathbf{t}}=\{\mathbf{x}\in \mathrm{N}^4|\mathbf{t}=A\mathbf{x}\}$$

$$A=\begin{pmatrix}1&1&0&0\\0&0&1&1\\1&0&1&0\\0&1&0&1\end{pmatrix},x=\begin{pmatrix}x_{11}\\x_{12}\\x_{21}\\x_{22}\end{pmatrix},t=\begin{pmatrix}x_{1+}\\x_{2+}\\x_{+1}\\x_{+2}\end{pmatrix}$$

ここで, A は配置行列, $\mathbf{x}$ は分割表のセル頻度, $\mathbf{t}$ は周辺頻度を表現したベクトルである. また move の集合 $M(A)$ は以下で定義される.

$$M(A)=\{\mathbf{z}\in \mathrm{Z}^4|\mathbf{0}=A\mathbf{z}\}=\ker(A)\cap \mathrm{Z}^4.$$

このとき $A(\mathbf{x}+\mathbf{z})=A\mathbf{x}=\mathbf{t}$ であるため, move を分割表に作用させても, 周辺頻度 $\mathbf{t}$ は変化しない. $\mathbf{z}=(1,-1,-1,1)$ というように, 分割表の対角的なセル頻度に対し, $\{1,-1\}$ を作用させる move を自明な move と呼ぶ. 一方, $M(A)$ の部分集合 B がマルコフ基底であるとは,

$\forall\mathbf{t}\in \mathrm{N}^{\mathrm{p}},\forall\mathbf{x},\mathbf{y}\in F_{\mathbf{t}}$ に対して,

$\mathbf{y}=\mathbf{x}+\sum_{i=1}^{\mathrm{N}}\varepsilon_i z_i,\ \mathbf{x}+\sum_{i=1}^{\mathrm{n}}\varepsilon_i z_i\in F_{\mathbf{t}},\ n=1,2,\cdots,\mathrm{N}$

を満たすような $\mathrm{N}>0, z_i\in B, \varepsilon_i\in\{-1,1\}, i=1,2,\cdots,\mathrm{N}$ が存在することをいう. I を添え字集合として, セル頻度 $\{x_i\}_{(i\in I)}$ と周辺頻度 $\mathbf{t}$ に対応する不定元集合を, それぞれ $U=\{u_i\}_{(i\in I)}, V=\{v_1,v_2,\cdots,v_d\}$ として, 以下にトーリックイデアル I_A を定義する.

$$I_A=<\{u^{z^+}-u^{z^-}|z\in M(A)\}>$$

ただし $z_i^+=\max\{0,z_i\}, z_i^-=-\min\{0,z_i\}$ である. ここで配置行列 A に対するマルコフ基底は, トーリックイデアル I_A の生成系と同値である [2]. また配置行列 A に関する環の準同型写像 $\phi_A: k[U]\to k[V]$ を $u_i\mapsto v_1^{a_{1i}}v_2^{a_{i2}}\cdots v_d^{a_{dj}}$ と定義すれば, 以下の関係が成り立つ.

$$I_A=\{f\in k[U]|\phi_A(f)=0\}$$

すなわちトーリックイデアル I_A は準同型写像 ϕ_A の核 $\ker(\phi_A)$ と同値で, マルコフ基底は $\ker(\phi_A)$ の生成系となる. 具体的にマルコフ基底は以下の定理から計算される [2].

定理 1

多項式環 $k[U,V]$ 上のイデアル $I_A^* =< -\phi_A(u_i) + u_i, i \in I >$ は $I_A = I_A^* \cap k[U]$ を満たし, $V > U$ を満たす適当な項順序を定義すれば, I_A^* の被約グレブナー基底 G^* が計算できる. このときトーリックイデアル I_A の被約グレブナー基底 G は $G = G^* \cap k[U]$ である.

3. 主定理

第 3 節では, 自明なマルコフ基底を持つ 2×2 分割表, $2 \times 2 \times 2$ 分割表に対し, 周辺頻度に関する不定元を消去したとき, マルコフ基底が被約グレブナー基底として得られない場合を, 定理として与える. またその際の配置行列に対応する行列を A' と定義する.

定理 2

2×2 分割表の不定元集合 V に含まれる 2 つ以上の任意の元を消去しても, マルコフ基底は被約グレブナー基底に含まれない.

定理 2 の証明
2 つ以上の不定元を消去したとき, 以下に示した 2 項式の少なくともどちらか一方は, イデアル $I_{A'}^*$ に含まれる.

$$f_1 = u_k - 1, f_2 = u_k - u_l \quad (1 \leq k, l \leq 4)$$

このとき 2 項式 f_1, f_2 のイニシャル単項式は共に $in_<(f_1) = in_<(f_2) = u_k \in in_<(I_{A'}^*)$ であり, グレブナー基底の定義より $in_<(g) = u_k$ なる 2 項式 g がグレブナー基底に含まれる. しかしマルコフ基底 $u_1u_4 - u_2u_3$ の単項式は $in_<(g)$ によって割り切られてしまうため, 被約グレブナー基底の定義から, マルコフ基底は被約グレブナー基底に含まれない. 定理 2 より, 直ちに系 2-1 が得られる.

系 2-1

不定元集合 V に含まれる元を消去し, 以下が満たされたとき, イデアル $I_{A'}^*$ の被約グレブナー基底にマルコフ基底 $u^{z^+} - u^{z^-}$ は含まれない.

$$^\exists g \in I_{A'}^* \quad \text{s.t.} \quad in_<(g) | u^{z^+} \text{ or } in_<(g) | u^{z^-}$$

$2 \times 2 \times 2$ 分割表に関して, 以下の定理を与える.

定理 3

$2 \times 2 \times 2$ 分割表の不定元集合 V に含まれる 6 つ以上の任意の元を消去しても, マルコフ基底は被約グレブナー基底に含まれない.

定理 3 の略証
証明は定理 1 と同様である. 不定元集合 V の 6 つの元を消去する選び方は 924 通りであり, すべての組み合わせにおいて, 系 2-1 を考えることで定理 3 を示すことができる. 実際にはイデアル $I_{A'}^*$ が持つ対称性を利用し, 計算する.

4. マルコフ基底の計算

第 4 節では, 自明なマルコフ基底を持つ分割表に関し, マルコフ基底が被約グレブナー基底に含まれない, 不定元 $v_i \in V$ を消去する組み合わせを示した. しかしマルコフ基底が被約グレブナー基底に含まれる不定元の組み合わせも存在し, そのとき計算される被約グレブナー基底の数は減少する. そこで, $2 \times 2 \times 2$ 分割表を対象として, マルコフ基底が得られるような不定元を消去する組み合わせの一例と, そのとき得られる被約グレブナー基底の数を以下に示す. 尚, 計算には Risa-Asir を用いた.

表 1 対応表

周辺頻度	不定元
x_{+00}	v_1
x_{+01}	v_2
x_{+10}	v_3
x_{+11}	v_4
x_{0+0}	v_5
x_{0+1}	v_6
x_{1+0}	v_7
x_{1+1}	v_8
x_{00+}	v_9
x_{01+}	v_{10}
x_{10+}	v_{11}
x_{11+}	v_{12}

表 2 実験結果

消去する不定元	基底の数
なし	46
v_1	41
v_1, v_2	36
v_1, v_5	30
v_1, v_2, v_5	30
v_1, v_5, v_{11}	21
v_1, v_2, v_5, v_7	22
v_2, v_3, v_5, v_8	20
v_2, v_3, v_5, v_8, v_9	14

不定元を消去せずマルコフ基底を計算したとき, 被約グレブナー基底の数は 46 個であった. 本実験において, 最も被約グレブナー基底の数が少なく, かつマルコフ基底が得られる不定元の組み合わせは v_2, v_3, v_5, v_8, v_9 を消去した場合であり, 被約グレブナー基底の数は 14 個であった.

5. まとめ

本稿では, 自明なマルコフ基底を持つ分割表に対し, 被約グレブナー基底にマルコフ基底が含まれないような, 消去する不定元の組み合わせについて定理を与えた. またマルコフ基底が含まれるような不定元の組み合わせの一例と, 対応する被約グレブナー基底の数を示し, マルコフ基底を保持しつつ冗長な被約グレブナー基底を削ることが可能であることを確認した. 非自明なマルコフ基底を持つ分割表において, 同様の実験と考察を行うことが今後の課題である.

参考文献

[1] S.J. Haberman, “A warning on the use of chi-squared statistics with frequency tables with small expected cell counts”, Journal of the American Statistical Association, No.83, pp555-560, (1988)
[2] P.Diaconis and B.Sturmfels, “Algebraic algorithms for sampling from conditional distributions”, Annals of Statistics, No. 26, pp363-397, (1998)
[3] S.Aoki and A.Takemura, Minimal invariant Markov basic for sampling contingency tables with fixed marginal”, Annals of the Institute of Statistical Mathematics, No.60, pp229-256, (2008)