サンプリングと計算代数

間野 修平　数理・推論研究系 准教授

0　本ポスターの内容

与えられた確率分布に従う確率変数をシミュレートするプログラムをサンプラーといいます．本ポスターでは，微分作用素環のグレブナー基底を用いることで，カウントデータの標準的な統計モデルであるトーリックモデルについては原理的にサンプリングが可能であることを紹介しています．本発表は高山信毅氏(神戸大学)との共同研究を含みます．

1　トーリックモデル

トーリックモデルはカウントデータの標準的な統計モデルです．変量の組み合わせが定める状態をセルとよびます．トーリックモデルは標本点がセルに入る確率の対数が変量の周辺効果と交互作用について線形で，単体的複体の各要素に対応する項をもつモデルです．特に，ファセット（包含関係について極大な要素）がグラフの極大クリーク（完全部分グラフ）に対応するモデルをグラフィカルモデルといいます．

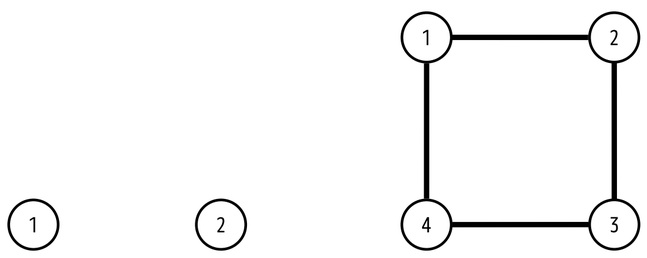

図1 二元分割表の独立モデル$\{\phi, \{1\}, \{2\}\}$（左）とサイクルモデル$\{\phi, \{1\}, \{2\}, \{3\}, \{4\}, \{1,2\}, \{2,3\}, \{3,4\}, \{1,4\}\}$

グラフィカルモデルでないけれどもよく知られたトーリックモデルに三元分割表の無三因子交互作用モデルがあります．

定義 1. 単体的複体Γに付随するトーリックモデルとは，母数$\theta^{(F)}_{i_F} > 0$を用いて表される離散確率分布族

$$\mathcal{M}_A = \left\{ p \in \Delta_{|\mathcal{R}|-1};\, p_i = \frac{1}{Z_i(\theta)} \prod_{F \in \mathrm{facet}(\Gamma)} \theta^{(F)}_{i_F},\ \text{for all } i \in \mathcal{R} \right\}$$

をいう．$Z_i(\theta)$は正規化定数，Δ_{m-1}は$m-1$単体である．

例 1 (二元分割表の独立モデル)**.** セル(i_1, i_2)のカウントの確率$p_{i_1 i_2}$は母数$\log \theta^{(1)}_{i_1} = \beta^{(1)}_{i_1}$, $\log \theta^{(2)}_{i_2} = \beta^{(2)}_{i_2}$により$\log p_{i_1 i_2} = \mu + \beta^{(1)}_{i_1} + \beta^{(2)}_{i_2} \Leftrightarrow p_{i_1 i_2} \propto \theta^{(1)}_{i_1} \theta^{(2)}_{i_2}$と表される．

モデル$\mathcal{M}_A$を

$$p_i(\theta) = \frac{1}{Z_i(\theta)} \prod_{j=1}^{d} \theta_j^{a_{ji}},$$

と表します．$A = (a_{ji})$は$d \times m$行列，行はファセット，列は状態です．多項抽出の標本$c \in \mathbb{N}_0^m$について，Acは最小十分統計量，$\mathcal{F}_b(A) = \{c : Ac = b\}$は$b$-ファイバーとよばれます．

例 2 (2×2分割表の独立モデル)**.**

c_{11}	c_{12}	$c_{1\cdot}$
c_{21}	c_{22}	$c_{2\cdot}$
$c_{\cdot 1}$	$c_{\cdot 2}$	$\lvert c \rvert$

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix},\quad c = \begin{pmatrix} c_{11} \\ c_{12} \\ c_{21} \\ c_{22} \end{pmatrix},\quad b = \begin{pmatrix} c_{1\cdot} \\ c_{2\cdot} \\ c_{\cdot 1} \\ c_{\cdot 2} \end{pmatrix}$$

2　サンプラー

あるトーリックモデルを帰無仮説とし，より多くの項を含むトーリックモデルを対立仮説とするとき，帰無仮説の母数の十分統計量で条件つけた分布は

$$\mathbb{P}(C = c | AC = b) = \frac{1}{Z_A(b; p)} \frac{p^c}{c!},\quad p^c = \prod_{i=1}^{m} p_i^{c_i},\quad c! = \prod_{i=1}^{m} c_i!$$

となります．pは帰無仮説が含まない母数の関数で，帰無仮説の下で$p = 1$です．この分布は規格化定数

$$Z_A(b; p) = \sum_{c \in \mathcal{F}_b(A)} \frac{p^c}{c!}$$

がGelfandら(1990)により定義されたA超幾何多項式であることから，A超幾何分布とよばれます(Takayama et al. 2018).

例 3 (2×2分割表)**.** 帰無仮説(図1左)の下での分布は超幾何分布，対立仮説(フルモデル)の下での分布は一般化超幾何分布で，$p_1 = p_{11}p_{22}/(p_{12}p_{21})$はオッズ比．規格化定数はGaussの超幾何多項式．

あるモデルから観察されたデータ以上に稀なデータが現れる確率を裾確率とよび，検定には裾確率が必要です．裾確率が閉じた形で与えられないときの評価はサンプラーの典型的な用途の一つです．例えば，二元分割表の独立モデルからのサンプラーは壺モデルにより表現できることはよく知られています．しかし，二元分割表を含む分解可能モデルとよばれる特別なグラフィカルモデルを帰無仮説とする場合以外のサンプリングは難しいと考えられてきたため，多項式環のグレブナー基底を用いてb-ファイバーに既約なマルコフ連鎖を定めて，マルコフ連鎖モンテカルロとよばれる近似的なサンプラーを構成する研究が行われてきました(Diaconis, Sturmfels 1998)．ところが，微分作用素環のグレブナー基底を利用すれば，任意のトーリックモデルについて帰無仮説に限らずサンプリングが可能になることを示しました．

アルゴリズム 1 (M 2017)**.** 行列A, 十分統計量bのA超幾何分布からのサンプリング．十分統計量がβのときのセルiのカウントの期待値を$e(\beta; i)$，Aの第i列ベクトルをa_iとする．

1. Pick $i \in \{1, ..., m\}$ with $\mathbb{P}(T_1 = i) = e(b; i)/|c|$.
2. For $i = 2, ..., |c|$, pick $j \in \{1, ..., m\}$ with
$$\mathbb{P}(T_i = j | t_1, ..., t_{i-1}) = \frac{e(b - (a_{t_1} + \cdots + a_{t_{i-1}}); j)}{|c| - i + 1}.$$

注意 1. 期待値はA超幾何多項式で与えられ，その漸化式により逐次評価する．漸化式は任意のAについて超幾何微分方程式系が定めるグレブナー基底を用いて得られる．

参考文献

Shuhei Mano (2018) *Partitions, Hypergeometric Systems, and Dirichlet Processes in Statistics*, Springer

高山信毅，間野修平 (2019) 計算代数のdirect samplerへの応用．京都大学数理解析研究所講究録