超幾何系によるDirect Sampler

間野 修平　数理・推論研究系 准教授

0　本ポスターの内容

計数データ解析では，仮説検定や事後分布の評価など，与えられた分布に従う標本を生成する必要があることが多いです．多くの場合，正規化定数を計算することは可能な標本の数が膨大なため困難として，正規化定数の計算を必要としないけれども近似的な生成法が使われます．しかし，その説明に現れる例は単純なので，本当に正規化定数を計算することは難しいのだろうか，と疑問に思われるかも知れません．

本ポスターは，典型的な計数データ解析に現れる，対数アフィンモデルからの交換可能な標本の条件付き分布については，原理的に，可能な標本を数え上げることなく正規化定数を計算できることを指摘し，与えられた分布に正確に従う生成法(direct sampler)を紹介しています．

1　対数アフィンモデル

セルがm通りあり，t_iを大きさnの標本のi番目の標本点とします．標本$(t_1,...,t_n)$の計数ベクトルを$(c_1,...,c_m)$, $c_i := \#\{j; t_j = i\}$ とします．条件$c_1+\cdots+c_m = n$を斉次性とよびます．標本が交換可能列であれば

$$\mathbb{P}(C_1 = c_1,...,C_m = c_m) = \frac{n!}{c_1!\cdots c_m!}\mathbb{P}(T_1 = t_1,...,T_n = t_n)$$

なので，母数$(p_1(\xi),...,p_m(\xi))$の多項分布とし，対数アフィンモデル

$$\log p_i(\xi) = \sum_{j=1}^{d+g} a_{ji}\xi_j - \phi_i(\xi)$$

を仮定すると，

$$\mathbb{P}(C = c) = \frac{n!}{c!}\exp\left\{\sum_{j=1}^{d+g}\xi_j b_j - \sum_{i=1}^{m}\phi_i(\xi)\right\}, \qquad c! := \prod_{i=1}^{m} c_i!$$

の形になります．ここで，$(b_1,...,b_{d+g})$は十分統計量で，$b_j = \sum_{i=1}^{m} a_{ji}c_i$, $j = 1,...,d+g$を満たしますが，$j = 1,...,d$の条件を行列Aとベクトルbにより$Ac = b$と表すと，条件付き分布は

$$\mathbb{P}(C = c|AC = b) \propto \frac{1}{c!}\exp\left\{\sum_{j=d+1}^{d+g}\xi_j b_j\right\} = \frac{x^c}{c!}, \qquad \log x_i := \sum_{j=d+1}^{d+g} a_{ji}\xi_j$$

となります．正規化定数がA超幾何多項式とよばれることから，この分布はA超幾何分布とよばれます(Takayama, Kuriki, Takemura, 2018).

2　Markov Chain Monte Carlo

有限集合の状態空間に値をとる，既約，非周期的，対称なMarkov連鎖の推移確率行列$R = (r_{ij})$に対し，

$$p_{ij} = r_{ij}\min\left\{1, \frac{\pi_j}{\pi_i}\right\}, \qquad i \neq j, \qquad p_{ii} = 1 - \sum_{j\neq i} p_{ij}$$

を推移確率行列とするMarkov連鎖は一意な定常分布πを持ちます．

アルゴリズム 1 (Metropolis et al. 1953) 分布πからの抽出．

1. 初期標本$c^{(0)}$を抽出し，$t = 0$とする．

2. 候補c'を推移確率行列Rに従い抽出する．

3. 候補c'を確率

$$\min\left\{1, \frac{\pi_{c'}}{\pi_{c^{(t)}}}\right\}.$$

で受理し，$c^{(t+1)} = c'$とする．棄却されれば，$c^{(t+1)} = c^{(t)}$ とする．

4. tに*1*を加えて*2*に戻る．

分布πを定常分布として実現するので，次の原理的な欠点があります．

- 定常に達するまでに時間がかかり，確かめるのも困難
- 標本列の自己相関

Metropolisアルゴリズムは，正規化定数の計算を必要とせず，汎用性がありますが，個々の分布については，direct samplerを利用できるなら，その方が望ましいかも知れません．

3　Direct Sampler

可能な標本を単項式に対応させると，正規化定数は斉次多項式で，それを解とする多変数線形微分方程式系（超幾何系）の解の基底のGröbner基底による標準形が，解の基底の漸化式を定めます．漸化式を解くことで，数え上げを避けて計算することができます（差分ホロノミック勾配法とよばれます）．さらに，斉次性から従う漸化式が単項式に値をとるMarkov連鎖の推移規則を定めます．その経路は標本に対応する単項式を与え，各経路が現れる確率はA超幾何分布に従います．

アルゴリズム 2 (M 2017) 行列A, ベクトルbのA超幾何分布からの抽出．セルiの期待値を$e(b;i)$，Aの第i列ベクトルをa_iとする．

1. $\mathbb{P}(T_1 = i) = \dfrac{e(b;i)}{n}$.

2. $i = 2,...,n$に対し，$\mathbb{P}(T_i = j|t_1,...,t_{i-1}) = \dfrac{e(b-(a_{t_1}+\cdots+a_{t_{i-1}});j)}{n-i+1}$.

4　分割表の例

2×2分割表

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

を考えます．標準的なモデルは対数アフィンモデルです．普通はオッズ比ωに興味があるので，興味のない母数の十分統計量である周辺和で条件付けた一般化超幾何分布

$$\mathbb{P}(C_{11} = c_{11}|c_{\cdot\cdot}, c_{1\cdot}, c_{\cdot 1}) \propto \frac{\omega^{c_{11}}}{c_{11}!c_{12}!c_{21}!c_{22}!}.$$

を考えます．正規化定数はGaussの超幾何多項式に比例します．行と列の独立性($\omega = 1$)の検定は，Fisherの正確確率検定とよばれます．第13次日本人の国民性調査(2013)の#2.11bへの20歳代の結果で検証しました．

1. 仕事や遊びなどで自分の可能性をためすために，できるだけ多くの経験をしたい．
2. わずらわしいことはなるべく避けて，平穏無事に暮らしたい．

	1	2	計
男	53	19	72
女	56	31	87
	109	50	159

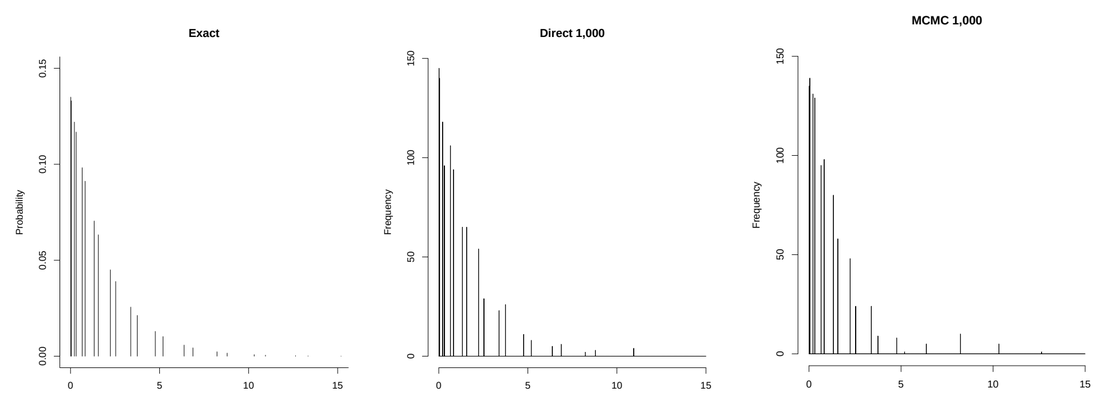

図1 分割表のχ^2値の真の分布と，direct sampler, MCMCにより1,000回生成したχ^2値のヒストグラム．p値はそれぞれ0.233, 0.236, 0.193.

参考文献

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