2003年日本オペレーションズ・リサーチ学会
春季研究発表会

$2 \times n$分割表のPerfect Sampling

東京大学　*来嶋秀治　KIJIMA Shuji
01605000　東京大学　松井知己　MATSUI Tomomi

1. はじめに

2元分割表は正の整数からなる行和と列和を持ち、表中の値として非負整数をとる表（行列）である。2元分割表は医療統計の分野などで統計データを扱うのに用いられる。分割表に対する統計学的興味の一つに行と列の相関関係があげられる。行と列の独立性の検定にχ^2検定があるが、周辺和の中に小さな値があると、χ^2検定の当てはまりが悪くなるため、正確検定が行われる。

正確検定に用いるp値の計算には、周辺和を満たす分割表を全列挙する方法があるが、分割表の全列挙は一般に困難なため、MCMC（マルコフ連鎖モンテカルロ）法がよく用いられる。MCMC法はモンテカルロ法を実行する際、マルコフ連鎖の定常分布からのサンプルを用いる手法である。しかし、定常分布は無限回の推移の後に得られるため、定常分布からのサンプリングは一般に困難である。本報告では$2 \times n$分割表の一様標本抽出法にサンドイッチングの技法を応用したCFTP(Coupling From The Past)理論を用い、Perfect Sampling（厳密に一様な分布からの標本抽出）を行う手法を提案する。

2. マルコフ連鎖と標本抽出アルゴリズム

整数（非負整数、正整数）全体の集合をそれぞれ$\mathbb{Z}$（$\mathbb{Z}_+$、$\mathbb{Z}_{++}$）で表すことにする。ベクトル$r = (r_1, r_2) \in \mathbb{Z}_{++}^2$と$s = (s_1, \ldots, s_n) \in \mathbb{Z}_{++}^n$は正整数$N \in \mathbb{Z}_{++}$に対して、$\sum_{i=1}^{2} r_i = \sum_{j=1}^{n} s_j = N$を満たすとする。行和$r$および列和$s$をもち、非負整数を表値にとる2行$n$列の2元分割表全体の集合$\Sigma_{r,s}$を、

$$\Sigma_{r,s} \stackrel{\text{def.}}{=} \left\{ X \in \mathbb{Z}_+^{2\times n} \;\middle|\; \sum_{j=1}^{n} X(i,j) = r_i \quad (1 \le \forall i \le 2), \quad \sum_{i=1}^{2} X(i,j) = s_j \quad (1 \le \forall j \le n) \right\}$$

で定義する。但し、$X(i,j)$は分割表Xのi行j列の値を表す。集合$\Sigma_{r,s}$を状態空間にもつマルコフ連鎖$\mathcal{M}$を定義する。マルコフ連鎖$\mathcal{M}$の推移$X^t \mapsto X^{t+1}$ $(t \in \mathbb{Z})$は一様実数乱数$\lambda \in [1, n)$が与えられた時、以下のように定義される。

Step 1: i列と$i+1$列$(i = 1, \ldots, n-1)$について、$a = X^t(1,i) + X^t(1,i+1)$、$b = X^t(2,i) + X^t(2,i+1)$として、$\theta_{X^t}(i) \stackrel{\text{def.}}{=} \min\{a, b, s_i, s_{i+1}\} + 1$を定義する。

Step 2: 乱数$\lambda \in [1, n)$に対して、$\lfloor\lambda\rfloor$列と$\lfloor\lambda\rfloor + 1$列を選び、$\phi_{X^t}(\lambda) \stackrel{\text{def.}}{=} \lfloor\theta_{X^t}(\lfloor\lambda\rfloor)\,(\lambda - \lfloor\lambda\rfloor)\rfloor$として、$X^{t+1} \in \Sigma_{r,s}$を

$$\begin{aligned} X^{t+1}(1,j) &= \begin{cases} \phi_{X^t}(\lambda) + \min\{a, s_i\} - \min\{a, b, s_i, s_{i+1}\} & (j = \lfloor\lambda\rfloor), \\ a - X^{t+1}(1, j-1) & (j = \lfloor\lambda\rfloor + 1), \\ X^t(1,j) & (\text{otherwise}), \end{cases} \\ X^{t+1}(2,j) &= s_j - X^{t+1}(1,j), \end{aligned}$$

と置く。

このマルコフ連鎖$\mathcal{M}$は、明らかにエルゴード性を満たす。また$\mathcal{M}$はdetailed balance equationsを満たし、定常分布として一様分布をもつ。

今、分割表$X_U, X_L \in \Sigma_{r,s}$をそれぞれ、北西隅のルールおよび南西隅のルールに従ってできる表とする。すなわち、X_Uは$X(1,1) = s_1$から順に$X(1,2) = s_2, X(1,3) = s_3, \ldots$と、$\sum_{j=1}^{k} X(1,j) = r_1$となるまで1行目を決め、残りの表値は周辺和を満たすように一意に決める。同様にX_Lは2行目の1列目から表値を決める。すな

© 日本オペレーションズ・リサーチ学会. 無断複写・複製・転載を禁ず.

わち、

$$X_{\mathrm{U}} \stackrel{\text{def.}}{=} \left(X(i,j) \in \mathbb{Z}_+ \;\middle|\; \exists k,\ r_1 = \sum_{j=1}^{k} X(1,j) \le \sum_{j=1}^{k} s_j,\ X(2,j) = 0\ (j = 1,\ldots,k-1) \right),$$

$$X_{\mathrm{L}} \stackrel{\text{def.}}{=} \left(X(i,j) \in \mathbb{Z}_+ \;\middle|\; \exists k,\ r_2 = \sum_{j=1}^{k} X(2,j) \le \sum_{j=1}^{k} s_j,\ X(1,j) = 0\ (j = 1,\ldots,k-1) \right),$$

となる。それぞれ X_{U} を NW(North-West) 表、X_{L} を SW(South-West) 表と呼ぶことにする。
マルコフ連鎖 $\mathcal{M}$ の定常分布からの標本抽出アルゴリズム P を以下のように定義する。

Step 1: シミュレーション開始時刻を $t = -1$ とし、$\boldsymbol{\lambda}_t$ を空列とする。

Step 2: 一様乱数 $\lambda(t) \in [1,n)$ を生成して $\boldsymbol{\lambda}_t$ の先頭に挿入し、乱数列 $\boldsymbol{\lambda}_t := (\lambda(t), \lambda(t+1), \ldots, \lambda(-1))$ とする。

Step 3: 時刻 t における 2 つの状態 $Y_{\mathrm{U}}, Y_{\mathrm{L}} \in \Sigma_{\boldsymbol{r},\boldsymbol{s}}$ を $Y_{\mathrm{U}}^t := X_{\mathrm{U}}$、$Y_{\mathrm{L}}^t := X_{\mathrm{L}}$ とする。

Step 4: $Y_{\mathrm{U}}, Y_{\mathrm{L}}$ を初期状態とし、数列 $\boldsymbol{\lambda}_t = (\lambda(t), \lambda(t+1), \ldots, \lambda(-1))$ を用いて、時刻 t から 0 までそれぞれマルコフ連鎖 $\mathcal{M}$ を推移させる。

a) もし、$Y_{\mathrm{U}}^0, Y_{\mathrm{L}}^0$ が時刻 0 で一致して $Y_{\mathrm{U}}^0 = Y_{\mathrm{L}}^0$ となっていれば、$X_0 := Y_{\mathrm{U}}^0 = Y_{\mathrm{L}}^0$ を標本とする。

b) 一致していなければ $t := t - 1$ として Step 2 にもどる。

この標本抽出法は厳密に一様分布に従う。すなわち以下の定理が成り立つ。

定理 1 *アルゴリズム Pからの標本は厳密に一様分布に従う。*

3. CFTP とサンドイッチング

CFTP 理論は 1996 年 Prop and Willson によって提案された。CFTP の重要な主張は以下の定理である。

定理 2 *有限離散空間 Ω 上のマルコフ連鎖 MC はエルゴード性を満たすとする。Ω 中のすべての要素について、時刻 $t < 0$ から同一の一様乱数列 $\lambda \in [0,1)^t$ を用いて MC にしたがって推移させる。この時、時刻 0 の状態がすべて一致 (coalescence) して、$X(\in \Omega)$ となっていれば、X はマルコフ連鎖 MC の定常分布に従って得られた標本である。*

したがって CFTP 理論によって厳密に定常分布に従う標本を得ることが出来る。しかし、CFTP 理論では全状態を初期状態とするマルコフ連鎖の推移を追跡する必要があり、全状態からのシミュレーションは現実には不可能である。CFTP の技法の 1 つとして、サンドイッチングがある。これは複数の特定の状態からの推移が一致することが全状態からの推移が一致することの十分条件になっていれば、その特定の状態についてのみシミュレーションをすれば良いというアイデアである。したがってサンドイッチングが適用できるマルコフ連鎖に対して、その特定の状態からの CFTP によって、厳密に定常分布に従う標本を得ることが出来る。

我々は $2 \times n$ 分割表のマルコフ連鎖 $\mathcal{M}$ に対してサンドイッチングが適用できることを示した。それが以下の定理である。

定理 3 *マルコフ連鎖 $\mathcal{M}$ について、時刻 $t < 0$ における状態 $X_{\mathrm{U}}, X_{\mathrm{L}} \in \Sigma_{\boldsymbol{r},\boldsymbol{s}}$ から、乱数列 $\boldsymbol{\lambda} \in [1,n)^t$ を用いて得られる推移が、時刻 0 において一致して $X_0(\in \Sigma_{\boldsymbol{r},\boldsymbol{s}})$ となるならば、全状態からのマルコフ連鎖は時刻 0 において状態 X_0 を取る。*

また離散化 Dirichlet 分布に対しても、同様の手法を用いて perfect sampling することができる。

参考文献

[1] O. Häggström, *"Finite Markov Chains and Algorithmic Application,"* London Mathematical Society, Student Texts 52, Cambridge University Press, 2002.

[2] J. Propp and D. Wilson, "Exact sampling with coupled Markov chains and applications to statistical mechanics," *Random Structures and Algorithms,* 9 (1996), pp. 232–252.

© 日本オペレーションズ・リサーチ学会. 無断複写・複製・転載を禁ず.