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

Dirichlet 分布に従う多項式時間近似サンプリング法

01605000 東京大学 *松井知己
北陸先端科学技術大学院大学 元木光雄
東京女子医科大学附属 膠原病リウマチ痛風センター 鎌谷直之

1. はじめに

近年, 個人毎の遺伝子配列の違い (多型) をもとに, 遺伝子配列から疾患関連遺伝子を探索する手法が広く研究されている [1]. これに対し隠れマルコフモデルを用いた様々な手法が提案されているが, その際用いる事前分布として Dirchlet 分布がしばしば用いられる. 本稿では, マルコフ連鎖を用いた Dirichlet 分布に従うサンプリング手法を提案する. また提案するマルコフ連鎖の推移回数について, その mixing time が多項式時間であることを, path coupling method を用いて示す.

2. マルコフ連鎖の提案

Dirichlet 分布は確率変数 $P_1, P_2, \ldots, P_n$ を持つ確率分布である. 本論文では, $n \geq 2$ と仮定する. 確率変数 $P_1, P_2, \ldots, P_n$ は, $P_1 + \cdots + P_n = 1, P_1, P_2, \ldots, P_n > 0$ を満たす. また母数として n 個の非負実数 $u_1, \ldots, u_n$ を取る. 確率密度関数は $\frac{\Gamma(\sum_{i=1}^n u_i)}{\prod_{i=1}^n \Gamma(u_i)} \prod_{i=1}^n p_i^{u_i-1}$ で与えられる. ただし $\Gamma(u)$ はガンマ関数である. 以下では確率変数空間を離散化して扱う. そのために, 与えられた正整数 Δ に対し

$$\Omega \stackrel{\text{def.}}{=} \{X = (X_1, X_2, \ldots, X_n) \in \mathbb{Z}^n \mid X_i > 0\ (\forall i),\ X_1 + \cdots + X_n = \Delta\}$$

という確率変数空間を導入する [2]. そして事象 $X = (X_1, \ldots, X_n) \in \Omega$ が起きる確率を $g(X) \stackrel{\text{def.}}{=} C_\Delta \prod_{i=1}^n (X_i/\Delta)^{u_i-1}$ とする. ただし C_Δ は, 確率の総和を 1 とするための正規化定数 (分配関数) である.

任意の 2 以上の整数 b に対し, $\Omega(b) = \{(Y_1, Y_2) \in \mathbb{Z}^2 \mid Y_1, Y_2 > 0,\ Y_1 + Y_2 = b\}$ とする. $\Omega(b)$ 上の分布関数 $f_b(Y_1, Y_2 \mid u_i, u_j)$ は母数 $u_i, u_j \geq 0$ を持つ関数で, $f_b(Y_1, Y_2 \mid u_i, u_j) \stackrel{\text{def.}}{=} C(u_i, u_j, b) Y_1^{u_i-1} Y_2^{u_j-1}$ で定義されるとする, ただし $C(u_i, u_j, b)$ は総和が 1 となるための定数 (分配関数) である.

Ω 上のマルコフ連鎖 $\mathcal{M}$ の推移 $X^t \mapsto X^{t+1}$ は次のように定義される.

Step 1: 相異なる添え字の対 $\{i, j\} \subseteq \{1, 2, \ldots, n\}$ をランダムに選ぶ.
Step 2: $b := X_i^t + X_j^t$ とする. 確率変数 $(Y_1, Y_2) \in \Omega(b)$ を, $f_b(Y_1, Y_2 \mid u_i, u_j)$ に従って発生させる.
Step 3: 状態 X^{t+1} を以下のように定める.

$$X_k^{t+1} = \begin{cases} Y_1 & (k = i), \\ Y_2 & (k = j), \\ X_k^t & (\text{otherwise}). \end{cases}$$

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 Ω 上の分布関数 $g(X)$ は, $\mathcal{M}$ に関する detailed balance equations を満たすので, $\mathcal{M}$ の定常分布は $g(X)$ となる. この マルコフ連鎖 において, 次の定理が成り立つ.

定理： マルコフ連鎖 $\mathcal{M}$ の mixing time $\tau(\varepsilon)$ は, $\tau(\varepsilon) \leq (1/2)n(n-1)\ln((\Delta - n)\varepsilon^{-1})$ を満たす.

ただし, mixing time $\tau(\varepsilon)$ は以下のように定義される.

$$\tau(\varepsilon) \stackrel{\text{def.}}{=} \max_{x \in \Omega} \min \left\{ t \,\middle|\, \forall t' \geq t,\ (1/2) \sum_{x' \in \Omega} \left| g(x') - \Pr[X^0 = x \text{ and } X^{t'} = x'] \right| \leq \varepsilon \right\}.$$

以下では, path coupling method を用いて上記の定理の証明を行う.

3. Mixing Time の算定

以下では, path coupling method に用いる joint process を定義する. まず Ω を頂点集合とする無向グラフ $G = (\Omega, E)$ を導入する. 状態対 $\{x, y\}$ が G 上で隣接する必要十分条件は, $\|x - y\|_1 \stackrel{\text{def.}}{=}$

$(|x_1-y_1|+\cdots+|x_n-y_n|)=2$ とし, $E \stackrel{\text{def.}}{=} \{\{\boldsymbol{x},\boldsymbol{y}\} \mid \boldsymbol{x},\boldsymbol{y}\in\Omega, \|\boldsymbol{x}-\boldsymbol{y}\|_1=2\}$ と定義する. グラフ G は連結である. 任意の状態の対 $\{\boldsymbol{x},\boldsymbol{y}\}\in E$ に対し, 以下で推移規則を定める. 一般性を失う事無く, $x_1=y_1+1, x_2=y_2-1, x_3=y_3,\ldots,x_n=y_n$ と仮定する事ができる. このとき joint process $(X,Y)\mapsto(X',Y')$ を以下のように定義する.

Step 1: 相異なる添え字の対 $\{i,j\}$ をランダムに選ぶ.

Step 2: 任意の $i'\in\{1,2,\ldots,n\}\setminus\{i,j\}$ について, $X'_{i'}=X_{i'}$, $Y'_{i'}=Y_{i'}$ とする. 添え字 i,j については, $((X'_i,X'_j),(Y'_i,Y'_j))$ を $\Omega(X_i+X_j)\times\Omega(Y_i+Y_j)$ から, 以下に定める規則に従って選ぶ.

(Case 1) $\{1,2\}\cap\{i,j\}=\emptyset$ の場合. このとき等式 $X_i+X_j=Y_i+Y_j$ が成り立つ. Step 2 において, まず最初に (X'_i,X'_j) を確率分布 $f_{(X_i+X_j)}(X'_i,X'_j\mid u_i,u_j)$ に従って選び, 次に $(Y'_i,Y'_j)=(X'_i,X'_j)$ とする. このとき, 推移後の状態対は $(X',Y')\in E$ を満たす.

(Case 2) $\{1,2\}=\{i,j\}$ の場合. Step 2 において, Case 1 と同じ操作を行う. このとき推移後の状態対は $X'=Y'$ を満たす.

(Case 3) $\{1,2\}\cap\{i,j\}=\{2\}$ の場合. 一般性を失う事なく $i=2$ と仮定することができる. $b=X_i+X_j$ とする. 明らかに $Y_i+Y_j=b+1$ が成り立つ. Step 2 で用いる $\Omega(b)\times\Omega(b+1)$ 上の確率を以下に定義する. 確率が正の値を取りうるのは, 以下の集合

$$\{((1,b-1),(1,b)),\cdots,((b-1,1),(b-1,2)),((1,b-1),(2,b-1)),\cdots((b-1,1),(b,1))\}$$

中の要素だけである. それぞれの確率は, $k=1,2,\ldots,b-1$ に対し,

$$\begin{aligned}\Pr[((X'_i,X'_j),(Y'_i,Y'_j))=((k,b-k),(k+1,b-k))] &= C_b\textstyle\sum_{l=1}^{k} l^{u_i-1}(b-l)^{u_j-1}-C_{b+1}\sum_{l=1}^{k} l^{u_i-1}(b-l+1)^{u_j-1}\\ \Pr[((X'_i,X'_j),(Y'_i,Y'_j))=((k,b-k),(k,b-k+1))] &= C_{b+1}\textstyle\sum_{l=1}^{k} l^{u_i-1}(b-l+1)^{u_j-1}-C_b\sum_{l=1}^{k-1} l^{u_i-1}(b-l)^{u_j-1}\end{aligned}$$

によって定める. ただし $C_b=C(u_i,u_j,b)$, $C_{b+1}=C(u_i,u_j,b+1)$ である. 上記推移確率の値の非負性は以下の補題で保証される.

補題 $\forall u_i,\forall u_j\geq 0, \forall k\in\{1,2,\ldots,b-1\}$ に対し,

$$\begin{aligned}&\Pr[((X'_i,X'_j),(Y'_i,Y'_j))=((k,b-k),(k+1,b-k))],\\ &\Pr[((X'_i,X'_j),(Y'_i,Y'_j))=((k,b-k),(k,b-k+1))],\end{aligned}$$

は全て非負である.

この joint process の周辺分布が, マルコフ連鎖 $\mathcal{M}$ の推移確率となっていることは明らかである. 推移確率が正の値を持つ (X',Y') の対では, $\{X',Y'\}\in E$ が成り立っている.

(Case 4) $\{1,2\}\cap\{i,j\}=\{1\}$ の場合. (Case 3) の手続きにおいて添え字 1 と 2 を交換して得られる手続きで (X',Y') を定める.

定理の証明: 前節で定義したグラフ $G=(\Omega,E)$ において, 任意の状態対 $(\boldsymbol{x},\boldsymbol{y})\in\Omega^2$ に対し, 距離 $\mathrm{d}(\boldsymbol{x},\boldsymbol{y})$ を $\boldsymbol{x}$ から $\boldsymbol{y}$ への G 上の最短路の長さとする. グラフ G の直径は $\Delta-n$ となる.

(Case 1),(Case 3),(Case 4) では, joint process による推移後の状態対の距離は 1 であり, (Case 2) では, 推移後の状態対の距離が 0 となる. (Case 2) の生起確率は $2/(n(n-1))$ である事より, joint process による推移後の状態対の距離の期待値 $\mathrm{E}[\mathrm{d}(X',Y')]$ は $1-2/(n(n-1))$ となる. path coupling theorem [3] により, 状態間の距離の期待値が β 倍になるとき, mixing rate $\tau(\varepsilon)$ は $(1-\beta)^{-1}\ln((\Delta-n)\varepsilon^{-1})$ 以下となる. $\beta=1-2/(n(n-1))$ と置くと定理の主張が導かれる. □

参考文献

[1] 鎌谷直之編著, 「ポストゲノム時代の遺伝統計学」, 羊土社, 2001.

[2] 元木光雄, 鎌谷直之, 「Dirichlet 分布に従う多項分布の母数の Makov chain を用いた approximate sampler」, 2001 年冬の LA シンポジウム.

[3] R. Bubley Randomized Algorithms : Approximation, Generation, and Counting, Springer-Verlag, New York, 2001.