Dirichlet過程とPoisson-Dirichlet分布について

半田賢司（佐賀大学）

題材

Dirichlet過程 [Ferguson,1973]

Dirichlet分布（ベータ分布の多次元化）の無限次元化

Poisson-Dirichlet分布 [Kingman,1975]

Dirichlet分布の順序統計量の極限分布

内容

いくつかの基本的な性質，**ランダム測度**や**点過程**としての特徴づけ

ランダム分割や他の文脈（含む集団遺伝学）との関わりについて

Dirichlet分布の定義と一つの表現

$n \geq 2$, $a_1, \ldots, a_n > 0$: **所与**

Dirichlet分布 $D(a_1, \ldots, a_n)$

$$\frac{\Gamma(a_1 + \cdots + a_n)}{\Gamma(a_1) \cdots \Gamma(a_n)} p_1^{a_1 - 1} \cdots p_n^{a_n - 1} dp_1 \cdots dp_{n-1}$$

on

$$\Delta_{n-1} = \left\{ (p_1, \ldots, p_n) : p_i \geq 0 (i = 1, \ldots, n), \sum_{i=1}^{n} p_i = 1 \right\}$$

(注意) $n = 2$**のとき,** $D(a_1, a_2) = \mathrm{Beta}(a_1, a_2)$

事実 **[ガンマ分布との関係]** $\{X_i\}_{i=1}^n$ $\perp\!\!\!\perp$ **(独立)**
$X_i \sim \mathrm{Gamma}(a_i) = x^{a_i - 1} e^{-x} dx / \Gamma(a_i)$ **on** $(0, \infty)$ **ならば**

$$(X_1 + \cdots + X_n) \perp\!\!\!\perp \left\{ \frac{X_i}{X_1 + \cdots + X_n} \right\}_{i=1}^{n} \sim D(a_1, \ldots, a_n)$$

Dirichlet過程の定義 [Ferguson,1973]

$(\mathcal{X}, \mathcal{B})$：可測空間，　ν：(0でない) 有限測度 on $(\mathcal{X}, \mathcal{B})$

パラメータνのDirichlet過程(Dirichlet乱測度ともいう)：

ランダムな確率測度M on $\mathcal{X}$ で，$\mathcal{X}$の任意の有限分割$B_1, \ldots, B_n$に対し

$$(M(B_1), \ldots, M(B_n)) \sim D(\nu(B_1), \ldots, \nu(B_n))$$

記号　このMの確率法則を$\mathcal{D}(\nu)$ または$\mathcal{D}(\mathcal{X}, \nu)$で表す．

(注意) $\mathcal{X} = \{x_1, \ldots, x_n\} (n = \sharp\mathcal{X})$のとき，

$$(P_1, \ldots, P_n) \sim D(a_1, \ldots, a_n) \qquad \text{ならば}$$

$$P_1\delta_{x_1} + \cdots + P_n\delta_{x_n} \sim \mathcal{D}(a_1\delta_{x_1} + \cdots + a_n\delta_{x_n}).$$

ただし，δ_xはxに集中したデルタ分布．

Dirichlet過程の事後分布

$n \geq 1$，M：ランダムな確率測度 on $\mathcal{X}$

定義　$\mathcal{X}$-値確率変数 $X_1, \ldots, X_n$ が M からの大きさ n の**サンプル**であるとは

$$((X_1, \ldots, X_n) \mid M) \sim M^{\otimes n}$$

定理 **[Ferguson,1973]**

$M \sim \mathcal{D}(\nu)$
$X_1, \ldots, X_n$：M からの大きさ n のサンプル

とするとき，

$$(M \mid (X_1, \ldots, X_n)) \sim \mathcal{D}\left(\nu + \delta_{X_1} + \cdots + \delta_{X_n}\right).$$

(証明のエッセンス)　$n = 1$ の場合を示せばよいが，それは次に帰着：

$$p_i D(a_1, \ldots, a_n)(dp_1 \cdots dp_n) \propto D(a_1, \ldots, a_i + 1, \ldots, a_n)(dp_1 \cdots dp_n)$$

Dirichlet過程の特徴づけ（1）

$\langle \mu, f \rangle$ ： 測度μに関するfの積分，$\theta = \nu(\mathcal{X}) > 0$

定理 **[Cifarelli-Regazzini,1990]**

$M \sim \mathcal{D}(\nu)$ であるための必要十分条件は

$$E\left[\langle M, f\rangle^{-\theta}\right] = e^{-\langle \nu, \log f\rangle}, \quad \forall f > 0 \text{ with } |\langle \nu, \log f\rangle| < \infty$$

（証明のエッセンス） 階段関数$f = \sum_{i=1}^{n} t_i 1_{B_i}$（ただし$t_1, \ldots, t_n > 0$，$B_1, \ldots, B_n$：$\mathcal{X}$の分割，$1_B$は$B$の定義関数）に対し，上の式は次と同じ：

$$E\left[(t_1 M(B_1) + \cdots + t_n M(B_n))^{-\theta}\right] = t_1^{-\nu(B_1)} \cdots t_n^{-\nu(B_n)}$$

右辺で $t_i^{-a} = \int_0^\infty dz_i z_i^{a-1} e^{-t_i z_i}/\Gamma(a)$ を代入し，$dz_1 \cdots dz_n$において

変数変換 $z := z_1 + \cdots + z_n, \quad p_i := z_i/z \ (1 \le i \le n-1)$ **を行う．**

Dirichlet過程の特徴づけ（2）

$f : \mathcal{X} \to (0, \infty)$（$\log f$: 有界）により$\mathcal{X}$上の確率測度全体$\mathcal{P}(\mathcal{X})$上の変換

$$\mathcal{M}_f : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X}), \ \ \mu \mapsto \langle \mu, f \rangle^{-1} f d\mu$$

が導かれ，$\mathcal{M}_f \circ \mathcal{M}_g = \mathcal{M}_{fg}$が成立．

定理 (準不変性) [H,2001] $\theta = \nu(\mathcal{X}) > 0$ とおく．
$\mathcal{D}(\nu)$と像測度$\mathcal{D}(\nu) \circ \mathcal{M}_f := \mathcal{D}(\nu) \circ (\mathcal{M}_{1/f})^{-1}$は互いに絶対連続で

$$\frac{d(\mathcal{D}(\nu) \circ \mathcal{M}_f)}{d\mathcal{D}(\nu)}(\mu) = e^{\langle \nu, \log f \rangle} \langle \mu, f \rangle^{-\theta}, \qquad \mathcal{D}(\nu)\text{-a.e.}\mu$$

(注意) 両辺を$\mathcal{D}(\nu)$で積分$\Longrightarrow$ [Cifarelli-Regazzini,1990]の等式

準不変性の公式 $\Longleftrightarrow$ 事後分布の公式

$M \sim \mathcal{D}(\nu)$, $\theta = \nu(\mathcal{X}) > 0$, $\quad X_1$：M からの大きさ１のサンプル

準不変性公式の言い換え： $\forall \Phi : \mathcal{P}(\mathcal{X}) \to [0, \infty)$,

$$E\left[e^{\langle \nu, \log f \rangle} \langle M, f \rangle^{-\theta} \Phi\left(\frac{f dM}{\langle M, f \rangle}\right)\right] = E\left[\Phi(M)\right] \qquad \textbf{(QI)}$$

事後分布公式$(n = 1)$**の言い換え**： $\forall g : \mathcal{X} \to [0, \infty)$,

$$E\left[\langle M, f \rangle^{-(\theta+1)} g(X_1)\right] = E\left[e^{-\langle \nu, \log f \rangle - \log f(X_1)} g(X_1)\right]$$

$(X_1|M) \sim M$ により，各辺それぞれ書き換えると

$$E\left[\langle M, f \rangle^{-(\theta+1)} \langle M, g \rangle\right] = E\left[e^{-\langle \nu, \log f \rangle} \langle M, g/f \rangle\right]$$

これは **(QI)** で $\quad \Phi(M) = \langle M, g/f \rangle \quad$ と取ったものに等しい．

準不変性の公式 $\Longleftrightarrow$ 部分積分公式

準不変性公式の言い換え（別の型）：

$$E\left[e^{\langle \nu, \log f\rangle}\langle M, f\rangle^{-\theta}\Phi(M)\right] = E\left[\Phi\left(\frac{f^{-1}dM}{\langle M, f^{-1}\rangle}\right)\right]$$

において， $f \mapsto e^{tf}$ **（** f**:有界) と置き換えて** $(d/dt)|_{t=0}$ **を計算すると**
部分積分公式：‘良い’関数 Φ **に対して**

$$E\left[(\langle \nu, f\rangle - \theta\langle M, f\rangle)\Phi(M)\right] = -E\left[\langle M, f\frac{\delta\Phi}{\delta M}\rangle - \langle M, f\rangle\langle M, \frac{\delta\Phi}{\delta M}\rangle\right] \quad \textbf{(IP)}$$

ただし， $\dfrac{\delta\Phi}{\delta M}(x) = \left.\dfrac{d}{dt}\right|_{t=0}\Phi(M + t\delta_x)$

（注意）さらに（IP） $\Longrightarrow$ **「**$\mathcal{D}(\nu)$ **は** $\mathcal{P}(\mathcal{X})$ **上のある拡散過程*の定常分布」**

*集団遺伝学における「**無限個中立アレルモデル**」
($\mathcal{X}$：‘タイプ’空間，θ：突然変異率，$\theta^{-1}\nu$：変異後のタイプの分布)

Dirichlet過程の‘離散的’表現

(ガンマ過程のジャンプに付随した) Poisson点過程による構成

定理 **[Ferguson,1973]**, **[Kingman,1975]** $\theta := \nu(\mathcal{X}) > 0$

$\{Y_i\}_{i=1}^{\infty}$: $\mathcal{X}$-**値i.i.d.,** $Y_i \sim \theta^{-1}\nu$

$\xi = \sum_{i=1}^{\infty} \delta_{Z_i}$: **Poisson点過程，平均測度** $\Lambda_\theta(dz) = \theta \dfrac{e^{-z}}{z} 1_{(0,\infty)}(z)dz$

i.e., $(0,\infty)$**の任意の有限分割** $A_1, \ldots, A_n$ **に対し**

$$(\xi(A_1), \ldots, \xi(A_n)) \sim \mathrm{Poisson}(\Lambda_\theta(A_1)) \otimes \cdots \otimes \mathrm{Poisson}(\Lambda_\theta(A_n))$$

とし，この２つは独立であると仮定すれば

$$M := \sum_{i=1}^{\infty} \frac{Z_i}{Z} \delta_{Y_i} \sim \mathcal{D}(\nu) \quad \textbf{ただし } Z = \sum_{i=1}^{\infty} Z_i \sim \mathrm{Gamma}(\theta)$$

(注意) (i) $M \perp\!\!\!\perp Z$. **(ii)** $\sum_{i=1}^{\infty} Z_i \delta_{Y_i}$ **は平均測度**ν**のガンマ乱測度.**

Poisson-Dirichlet分布（定義）

Dirichlet過程の表現（再掲）

$$M = \sum_{i=1}^{\infty} \frac{Z_i}{Z} \delta_{Y_i} \quad \text{において} \quad \{Z_i\}_{i=1}^{\infty} \perp\!\!\!\perp \{Y_i\}_{i=1}^{\infty} \ (: \text{ i.i.d.})$$

そこで，‘simplicial part’ $\{Z_i/Z\}_{i=1}^{\infty}$ の確率法則を調べる．ただし，

$$\sum \delta_{Z_i} \sim \text{Poisson}((0,\infty), \theta z^{-1} e^{-z} dz)$$

定義 $Z_1, Z_2, \ldots$ **を大きい順に並べたものを** $Z_{(1)} \geq Z_{(2)} \geq \cdots$ **と書くとき，**
正規化された無限列 $(Z_{(1)}/Z, Z_{(2)}/Z, \ldots)$ **が導く**

$$\nabla_\infty = \left\{ (p_1, p_2, \ldots) : p_1 \geq p_2 \geq \cdots \geq 0, \sum_{i=1}^{\infty} p_i = 1 \right\}$$

上の分布をパラメータ θ **の Poisson-Dirichlet分布 とよぶ．**

記号 $\text{PD}(\theta)$

Poisson-Dirichlet分布（導出）

定理 (Poisson-Dirichlet極限) [Kingman,1975], [彼の本,1993]
$n = 1, 2, \ldots$に対し

$$(P_1^{(n)}, \ldots, P_n^{(n)}) \sim D(a_1^{(n)}, \ldots, a_n^{(n)})$$

とする．もし$n \to \infty$のとき

$$\begin{cases} a_1^{(n)} + \cdots + a_n^{(n)} \to \theta > 0 \\ \max\{a_1^{(n)}, \ldots, a_n^{(n)}\} \to 0 \end{cases}$$

[cf. Poisson極限の場合]

ならば，順序統計量

$$(P_{(1)}^{(n)}, \ldots, P_{(n)}^{(n)}, 0, 0, \ldots) \text{の分布} \to \mathrm{PD}(\theta) \quad \text{[弱収束]}$$

(他の導出例) ランダム置換におけるサイクル長 [Vershik-Schmidt,1977]
素因数分解における「大きな因子」の対数 [Billingsley,1972]

ランダム離散確率測度が導くランダム分割

$M = \sum_i P_i \delta_{Y_i}$：ランダム離散確率測度
$X_1, \ldots, X_n$：M からの大きさ n のサンプル
} とする．

もし，$\{Y_i\}$：i.i.d.，Y_i の分布が連続（point massなし）であれば，
n の任意の整数分割 $n = n_1 + \ldots + n_k (n_j \in \mathrm{N})$ に対して

$$P\left(X_1, \ldots, X_n \text{の類別の大きさが分割} n = n_1 + \cdots + n_k \text{を与える} \mid M\right)$$

$$= \frac{n!}{n_1! \cdots n_k! \prod_{l=1}^{n} \sharp\{j : n_j = l\}!} \sum_{\substack{i_1, \ldots, i_k \\ \text{distinct}}} P_{i_1}^{n_1} \cdots P_{i_k}^{n_k}$$

（注意）$Y_i \neq Y_j (i \neq j)$ a.s.

Poisson-Dirichlet分布の表現

定義(Residual allocation model(RAM)) $\theta > 0$: **所与**

$\{W_i\}_{i=1}^{\infty}$ **:i.i.d.,** $W_i \sim \mathrm{Beta}(1, \theta)$

"stick breaking"
$$\begin{cases} V_1 := W_1 \\ V_2 := (1 - W_1)W_2 \\ V_3 := (1 - W_1)(1 - W_2)W_3 \\ \cdots \end{cases}$$

(注意) $\sum_{i=1}^{\infty} V_i = 1$**, a.s. であり,** $(V_1, V_2, \ldots)$**の分布はGEM分布と呼ばれる.**

定理 **[Patil-Taillie,1977]** $(V_{(1)}, V_{(2)}, \ldots) \sim \mathrm{PD}(\theta)$

(証明のエッセンス:有限次元近似) **[Donnelly-Joyce,1989]**

$(P_1, \ldots, P_n) \sim D(a, \ldots, a)$**に対して "stick breaking" の逆変換により**

$(W_1, \ldots, W_n)$**を定めると,** $\{W_i\}_{i=1}^{n}$ **⫫,** $W_i \sim \mathrm{Beta}(a + 1, (n - i)a)$**.**

そこで, $n \to \infty$**,** $na \to \theta$ **とする.**

Poisson-Dirichlet点過程（定義）

定義: $(P_1, P_2, \ldots) \sim \mathrm{PD}(\theta)$ **のとき，区間** $(0,1)$ **上の点過程** $\sum_{i=1}^{\infty} \delta_{P_i}$ **を**
パラメータ θ **の** Poisson-Dirichlet **点過程と呼ぶ.**

(注意) $\sum \delta_{P_i} \overset{\text{law}}{=} \sum \delta_{Z_i/Z} \overset{\text{law}}{=} \sum \delta_{V_i}$

定理 (確率母（汎）関数) $(P_1, P_2, \ldots) \sim \mathrm{PD}(\theta)$, $g : (0,\infty) \to \mathbb{C}$ **とする.**
$\int_0^\infty |g(z)-1| e^{-\lambda z} z^{-1} dz < \infty$ **なる** $\lambda > 0$ **に対して**

$$\frac{\lambda^\theta}{\Gamma(\theta)} \int_0^\infty ds s^{\theta-1} e^{-\lambda s} E\left[\prod_{i=1}^{\infty} g(sP_i)\right] = \exp\left(\theta \int_0^\infty \frac{dz}{z} e^{-\lambda z}(g(z)-1)\right)$$

(証明) $\{Z_i/Z\}_{i=1}^{\infty} \perp\!\!\!\perp Z \sim \mathrm{Gamma}(\theta)$
および $\sum \delta_{Z_i} \sim \mathrm{Poisson}((0,\infty), \theta z^{-1} e^{-z} dz)$ **を組み合わせる.**

Poisson-Dirichlet点過程（特性量）

定義: $k \geq 1$ とする．R 上の点過程 $\sum \delta_{P_i}$ の k 次**相関関数** q_k とは，次を満たす k 変数関数 q_k である：

$$(f, q_k)_{L^2(\mathbf{R}^k)} = E\left[\sum_{\substack{i_1,\ldots,i_k \\ \text{distinct}}} f(P_{i_1}, \ldots, P_{i_k})\right], \quad \forall f : \mathbf{R}^k \to [0, \infty)$$

（注意）q_k は母関数 $E[\prod_i (1 + t\phi(P_i))]$ における t^k の係数を記述する．

Wattersonの公式(1976) $(P_1, P_2, \ldots) \sim \mathrm{PD}(\theta)$ のとき，

$$q_k(p_1, \ldots, p_k) = \theta^k (p_1 \cdots p_k)^{-1} p_{k+1}^{\theta-1} \mathbf{1}_{\Delta_k}(p_1, \ldots, p_{k+1})$$

（注意）$f(p_1, \ldots, p_k) = p_1^{n_1} \cdots p_k^{n_k}$ ととる $\Longrightarrow$ Ewens抽出公式

Poisson-Dirichlet点過程（点過程論の活用）

これら特性量に関する結果の応用：$(P_1, P_2, \ldots) \sim \mathrm{PD}(\theta)$ **に対する**

- **有限次元密度** $P(P_1 \in dp_1, \ldots, P_n \in dp_n)$ **の表示** [Watterson,1976]
- **モーメント公式** [Griffiths,1979]
- $\theta \to \infty$ **のときのスケール極限** [Griffiths,1979], etc.

を再現可能．

2パラメータ版 $\mathrm{PD}(\alpha, \theta)$ **(**$0 \le \alpha < 1, \theta > -\alpha$**．** [Pitman-Yor,1997]**)**
に対しても拡張した結果が得られる． [H,2009]
ただし， $\mathrm{PD}(\alpha, \theta)$ **は，次の** $\{W_i\}$ **から定まるRAM に対応する：**

$$\{W_i\}_{i=1}^{\infty} \perp\!\!\!\perp, \quad W_i \sim \mathrm{Beta}(1-\alpha, \theta + \alpha i)$$

(注意) $\mathrm{PD}(0, \theta) = \mathrm{PD}(\theta)$