Poisson-Dirichlet Distribution With Small Mutation Rate

Shui Feng
McMaster University and Beijing Normal University

July, 2008

- Poisson-Dirichlet Distribution
- Connection to Population Genetics
- Asymptotic Results: Large Mutation Rate
- Asymptotic Results: Small Mutation Rate

Poisson-Dirichlet Distribution

Let $\{\tau_t, t \geq 0\}$ be a Gamma subordinator with Lévy measure

$$\Lambda(dx) = x^{-1}e^{-x}dx, x \geq 0.$$

Let $V_1(\theta) \geq V_2(\theta) \geq \cdots$ be the ranked jump sizes of $\{\tau_t, 0 \leq t \leq \theta\}$ for $\theta > 0$ and define

$$P_i(\theta) = \frac{V_i(\theta)}{\tau_\theta}, i = 1, \ldots.$$

The law Π_θ of $(P_1(\theta), P_2(\theta), ...)$ is called the **Poisson-Dirichlet distribution** with parameter θ.

Connection to Population genetics

Poisson-Dirichlet Distribution appears in many different contexts including Bayesian statistics, number theory, combinatorics, and population genetics. In the context of population genetics, the distribution describes the equilibrium proportions of different alleles in the infinitely many neutral alleles model with generator

$$L = \frac{1}{2}\sum_{i,j=1}^{\infty} x_i(\delta_{ij} - x_j)\frac{\partial^2}{\partial x_i \partial x_j} - \frac{\theta}{2}\sum_{i=1}^{\infty} x_i \frac{\partial}{\partial x_i}.$$

The parameter θ represents the scaled **mutation rate**. The connection discussed here is part of the so-called neutral theory, i.e., no individual has selective advantage.

Asymptotic Results: Large θ

Law of Large Numbers

In population genetics, $\theta = 4N_e u$ with u being the individual mutation rate and N_e the effective population size. Hence for fixed u, the limiting procedure of θ approaching infinity is equivalent to effective population size getting large.

WLLN: $\lim_{\theta \to \infty}(P_1(\theta), P_2(\theta), ...) = (0, 0, ...)$.

But with probability one and for all θ,

$$\sum_{i=1}^{\infty} P_i(\theta) = 1$$

Fluctuations

For each $r \geq 1$, consider random variables $Y_1, \ldots, Y_r$ such that

$Y_1 \sim e^{-y-e^{-y}}$, i.e., Y_1 has Gumbel distribution,

$Y_k \sim \frac{1}{(k-1)!} \exp\{-(ky + e^{-y})\}$,

$(Y_1, \ldots, Y_r) \sim \exp\{-(y_1 + \cdots + y_r) - e^{-y_r}\}, y_1 \geq y_2 \geq \cdots \geq y_r$.

Set $\beta(\theta) = \log\theta - \log\log\theta$.

Theorem 1. (Griffiths (79)) *For each* $r \geq 1$,

$$(\theta P_1(\theta) - \beta(\theta), ..., \theta P_r(\theta) - \beta(\theta)) \Rightarrow (Y_1, ..., Y_r)$$

when θ *goes to infinity.*

It follows from the theorem that

$$P_k(\theta) \approx \frac{Y_k}{\theta} + \frac{\log\theta}{\theta} - \frac{\log\log\theta}{\theta}.$$

Let $Z_i(\theta) = e^{-(\theta P_i(\theta) - \beta(\theta))}$ and $Z_i = e^{-Y_i}$. Then each Z_i is a $Gamma(i, 1)$ random variable and $(Z_1, \ldots, Z_r)$ has a joint density function

$$h(z_1, \ldots, z_r) = e^{-z_r}, 0 \leq z_1 \leq \cdots \leq z_r.$$

By continuous mapping theorem, for every $r \geq 1$,

$$(Z_1(\theta), \ldots, Z_r(\theta)) \Rightarrow (Z_1, ..., Z_r).$$

Large Deviations

Set

$$\nabla = \{(p_1, \cdots, p_2, \ldots) : p_1 \geq p_2 \geq \cdots \geq 0, \sum_{i=1}^{\infty} p_i \leq 1\}.$$

Theorem 2. (Dawson and F(06)). *The family* $\{\Pi_\theta : \theta > 0\}$ *on space* ∇ *satisfies a LDP with speed* θ *and rate function*

$$I(p_1, p_2, ...) = \begin{cases} \log \frac{1}{1-\sum_{i=1}^{\infty} p_i}, & \sum_{i=1}^{\infty} p_i < 1 \\ \infty, & \textit{else.} \end{cases}$$

Moderate Deviations

Choosing $a(\theta)$ such that

$$\lim_{\theta\to\infty} a(\theta) = \infty, \ \lim_{\theta\to\infty} \frac{a(\theta)}{\theta} = 0.$$

Then

$$a(\theta)[(P_1(\theta), P_2(\theta), \ldots) - \frac{\beta(\theta)}{\theta}(1, 1, \ldots)] \Rightarrow (0, 0, \ldots).$$

Large deviations correspond to $a(\theta) = 1$ or a constant, and fluctuation corresponds to $a(\theta) = \theta$.

Theorem 3. (F and Gao (2008)). *For above $a(\theta)$, the family $\{a(\theta)[(P_1(\theta), P_2(\theta), \ldots) - \frac{\beta(\theta)}{\theta}(1, 1, \ldots)] : \theta > 0\}$ on space R^∞ satisfies an LDP with speed $\theta/a(\theta)$ and rate function*

$$I(x_1, x_2, \ldots) = \begin{cases} \sum_{i=1}^{\infty} x_i, & \sum_{i=1}^{\infty} x_i < \infty, x_1 \geq \cdots \geq 0 \\ \infty, & \text{otherwise.} \end{cases}$$

Asymptotic Results: Small θ

Motivation

- Critical phenomenon of the infinitely many neutral alleles model by Schmuland
- Superprocess near extinction by Tribe
- Understanding the role of mutation

Law of "Large" Numbers

WLLN: $\lim_{\theta \to 0}(P_1(\theta), P_2(\theta), ...) = (1, 0, ...)$.

For any $n \geq 1$, set

$$\nabla_n = \{(p_1, ..., p_n, 0, 0, ...) \in \nabla : \sum_{i=1}^{n} p_i = 1\}$$

and

$$C = \bigcup_{n=1}^{\infty} \nabla_n.$$

Large Deviations

Theorem 4. $(\mathrm{F}(08))$. *The family* $\{\Pi_\theta : \theta > 0\}$ *on space* ∇ *satisfies a LDP as* θ *approaches zero with speed* $-\log\theta$ *and rate function*

$$S(p_1, p_2, ...) = \begin{cases} 0, & \mathbf{p} \in \nabla_1 \\ n-1, & \mathbf{p} \in \nabla_n, p_n > 0, n \geq 2 \\ \infty, & \mathbf{p} \notin C \end{cases}$$

Define

$$A = \{(p_1, p_2, \ldots) \in \nabla : \sum_{i=1}^{\infty} p_i < 1\},$$

and

$$B = \{(p_1, p_2, \ldots) \in \nabla : p_i > 0, i = 1, ...; \sum_{j=1}^{\infty} p_j = 1\}.$$

Then A, B, C are disjoint and $\nabla = A \cup B \cup C$. It is worth noting that

A = effective domain of large mutation LDP,

C = effective domain of small mutation LDP,

B = the space where Poisson-Dirichlet concentrates for all positive θ.

Selection Impact

For $\lambda > 0$, set

$$\mathbf{p} = (p_1, p_2, ...), \ H(\mathbf{p}) = \sum_{i=1}^{\infty} p_i^2,$$

$$C_H = \frac{1}{\int_{\nabla} e^{-2\lambda |\log \theta| H(\mathbf{p})} \Pi_\theta(d\mathbf{p})}.$$

Define a new probability $\Pi_{\lambda,H,\theta}$ on ∇, called the Poisson-Dirichlet distribution with overdominant selection, as

$$\Pi_{\lambda,H,\theta}(d\mathbf{q}) = C_H e^{-2\lambda |\log \theta| H(\mathbf{q})} \Pi_\theta(d\mathbf{q}).$$

The function H is the homozygosity, and under the overdominant selection heterozygote has advantage over homozygote.

Theorem 5. $(\mathrm{F}(2008))$. *The family* $\{\Pi_{\lambda,H,\theta} : \theta > 0\}$ *on space* ∇ *satisfies an LDP as* θ *goes to zero with speed* $-\log\theta$ *and rate function*

$$S_{\lambda,H}(\mathbf{p}) = S(\mathbf{p}) + 2\lambda H(\mathbf{p}) - \inf\{\frac{2\lambda + n(n-1)}{n} : n \geq 1\}.$$

For $m \geq 1$, set $\lambda_m = \frac{m(m+1)}{2}$. Then for $\lambda = \lambda_m$, the rate function is zero at two points $(\frac{1}{m}, ..., \frac{1}{m}, ...)$ and $(\frac{1}{m+1}, ..., \frac{1}{m+1}, ...)$; for λ in $(\lambda_m, \lambda_{m+1})$, the rate function has a unique zero point $(\frac{1}{m+1}, ..., \frac{1}{m+1}, ...)$.

$$\Pi_{\lambda,H,\theta} \Rightarrow \begin{cases} \delta_{(1,0,\ldots)}, & 2\lambda \in [0,\ 1 \cdot 2) \\ \delta_{(\frac{1}{2},\frac{1}{2},0,\ldots)}, & 2\lambda \in (1 \cdot 2,\ 2 \cdot 3) \\ \delta_{(\frac{1}{m+1},\ldots,\frac{1}{m+1},0,\ \ldots)}, & \lambda \in (\lambda_m, \lambda_{m+1}), m > 2. \end{cases}$$

For $\lambda = \lambda_m$, the limit will concentrate on $(\frac{1}{m}, ..., \frac{1}{m}, 0, \ldots)$ $(\frac{1}{m+1}, ..., \frac{1}{m+1}, 0, \ldots)$. It is worth noting that the critical point λ_m is the transition rate of Kingman's coalescent!