Asymptotic Behavior of Poisson-Dirichlet Distribution and Dirichlet Process with Two Parameters

Shui Feng
McMaster University and Beijing Normal University

- Poisson-Dirichlet Distribution
- An Example
- Dirichlet Process
- Asymptotic Results
- Two Parameter Generalization
- Perman's Formula
- Subordinator Representation
- Asymptotic Results for Two-parameter Model

Poisson-Dirichlet Distribution

Let $U_1, U_2, \cdots$ be a sequence of i.i.d. random variables with common distribution $Beta(1, \theta)$, i.e., density function of U_1 is

$$f(x) = \theta(1-x)^{\theta-1}, 0 \leq x \leq 1.$$

Define

$$X_1 = U_1, X_i = U_i(1-U_1)\cdots(1-U_{i-1}), i > 1$$

and let $(P_1(\theta), P_2(\theta), ...)$ be the decreasing order of $\{X_i : i \geq 1\}$. The law Π_θ of $(P_1(\theta), P_2(\theta), ...)$ is called the **Poisson-Dirichlet distribution** with parameter θ and $\{X_i : i \geq 1\}$ is called the GEM representation of Π_θ.

An Example

For each integer $n \geq 1$, let N_n be chosen at random from $1, 2, ..., n$. Consider the prime factorization of $N_n = \Pi_p p^{C_p(n)}$, and $\{C_p(n)\}$ is the multiples of p. Let

$K = \sum_p C_p(n)$,

$\tilde{C}_i(n) = i$th biggest component of $\{C_p(n)\}, i = 1, ..., K$,

$\tilde{C}_i(n) = 1, i \geq K,$

$L_i(n) = \log \tilde{C}_i(n), i \geq 1.$

Theorem 1.

$$(\frac{L_1(n)}{\log n}, \frac{L_2(n)}{\log n} ...) \to (P_1(1), P_2(1), ...), \text{ as } n \to \infty.$$

Dirichlet Process

Let $\xi_k, k = 1, \ldots$ be a sequence of i.i.d. random variables with common diffusive distribution ν on $[0,1]$, i.e., $\nu(x) = 0$ for every x in $[0,1]$. Set

$$\Xi_{\theta,\nu} = \sum_{k=1}^{\infty} P_k(\theta)\delta_{\xi_k}.$$

We call the law of $\Xi_{\theta,\nu}$ Dirichlet process, denoted by $Dirichlet(\theta, \nu)$.

Asymptotic Results

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

WLLN: $\lim_{\theta \to \infty} \Xi_{\theta,\nu} = \nu$.

Fluctuations

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

$Y_1 \sim e^{-y_1 - e^{-y_1}}$, 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 2. (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

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

Noting that for each $k \geq 1$, $E[X_k] = (\frac{\theta}{1+\theta})^{k-1}\frac{1}{\theta+1} \sim \frac{1}{\theta}$. Hence the ordering increases the value of $P_k(\theta)$ by a factor of $\log\theta$.

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

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

Theorem 3. (Dawson and F(06)). *The family of* $\{\Pi_\theta : \theta > 0\}$ *on space* ∇ *satisfies an LDP with speed* θ *and rate function*

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

Let $M_1(\nabla)$ be the space of probability measures on ∇. Consider the following function on $M_1(\nabla)$:

$$S(\cdot) = H(\nu|\cdot),$$

where for each μ in $M_1(\nabla)$, $H(\nu|\mu)$ is the relative entropy of ν with respect to μ.

Theorem 4. (Lynch and Sethuraman(87),Dawson and F(01)). *Assume that the support of* ν *is* $[0,1]$. *Then the family of* $\{Dirichlet(\theta,\nu) : \theta > 0\}$ *on space* $M_1(\nabla)$ *satisfies an LDP with speed* θ *and rate function* $S(\mu)$.

Two-Parameter Generalizations

For any α in $(0,1)$ and $\theta > -\alpha$, let $V_k, k = 1, 2, ...$, be a sequence of independent random variables such that V_k has $Beta(1-\alpha, \theta + k\alpha)$ distribution. Set

$$X_1^{\theta,\alpha} = V_1, \;\; X_n^{\theta,\alpha} = (1 - V_1)\cdots(1 - V_{n-1})V_n, \;\; n \geq 2.$$

Let $\mathbf{P}(\alpha, \theta) = (P_1(\alpha, \theta), P_2(\alpha, \theta), ...)$ denote $(X_1^{\theta,\alpha}, X_2^{\theta,\alpha}, ...)$ in descending order. The law of $\mathbf{P}(\alpha, \theta)$ is called the two-parameter Poisson-Dirichlet distribution, and is denoted by $PD(\alpha, \theta)$.

Let $\xi_k, k = 1, \ldots$ be a sequence of i.i.d. random variables with common diffusive distribution ν on $[0, 1]$, i.e., $\nu(x) = 0$ for every x in $[0, 1]$. Set

$$\Xi_{\theta,\alpha,\nu} = \sum_{k=1}^{\infty} P_k(\alpha, \theta)\delta_{\xi_k}.$$

We call the law of $\Xi_{\theta,\alpha,\nu}$ the two-parameter Dirichlet process, denoted by $Dirichlet(\theta, \alpha, \nu)$.

The two-parameter Poisson-Dirichlet is the most general distribution whose GEM representation is invariant under a procedure called *size-biased permutation.*

Question: What is the impact of α on the asymptotic behavior of the two-parameter model?

Perman's Formula

For $0 \le \alpha < 1$ and any constant $C > 0, \beta > 0$, let

$$h(x) = \alpha C x^{-(\alpha+1)}, x > 0,$$

and

$$c_{\alpha,\beta} = \frac{\Gamma(\beta+1)(C\Gamma(1-\alpha))^{\beta/\alpha}}{\Gamma(\beta/\alpha+1)}.$$

Let $\psi(t)$ be a density function over $(0, \infty)$ such that for all $\beta > -\alpha$

$$\int_0^\infty t^{-\beta}\psi(t)dt = \frac{1}{c_{\alpha,\beta}}.$$

Let $\{\tau_s : s \geq 0\}$ be the stable subordinator with index α. Then $\psi(t)$ is the density function of τ_1.

Set

$$\psi_1(t,p) = h(tp)t\psi(t\bar{p}), t > 0, 0 < p < 1, \bar{p} = 1 - p$$
$$\psi_{n+1}(t,p) = \begin{cases} h(tp)t \int_{p/\bar{p}}^{1} \psi_n(t\bar{p},q)d\,q, & p \leq 1/(n+1) \\ 0, & \text{else.} \end{cases}$$

Lemma 5. (Perman's Formula) *For each* $k \geq 1$*, let* $f(p_1,...,p_k)$ *denote the joint density function of* $(P_1(\alpha,\theta),...,P_k(\alpha,\theta))$*. Then*

$$f(p_1,...,p_k) = c_{\alpha,\theta}\int_0^\infty t^{-\theta} g_k(t,p_1,...,p_k)d\,t,$$

where for $k \geq 2, t > 0,\ 0 < p_k < \cdots < p_1, \sum_{i=1}^{k} p_i < 1,$ *and* $\hat{p}_k = 1 - p_1 - \cdots - p_{k-1}$,

$$g_k(t, p_1, ..., p_k) = \frac{t^{k-1} h(tp_1) \cdots h(tp_{k-1})}{\hat{p}_k} g_1(t\hat{p}_k, \frac{p_k}{\hat{p}_k})$$

and

$$g_1(t, p) = \sum_{n=1}^{\infty} (-1)^{n+1} \psi_n(t, p).$$

Subordinator Representation

Let $\{\sigma(t) : t \geq 0, \sigma_0 = 0\}$ be a subordinator with Lévy measure $x^{-(1+\alpha)} e^{-x} d\,x$, $x > 0$, and $\{\tau(t) : t \geq 0, \tau_0 = 0\}$ be a gamma subordinator that is independent of $\{\sigma_t : t \geq 0, \sigma_0 = 0\}$ and has Lévy measure $x^{-1} e^{-x} d\,x$, $x > 0$.

Lemma 6. (Pitman and Yor) *Let*

$$\gamma(\alpha, \theta) = \frac{\alpha \tau(\frac{\theta}{\alpha})}{\Gamma(1-\alpha)}.$$

For each $n \geq 1$, and each partition $0 < t_1 < \cdots < t_n = 1$ of E, let

$A_i = (t_{i-1}, t_i]$ *for* $i = 2, ..., n$, $A_1 = [0, t_1]$, *and* $a_j = \nu(A_j)$. *Set*

$$Y_{\alpha,\theta}(t) = \sigma(\gamma(\alpha,\theta)t), t \geq 0.$$

Then the distribution of $(\Xi_{\theta,\alpha,\nu}(A_1), ..., \Xi_{\theta,\alpha,\nu}(A_n))$ *is the same as the distribution of*

$$(\frac{Y_{\alpha,\theta}(a_1)}{Y_{\alpha,\theta}(1)}, ..., \frac{Y_{\alpha,\theta}(\sum_{j=1}^{n} a_j) - Y_{\alpha,\theta}(\sum_{j=1}^{n-1} a_j)}{Y_{\alpha,\theta}(1)}).$$

Asymptotic Results for Two-parameter Model

Law of Large Numbers

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

WLLN: $\lim_{\theta\to\infty}\Xi_{\theta,\alpha,\nu} = \nu$.

Fluctuations

For each $r \geq 1$, let $\infty > Y_1 > Y_2 > \cdots > Y_r > -\infty$ be as before.Set

$$\beta(\alpha,\theta) = \log\theta - (\alpha+1)\log\log\theta - \log\Gamma(1-\alpha).$$

Theorem 7. $(\mathrm{Handa(07)})$ *For each $r \geq 1$,*

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

when θ goes to infinity.

Thus

$$P_k(\alpha,\theta) \approx \frac{Y_k}{\theta} + \frac{\log\theta}{\theta} - \frac{(\alpha+1)\log\log\theta}{\theta} - \frac{\log\Gamma(1-\alpha)}{\theta}.$$

Large Deviations

Theorem 8. (F(07)). *The family of* $\{PD(\alpha,\theta):\theta>0\}$ *on space* ∇ *satisfies an LDP with speed* θ *and rate function*

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

Thus the parameter α has no impact on the LDP in this case.

For each μ in $M_1(\nabla)$, set

$$S_\alpha(\mu) = \sup_{f>0, f\in C_b(E)} \{\frac{1}{\alpha}\log(\int (f(x))^\alpha \nu(dx)) + 1 - \int f(x)\mu(dx)\}.$$

Theorem 9. $(\mathrm{F}(07))$. *Assume that the support of ν is $[0,1]$. Then the family of $\{Dirichlet(\theta,\alpha,\nu) : \theta > 0\}$ on space $M_1(\nabla)$ satisfies an LDP with speed θ and rate function $S_\alpha(\mu)$.*

It is not clear whether S_α converges to $S(\mu)$ as α approaches zero.