Moderate Deviations for Poisson-Dirichlet Distribution

Fuqing Gao
Wuhan University

This talk is based on joint work with Shui Feng

Outline

- Poisson-Dirichlet Distribution

  This section introduces a definition and some properties of Poisson-Dirichlet Distribution.

- Moderate Deviations for Poisson-Dirichlet Distribution

  The moderate deviation principle (MDP) for Poisson-Dirichlet distribution is presented in this section.

- Moderate Deviations for Homozygosity

  In this section we discuss moderate deviations for homozygosity.

Poisson-Dirichlet Distribution

- Definition.

  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.

- Poisson Process Representation.

  Consider a nonhomogeneous Poisson process Ξ_θ with mean measure density
  $$\theta u^{-1} e^{-u}, u > 0.$$
  Let $\sigma_1 \geq \sigma_2 \geq \cdots$ be the points of Ξ_θ in descending order, and $\sigma = \sum_{i=1}^\infty \sigma_i$. Set

  $$\mathbf{P}(\theta) = \left(\frac{\sigma_1}{\sigma}, \frac{\sigma_2}{\sigma}, \cdots \right).$$

  Then it is known that $\mathbf{P}(\theta)$ and σ are independent, and σ is a $Gamma(\theta, 1)$ random variable. The law of $\mathbf{P}(\theta)$ is also **Poisson-Dirichlet distribution** with parameter θ.

- Connection to Population Genetics.

  Poisson-Dirichlet Distribution appears in many different contexts including Bayesian statistics, number theory (prime number representation), and population genetics. In the context of population genetics, the distribution describes the equilibrium proportions of different alleles in the infinitely many alleles model. The parameter θ represents the **mutation rate**.

  The limiting procedure of large θ is equivalent to population size getting large. When the population grows, the proportion of types will become small.

- Law of Large Number. $\lim_{\theta\to\infty}(P_1(\theta), P_2(\theta), ...) = (0, 0, ...)$.

- Fluctuation Theorem $(\mathrm{Griffiths}\ (1979))$. For each $r \geq 1$, let $\infty > Y_1 > Y_2 > \cdots > Y_r > -\infty$ have a joint distribution with density

  $$\exp\{-(y_1 + \cdots + y_r) - e^{-y_r}\}.$$

  Set

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

  Then for each $r \geq 1$,

  $$(\theta P_1(\theta) - \beta(\theta), \cdots, \theta P_r(\theta) - \beta(\theta)) \Rightarrow (Y_1, ..., Y_r) \qquad as \;\; \theta \to \infty.$$

Now replacing the scale θ with a scale $a(\theta)$ such that

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

Then by the fluctuation Theorem, we have

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

Question: How fast does $\frac{a(\theta)}{\theta}(\theta P_k(\theta) - \beta(\theta))$ converge to zero?

When $a(\theta)$ is a constant, the question is a large deviation problem.

The next result is obtained by Dawson and Feng (2006,AAP).

- Large Deviations $(\text{Dawson and Feng, 2006})$. The family of $\{\Pi_\theta : \theta > 0\}$ satisfies an LDP with speed θ and rate function

$$I(p_1, p_2, ...) = \log \frac{1}{1 - \sum_{k=1}^{\infty} p_k}.$$

In particular, for each $n \geq 1$, the family $\{P_n(\theta) : \theta > 0\}$ satisfies an LDP with rate function

$$I_n(p) = \log \frac{1}{1 - np}.$$

If $a(\theta)$ satisfies

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

the question is a moderate deviation problem. Moderate deviation result lies between a large deviation result and a fluctuation result.

The following result is the MDP for Poisson-Dirichlet distribution.

Moderate Deviations for Poisson-Dirichlet Distribution

- Moderate deviations

Theorem 1. (Feng and Gao (2007)). *For above $a(\theta)$, the family $\{\frac{a(\theta)}{\theta}[\theta(P_1(\theta), P_2(\theta), \ldots) - \beta(\theta)(1, 1, \ldots)] : \theta > 0\}$ on space R^∞ satisfies an LDP with speed $\theta/a(\theta)$ and rate function*

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

- Sketch of the Proof.

  Step 1. To establish the LDP for $\frac{a(\theta)}{\theta}(\sigma_1-\beta(\theta),\cdots,\sigma_n-\beta(\theta))$ by analyzing asymptotic behavior of the density function of $\frac{a(\theta)}{\theta}(\sigma_1-\beta(\theta),\cdots,\sigma_1-\beta(\theta))$.

  Step 2. To prove that $\frac{a(\theta)}{\theta}(\theta P_1(\theta)-\beta(\theta),\cdots,\theta P_n(\theta)-\beta(\theta))$ and $\frac{a(\theta)}{\theta}(\sigma_1-\beta(\theta),\cdots,\sigma_n-\beta(\theta))$ are exponential equivalent for all $n\geq 1$.

  Step 3. To derive the LDP for $\{\frac{a(\theta)}{\theta}[\theta(P_1(\theta),P_2(\theta),\ldots)-\beta(\theta)(1,1,\ldots)]:\theta>0\}$ by the projective limit theorem.

Moderate Deviations for Homozygosity

- Homozygosity

  For $m \geq 2$,

$$H_m(\theta) = H_m(P_1(\theta), P_2(\theta), \ldots) = \sum_{i=1}^{\infty} P_i^m(\theta)$$

is called the homozygosity of order m. Set

$$\begin{aligned} Z_m(\theta) &= \sqrt{\theta}[\frac{\theta^{m-1}}{\Gamma(m)} H_m(\theta) - 1] \\ &= \sqrt{\theta}\frac{\theta^{m-1}}{\Gamma(m)}[H_m(\theta) - \frac{\Gamma(m)}{\theta^{m-1}}]. \end{aligned}$$

- CLT for Homozygosity (Griffiths (1979), Joyce, Krone and Kurtz (2002,2003))

$$Z_m(\theta) \Longrightarrow N(0, \frac{\Gamma(2m)}{(\Gamma(m))^2} - m^2).$$

- LDP for Homozygosity (Dawson and Feng (2006)). The family of the laws of $H_m(\theta)$ on space $[0,1]$ satisfies an LDP with speed θ and rate function

$$I(x) = \log \frac{1}{1 - x^{1/m}}.$$

- MDP for Homozygosity

Theorem 2. (Feng and Gao (2007)). *Let* $a(\theta)$ *satisfy*

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

and

$$\liminf_{\theta\to\infty} \frac{a^{1-\varepsilon}(\theta)}{\theta^{(m-1)/(2m-1)}} > 0,$$

for some $0 < \epsilon < 1/(2m-1)$. *Then the family* $a(\theta)\left(\frac{\theta^{m-1}}{\Gamma(m)} H_m(\theta) - 1\right)$ *satisfies an LDP with speed* $\frac{a^2(\theta)}{\theta}$ *and rate function*

$$\frac{z^2}{2(\Gamma(2m)/\Gamma(m)^2 - m^2)}.$$

The main idea of the proof is to explore the connection between homozygosity and Poisson process, and apply Campbell's theorem. The difficulty here is that the exponential moment is not finite, and thus a truncation procedure is used.

Choosing the scaling factor of $a(\theta) = \theta^{\gamma}$. Then the MDP obtained here requires that γ is between $\frac{m-1}{2m-1}$ and $\frac{1}{2}$.

It is natural to ask what happens for $\gamma \leq \frac{m-1}{2m-1}$.

We can prove that the LDP for $(a(\theta)(\frac{\theta^{m-1}}{\Gamma(m)}H_m(\theta) - 1), \frac{a^2(\theta)}{\theta}, I(x))$ and the compactness of $I(x)$ imply $\gamma \geq \frac{m-1}{2m-1}$.

Consider the case of $m = 2$ and set

$$Z^r(\theta) = \theta^r[H_2(\theta) - \frac{1}{\theta}].$$

Then the CLT corresponds to $r = \frac{3}{2}$. The MDP corresponds to r in $(\frac{4}{3}, \frac{3}{2})$. The LDP obtained corresponds $r = 0$. $(\frac{4}{3}, \frac{3}{2})$ is the range for Gaussian MDP and the MDP for r in $(0, \frac{4}{3}]$ will be non-Gaussian.

Further guess: a further phase transition may occur in the neighborhood of 1.

- Sketch proof of Theorem 2.

Step 1. Choose a positive function $\gamma(\theta) \to \infty$ that grows faster than a positive power of θ such that

$$\lim_{\theta\to\infty} \frac{\gamma(\theta)}{a^{(l-2)/(m-1)l}(\theta)} = 0, \qquad \lim_{\theta\to\infty} \frac{a^2(\theta)\gamma(\theta)}{\theta} = \infty.$$

Set

$$\tilde{G}_\theta^{(1)} = \sum_{j=1}^{\infty} \sigma_j I_{\{\sigma_j \le \gamma(\theta)\}}, \qquad \tilde{G}_\theta^{(m)} = \sum_{j=1}^{\infty} \sigma_j^m I_{\{\sigma_j \le \gamma(\theta)\}},$$

and

$$\tilde{G}_\theta = (\tilde{G}_\theta^{(1)} - E(\tilde{G}_\theta^{(1)}), \tilde{G}_\theta^{(m)} - E(\tilde{G}_\theta^{(m)})).$$

By Gärtner-Ellis theorem, it is easy to get that $\left(\frac{a(\theta)}{\theta}\tilde{G}_\theta, \frac{a^2(\theta)}{\theta}, \Lambda^*\right)$ satisfies LDP, where

$$\Lambda^*(x,y) := \frac{1}{2(\Gamma(2m) - \Gamma(m+1)^2)}(\Gamma(2m)x^2 - 2\Gamma(m+1)xy + y^2),$$

Step 2. Set

$$G_\theta^{(m)} = \sum_{j=1}^{\infty} \sigma_j^m, \qquad G_\theta = (\sigma - \theta, G_\theta^{(m)} - \Gamma(m)\theta).$$

Then the family $\frac{a(\theta)}{\theta}G_\theta$ and the family $\frac{a(\theta)}{\theta}\tilde{G}_\theta$ are exponential equivalent, and so the family $\frac{a(\theta)}{\theta}G_\theta$ satisfies a LDP with speed $\frac{a^2(\theta)}{\theta}$ and rate function $\Lambda^*(x,y)$.

Step 3. By the step 2 and the contraction principle, we can complete the proof of Theorem 2.

Thanks for your attention