Asymptotic Behaviour of Homozygosity

Shui Feng

McMaster University

shuifeng@mcmaster.ca

AMS Fall Central Sectional Meeting, Ann Arbor, Michigan

October 20-21, 2018

Outline

1. Homozygosity
   - Definition
   - LLN and Gaussian Limit
2. Large Deviations
   - Deviation From Zero
   - Deviation From One
   - Link Between the Two Rate Functions
3. Main Steps of Proof

Definition

Let $\gamma(t)$ denote the gamma subordinator with Lévy measure

$$\Lambda(d\,x) = x^{-1} e^{-x} d\,x, \quad x > 0.$$

For any $\theta > 0$, let $J_1(\theta) \geq J_2(\theta) \geq \cdots$ denote the jump sizes of $\gamma(t)$ over the interval $[0, \theta]$ in descending order. If we set $P_i(\theta) = J_i(\theta)/\gamma(\theta), i \geq 1$, then the law of

$$\mathbf{P}(\theta) = (\mathbf{P}_1(\theta), \mathbf{P}_2(\theta), \ldots)$$

is Kingman's Poisson-Dirichlet distribution $\mathbf{PD}(\theta)$. It is a probability on the infinite-dimensional simplex

$$\nabla_\infty = \{\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2, \ldots) : \mathbf{p}_1 \geq \mathbf{p}_2 \geq \cdots \geq 0, \sum_{\mathbf{i}=1}^{\infty} \mathbf{p_i} \leq 1\}.$$

Definition

For any integer $\mathbf{m} \geq 2$, the function

$$\mathbf{H}(\mathbf{p}; \mathbf{m}) = \sum_{\mathbf{i}=1}^{\infty} \mathbf{p}_{\mathbf{i}}^{\mathbf{m}}, \quad \mathbf{p} \in \nabla_{\infty}$$

is loosely called the homozygosity of order $\mathbf{m}$. The name is taken from population genetics where the homozygosity corresponds to $\mathbf{m} = 2$. It represents the probability that all samples are of the same type when a random sample of size $\mathbf{m}$ is selected from the population.

Definition

The function is closely associated with the Shannon entropy in communication, the Herfindahl-Hirschmam index in economics, and the Gini-Simpson index in ecology. It provides a measure of concentration of the population in terms of individual types with large values corresponding to higher concentration.

Definition

If the proportions are random, then homozygosity becomes a random variable.

Assume that the proportions of individual types follow distribution $\mathbf{PD}(\theta)$. The focus of this talk will be the random homozygosity

$$\mathbf{H}(\mathbf{P}(\theta); \mathbf{m}).$$

LLN

Question: What is the asymptotic behaviour of $\mathbf{H}(\mathbf{P}(\theta); \mathbf{m})$ when θ tends to infinity?

LLN:

$$\mathbf{H}(\mathbf{P}(\theta); \mathbf{m}) \to 0 \text{ in probability}, \ \theta \to \infty.$$

$$\frac{\theta^{\mathbf{m}-1}}{\Gamma(\mathbf{m})}\mathbf{H}(\mathbf{P}(\theta); \mathbf{m}) \to 1 \text{ in probability}, \ \theta \to \infty.$$

Gaussian Limit

Theorem (Joyce, Krone and Kurtz (02))

$$\sqrt{\theta}[\frac{\theta^{\mathbf{m}-1}}{\Gamma(\mathbf{m})}\mathbf{H}(\mathbf{P}(\theta);\mathbf{m}) - 1] \Rightarrow \mathbf{Z_m}$$

where $\mathbf{Z_m}$ *is a normal random variable with mean zero and variance*

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

Gaussian Limit

$$\mathbf{H}(\mathbf{P}(\theta);\mathbf{m}) \approx \frac{\Gamma(\mathbf{m})}{\theta^{\mathbf{m}-1}} + \frac{\Gamma(\mathbf{m})}{\theta^{\mathbf{m}-1/2}}\mathbf{Z}_{\mathbf{m}}$$

and

$$\frac{\theta^{\mathbf{m}-1}}{\Gamma(\mathbf{m})}\mathbf{H}(\mathbf{P}(\theta);\mathbf{m}) \approx 1 + \theta^{-1/2}\mathbf{Z}_{\mathbf{m}}$$

It is natural to investigate more refined structures associated with the limits

$$\mathbf{H}(\mathbf{P}(\theta);\mathbf{m}) \to 0,\ \ \theta \to \infty$$

and

$$\frac{\theta^{\mathbf{m}-1}}{\Gamma(\mathbf{m})}\mathbf{H}(\mathbf{P}(\theta);\mathbf{m}) \to 1,\ \ \theta \to \infty.$$

Large Deviations From Zero

Theorem (Dawson and F (06))

The family $\{\mathbf{H}(\mathbf{P}(\theta);\mathbf{m}) : \theta > 0\}$ *satisfies a LDP with speed* θ *and rate function*

$$\mathbf{I}(\mathbf{y}) = \begin{cases} \log \frac{1}{1-\mathbf{y}^{1/\mathbf{m}}}, & \mathbf{y} \in [0,1] \\ \infty, & \textit{else.} \end{cases}$$

Moderate Deviations

Let $\mathbf{a}(\theta)$ satisfy

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

and

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

for some ϵ in $(0, \frac{1}{2\mathbf{m}-1})$.

Moderate Deviations

Theorem (Gao and F (08))

The family $\mathbf{a}(\theta)\left(\frac{\theta^{\mathbf{m}-1}}{\Gamma(\mathbf{m})}\mathbf{H}(\mathbf{P}(\theta),\mathbf{m})-1\right)$ *satisfies a LDP with speed* $\frac{\mathbf{a}^2(\theta)}{\theta}$ *and rate function* $\frac{\mathbf{x}^2}{2(\Gamma(2\mathbf{m})/\Gamma(\mathbf{m})^2-\mathbf{m}^2)}$, $\mathbf{x}\in\mathbf{R}$.

Remark

Let $\mathbf{a}(\theta) = \theta^{\delta}$. Then moderate deviation holds for

$$\theta^{\delta}\left(\frac{\theta^{\mathbf{m}-1}}{\Gamma(\mathbf{m})}\mathbf{H}(\mathbf{P}(\theta), \mathbf{m}) - 1\right)$$

if and only if $\delta \in (\frac{\mathbf{m}-1}{2\mathbf{m}-1}, \frac{1}{2})$.

This indicates a significant departure from the Gaussian regime when δ is between 0 and $\frac{\mathbf{m}-1}{2\mathbf{m}-1}$.

Large Deviations From One

The case $\delta = 0$ corresponds to the large deviations of

$$\frac{\theta^{\mathbf{m}-1}}{\Gamma(\mathbf{m})}\mathbf{H}(\mathbf{P}(\theta), \mathbf{m})$$

from one.

Fundamental Differences From LDP for $\mathbf{H}(\mathbf{P}(\theta), \mathbf{m})$

- The state space is no longer compact
- Exponential tightness is not free
- Do not have exponential moment in the neighbourhood of zero

Large Deviations From One

Theorem (Dawson and F(16))

A large deviation principle holds for $\frac{\theta^{\mathbf{m}-1}}{\Gamma(\mathbf{m})}\mathbf{H}(\mathbf{P}(\theta);\mathbf{m})$ *as* θ *converges to infinity on space* $\mathbf{R}$ *with speed* $\theta^{1/\mathbf{m}}$ *and good rate function*

$$\mathbf{S}(\mathbf{x}) = \begin{cases} [\Gamma(\mathbf{m})(\mathbf{x}-1)]^{1/\mathbf{m}}, & \mathbf{x} \geq 1, \\ +\infty, & \textbf{otherwise}. \end{cases}$$

Note: The scale of deviations for $\mathbf{x} < 1$ is different from that of $\mathbf{x} > 1$.

Link Between the Two Rate Functions

Question: Can one derive the LDP for $\frac{\theta^{\mathbf{m}-1}}{\Gamma(\mathbf{m})}\mathbf{H}(\mathbf{P}(\theta);\mathbf{m})$ from the LDP for $\mathbf{H}(\mathbf{P}(\theta);\mathbf{m})$ or vice versa?

Answer: ??

Link Between the Two Rate Functions

Recall that the LDP for $\mathbf{H}(\mathbf{P}(\theta);\mathbf{m})$ has speed θ and rate function

$$\mathbf{I}(\mathbf{y}) = \begin{cases} -\log(1-\mathbf{y}^{1/\mathbf{m}}), & \mathbf{y} \in [0,1] \\ \infty, & \textbf{otherwise} \end{cases}$$

Since $\mathbf{H}(\mathbf{P}(\theta);\mathbf{m})$ and $\mathbf{H}(\mathbf{P}(\theta);\mathbf{m}) - \frac{\Gamma(\mathbf{m})}{\theta^{\mathbf{m}-1}}$ are exponentially equivalent, the same LDP holds for $\mathbf{H}(\mathbf{P}(\theta);\mathbf{m}) - \frac{\Gamma(\mathbf{m})}{\theta^{\mathbf{m}-1}}$.

Link Between the Two Rate Functions

Write $\frac{\theta^{\mathbf{m}-1}}{\Gamma(\mathbf{m})}\mathbf{H}(\mathbf{P}(\theta);\mathbf{m})$ as

$$\frac{\theta^{\mathbf{m}-1}}{\Gamma(\mathbf{m})}[\mathbf{H}(\mathbf{P}(\theta);\mathbf{m}) - \frac{\Gamma(\mathbf{m})}{\theta^{\mathbf{m}-1}}] + 1.$$

For $\mathbf{x} \in [1, \infty)$ and $\frac{\theta^{\mathbf{m}-1}}{\Gamma(\mathbf{m})}\mathbf{H}(\mathbf{P}(\theta);\mathbf{m}) = \mathbf{x}$, let $\mathbf{y} = \frac{\Gamma(\mathbf{m})}{\theta^{\mathbf{m}-1}}(\mathbf{x} - 1)$. Then

$$\begin{aligned}\exp\{-\theta\mathbf{I}(\mathbf{y})\} &= \exp\{-\theta^{1/\mathbf{m}+\mathbf{m}/(\mathbf{m}-1)}\log\frac{1}{1-(\frac{\Gamma(\mathbf{m})}{\theta^{\mathbf{m}-1}}(\mathbf{x}-1))^{1/\mathbf{m}}}\} \\ &\approx \exp\{-\theta^{1/\mathbf{m}}\mathbf{S}(\mathbf{x})\}.\end{aligned}$$

Main Steps of Proof

Step 1

Showing that LDP for general θ is equivalent to θ being integers.

Step 2

For integer θ, find a new representation of $\frac{\theta^{\mathbf{m}-1}}{\Gamma(\mathbf{m})}\mathbf{H}(\mathbf{P}(\theta);\mathbf{m})$ as

$$\frac{\theta^{\mathbf{m}-1}}{\Gamma(\mathbf{m})}\mathbf{H}(\mathbf{P}(\theta);\mathbf{m}) = \frac{\theta^{\mathbf{m}-1}}{\Gamma(\mathbf{m})}[\frac{1}{\gamma^{\mathbf{m}}(\theta)}\sum_{\mathbf{k}=1}^{\theta}\mathbf{W}_{\mathbf{k}}^{\mathbf{m}}\mathbf{H}_{\mathbf{k}}]$$

where $\mathbf{W}_1, \ldots, \mathbf{W}_\theta$ are independent copies of $\gamma(1)$, and independently, $\mathbf{H}_1, \ldots, \mathbf{H}_\theta$ are independent copies of $\mathbf{H}(\mathbf{P}(1);\mathbf{m})$.

Main Steps of Proof

Step 3

Exploring the independence and the LDP for gamma distribution to verify that the LDP for $\frac{\theta^{\mathbf{m}-1}}{\Gamma(\mathbf{m})}\mathbf{H}(\mathbf{P}(\theta);\mathbf{m})$ is equivalent to the LDP for

$$\frac{1}{\Gamma(\mathbf{m})\theta}\sum_{\mathbf{k}=1}^{\theta}\mathbf{W}_{\mathbf{k}}^{\mathbf{m}}\mathbf{H}_{\mathbf{k}}$$

Step 4

Applying Cramér's theorem for $\mathbf{x} < 1$.

Step 5

Applying Nagaev's result for $\mathbf{x} > 1$.

Generalizations

What about other random distributions?

References

- D.A. Dawson and S. Feng (2006). Asymptotic behavior of Poisson-Dirichlet distribution for large mutation rate. *Ann. Appl. Probab.* Vol. 16, No.2, 562–582.
- D.A. Dawson and S. Feng (2016). Large deviations for homozygosity. *Electron. Commun. Probab.* Vol. 21, no. 1, 1–8.
- A. Depperschmidt, P. Pfaffelhuber, and A. Scheuringer (2015). Some large deviations in Kingman's coalescent. *Electron. Commun. Probab.* **20**, 1–14.
- S. Feng and F.Q. Gao (2008). Moderate deviations for Poisson–Dirichlet distribution. *Ann. Appl. Probab.* **18**, No. 5, 1794–1824.
- R.C. Griffiths (1979). On the distribution of allele frequencies in a diffusion model. *Theor. Pop. Biol.* **15**, 140–158.
- P. Joyce, S.M. Krone, and T.G. Kurtz (2002). Gaussian limits associated with the Poisson–Dirichlet distribution and the Ewens sampling formula. *Ann. Appl. Probab.* **12**, No. 1, 101–124.
- S.V. Nagaev (1969). Integral limit theorems taking large deviations into account when Cramér's condition does not hold,I. *Theory of Probability and Its Applications*,14(1):51–64.

Thanks!