Diversity Indexes

Shui Feng

McMaster University

shuifeng@mcmaster.ca

The 13th Workshop on Markov Processes and Related Topic

Wuhan University and Beijing Normal University

July 17-21, 2017

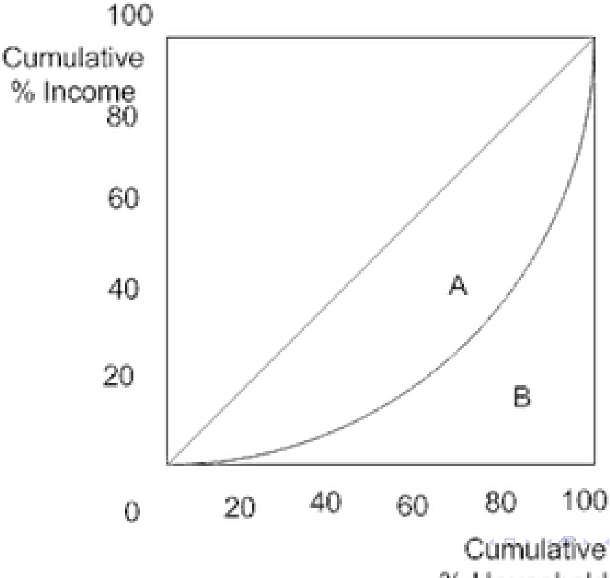
Outline

1. Indexes
   - Gini Index or Coefficient
   - Simpson Index
   - True Diversity Index
   - Shannon Entropy
2. Random Indexes
   - Equilibrium
   - Non-equilibrium
3. Mathematical Issues

Gini

Consider a population of individuals and their incomes. The graph consisting of points (x, y), where the bottom $x\%$ of the population make $y\%$ of the total income, is called the Lorenz curve.

Gini


100
Cumulative
% Income
80
60
40
20
0
20
40
60
80
100
A
B
Cumulative

Gini

Let A denote the area below the line of equality and above the curve and B the area below the curve. Then the Gini coefficient is

$$\frac{A}{A+B}.$$

A large value of Gini index corresponds to a high degree of concentration of wealth in a small percentage of the population. The most equal population has Gini index zero.

Simpson

Consider a population of individuals belonging to up to countable number of types labelled $1, 2, \ldots$. The proportion of type i is p_i. Then ([14])

$$\text{Simpson Index} = \sum_{i=1}^{N} p_i^2,$$

where N can be finite or infinite.

Clearly high degree of concentration corresponds to large value of Simpson Index. In population genetics, the index is called the homozygosity. It is also related to the Herfindahl-Hirschmam index ([10],[11]) in economics.

True Diversity

For any $r > 0, r \neq 1$, and any discrete distribution $\mathbf{p} = (p_1, p_2, \ldots)$, the True Diversity Index is defined as

$$D(\mathbf{p}; r) = \left(\sum_{i=1}^{N} p_i^r \right)^{1/(1-r)} .$$

$D(\mathbf{p}; r)$ represents the number of equally abundant types needed for the average proportional abundance of the types to equal that observed in the dataset of interest. In someway it can be viewed as the effective number of types.

The reciprocal of $D(\mathbf{p}; 2)$ is simply the Simpson index.

Shannon

The Shannon Index for a given $\mathbf{p}$ is given by

$$S(\mathbf{p}) = -\sum_{i=1}^{N} p_i \log p_i.$$

The case $r = 1$ is not defined for the true diversity index. But when N is finite one has

$$\lim_{r \to 1} D(\mathbf{p}; r) = D(\mathbf{p}; 1) = \exp\{-\sum_{i=1}^{N} p_i \log p_i\}.$$

Random Indexes

The diversity index becomes random when the discrete distribution $\mathbf{p}$ is replaced by random discrete distribution $\mathbf{P}$. Below are two constructions of random discrete distributions.

- Jumps of subordinators (equilibrium)
- Marginal distributions of measure-valued processes (non-equilibrium)

Equilibrium

A subordinator $\rho(t)$ is Lévy process with Lévy measure $\Lambda(d\,x), x > 0$. In the sequel, we assume all subordinators are drift free.

Example 1. The one-dimensional Poisson process N_t with parameter $\gamma > 0$ is a subordinator with Lévy measure $\Lambda(d\,x) = \gamma\delta_1(d\,x)$.

Example 2. For $\alpha \in (0,1)$, let $\Lambda(d\,x) = \frac{C\alpha}{\Gamma(1-\alpha)}x^{-(1+\alpha)}d\,, x > 0$. The corresponding subordinator $\rho(t)$ is called the stable subordinator with index α.

Equilibrium

Example 3. The subordinator $\{\rho(t) : t \geq 0\}$ is called a Gamma subordinator if its Lévy measure is

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

Example 4. The subordinator $\{\rho(t) : t \geq 0\}$ is a generalized Gamma subordinator with scale parameter one ([1]) if its Lévy measure is $\Lambda(d\,x) = \Gamma(1-\alpha)^{-1}x^{-(1+\alpha)}e^{-x}d\,x, \quad x > 0, 0 < \alpha < 1.$

Equilibrium

Given a subordinator $\rho(t)$, a fixed time T, let $J_1(T), J_2(T), \ldots$ denote the jump sizes of all jumps occurred between time zero and T. Assuming that the number of jumps is infinite and $\rho(T)$ is almost surely finite. Then a random discrete distribution can be constructed as

$$\mathbf{P}(\rho; T) = (P_1(\rho, T), P_2(\rho, T), \ldots) = (\frac{J_1(T)}{\rho(T)}, \frac{J_2(T)}{\rho(T)}, \ldots)$$

which arises as equilibrium distributions of some processes.

For any $r > 0$, the equilibrium random index discussed below has the form

$$H_r(\mathbf{P}(\rho; T)) = \sum_{i=1}^{\infty} P_i^r(\rho, T).$$

Non-equilibrium

Probability-valued processes offer a rich source of random discrete distributions that can be used to construct random indexes.

Example 1 (Wright-Fisher Diffusion). For any $N \geq 2$, let

$$\Delta_N = \{(p_1, \ldots, p_N) : 1 \leq p_i \leq 1, i = 1, \ldots, N, \sum_{k=1}^{N} p_k = 1\}.$$

The Wright-Fisher diffusion is a Δ_N-valued process $\mathbf{P}(t)$ with generator

$$\frac{1}{2}[\sum_{i,j=1}^{N} p_i(\delta_{ij} - p_j)\frac{\partial^2}{\partial p_i \partial p_j} + \theta \sum_{i=1}^{N}(\frac{1}{N-1} - \frac{N}{N-1}p_i)\frac{\partial}{\partial p_i}].$$

Example 2 (Infinitely-Many-Neutral-Alleles Model [5]). Let

$$\nabla_\infty = \{(p_1, p_2, \ldots) : p_1 \geq p_2 \geq \ldots \geq 0, \sum_{k=1}^{\infty} p_k = 1\}.$$

The Infinitely-Many-Neutral-Alleles Model is a ∇_∞-valued process $\mathbf{P}(t)$ with generator

$$\frac{1}{2}[\sum_{i,j=1}^{\infty} p_i(\delta_{ij} - p_j)\frac{\partial^2}{\partial p_i \partial p_j} - \sum_{i=1}^{\infty} \theta p_i \frac{\partial}{\partial p_i}].$$

Non-equilibrium

Example 3 (Petrov Diffusion [13]). This is a ∇_∞-valued process $\mathbf{P}(t)$ with generator

$$\frac{1}{2}[\sum_{i,j=1}^{\infty} p_i(\delta_{ij} - p_j)\frac{\partial^2}{\partial p_i \partial p_j} - \sum_{i=1}^{\infty}(\alpha + \theta p_i)\frac{\partial}{\partial p_i}].$$

Example 4 (GEM Process [8]). Let

$$\Delta = \{(x_1, x_2, \dots) : 1 \le x_i \le 1, i = 1, \dots, \sum_{k=1}^{\infty} x_k = 1\}.$$

The GEM process is a Δ-valued diffusion process $\mathbf{P}(t)$ with generator

$$\sum_{i,j=1}^{\infty} a_{ij}(\mathbf{x})\partial_{ij}^2 + \sum_{i=1}^{\infty} b_i(\mathbf{x})\partial_i,$$

where

$$a_{ij}(\mathbf{x}) := x_i x_j \sum_{k=1}^{i\wedge j} \frac{(\delta_{ki}(1 - \sum_{l=1}^{k-1} x_l) - x_k)(\delta_{kj}(1 - \sum_{l=1}^{k-1} x_l) - x_k)}{x_k(1 - \sum_{l=1}^{k} x_l)},$$

$$b_i(\mathbf{x}) := x_i \sum_{k=1}^{i} \frac{(\delta_{ik}(1 - \sum_{l=1}^{k-1} x_l) - x_k)(a_k(1 - \sum_{l=1}^{k-1} x_l) - (a_k + b_k)x_k)}{x_k(1 - \sum_{l=1}^{k} x_l)}.$$

and $a_k, b_k > 0$, $\inf_i b_i \ge \frac{1}{2}$.

Example 5 (Weak Interaction Process [7]). This is a Δ-valued diffusion process $\mathbf{P}(t)$ with generator

$$\sum_{k=1}^{\infty}\left[x_k(1-||\mathbf{x}||)\partial_k^2+(\alpha_k(1-||\mathbf{x}||)-\alpha_\infty x_k)\partial_k\right]$$

where $||\mathbf{x}||=\sum_{i=1}^{\infty}x_i,(\alpha_1,\alpha_2,\ldots,\alpha_\infty)\in\Delta$.

Math Issue

The random discrete distribution $\mathbf{P}$ usually depends on some parameters. The random diversity index can serve as a good estimator for some of these parameters. It is thus natural to consider

- The consistency or law of large numbers
- The confidence interval or fluctuation results such as CLT
- More refined asymptotic information such as moderate deviations and large deviations

Math Issue: One Fully Understood Case

Let $\rho(t)$ be the gamma subordinator, $r \geq 2$, and $T = \theta$. Asymptotic behaviour of the random diversity index

$$H_r(\mathbf{P}(\rho; \theta))$$

is known completely when θ converges to infinity.

LLN and CLT

LLN: $H_r(\mathbf{P}(\rho;\theta)) \to 0, \theta \to \infty$.

Gaussian Limit([9],[12]):

$$\sqrt{\theta}[\frac{\theta^{r-1}}{\Gamma(r)} H_r(\mathbf{P}(\rho;\theta)) - 1] \Rightarrow Z_r$$

where Z_r is a normal random variable with mean zero and variance

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

Large Deviations

Theorem (Dawson and F (06))

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

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

Moderate Deviations

Let $a(\theta)$ satisfy

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

Theorem (Gao and F (08))

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

Large Deviations

Theorem (Dawson and F(16))

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

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

Math Issue: Future Studies

- Other subordinators
- Non-equilibrium case

References

[1] A. Brix (1999). Generalized gamma measures and shot-noise Cox processes. *Adv. Appl. Probab.* **31**, 929–953.

[2] 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.

[3] D.A. Dawson and S. Feng (2016). Large deviations for homozygosity. *Electron. Commun. Probab.* Vol. 21, no. 1, 1–8.

[4] A. Depperschmidt, P. Pfaffelhuber, and A. Scheuringer (2015). Some large deviations in Kingman's coalescent. *Electron. Commun. Probab.* **20**, 1–14.

[5] S.N. Ethier and T.G. Kurtz (1981). The infinitely-many-neutral-alleles diffusion model. *Adv. Appl. Probab.* **13**, 429–452.

[6] S. Feng and F.Q. Gao (2008). Moderate deviations for Poisson–Dirichlet distribution. *Ann. Appl. Probab.* **18**, No. 5, 1794–1824.

[7] S. Feng, L. Miclo and F.-Y. Wang (2017). Poincaré inequality for Dirichlet distributions and infinite-dimensional generalizations. *ALEA, Lat. Am. J. Probab. Stat.*, Vol.14, 361–380.

References

[8] S. Feng and F-Y. Wang (2007) .A class of infinite-dimensional diffusion processes with connection to population genetics. *J. Appl. Probab.*, **44**, (2007), 938–949.

[9] R.C. Griffiths (1979). On the distribution of allele frequencies in a diffusion model. *Theor. Pop. Biol.* **15**, 140–158.

[10] O.C. Herfindahl (1950). Concentration in the U.S. Steel Industry. Unpublished doctoral dissertation, Columbia University.

[11] A.O. Hirschman. National power and the structure of foreign trade. University of California Press,Berkeley, 1945.

[12] 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.

[13] L.A. Petrov (2009). Two-parameter family of infinite-dimensional diffusions on the Kingman simplex. *Funct. Anal. Appl.* **43**, No. 4, 279–296.

[14] E.H. Simpson, E. H (1949). Measurement of diversity. *Nature*, **163**: 688-688.

The End