Truncation, the Lynden-Bell estimator, and galaxy data

1. Definitions

Suppose there are i.i.d. pairs (X_k, Y_k) of variables for $k = 1, ..., N$ where N is unknown to the observer. Within each pair, X_k and Y_k are independent positive real variables with distributions F and G respectively.

In the "truncation" or "left truncation" model, the specific restriction is that the pair of values (X_k, Y_k) is observed if and only if $Y_k \leq X_k$. Moreover, the index k is not observed. One wants to estimate F.

Suppose then that we observe (x_j, y_j), $j = 1, ..., n$, so that we observe a value of n and know about N at first that $N \geq n$. Recall the survival function corresponding to F, $S(x) \equiv 1 - F(x)$,

2. The Lynden-Bell estimator

Let ξ_i for $i = 1, ..., m$ be the distinct values of x_j. What is called the *Lynden-Bell* estimator of $S(x)$ is

$$\widehat{S}_n(x) = \prod_{\xi_i \leq x} \left(1 - \frac{r_i}{nC_n(\xi_i)}\right) \tag{1}$$

where r_i is the number of $j \leq n$ such that $x_j = \xi_i$ and

$$C_n(s) = \frac{1}{n} \sum_{j=1}^{n} 1_{\{y_j < s \leq x_j\}}.$$

These formulas are as given by Woodroofe (1985, (8)) and Chen et al. (1995, (1)), originating with Lynden-Bell (1971).

3. Absolute and apparent magnitudes for astronomical objects

Magnitudes were first assigned in ancient times to stars, with the brightest being assigned first magnitude, a next-brightest category second magnitude, and so on to the faintest stars visible to the naked eye under good conditions, 6th magnitude. In a modern urban area, only stars up to magnitude 3 or 4 can be seen by unaided eye due to ambient artificial light and air pollution.

With modern units, a difference of 1.0 in magnitude corresponds to a factor of $10^{0.4} \doteq 2.512$ in brightness. The "apparent magnitude" is the magnitude as measured from Earth or the vicinity (e.g. from the Hubble Space Telescope). The brightest star, Sirius, has an apparent magnitude -1.47. The *absolute magnitude* M is the apparent magnitude an object would have if seen at a distance of 10 parsecs, about 32.6 light years. Let m be the apparent magnitude of an object such as a galaxy. Then we have

$$M = m - 5(\log_{10} D_L - 1) \tag{2}$$

where D_L is the "luminosity distance" to the object, measured in parsecs.

For not too distant objects, D_L agrees with usual notions of distance. For the most distant known objects, some quasars, general relativistic effects complicate the evaluation of D_L. The Hubble relation is defined in terms of "proper distance" D_p and is given by

$$v = H_0 D_p \tag{3}$$

where H_0, the Hubble constant, is about 73 (km/sec)/(Mpc) by a current estimate, and where Mpc abbreviates megaparsec (10^6 parsecs). The velocity v of recession away from us is calculated from the redshift z by

$$v = cz/(1+z) \tag{4}$$

where c is the velocity of light, 299,795 km/sec. Combining (3) and (4) gives

$$D_p = \frac{v}{H_0} = d(z) := \frac{cz}{H_0(1+z)} \tag{5}$$

in megaparsecs, or 10^6 times that in parsecs. Approximating D_L by D_p for galaxies in the sample we'll consider, we then get from (2)

$$\begin{aligned} M &= m - 5\left(-1 + 6 + \log_{10}\left(d(z)\right)\right) \\ &= m - 25 - 5\left(\log_{10}\left(d(z)\right)\right). \end{aligned} \tag{6}$$

Let

$$L(z) := \log_{10}(d(z)).$$

Suppose given some observations (z_i, m_i) of redshift and apparent magnitude for the ith galaxy in a sample, $i = 1, \ldots, n$, where there is some largest (faintest) value m_c of apparent magnitude included in the sample. Let M_i be the absolute magnitude of the ith galaxy. Then from (2) we have $M_i + 25 + 5L(z_i) \le m_c$. Let $X_i := C \cdot 10^{-.4M_i}$, which is a measure of intrinsic brightness of the ith galaxy, where C is a constant depending on units in which brightness is measured. From (6) and $m_i \le m_c$ we have

$$X_i \ge Y_i := C \cdot 10^{-.4m_c + 10 + 2L(z_i)} = C_c 10^{2L(z_i)} \tag{7}$$

where $C_c = C \cdot 10^{-.4m_c+10}$ is a constant depending on m_c. Let $d_i := d(z_i)$ be the estimated distance to the ith galaxy. Then $Y_i = C_c d_i^2$, so (7) is equivalent to $X_i/d_i^2 \geq C_c$. The apparent brightness of an object with a given intrinsic brightness decreases with the square of the distance. The magnitude is a constant times the logarithm of the brightness, which involves a somewhat arbitrary choice of units, but the definitions taken together do fit with decrease of apparent brightness as inverse squared distance.

From estimating F, we may hope get a valid distribution for the upper tail of the distribution of X_i, the intrinsic brightness of the ith galaxy, or equivalently, of the distribution of absolute magnitudes which are most negative (lower tail of the distribution of absolute magnitudes). Estimation of the distributions for intrinsically faint objects is more difficult, as they can only be seen relatively nearby. Fainter than large galaxies are "dwarf" galaxies. Our own galaxy, the Milky Way, is estimated to have about 200 billion stars. It has several satellite dwarf galaxies, of which the largest, the Large Magellanic Cloud, contains about 30 billion stars. It is easily visible in the sky from the Southern Hemisphere. It is said to be intermediate in size between most dwarf galaxies and most galaxies seen at large distances. Smaller and smaller galaxies are being found. There are "hobbit galaxies" smaller than the originally known dwarf galaxies. In 2004, "ultra-compact dwarfs" were reported, which may contain "only" 100 million stars, beside having remarkably small diameter. The distribution of brightness of galaxies at the faint end seems very hard to estimate, and not feasible at all from samples of mainly bright, quite distant galaxies such as those in the direction of the Corona Borealis supercluster.

Should one also estimate G and possibly also N as the sources mention? I think for our data set, of galaxies in the direction of Abell 2067, estimating G would not be estimating anything general. From the Postman, Huchra, and Geller data selected from the the Corona Borealis supercluster, with $m_c = 15.7$, the distribution of z would be unimodal with a mode a little more than .07. From the Small et al. data, with $m_c = 19$, there is an additional mode around $z = .11$, and both modes are seen by limiting the direction to that of Abell 2067. In fact I focused on that direction after noticing there was a rather newly named cluster "A2067B," so that the physical A2067 is at z around 0.07 and A12067B has z around 0.11. Naturally, there are both foreground and background galaxies in the same direction. It seems to me there is no good reason to expect modes near the same values of z in other directions in the sky.

Another difficulty in estimating a distribution for the distribution of z_i, or equivalently for the distribution G of Y_i after the transformation in (7), is as follows. In (7) we also see the Y_i are proportional to d_i^2. Consider a relatively short range of values of d_i^2, say $a \leq d_i^2 \leq a+h$, where $0 < h << a$, or equivalently

$$\sqrt{a} \leq d_i \leq \sqrt{a+h} \doteq \sqrt{a} + \frac{h}{2\sqrt{a}}$$

by a short Taylor expansion. This gives a spherical shell of radius $\sqrt{a}$ and thickness $h/(2\sqrt{a})$, whose volume, if geometry is Euclidean, is approximately $4\pi a \cdot (h/(2\sqrt{a}) \doteq 2\pi h\sqrt{a}$, which increases with a. If galaxies are on average distributed equally in equal volumes of space, then the density of d_i would be increasing as $\sqrt{a}$, which would be unbounded and not normalizable. Actually space, although approximately Euclidean at small scales (except near quasars?) may not be at very large distances. The large-scale geometry of the universe (cosmology) could affect the distribution of z_i and d_i for galaxies.

The estimation of G might make more sense for the special class of objects (quasars of a certain type) that Lynden-Bell studied.

Similarly, the total number N of all galaxies, even in the direction of A2067, may be virtually unbounded, or at any rate, not reasonably estimable from a given sample with bounded apparent magnitude.

In (7), the values of m_c and so C_c depend on the study, for example m_c was 15.7 in Postman, Huchra and Geller, and is 19.0 in Small, Sargent and Hamilton. Thus the definition of Y_i depends on this m_c. In this kind of situation what would it mean to let (N and) n become very large? In fact, a larger sample might well come along with a larger m_c, so that the definition of Y_i would change.

REFERENCES

Chen, Kani, Chao, Min-Te, and Lo, Shaw-Hwa (1995). On strong uniform consistency of the Lynden-Bell estimator for truncated data. *Ann. Statist.* **23**, 440–449.

Lynden-Bell, D. (1971). A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monthly Notices Royal Astron. Soc.* **155**, 95–118.

Woodroofe, M. (1985). Estimating a distribution function with truncated data. *Ann. Statist.* **13**, 163–177.