$\log(N) - \log(S)$: A Measuring Stick for the Universe

P.D. Baines[1], I.S. Udaltsova[1], A. Zezas[2,3] and V.L. Kashyap[3]

[1]Department of Statistics, UC Davis, [2]University of Crete and [3]Harvard-Smithsonian Center for Astrophysics

1. Introduction

The number of sources as a function of flux ($\log(N) - \log(S)$) is an important tool for describing and investigating the properties of various types of source populations. In practice, observations intended to measure the flux distribution are subject to a number of natural and detector induced uncertainties and biases. The most important consequence of these effects is that a subset of the source population of interest will be unobserved. Since fainter sources are more likely to be unobserved, the missing data mechanism is *non-ignorable* and can lead to serious inferential bias unless accounted for.

We develop a Bayesian method for estimating: (i) the number of sources unobserved due to detector effects at a given sensitivity, (ii) the flux of all observed sources, and, (iii) the parameters of the $\log(N) - \log(S)$ curve (e.g., the slope if assuming a power law). By modeling the missing data mechanism we naturally correct for possible detection biases (e.g., Eddington bias) and obtain posterior distributions that account for detector uncertainties.

2. Data

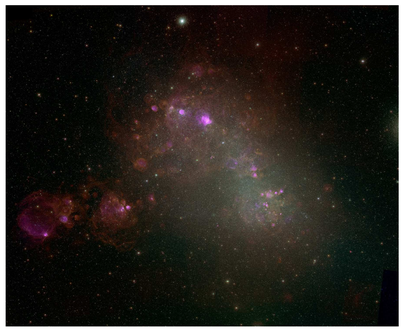

FIGURE 1: An optical image of the Small Magellanic Cloud (Anglo-Australian Telescope)

The Small Magellanic Cloud (SMC) is a nearby galaxy which is ideally suited for the study of young X-ray source populations. Previous observations with ROSAT (e.g. Haberl et al, 2000), ASCA (Yokogawa et al, 2003), RXTE (Laycock et al, 2005), XMM-Newton (e.g. Saski et al, 2000) and Chandra (Zezas et al, 2003) have revealed a rich population of X-ray binary pulsars (XBPs). The X-ray emission of these objects is produced by gas from a companion star that is falling onto the pulsar. The pilot data are based on observations of 26 X-ray emitting pulsars in the SMC observed with the Chandra ACIS-I detector.

3. The $\log(N) - \log(S)$ Curve

The starting point for most $\log(N) - \log(S)$ analyses is the *Power law* model:

$$N(>S) \;=\; \sum_{i=1}^{N} I_{\{S_i>S\}} \;\approx\; \alpha S^{-\theta},\; S > S_{min}$$

which specifies the (unnormalized) survival function $N(>S)$, as a function of the flux S. Taking the logarithm of both sides gives the linear $\log(N) - \log(S)$ relationship. The power-law relationship defines the marginal survival function of the population, and the marginal distribution of flux can be seen to be a Pareto distribution:

$$S_i | S_{min}, \theta \overset{iid}{\sim} Pareto\,(\theta, S_{min})\,, \qquad i = 1, \ldots, N.$$

The analyst must specify S_{min}, a threshold above which we seek to estimate θ.

Overview

The goals of this work are to:

- Develop a methodology for estimating the $\log(N) - \log(S)$ relationship in a wide range of source populations
- Account for all sources of observational uncertainty, bias and possible missing data in astronomical observations
- Allow for extensions to non power-law source populations

Our method provides:

- Fast computation that is insensitive to the number of missing sources, thus allowing applications to faint source populations
- Realistic uncertainty estimates on the estimated $\log(N) - \log(S)$
- The ability to test for breaks in the power law relationship

4. Model Specification

The total number of sources (unobserved and observed), denoted by N, is modeled as:

$$N \sim NegBinom\,(\alpha, \beta)\,, \qquad \text{with shape } \alpha \text{ and scale } \beta.$$

Next, we describe the observational process and detector effects. We observe photon counts contaminated with background noise and other detector effects, $Y_i^{tot} = Y_i^{src} + Y_i^{bkg}$,

$$Y_i^{src} | S_i, L_i, E_i \overset{iid}{\sim} Pois\,(\lambda(S_i, L_i, E_i))\,, \qquad Y_i^{bkg} | L_i, E_i \overset{iid}{\sim} Pois\,(k(L_i, E_i))\,.$$

The functions λ and k represent the intensity of source and background, respectively, for a given flux S_i, location L_i and effective exposure time E_i.

The probability of a source being detected, $g\,(S_i, L_i, E_i)$, is determined by the detector sensitivity, background and detection method. The marginal detection probability as a function of θ is defined as $\pi(\theta) = \int g(S_i, L_i, E_i) \cdot p(S_i|\theta) \cdot p(L_i, E_i)\; dS_i\; dE_i\; dL_i$. The prior on θ is assumed to be: $\theta \sim Gamma(a, b)$.

The posterior distribution, marginalizing over the unobserved fluxes, can be shown to be:

$$p\left(N, \theta, S_{obs} Y_{obs}^{src} | n, Y_{obs}^{tot}\right)$$
$$\propto p\,(N) \cdot p\,(\theta) \cdot p\,(n|N, \theta) \cdot p\,(S_{obs}|n, \theta) \cdot p\left(Y_{obs}^{tot}|n, S_{obs}\right) \cdot p\left(Y_{obs}^{src}|n, Y_{obs}^{tot}, S_{obs}\right).$$

5. Computational Details

The Gibbs sampler consists of four steps:

$$\left[Y_{obs}^{src}|n, Y_{obs}^{tot}, S_{obs}\right], \quad \left[S_{obs}|n, Y_{obs}^{tot}, Y_{obs}^{src}, \theta\right], \quad [\theta|n, N, S_{obs}], \quad [N|n, \theta]\,.$$

- Sample the observed photon counts $Y_{obs,i}^{src}|n, Y_{obs,i}^{tot}, S_{obs,i} \sim Binom\left(Y_{obs,i}^{tot}, \frac{\lambda(S_{obs,i}, L_{obs,i}, E_{obs,i})}{\lambda(S_{obs,i}, L_{obs,i}, E_{obs,i}) + k}\right), i = 1, \ldots, n.$
- Sample the fluxes $S_{obs,i}, i = 1, \ldots, n$ with Metropolis-Hastings algorithm, using a $t-$ or normal proposal distribution.
- Sample the power-law slope θ using the Metropolis-Hastings algorithm, with a $t-$ or normal proposal distribution.
- Compute the posterior distribution for the total number of sources, N, using numerical integration. Sample from the resulting distribution.
- The marginal detection probability $\pi(\theta)$ is pre-computed via the numerical integration.

6. Results

Figure 2, below, shows the uncertainties in the source fluxes and a visual representation of the power-law relationship and the posterior estimates for power-law slope and the total number of sources.

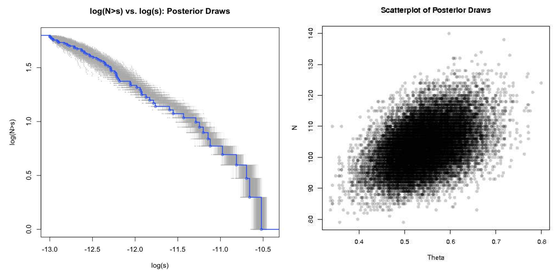

FIGURE 2: [Left] Simulated $\log(N) - \log(S)$. Posterior draws (gray), truth (blue). [Right] Bivariate plot of posterior draws of N and θ.

Empirical results comparing the MSE of different estimators for N and θ for simulated data are presented below. The posterior median is seen to be the preferred estimator.

MSE	N		θ	
Level of Exposure	Median	Mean	Median	Mean
Low	215.96	291.82	0.05439	0.07481
Medium	121.26	168.91	0.05558	0.07407
High	68.23	95.36	0.04578	0.05987

Figure 3 shows the results of our analysis for the SMC data. We note that there is evidence of a possible break in the power-law. In previous work, Zezas et al. (2003) estimated a power-law slope of $\hat{\theta} = 0.45$. The posterior median from our analysis is $\theta = 0.38$, with the 95% posterior interval consistent with competing estimators. We note that given the possible non-linearity of the $\log(N) - \log(S)$, more work is needed to allow for a broken power-law or more general parametric forms.

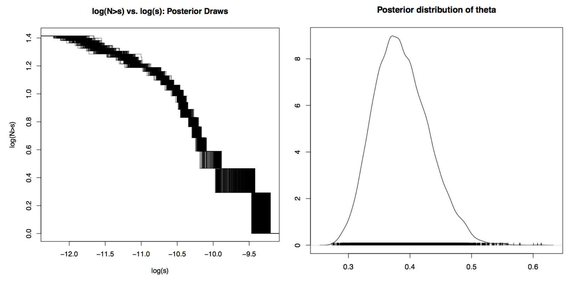

FIGURE 3: Log(N)-Log(S) from posterior draws of flux S [left], posterior of θ [right] based on SMC data.

7. Future Work & References

- Extension to broken power laws and more general parametric forms for $\log(N) - \log(S)$
- Generalization of detection probability curves to depend on more source-specific quantities
- Quantification of missing data impact under common observation conditions

References:

- A. Zezas et al. (2004) Chandra survey of the 'Bar' region of the SMC *Revista Mexicana de Astronoma y Astrofsica (Serie de Conferencias)* Vol. 20. IAU Colloquium 194, pp. 205-205.
- R.J.A. Little, D.B. Rubin. (2002) *Statistical analysis with missing data*, Wiley.