Bayesian Estimation of $\log N - \log S$

Paul Baines, Irina Udaltsova

Department of Statistics
University of California, Davis

July 12th, 2011

Introduction

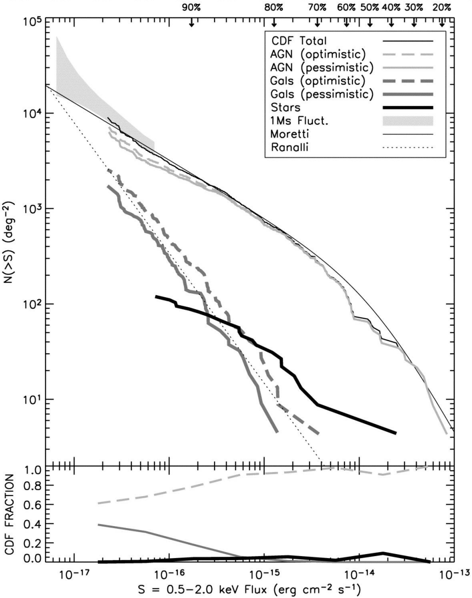
What is ‘$\log N - \log S$’?

- Cumulative number of sources detectable at a given sensitivity
- Defined as:

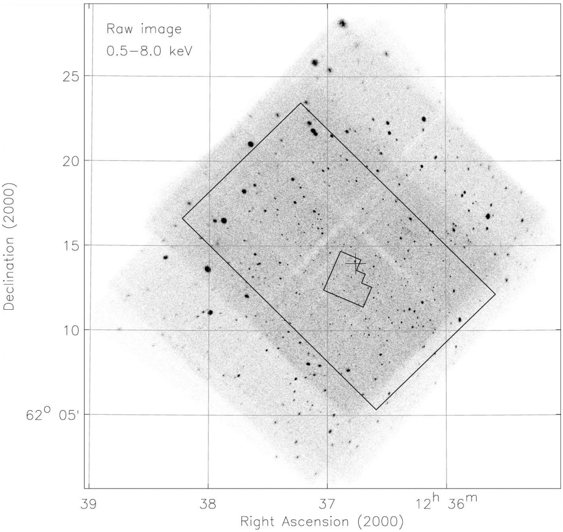
$$N(> S) = \sum_i I_{\{S_i > S\}}$$

  i.e., the number of sources brighter than a threshold.
- Considering the distribution of sources, this is related to the survival function i.e., N(>S)=N×(1-F(S))
- ‘$\log N - \log S$’ refers to the relationship beween (or plot of) $\log_{10} N(> S)$ and $\log_{10} S$.
- Why do we care?

  Constrains evolutionary models, dark matter distribution etc.



90%
80%
70%
60%
50%
40%
30%
20%
CDF Total
AGN (optimistic)
AGN (pessimistic)
Gals (optimistic)
Gals (pessimistic)
Stars
1Ms Fluct.
Moretti
Ranalli
N(>S) (deg^{-2})
CDF FRACTION
S = 0.5–2.0 keV Flux (erg cm^{-2} s^{-1})


Raw image
0.5–8.0 keV
25
20
15
10
62° 05'
Declination (2000)
39
38
37
12h 36m
Right Ascension (2000)

Inferential Process

To infer the $\log N - \log S$ relationship there are a few steps:

1. Collect raw data images
2. Run a detection algorithm to extract 'sources'
3. Produce a dataset describing the 'sources' (and uncertainty about them)
4. Infer the $\log N - \log S$ distribution from this dataset

Our analysis is focused on the final step – accounting for some (but not all) of the detector-induced uncertainties...

Inferential Process

To infer the $\log N - \log S$ relationship there are a few steps:

1. Collect raw data images
2. Run a detection algorithm to extract 'sources'
3. Produce a dataset describing the 'sources' (and uncertainty about them)
4. Infer the $\log N - \log S$ distribution from this dataset

Our analysis is focused on the final step – accounting for some (but not all) of the detector-induced uncertainties...

Adding further layers to the analysis to start with raw images is possible but that is for a later time...

The Data

The data is essentially just a list of photon counts – with some extra information about the background and detector properties.

Src_ID	Count	Src_area	Bkg	Off_axis	Eff_area
2	270	1720	3.16	4.98	734813.1074
3	117	96	0.19	5.72	670916.3154
7	33	396	0.61	6.17	670916.3154
18	7	128	0.22	6.34	319483.9597
19	12	604	0.96	4.51	670916.3154

Problems

The photon counts do not directly correspond to the source fluxes:

1. Background contamination
2. Natural (Poisson) variability
3. Detector efficiencies (PSF etc.)

Not all sources in the population will be detected:

1. Low intensity sources
2. Close to the limit: background, natural variability and detection probabilities are important.

Key Ideas

Our goals:

- Provide a complete analysis, accounting for all detector effects (especially those leading to unobserved sources)
- Allow for the incorporation of prior information
- Investigate parametric forms (testing) for $\log N - \log S$ (e.g., broken power-laws)
- Investigate the data-prior inferential limit
  (e.g., for which S^*_{min} does the information come primarily from the model and not the data)

Missing Data Overview

There are many potential causes of missing data in astronomical data:

- Background contamination (e.g., total=source+background)
- Low-count sources (below detection threshold)
- Detector schedules (source not within detector range)
- Foreground contamination (objects between the source and detector)
- etc.

Some are more problematic than others...

Missing Data Mechanisms

In the nicest possible case, if the particular data that is missing does not depend on any unobserved values then we can essentially ignore the missing data.

In this context, whether a source is observed is a function of its source count (intensity) – which is unobserved for unobserved sources. This missing data mechanism is non-ignorable, and needs to be carefully accounted for in the analysis.

The Model

$$N \sim NegBinom\left(\alpha, \beta\right),$$

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

$$\theta \sim \mathrm{Gamma}(a, b),$$

$$Y_i^{src} | S_i, L_i, E_i \overset{iid}{\sim} Pois\left(\lambda(S_i, L_i, E_i)\right),$$

$$Y_i^{bkg} | L_i, E_i \overset{iid}{\sim} Pois\left(k(L_i, E_i)\right),$$

$$I_i \sim \mathrm{Bernoulli}\left(g\left(S_i, L_i, E_i\right)\right).$$

The Model

It turns out that in many contexts there is strong theory that expects the $\log N - \log S$ to obey a *Power law*:

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

Taking the logarithm 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:

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

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

The Model cont...

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

$$N \sim NegBinom\,(\alpha, \beta)\,,$$

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 Model cont...

The probability of a source being detected, $g\left(S_i, L_i, E_i\right)$, 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 Model cont...

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

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

The Model

$$N \sim NegBinom\left(\alpha, \beta\right),$$

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

$$\theta \sim \mathrm{Gamma}(a, b),$$

$$Y_i^{src} | S_i, L_i, E_i \overset{iid}{\sim} Pois\left(\lambda(S_i, L_i, E_i)\right),$$

$$Y_i^{bkg} | L_i, E_i \overset{iid}{\sim} Pois\left(k(L_i, E_i)\right),$$

$$I_i \sim \mathrm{Bernoulli}\left(g\left(S_i, L_i, E_i\right)\right).$$

Computational Details

The Gibbs sampler consists of four steps:

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

- Sample the observed photon counts:

  $$Y^{src}_{obs,i}|n, Y^{tot}_{obs,i}, S_{obs,i} \sim Binom\left(Y^{tot}_{obs,i}, \frac{\lambda(S_{obs,i}, L_{obs,i}, E_{obs,i})}{\lambda(S_{obs,i}, L_{obs,i}, E_{obs,i}) + k}\right),$$

  for $i = 1, \ldots, n$.
- Sample the fluxes $S_{obs,i}, i = 1, \ldots, n$ (MH using a $t-$proposal).
- Sample the power-law slope θ (MH using a $t-$proposal).
- Compute the posterior distribution for the total number of sources, N, using numerical integration:

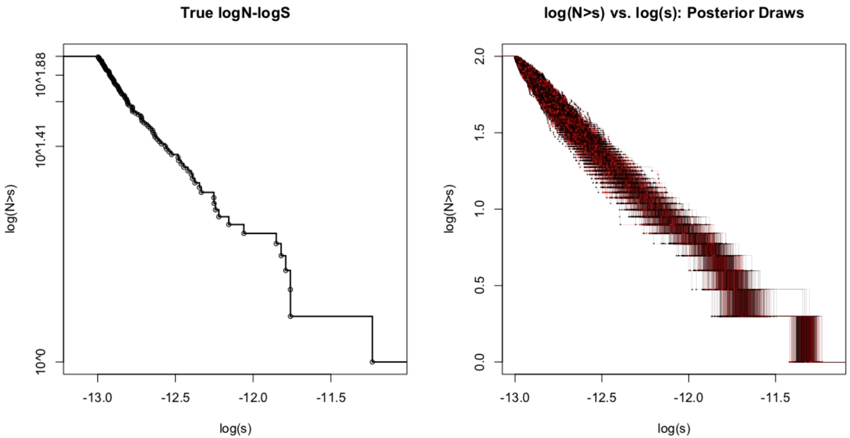
  $$p(N|n, \theta) \propto \frac{\Gamma(N + \alpha)}{\Gamma(N - n + 1)} \left(\frac{1 - \pi(\theta)}{\beta + 1}\right)^N \mathbb{I}\{N \geq n\}$$

  Note: The (prior) marginal detection probability $\pi(\theta)$ is pre-computed via the numerical integration.

Computational Notes

Some important things to note:

- Computation is fast (secs), and insensitive to the number of missing sources
- The fluxes of the missing sources are never imputed (only the *number* of missing sources)
- Most steps are not in closed form $\Rightarrow$ changing (some) assumptions has little computation impact
- Broken power law (or other forms) can be implemented by changing only one of the steps
- Fluxes of missing sources can (optionally) be imputed to produce posterior draws of a 'corrected' $\log N - \log S$



RED = Missing sources, BLACK = Observed sources.

Future Work

We currently do not include:

1. False sources (allowing that 'observed' sources might actually be background/artificial)
2. Spatially varying detection probabilities (straightforward, needs implementing)

Simulated Example

Assume parameter setting:

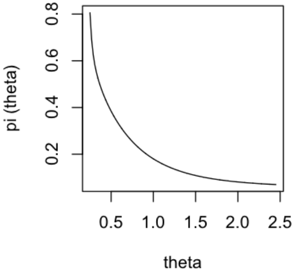
- $N \sim NegBinom(\alpha, \beta)$, where $\alpha = 200 =$ shape, $\beta = 2 =$ scale
- $\theta \sim Gamma(a, b)$, where $a = 20 =$ shape, $b = 1/20 =$ scale
- $S_i|\theta \sim Pareto(\theta, S_{min})$, where $S_{min} = 10^{-13}$
- $Y_i^{src}|S_i, L_i, E_i \sim Pois(\lambda(S_i, L_i, E_i))$
- $Y_i^{bkg}|S_i, L_i, E_i \sim Pois(k(L_i, E_i))$
- $\lambda = \frac{S_i \cdot E_i}{\gamma}$, where effective area $E_i \in (1,000, 100,000)$, and the energy per photon $\gamma = 1.6 \times 10^{-9}$
- $k_i = z \cdot E_i$, where the rate of background photon count intensity per million seconds $z = 0.0005$
- $n_{iter} = 21,000$, Burnin $= 1000$

Simulated Example cont. . .

Detection probability:

- $g(\lambda, k) = 1.0 - a_0 \cdot (\lambda + k)^{a_1} \cdot e^{a_2 \cdot (\lambda + k)}$, where
  $a_0 = 11.12, a_1 = -0.83, a_2 = -0.43$

Marginal detection probability:



Empirical results of MCMC sampler

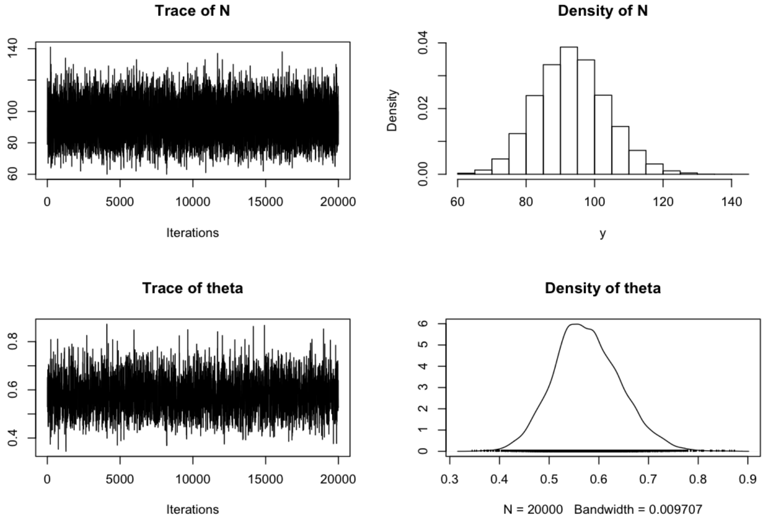
The actual coverage of nominal percentiles for all parameters for simulated data, for $M = 200$ validation datasets:

Coverage Percentile	50%	80%	90%	95%	98%	99%	99.9%
N	0.55	0.83	0.90	0.96	0.98	0.99	1.00
θ	0.50	0.82	0.92	0.97	0.99	0.99	1.00
all S_{obs}	0.51	0.81	0.90	0.95	0.98	0.99	1.00

Mean squared error of different estimators for N and θ for simulated data.

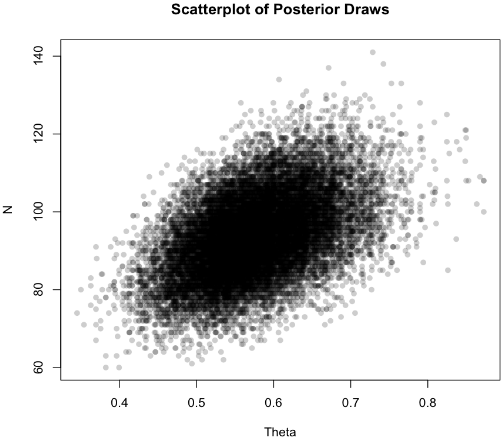
MSE	N		θ	
Level of Effective Area	Median	Mean	Median	Mean
Low	215.96	291.82	0.05439	0.07481
Medium	121.26	168.91	0.05558	0.07407
High	68.23	95.36	0.04578	0.05987

MCMC Draws



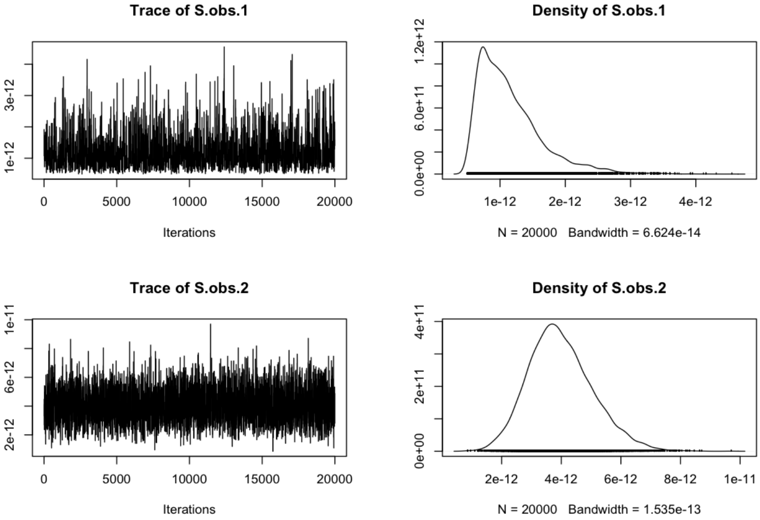
Trace of N
140
100
80
60
0
5000
10000
15000
20000
Iterations
Density of N
Density
0.04
0.02
0.00
60
80
100
120
140
y
Trace of theta
0.8
0.6
0.4
0
5000
10000
15000
20000
Iterations
Density of theta
6
5
4
3
2
1
0
0.3
0.4
0.5
0.6
0.7
0.8
0.9
N = 20000 Bandwidth = 0.009707

Posterior Correlations

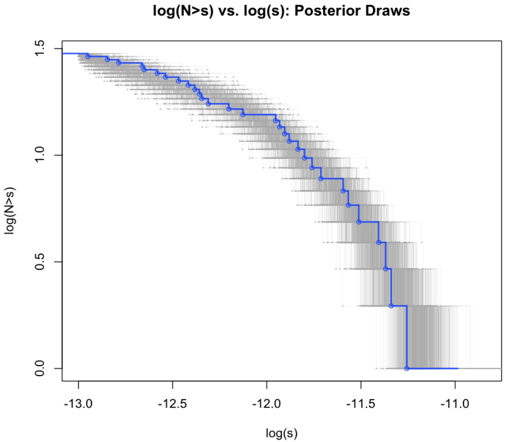


Posterior estimates for the power-law slope and the total number of sources.

MCMC Draws


Trace of S.obs.1
3e-12
1e-12
0
5000
10000
15000
20000
Iterations
Density of S.obs.1
1.2e+12
6.0e+11
0.0e+00
1e-12
2e-12
3e-12
4e-12
N = 20000 Bandwidth = 6.624e-14
Trace of S.obs.2
1e-11
6e-12
2e-12
0
5000
10000
15000
20000
Iterations
Density of S.obs.2
4e+11
2e+11
0e+00
2e-12
4e-12
6e-12
8e-12
1e-11
N = 20000 Bandwidth = 1.535e-13

Simulated $\log N - \log S$

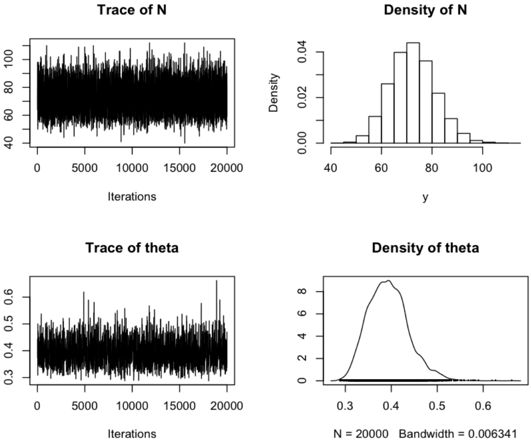


Uncertainties in source fluxes and a display of the power-law relationship. Posterior draws (gray), truth (blue).

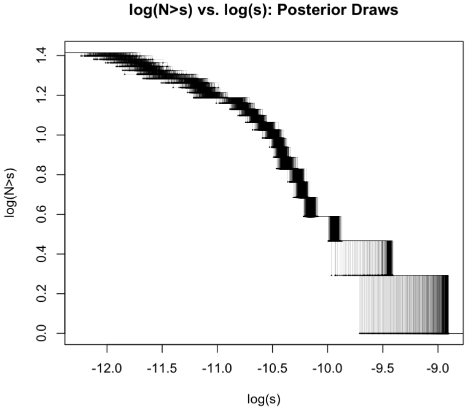
The Data

Src_ID	Count	Src_area	Bkg	Off_axis	Eff_area
2	270	1720	3.16	4.98	734813.1074
3	117	96	0.19	5.72	670916.3154
7	33	396	0.61	6.17	670916.3154
18	7	128	0.22	6.34	319483.9597
19	12	604	0.96	4.51	670916.3154

POSTERIORS OF PARAMETERS N AND θ



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.



Note: This a posterior plot for the *observed sources only*
(the 'corrected' plot would be more useful. . .)

Evidence of a possible break in the power-law in the observed $\log N - \log S$.
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.

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.