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

Paul D. Baines

Department of Statistics
University of California, Davis

May 10th, 2013

Introduction

Project Goals

Develop a comprehensive method to infer (properties of) the distribution of source fluxes for a wide variety source populations.

More generally, to also infer luminosity functions for source populations.

Collaborators: Irina Udaltsova (UCD), Andreas Zezas (University of Crete & CfA), Vinay Kashyap (CfA).

The Rationale for $\log N - \log S$ Fitting

Let $N(>S)$ denote the number of sources detectable to a sensitivity S i.e., $N(>S)$ is the empirical survival function of the flux distribution. In simple settings we expect:

$$\log_{10}\left(N(>S)\right) = \beta_0 + \beta_1 \log_{10}(S),$$

Cosmology complicates the anticipated linearity somewhat, but in many cases the relationship is approximately linear.

Primary Goal: Estimate β_1, the power law slope, while properly accounting for detector uncertainties and biases.

Note: There is uncertainty on both $x-$ and $y-$axes (i.e., N and s).

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' from the image
3. Produce a dataset describing the photon counts of all 'sources' (and uncertainty about them, background etc.)
4. Infer physical properties about the source population (e.g., 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...

Probabilistic Connection

Standard $\log(N) - \log(S)$ approaches make it difficult to coherently incorporate detector effects and uncertainties.

Probabilistic Connection: Under independent sampling, linearity on the $\log N - \log S$ scale is equivalent to the flux distribution being a Pareto distribution.

(Follows from log-linearity of the survival function)

The probabilistic representation for the flux distribution now allows for more rigorous analysis by embedding within a hierarchical model.

Beyond the Pareto

With Pareto flux distribution we obtain a linear relationship on the $log N - \log\left(N > S\right)$ scale. In general, with complete-data flux distribution G, we have:

$$S_i \overset{iid}{\sim} G \qquad \Rightarrow \qquad \log_{10}\left(1 - F_G(s)\right) := H\left(\log_{10}(s)\right). \qquad (1)$$

The function H is linear if and only if G is the Pareto distribution. Our framework will allow for flexible specification of the (parametrized) flux distribution.

Since linearity has both theoretical and empirical support, a commonly used generalization is a broken power-law:

$$\log_{10}\left(1 - F_G(s)\right) = \left\{ \begin{array}{ll} \alpha_0 - \theta_0 \log_{10}(s) & s \leq K \\ \alpha_1 - \theta_1 \log_{10}(s) & s > K \end{array} \right. , \qquad (2)$$

subject to a continuity constraint.

Broken Power-Law Modeling

The broken power-law in (2) can be represented as:

$$Y \sim \left[1 - \left(\frac{K}{S_{min}}\right)^{-\theta_0}\right] X_0 + \left(\frac{K}{S_{min}}\right)^{-\theta_0} X_1,$$

where:

$$X_0 \sim \mathrm{Truncated\text{-}Pareto}\left(S_{min}, \theta_0, K\right), \qquad X_1 \sim \mathrm{Pareto}\left(K, \theta_1\right).$$

The result is also an ‘if and only if’ result i.e., any distribution whose $\log N - \log S$ relationship is a broken power law, with M breakpoints, can be represented as a mixture of M truncated Pareto distributions and another (untruncated) Pareto distribution.

Physically Motivated Fitting

The insight from the probabilistic setting reveals that the broken power-law model has a number of unphysical properties (to be expected).

Notably, it requires an ‘initial source population’ to have a sharp cut-off, before yielding to a secondary source population present only above the threshold.

More physically realistic descriptions are also more natural statistically e.g.,

$$Y \sim \sum_{j=1}^{m} p_j X_j, \qquad \text{where:} \qquad X_j \sim \mathrm{Pareto}\left(S_{min}, \theta_j\right).$$

Note: the resulting $\log N - \log S$ plot will be curved!

Observational Challenges

The previous discussion centered around the flux distribution.

- We only observe photon counts from the source with intensity proportional to the flux
- There is background contamination for all sources
- Different sensitivities across the detector
- Some sources will not be observed to detector limitations
- We do not know how many sources there actually are
- Some 'sources' extracted from the image may not actually be sources

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

Broken power-law flux distribution (known break-points $\vec{C}$):

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

Source and background photon counts:

$$Y_i^{tot} | S_i, L_i, E_i \overset{\perp\!\!\!\perp}{\sim} Pois\left(\lambda(S_i, L_i, E_i) + k(B_i, L_i, E_i)\right), \qquad i = 1, \ldots, N,$$

Incompleteness, missing data indicators:

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

Prior distributions:

$$N \sim NegBinom\left(\alpha, \beta\right), \qquad p(B_i, L_i, E_i)$$
$$S_{min} \sim \mathrm{Gamma}(a_s, b_s), \qquad \theta \sim \mathrm{Gamma}(a_\theta, b_\theta).$$

Observed data: $Y_{obs} = \{(Y_i^{tot}, B_i, L_i, E_i) : i \in \mathcal{I}, |\mathcal{I}| = n\}$,
LVs/Missing Data: $Y_{mis} = \{(Y_i^{tot}, S_i, B_i, L_i, E_i) : i \notin \mathcal{I}\}, \{S_i : i \in \mathcal{I}\}$,
Parameters: $\Theta = \{N, \theta, S_{min}\}$.

The Model

Standard power-law flux distribution:

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

Source and background photon counts:

$$Y_i^{tot} | S_i, L_i, E_i \overset{\perp\!\!\!\perp}{\sim} Pois\left(\lambda(S_i, L_i, E_i) + k(B_i, L_i, E_i)\right), \qquad i = 1, \ldots, N,$$

Incompleteness, missing data indicators:

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

Prior distributions:

$$N \sim NegBinom\left(\alpha, \beta\right), \qquad p(B_i, L_i, E_i)$$

$$h(\vec{C}) \sim N(m, V), \qquad \theta_j \overset{\perp\!\!\!\perp}{\sim} \text{Gamma}(a_j, b_j), \qquad j = 1, \ldots, M.$$

Observed data: $Y_{obs} = \{(Y_i^{tot}, B_i, L_i, E_i) : i \in \mathcal{I}, |\mathcal{I}| = n\}$,
LVs/Missing Data: $Y_{mis} = \{(Y_i^{tot}, S_i, B_i, L_i, E_i) : i \notin \mathcal{I}\}, \{S_i : i \in \mathcal{I}\}$,
Parameters: $\Theta = \left\{N, \vec{\theta}, \vec{C}\right\}$.

Posterior Inference

Inference about θ, N and S is based on the observed data posterior distribution. Care must be taken with the variable dimension marginalization over the unobserved fluxes.

Computation is performed by Gibbs sampling.

Makes heavy use of the marginal probability of observing a source:

$$\pi(\theta, S_{min}) = \int g(S, B, L, E) \cdot p(S | S_{min}, \theta) \cdot p(B, L, E) dB\, dL\, dE\, dS$$

For broken power-law this is $2m$ dimensional where m is the number of ‘pieces’. It can be pre-computed but requires a sufficiently dense grid and careful interpolation in higher dimensions.

Model/Computational Notes

Things to note:

- The dimension of the missing data is unknown (care must be taken with conditioning)
- Incompleteness function g can take any form and is problem-specific
- $p(B_i, L_i, E_i)$ needs some care and can be tabulated or parametrized
- Prior parameters can be science-based or 'weakly informative'
- For single power-law models computation is fast, and insensitive to the number of missing sources
- Computation for the broken-power law model is slower
- Generalized mixtures of Pareto's (or other forms) require only minor modifications of general scheme

Paper Organization

Paper 1 (Single Pareto modeling only):

- Handling of incompleteness
- Handling of other detector effects (background, exposure maps, source location etc.)
- Incorporation of prior information
- Probabilistic/Bayesian modeling
- Model checking via posterior predictive checks

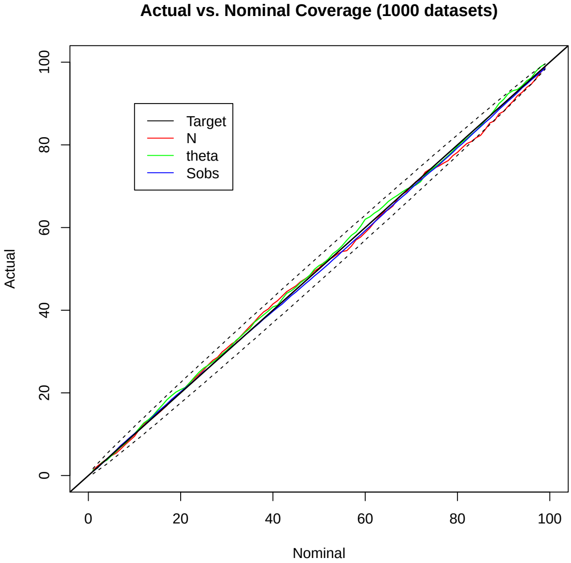
Paper 2 (Broken- and Mixture-Pareto extensions):

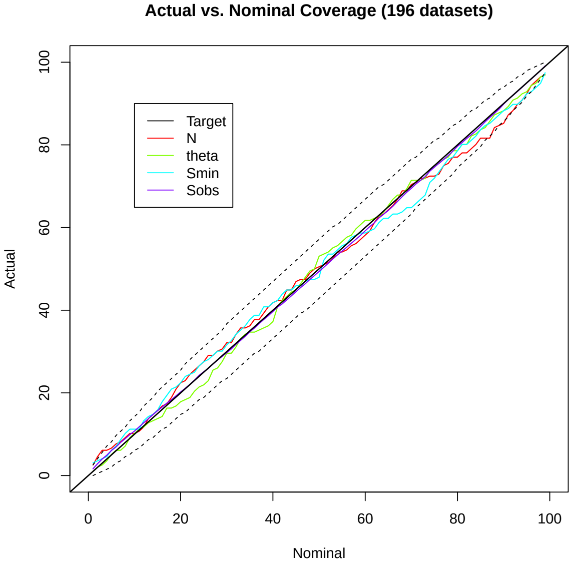
- Broken-Pareto modeling for $\log(N) - \log(S)$
- Mixture-Pareto modeling for $\log(N) - \log(S)$
- Model selection and model checking

Validation Details

Parameter specifications as follows:

- $N \sim \mathrm{NegBinom}(\alpha, \beta)$, where $\alpha = 200 =$ shape, $\beta = 2 =$ scale
- $\theta \sim \mathrm{Gamma}(a, b)$, 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 $E_i \in (1000, 100000)$, 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 $z = 0.0005$
- $n_{iter} = 250000$, Burnin $= 50000$


Actual vs. Nominal Coverage (1000 datasets)
Target
N
theta
Sobs
Actual
Nominal
0
20
40
60
80
100
0
20
40
60
80
100


Actual vs. Nominal Coverage (196 datasets)
Target
N
theta
Smin
Sobs
Actual
0
20
40
60
80
100
Nominal
0
20
40
60
80
100

Sensitivity Studies

There are a lot of cogs in the full model, so we have performed various sensitivity studies to investigate the impact of different priors, misspecification of the incompleteness function etc.

Sensitivity: θ prior

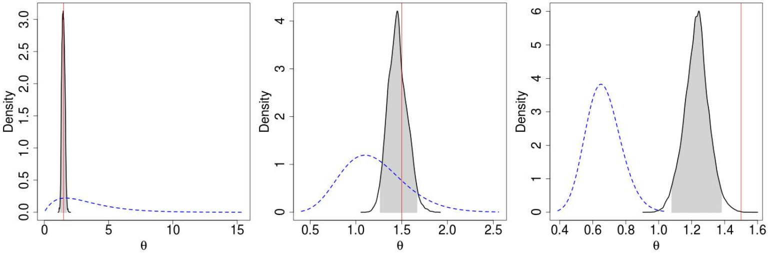


Figure:

(L) Weak prior and corresponding posterior for θ,
(C) a moderately informative prior for θ, and,
(R) a strongly informative but incorrect prior for θ.

SENSITIVITY: N PRIOR

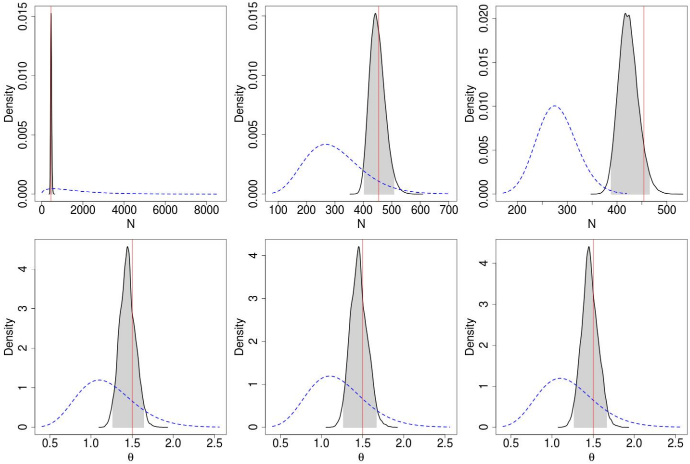


FIGURE: Prior and posterior distributions for N and θ. Clockwise from upper-left (top row) these represent a weak prior for N, a moderately informative prior for N and a strongly informative but incorrect prior for N. The bottom row shows the corresponding prior and posterior distributions for θ.

Fixed vs. Estimated S_{min}

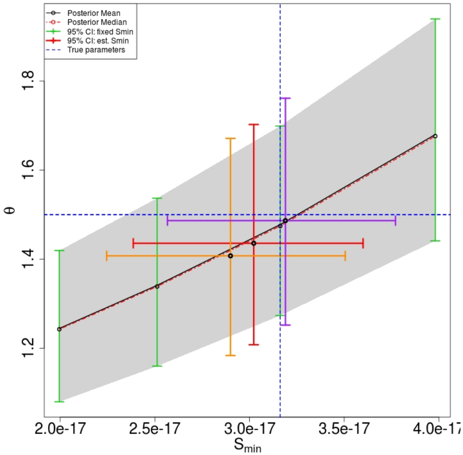


Figure: The grey regions provide the posterior 95% credible intervals for θ at the fixed S_{min} case, with the θ estimate in the center line. The cross intervals show posterior dispersion in both θ and S_{min} for varying priors on S_{min}.

FIXED S_{min} RESULTS

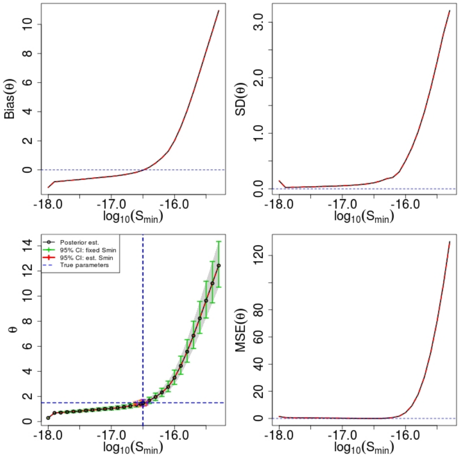


FIGURE: Sensitivity of S_{min} on estimate of θ. Under the fixed S_{min} scenarios, the plots show: (top-left) bias of θ, (top-right) standard deviation of θ, (mid-left) posterior regions and 95% credible intervals of θ, (mid-right) U-shape nature of MSE of θ.

SENSITIVITY TO MISSPECIFICATION OF g

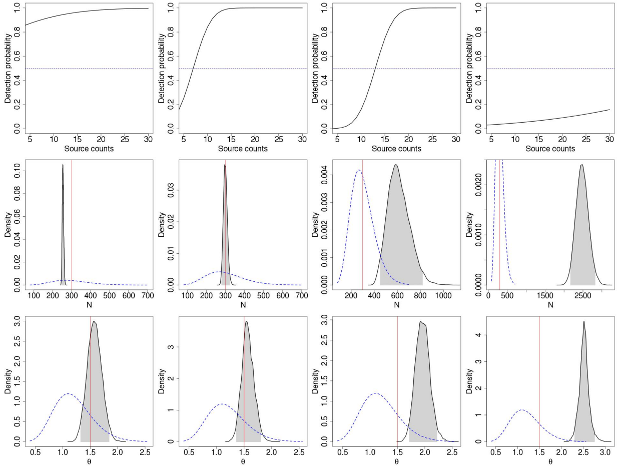


FIGURE: Top row: Four different incompleteness functions, Rows 2-3: Corresponding prior and posterior distributions of N and θ. The second column corresponds to the correct incompleteness function.

Posterior Predictive Diagnostics

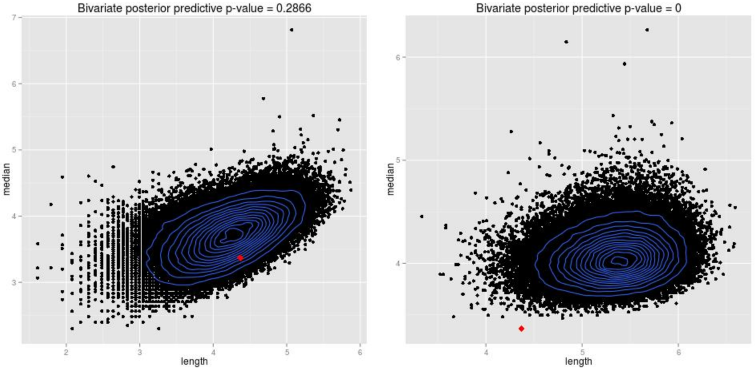


Figure: Bivariate posterior predictive scatterplot for the conditional model: (left) fitted S_{min} equal to truth, (right) fitted S_{min} larger to truth.

Application: CDF-N

- Chandra Deep Field North X-ray sources
- Subset of 225 sources < 8 arcmins
- Incompleteness function and priors in tabulated form
- Priors used are weakly informative
- S_{min} estimated from the data

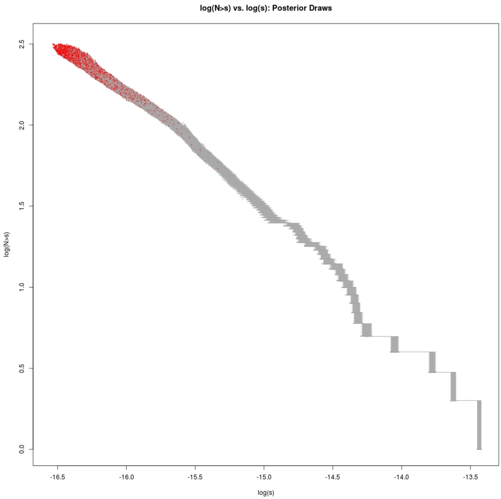


FIGURE: The $\log(N) - \log(S)$ plot for the CDFN data. Each line in the plot corresponds to a sample of fluxes for the complete source population from a single iteration of MCMC scheme with observed sources shown in grey and missing sources in red.

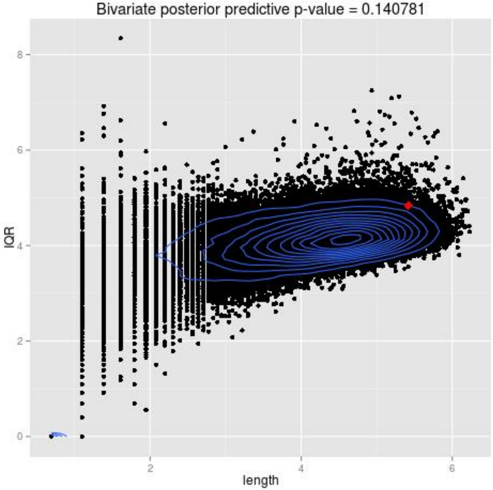


FIGURE: Posterior predictive plots for the single Pareto model fit to the CDF-N dataset.

CDF-N Results

The posterior predictive checks are passable but show a few issues ($p-$values from 0.03 to 0.41 for various features of the model fit).

Indications of possible breakpoint
(Broken-Pareto model gives better fit).

Estimates:

- $\hat{\theta} = 0.68$, $(0.59, 0.78)$ consistent with other analyses,
- $\hat{N} = 274$, $(256, 299)$ suggest completeness of $75\% - 87.5\%$.

Status & Future Work

Paper One:

- Draft available, almost complete
  (Switching data analysis to CDF-S from CDF-N)

Paper Two:

- Various simulations completed, more to do. . .
- Selection of number of 'pieces' for multiple power-law setting not investigated yet (can use Wong et. al (2013) as guide)
- Estimation of normalizing constants is tricky, other ideas?
- Real data: CDF-N, SMC

Future stuff:

- False sources (allowing that 'observed' sources might actually be background/artificial)
- Field contamination (allowing a mixture of a source population with known parameters)
- Extension to non-Poisson regimes

References

- Wong, R.K.W., Baines, P.D., Aue, A., Lee, T.C.M and Kashyap, V.K. (2013) Automatic Estimation of Flux Distributions of Astrophysical Source Populations, `http://arxiv.org/abs/1305.0979`.
- P.D. Baines, I.S. Udaltsova, A. Zezas, V.L. Kashyap (2011) Bayesian Estimation of $\log N - \log S$, *Proc. of Statistical Challenges in Modern Astronomy V*
- 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.