Bayesian Estimation of Flux Distributions

Paul D. Baines

Department of Statistics
University of California, Davis

August 21st, 2012

Introduction

Scientific Objectives

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

Statistical Objectives

- *Inference: Account for non-ignorable missing data (+more)*
- *Model Selection: Select the 'best' model for a given dataset*
- *Model Checking: Evaluate the adequacy of a given model*

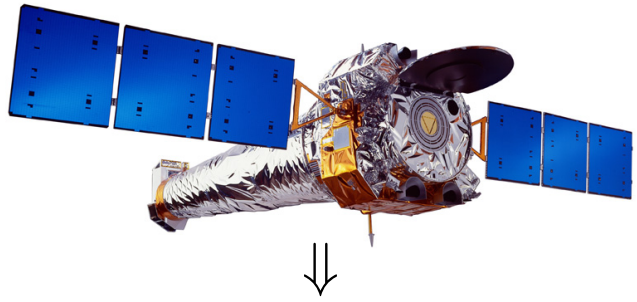
Collaborators:

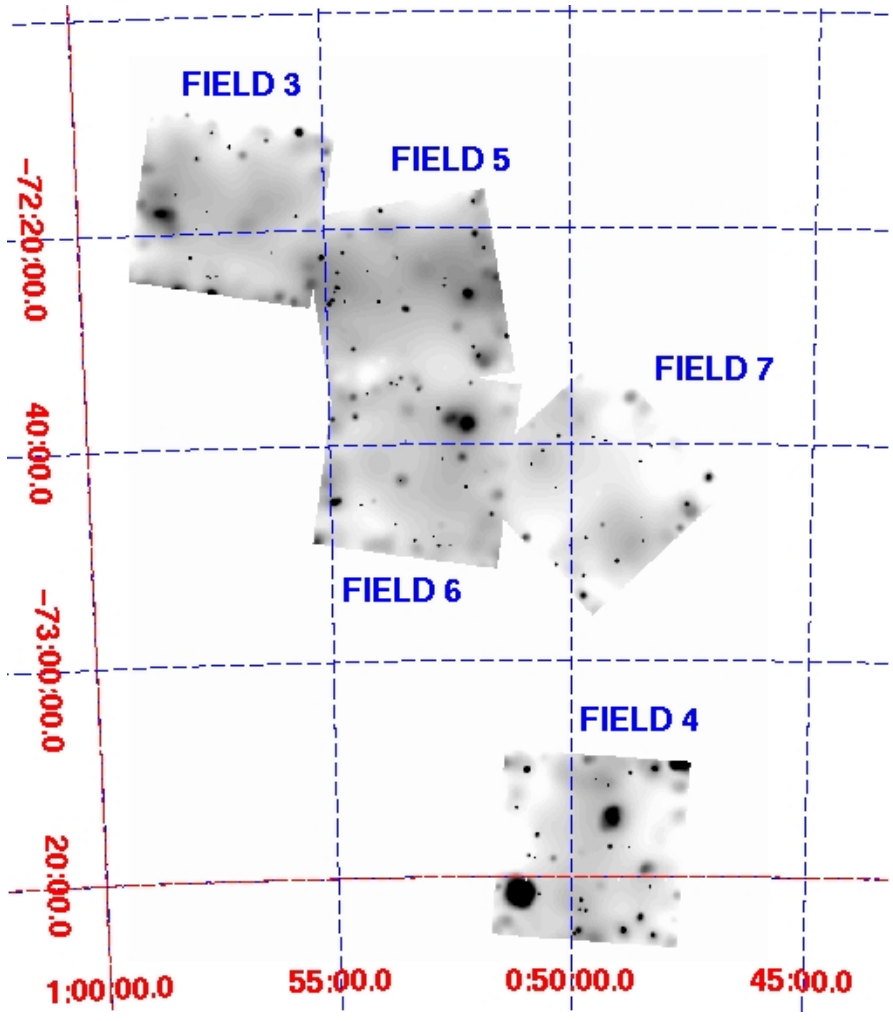
Irina Udaltsova (UCD),

Andreas Zezas (University of Crete & CfA),

Vinay Kashyap (CfA).

CHANDRA:




FIELD 3
FIELD 5
FIELD 7
FIELD 6
FIELD 4
-72:20:00.0
40:00.0
-73:00:00.0
20:00.0
1:00:00.0
55:00.0
0:50:00.0
45:00.0

Estimating Flux Distributions

Goal: Estimate the distribution of fluxes for the source population.

Knowing the specific relationship for different objects (e.g., stars, galaxies, pulsars) gives a lot of information about the underlying physics (e.g., the mass, age of galaxies).

Estimating Flux Distributions

Goal: Estimate the distribution of fluxes for the source population.

Knowing the specific relationship for different objects (e.g., stars, galaxies, pulsars) gives a lot of information about the underlying physics (e.g., the mass, age of galaxies).

Toy example: Uniformly distributed source population, same intrinsic luminosity L_0, then for telescopic sensitivity S, sources will be detectable to $d = \sqrt{L_0/4\pi S}$. The number of sources within this distance is then:

$$N(< d) = N(> S) = n_0 \left(\frac{4\pi}{3} d^3 \right) \propto S^{-3/2}$$

Therefore, the convention is to plot the log (base 10) of the cumulative number of sources as a function of log (base 10) flux.

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

In the simple case we have:

$$\log_{10}\left(N(>S)\right) = \alpha - \theta \log_{10}(S),$$

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

$$\log_{10}\left(1 - F_G(s)\right) = \begin{cases} \alpha_0 - \theta_0 \log_{10}(s) & K_0 < s < K_1 \\ \alpha_1 - \theta_1 \log_{10}(s) & K_1 < s < K_2 \\ \vdots & \vdots \\ \alpha_m - \theta_m \log_{10}(s) & K_m < s \end{cases}, \quad (1)$$

subject to continuity constraints.

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

In the simple case we have:

$$\log_{10}\left(N(>S)\right) = \alpha - \theta \log_{10}(S),$$

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

$$\log_{10}\left(1 - F_G(s)\right) = \begin{cases} \alpha_0 - \theta_0 \log_{10}(s) & K_0 < s < K_1 \\ \alpha_1 - \theta_1 \log_{10}(s) & K_1 < s < K_2 \\ \vdots & \vdots \\ \alpha_m - \theta_m \log_{10}(s) & K_m < s \end{cases}, \quad (1)$$

subject to continuity constraints.

Primary Goal: Estimate θ_j's (the power law slopes), while properly accounting for detector uncertainties and biases.

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

Probabilistic Connections

Under independent sampling, linearity on the $\log N - \log S$ scale is equivalent to the flux distribution being a Pareto distribution.
A piecewise linear $\log N - \log S$ also has a probabilistic analogue.

Probabilistic Connections

Under independent sampling, linearity on the $\log N - \log S$ scale is equivalent to the flux distribution being a Pareto distribution.
A piecewise linear $\log N - \log S$ also has a probabilistic analogue.

Theorem

Any distribution whose $\log N - \log S$ curve is a broken power law with M breakpoints, can be represented as a mixture of M truncated Pareto distributions and one (untruncated) Pareto distribution.

Probabilistic Connections

Under independent sampling, linearity on the $\log N - \log S$ scale is equivalent to the flux distribution being a Pareto distribution.
A piecewise linear $\log N - \log S$ also has a probabilistic analogue.

Theorem

Any distribution whose $\log N - \log S$ curve is a broken power law with M breakpoints, can be represented as a mixture of M truncated Pareto distributions and one (untruncated) Pareto distribution.

Example: A single break-point model is equivalent to:

$$S_i \overset{iid}{\sim} \mathrm{I}X_0 + (1 - \mathrm{I})X_1$$

$$\mathrm{I} \sim \mathrm{Binomial}\,(1; p)\,, \quad p = (K_1/K_0)^{-\theta_0}$$

$$X_0 \sim \text{Truncated-Pareto}\,(K_0, \theta_0, K_1)\,, \quad X_1 \sim \mathrm{Pareto}\,(K_1, \theta_1)\,.$$

For short, we denote $S_i \overset{iid}{\sim} \text{Broken-Pareto}(\theta, K, p)$.

Physically Motivated Fitting

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

Notably, it implies an 'initial source population' with a sharp cut-off, yielding to a secondary population above a threshold.

Physically Motivated Fitting

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

Notably, it implies an 'initial source population' with a sharp cut-off, yielding to a secondary population above a threshold.

The unphysical nature of the Broken Power-law can be relaxed quite easily, by removing the upper-truncation:

$$
\begin{aligned}
S_i &\overset{iid}{\sim} I_0 X_0 + I_1 X_1 + \cdots + I_m X_m \\
I_j &\sim \text{Multinomial}\left(1; p_0, p_1, \ldots, p_m\right), \\
X_j &\sim \text{Pareto}\left(K_j, \theta_j\right).
\end{aligned}
$$

For short, we denote $S_i \overset{iid}{\sim} \text{Mixture-Pareto}(\theta, K, p)$.

Note: The resulting logN-logS plot is no longer piecewise-linear.

Inference Roadmap

I **DETECTOR EFFECTS:**

II **INCOMPLETENESS:**

Inference Roadmap

I DETECTOR EFFECTS:

Photon counts do not directly correspond to the source fluxes:

1. Background contamination
2. Natural (Poisson) variability
3. Effective exposure, detector sensitivity etc.

II INCOMPLETENESS:

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.

Missingness is non-ignorable: whether or not a source is missing contains information about the parameters.

The Bayesian Hierarchical Model

Assumed power-law flux distribution:

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

Source and background photon counts:

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

Incompleteness, missing data indicators:

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

Prior distributions:

$$p(B_i, L_i, E_i | N), p(S_{min})$$
$$N \sim NegBinom\,(\alpha, \beta)\,,$$
$$\theta \sim \text{Gamma}(a, b).$$

The Bayesian Hierarchical Model

Assumed power-law flux distribution:

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

Source and background photon counts:

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

Incompleteness, missing data indicators:

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

Prior distributions:

$$p(B_i, L_i, E_i | N), p(S_{min}, \vec{C})$$
$$N \sim NegBinom\left(\alpha, \beta\right),$$
$$\theta_j \sim \text{Gamma}(a_j, b_j).$$

The Bayesian Hierarchical Model

Assumed power-law flux distribution:

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

Source and background photon counts:

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

Incompleteness, missing data indicators:

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

Prior distributions:

$$p(B_i, L_i, E_i | N), p(S_{min}, \vec{K}, \vec{p})$$
$$N \sim NegBinom\left(\alpha, \beta\right),$$
$$\theta_j \sim \text{Gamma}(a_j, b_j).$$

The Bayesian Hierarchical Models

Assumed power-law flux distribution:

$$S_i | S_{min}, \theta \vec{C}/\vec{K}, \vec{p} \overset{iid}{\sim} \begin{cases} \text{Pareto}\,(\theta, S_{min}) \\ \text{Broken-Pareto}\left(\vec{\theta}, S_{min}; \vec{C}\right) \\ \text{Mixture-Pareto}\left(\vec{\theta}, S_{min}; \vec{K}, \vec{p}\right) \end{cases} \qquad i = 1, \ldots, N$$

Source and background photon counts:

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

Incompleteness, missing data indicators:

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

Prior distributions:

$$p(B_i, L_i, E_i | N), \left\{p(S_{min}), p(S_{min}, \vec{C}), p(S_{min}, \vec{K}, \vec{p})\right\}$$

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

$$\{\theta \sim \text{Gamma}(a, b), \qquad \theta_j \sim \text{Gamma}(a_j, b_j), \qquad \theta_j \sim \text{Gamma}(a_j, b_j)\}$$

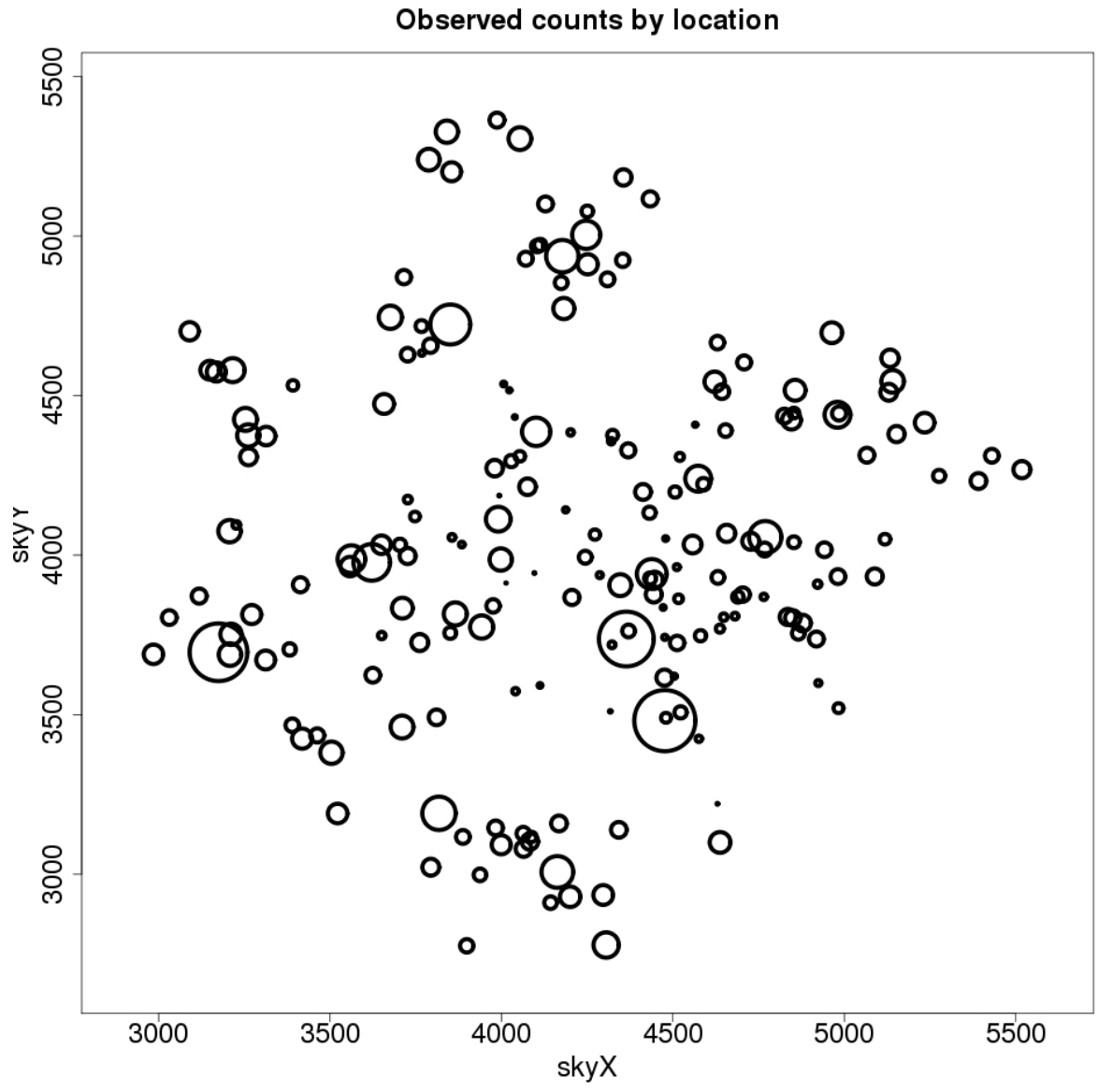
Model Overview

For all versions of the model (regular, broken and mixture-Pareto), inference about θ, N and S is based on the observed data posterior distribution.

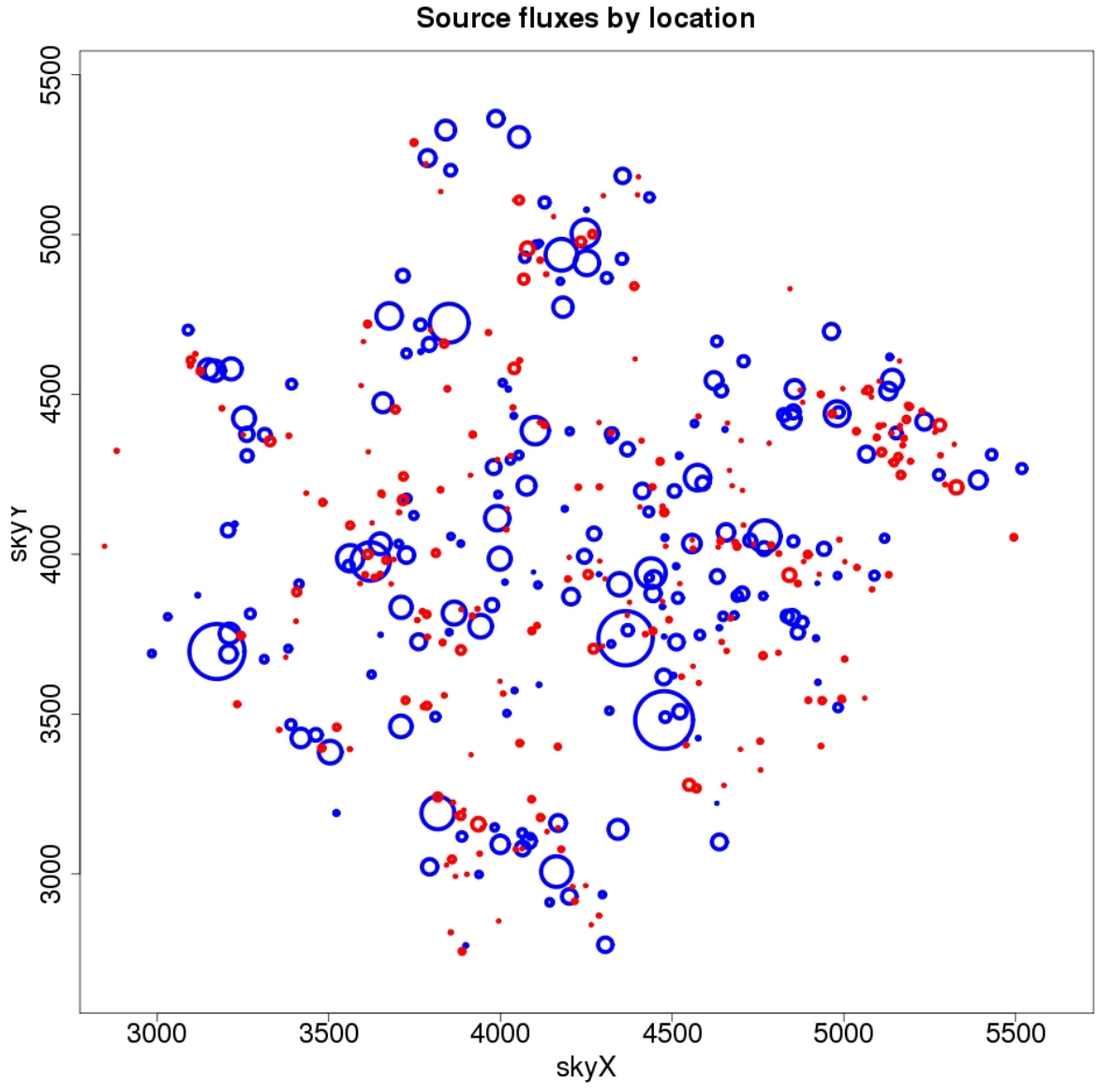
Computation is performed using MCMC.

▸ Skip MCMC Visualization

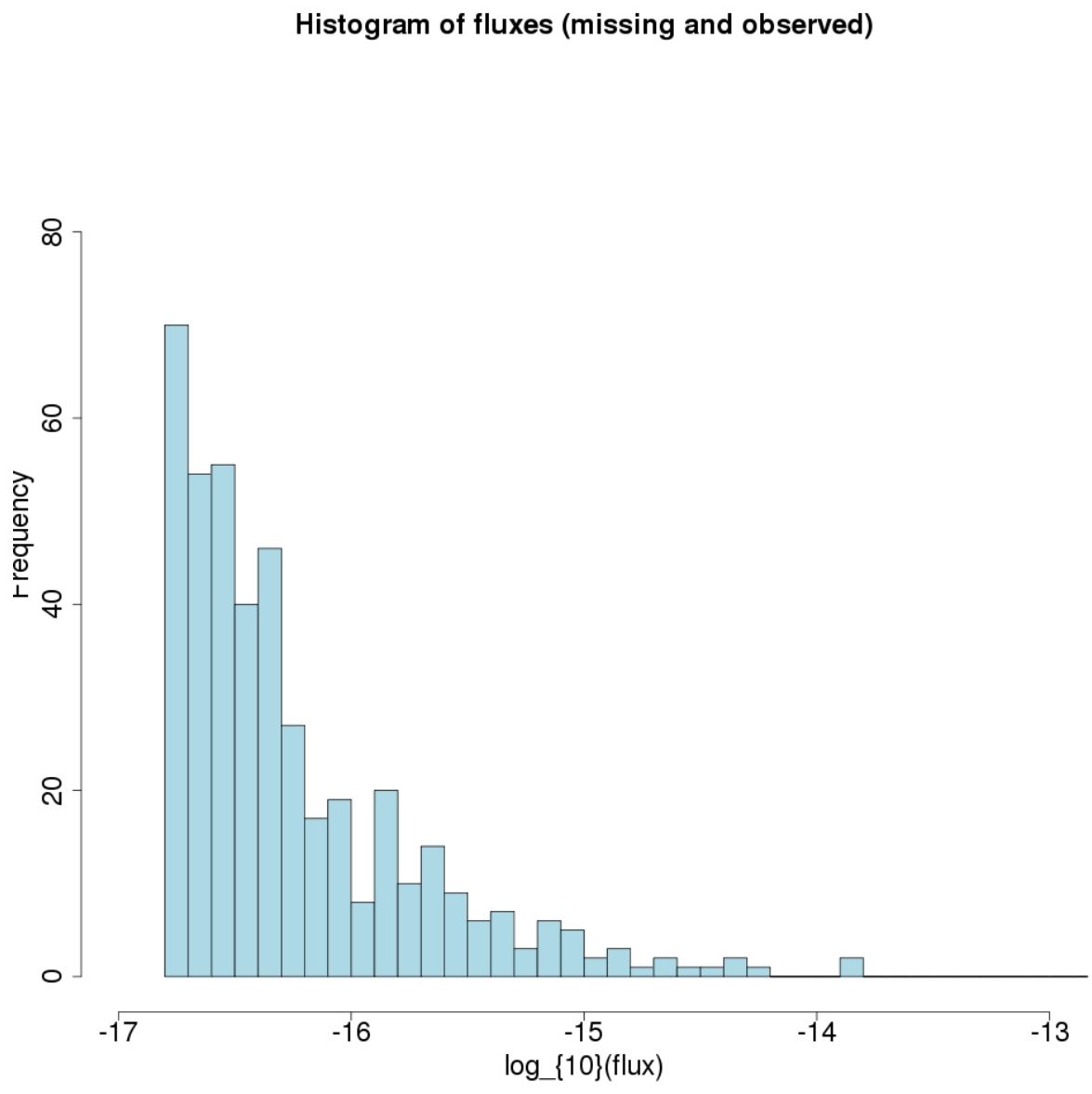
Counts by location (size proportional to # photons):



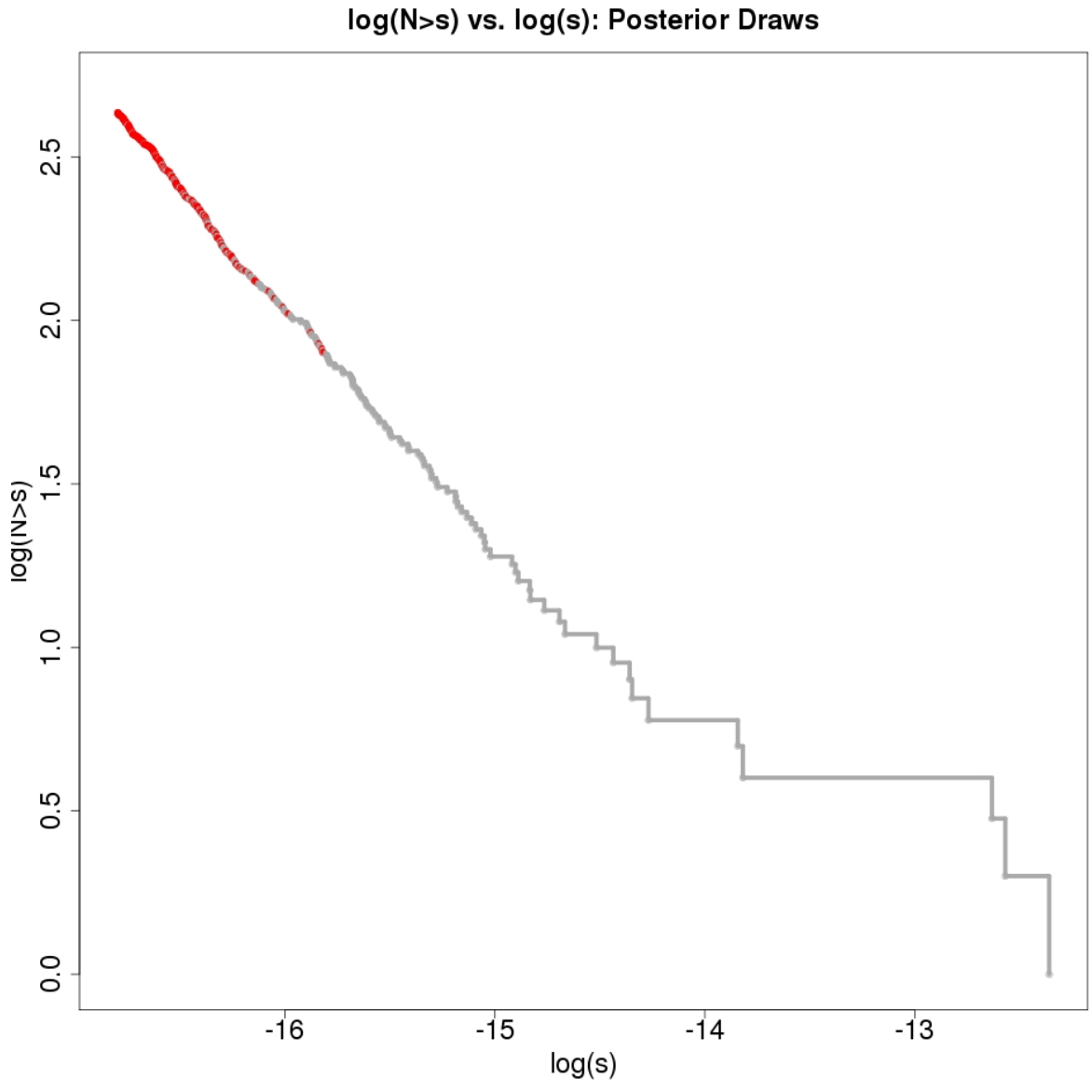
Flux by location (red=missing) size proportional to flux S_i):



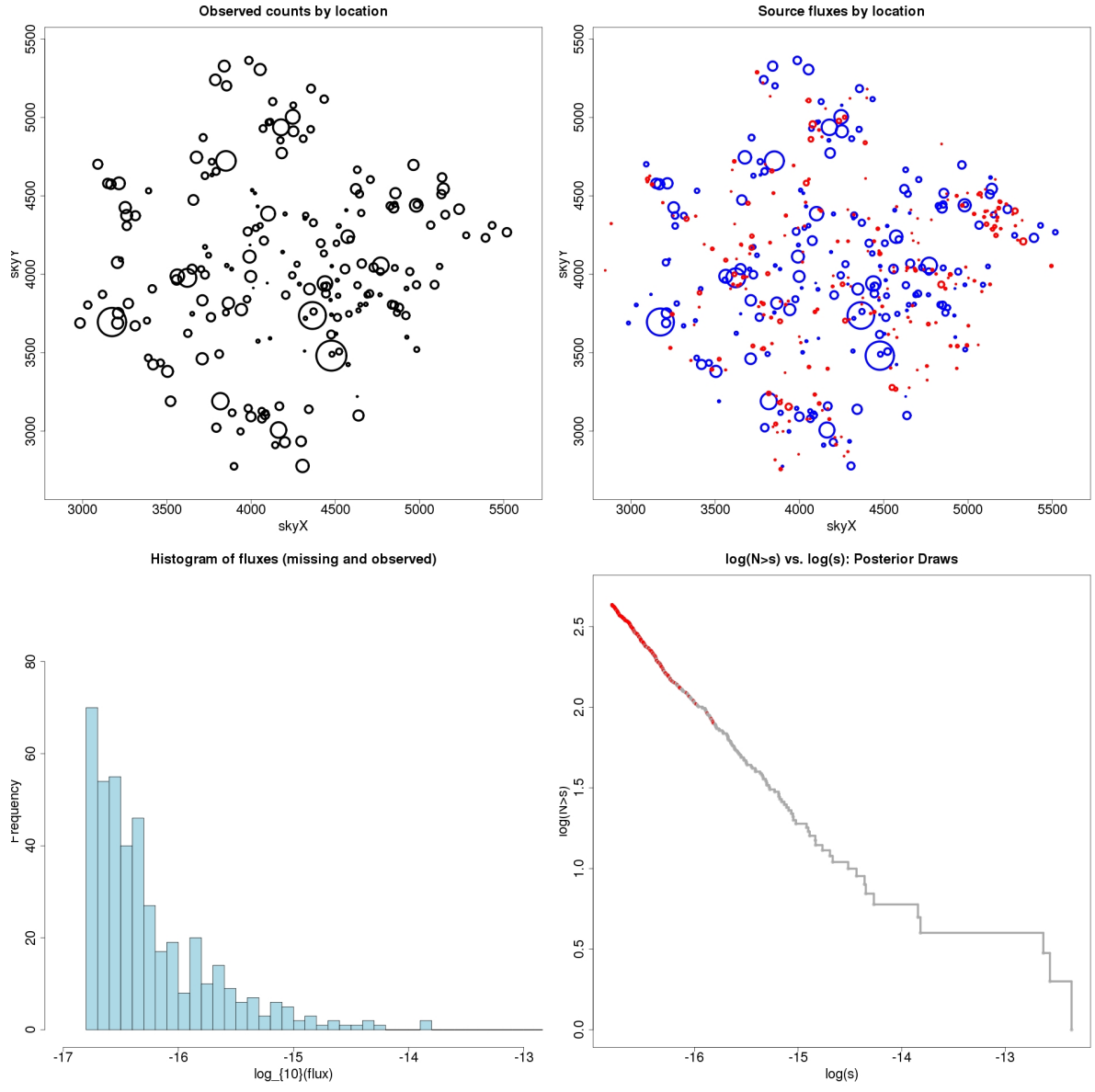
Histogram of flux distribution:



Complete-data logN-logS plot:



Visualizing Posterior Inference:



Application: Chandra Deep Field North

We now apply our method to a subset of the Chandra Deep Field North (CDFN) dataset.

- One of the deepest available X-ray surveys
- Tabulated observation-specific joint distribution of background, exposure map and off-axis
- 225 sources

Apply model selection criteria to select 'best' model. Verify model assumptions using ppc checks.

Statistical Objectives

- ✓ *Inference: Account for non-ignorable missing data (+more)*
- *Model Selection: Select the 'best' model for a given dataset*
- *Model Checking: Evaluate the adequacy of a given model*

Model Selection

Given an assortment of candidate models (e.g., single vs. Broken vs. Mixture-Pareto), we need a criteria to select the best model.

This allows us to address the most important question:
"Is there sufficient evidence of a 'break' in the logN-logS plot?"

We use a Bayesian model selection technique based on the Deviance Information Criterion (Spiegelhalter et al., 2002). Alternatives include Bayesian Predictive Information Criterion (Ando, 2007). DIC also has a model checking aspect.

In a simplified but realistic context (no incompleteness), $> 80\%$ classification success can be achieved (Wong, Baines, Lee, Aue; 2012).

Model Selection

How often can we recover the true number of breakpoints?

Model Selection

How often can we recover the true number of breakpoints?

For no background setting (Wong et al, 2012):

True B		$\hat{B}$			
		1	2	3	4
1	BIC	195	5	0	0
2	BIC	10	190	0	0
3	BIC	0	32	168	0

Table: Number of pieces $\hat{B}$ selected by BIC (True B=1,2,3)

With background but no incompleteness we obtain $\approx 80\%$ success.
With background, incompleteness and all effects. . . needs work ☺

CDFN: Model Selection

For CDFN, including incompleteness and all uncertainties, top candidates models (by DIC) are:

Model Type	K_0	K_1	$\lvert\theta_1 - \theta_0\rvert$	DIC
Broken-Pareto	-16.4	-15.59	0.27	3473.87
Broken-Pareto	-16.4	-15.68	0.24	3474.90
Broken-Pareto	-16.4	-15.77	0.22	3475.15
Regular	-16.4	—	—	3475.70

Data suggests a Broken Power-law, with $(\hat{\theta}_1, \hat{\theta}_2) = (0.60, 0.87)$.

CDFN: Model Selection

For CDFN, including incompleteness and all uncertainties, top candidates models (by DIC) are:

Model Type	K_0	K_1	$\lvert\theta_1 - \theta_0\rvert$	DIC
Broken-Pareto	-16.4	-15.59	0.27	3473.87
Broken-Pareto	-16.4	-15.68	0.24	3474.90
Broken-Pareto	-16.4	-15.77	0.22	3475.15
Regular	-16.4	—	—	3475.70

Data suggests a Broken Power-law, with $(\hat{\theta}_1, \hat{\theta}_2) = (0.60, 0.87)$.

Statistical Objectives

- ✓ *Inference:* *Account for non-ignorable missing data (+more)*
- ✓ *Model Selection:* *Select the 'best' model for a given dataset*
- ▶ *Model Checking:* *Evaluate the adequacy of a given model*

Model Checking

The posterior predictive $p-$value (Rubin, 1984), is a tool for assessing the adequacy of the model fit for Bayesian models based on the the posterior predictive distribution $p(y^*|y)$.

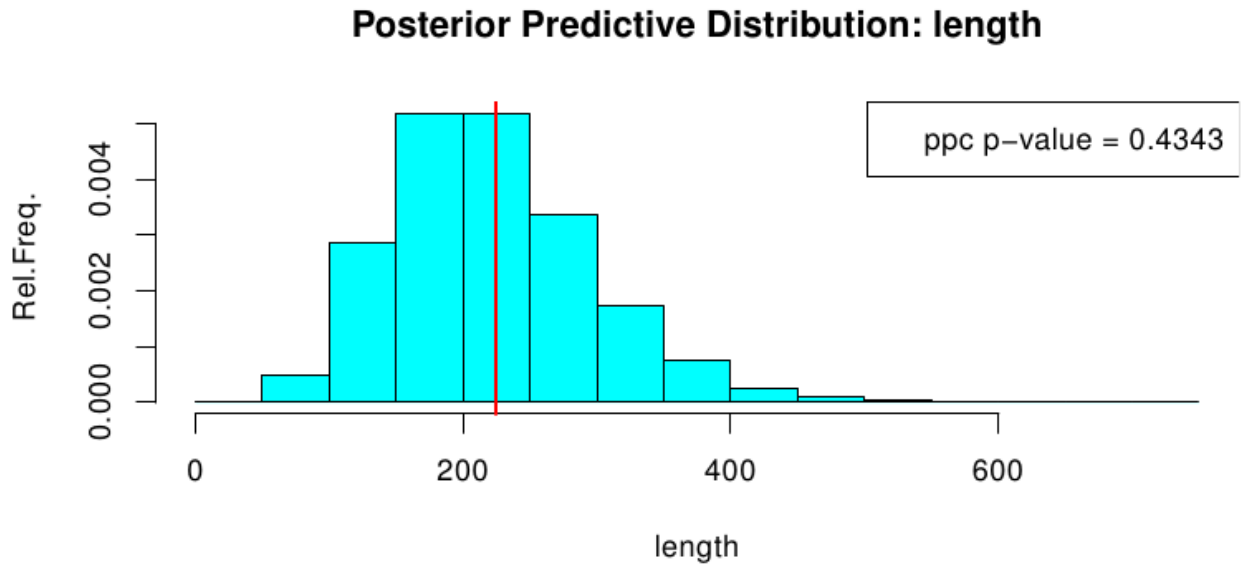
Consider testing the hypothesis:

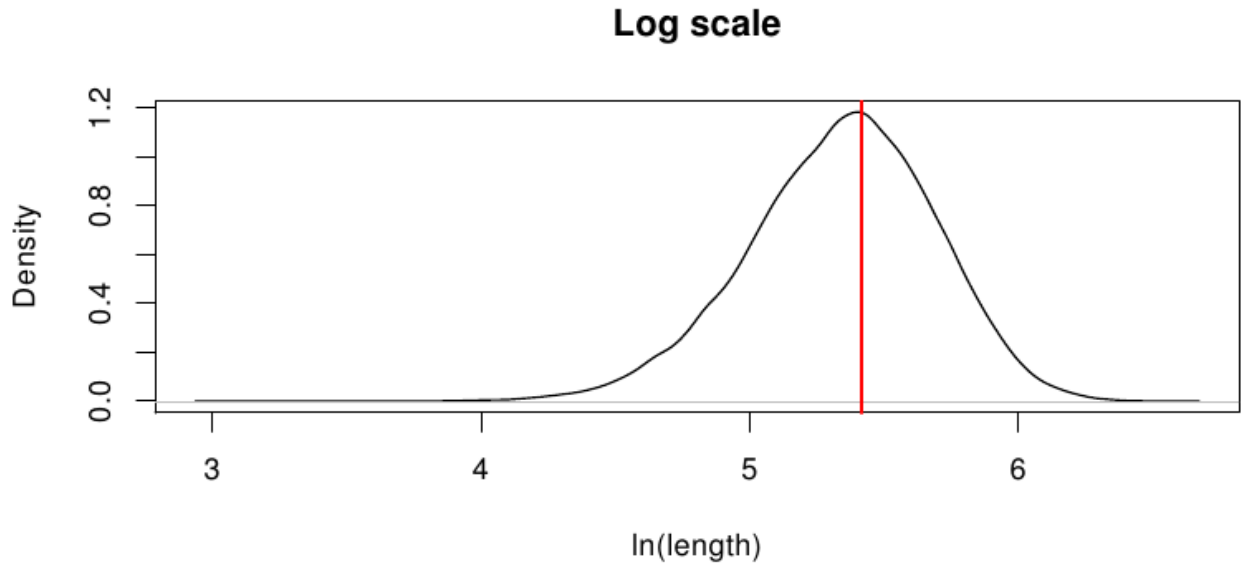
$$\begin{aligned} \mathcal{H}_0 &: \text{ The model is correctly specified , } vs., \\ \mathcal{H}_1 &: \text{ The model is not correctly specified .} \end{aligned}$$

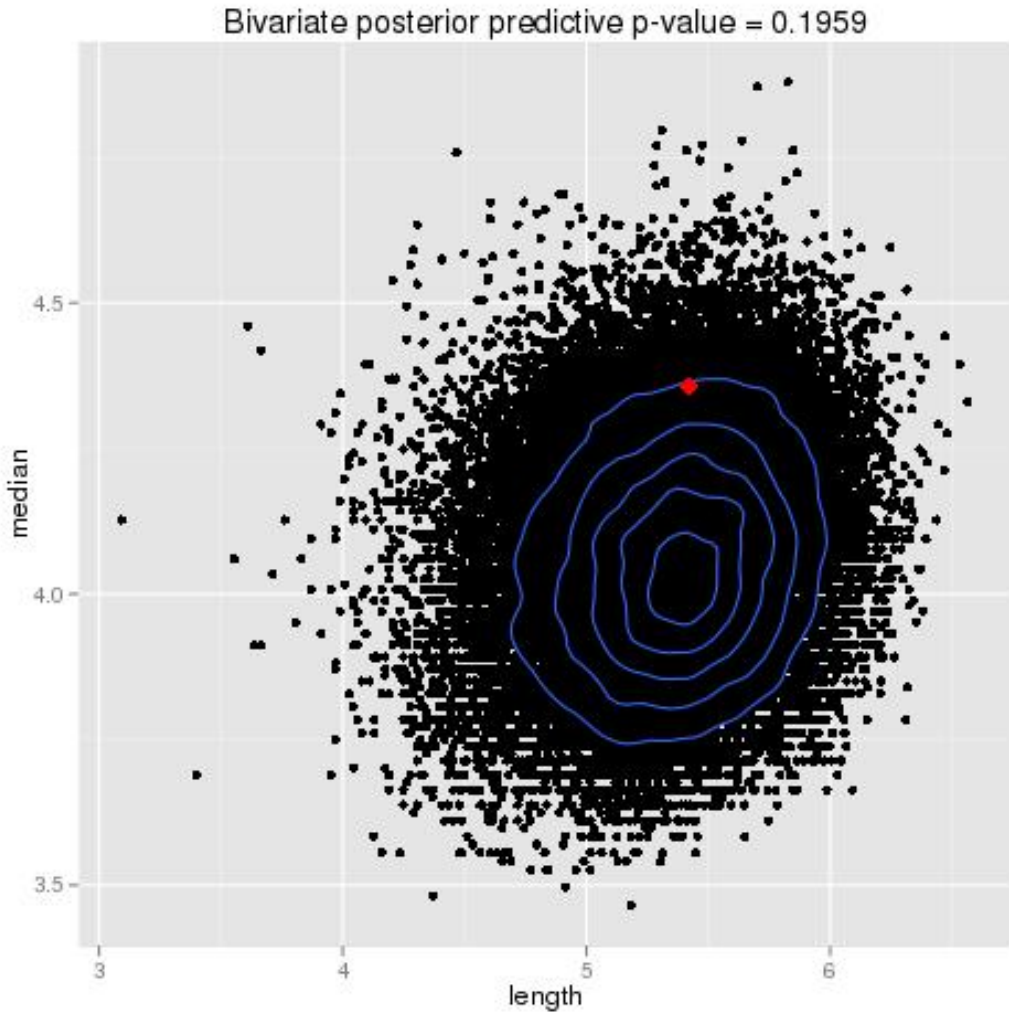
Select a test statistic $T(x)$ to perform the test, then we define the posterior predictive $p-$value to be:

$$p_b = \mathbb{P}\left(T(y^*) \geq T(y) | y, \mathcal{H}_0 \right).$$

Freedom of choice for $T(\cdot)$. Examples for the CDFN dataset...

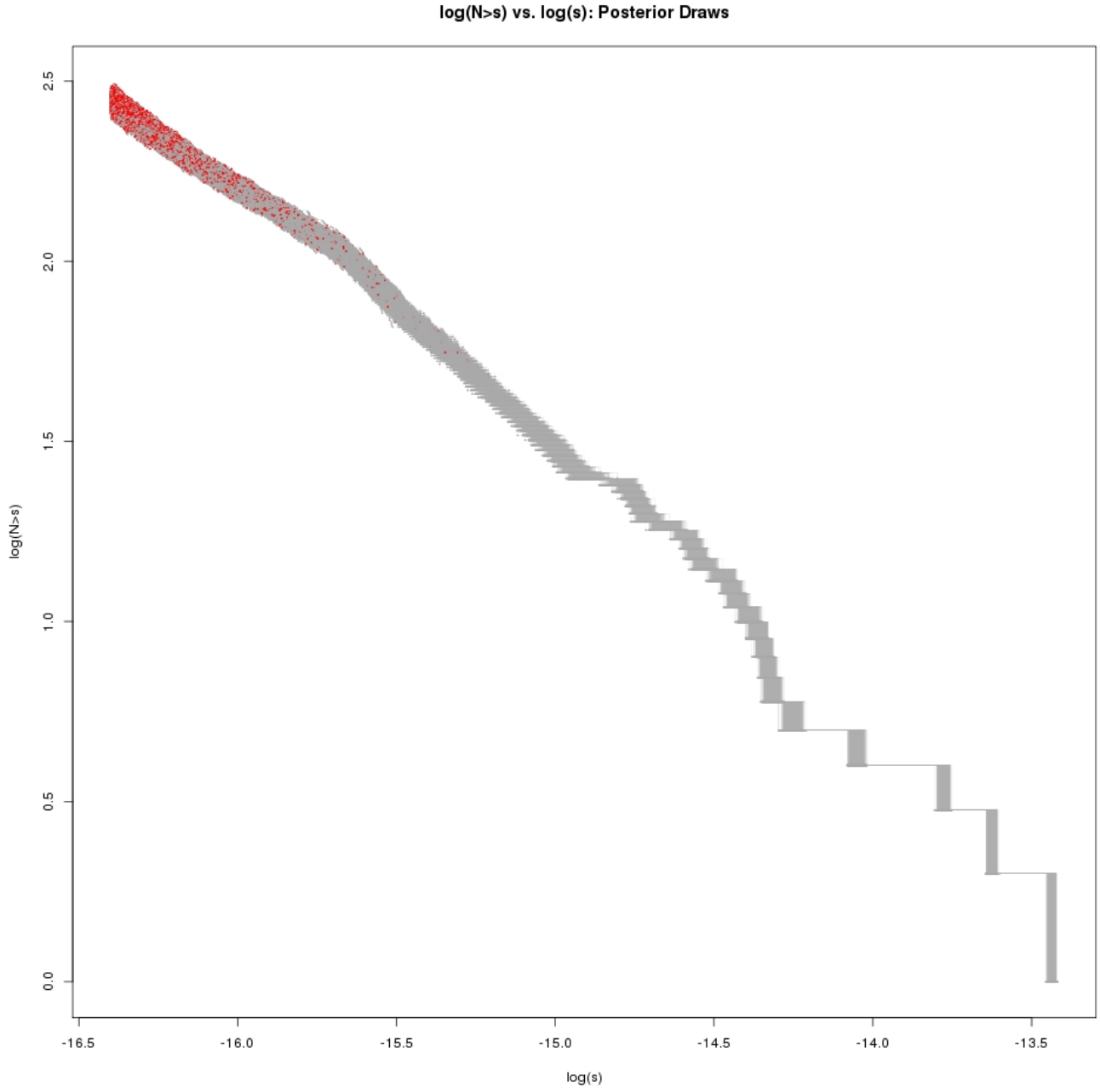

Posterior Predictive Distribution: length
ppc p−value = 0.4343
Rel.Freq.
0.000
0.002
0.004
0
200
400
600
length

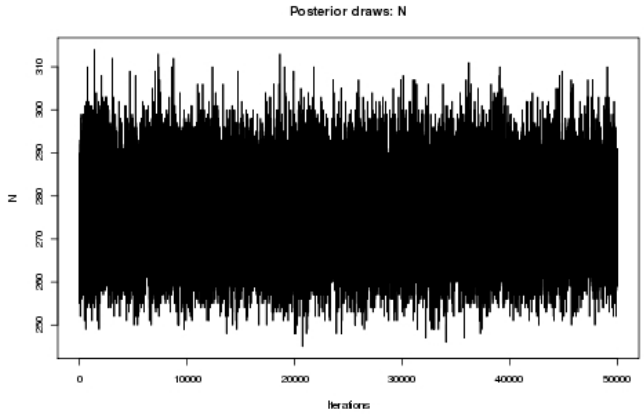

Log scale
Density
0.0
0.4
0.8
1.2
3
4
5
6
ln(length)

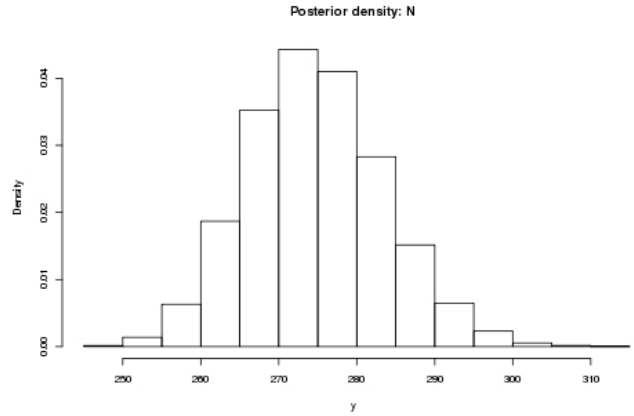

Bivariate posterior predictive p-value = 0.1959
median
length
4.5
4.0
3.5
3
4
5
6

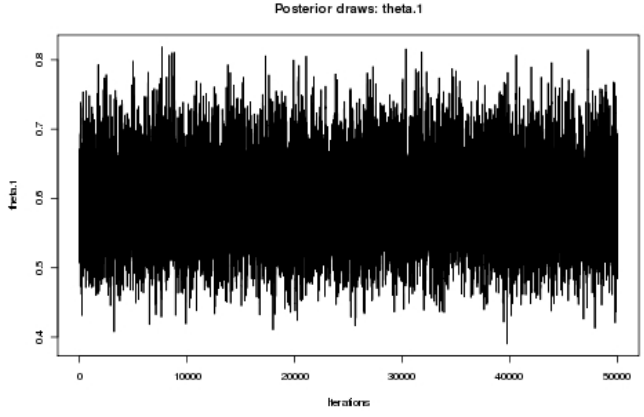
Statistical Objectives

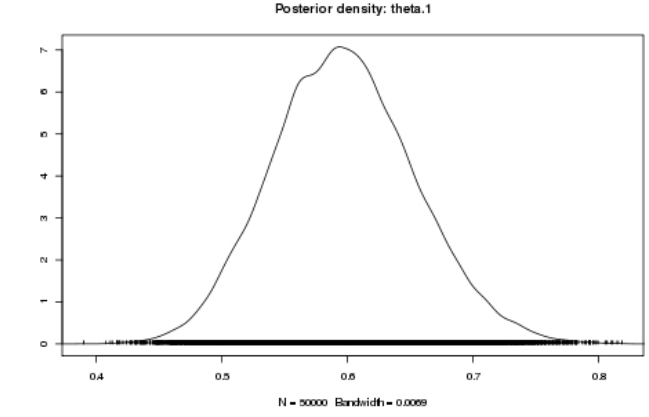
- ✓ *Inference: Account for non-ignorable missing data (+more)*
- ✓ *Model Selection: Select the 'best' model for a given dataset*
- ✓ *Model Checking: Evaluate the adequacy of a given model*

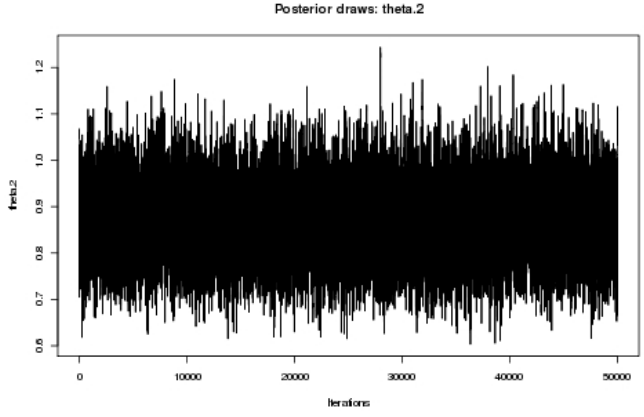

log(N>s) vs. log(s): Posterior Draws
log(N>s)
2.5
2.0
1.5
1.0
0.5
0.0
-16.5
-16.0
-15.5
-15.0
-14.5
-14.0
-13.5
log(s)

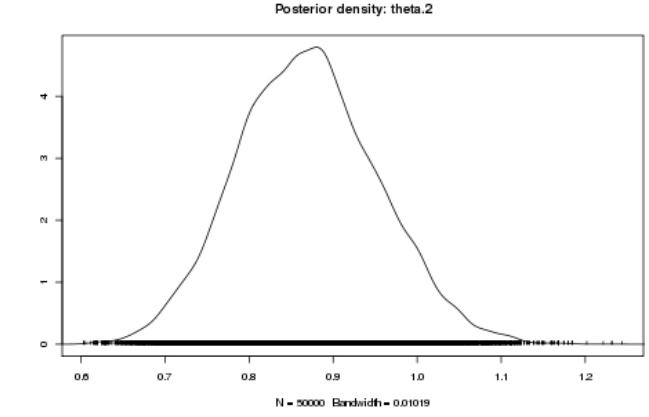

Posterior draws: N
N
Iterations


Posterior density: N
Density
y


Posterior draws: theta.1
theta.1
Iterations


Posterior density: theta.1
N = 50000 Bandwidth = 0.0069


Posterior draws: theta.2
theta.2
Iterations


Posterior density: theta.2
N = 50000 Bandwidth = 0.01019

Conclusions & Future Work

Conclusions:

1. Probabilistic insight allows us to build statistical procedures that correspond to more physically realistic models
2. Hierarchical modeling allows for us to account for multiple types of uncertainties
3. Flexible framework for computation (e.g., distributional assumptions for fluxes)
4. Provides a recipe for assessing goodness-of-fit
5. Provides a recipe for selecting between single and broken-pareto models
6. Explicity handles non-ignorable missing data

Future Work:

1. Break-point estimation for multiple power-law setting
2. Extension to non-Poisson regimes

References

- T. Ando (2007) Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models, *Biometrika*, 94, pp.443-458.
- 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*
- P.D. Baines, I.S. Udaltsova, A. Zezas, V.L. Kashyap (2012) Bayesian modeling of flux distributions: Estimation, Model Selection and Model Checking *(In prep.)*
- R.J.A. Little, D.B. Rubin. (2002) *Statistical analysis with missing data*, Wiley.
- D.B. Rubin (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12, pp.1151-1172.
- D.J. Spiegelhalter, N.G. Best, B.P.Carlin, A. van der Linde (2002) Bayesian measures of model complexity and fit. *JRRSB*,
- R.K.W. Wong, P.D. Baines, T.C.M. Lee, A. Aue (2012) Estimating astrophysical flux distributions using the IEM algorithm *(In Prep.)*