The Universe at Your Fingertips: Bayesian Modeling and Computation in Problems of Observational Cosmology

By

IRINA S. UDALTSOVA

B.S. (University of California, Davis) 2005
M.S. (University of California, Davis) 2007

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Statistics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

JIMING JIANG, Co-Chair

ETHAN ANDERES, Co-Chair

PAUL BAINES

Committee in Charge

2014

Contents

Irina S. Udaltsova
December 2014
Statistics

The Universe at Your Fingertips: Bayesian Modeling and Computation in Problems of Observational Cosmology

Abstract

In cosmology, the study of source populations is often conducted using the cumulative distribution of the number of sources detected at a given sensitivity. The resulting "log(N>S) – logS" distribution can be used to compare and evaluate theoretical models for source populations and their evolution. In practice, however, inferring properties of the source populations from cosmological observational data is complicated by the presence of detector-induced uncertainty and bias. This includes background contamination, uncertainty on both intensity and location of the observed sources, and, most challenging, the issue of non-detections or unobserved sources. Since the probability of a non-detection is a function of the unobserved flux, the missing data mechanism is non-ignorable. We present a Bayesian approach for inferring model parameters and the corrected log(N>S) – logS distribution for source populations. Our method extends existing work by allowing for joint estimation of both properties of the non-ignorable missing data process and the unknown number of unobserved sources. By correcting for the non-ignorable missing data mechanism and other detection phenomena, we are able to obtain corrected estimates of the flux distribution of partially observed source populations. We present a procedure for examining the goodness-of-fit of our hierarchical Bayesian model. Finally, we propose a novel approach for model selection in Bayesian settings.

Acknowledgments

I would like to thank many people who helped me through my graduate work. I am sure that without their support and encouragement, I would not be able to succeed. Growing up in a family of scientists may have inadvertently ingrained in me the value of a Ph.D. achievement. You grow as a person and as a researcher. I want to thank my family, Natalia Udaltsova, Vadim Barilko, Michael Udaltsov, Tatyana Udaltsova, Ariseny Barilko, and Shuriy Isaakovich, who have seen the beginning, middle, and the end of my Ph.D journey and supported me at all times during my change and growth.

I want to thank Professor Fushing Hsieh, who first suggested that I consider applying to graduate school, while being an undergraduate Statistics student at UC Davis. Thinking back now, I believe this was the best advice that dramatically changed my future. The work here allowed me to see the value in statistical computing, and I hope to continue work in this area. I want to thank Professors George Roussas and Frank Samaniego for recognizing my teaching abilities while my study as a Statistics graduate student. I also want to thank Professor Duncan Temple Lang for opening my interests in programming in `R` programming language. I also extend a great thanks to Sam Schmidt, who helped me in countless discussions about problems in cosmology. I feel very fortunate to have been part of the Statistics Department at UC Davis, which gave home away form home and allowed me to benefit from insights, experience, patience, and vast knowledge of every faculty member.

I owe huge thanks my dissertation advisers, Professors Jiming Jiang and Ethan Anderes, and my mentor Professor Paul Baines for allowing me to work on interesting research projects, for providing outstanding mentorship, necessary support, and discussions, and for allowing me to complete my study at UC Davis. I really appreciate that Jiming took me in as his student and allowed me to work in the area of astrostatistics, even though it was not a major part of his research. I also really appreciate the rigorous theoretical discussions I had with Ethan. I also extend my appreciation to Paul, who opened the opportunity to collaborate with talented astrophysicists Vinay Kashyap and Andreas Zezaz. My dissertation resulted from this collaboration. I thank Jiming for funding my research through NIH grant R01-GM085205A1 and NSF grants DMS-04-02824, DMS-08-06127 and SES-1121794.

Preface

«The purpose of computing is insight, not numbers.»

RICHARD HAMMING
— American mathematician

0.1. Astronomy and Statistics

0.1.1. Cosmology. The universe has fascinated human being for thousands of years. From the early writings of Copernicus about stars movement to every child's gaze into the dark starry night, we questioned, what is out there (in the Universe)? how does it work? and how did it come to be? To learn about the world beyond our planet Earth, we aided in advancement of technology, built tools for observation and data analysis, and postulated many theories about Universe. Through observation of the Cosmic Microwave Background (CMB), we were able to ascertain the Big Bang theory. Through observation of stars within our galaxy and galaxies beyond our possible reach, we developed theory of relativity and theory of quarks. The tendency of our Sun, the movement of planets, the existence of the dark matter are among many things that may or may not be directly observed. The study of celestial objects in and properties of our Universe became coined as cosmology.

The challenge in the field of cosmology is the space. We cannot travel across the Universe to measure it. Not only do we not have the technological development of space travel, we are physically prevented from directly touching and measuring our Sun - it is too hot and too large. Instead, we are limited to a view from Earth out to space. We construct better tools for observation and observe various areas of the sky or strongly resolved images of the same areas of the sky. The study of one Universe, one view of the sky, becomes a special science, that does not allow for replication of experiments. Thus, from the statistical point of view, ascertainment of good statistics is very difficult.

Of course, the sky observation does not come without complications. The measurements of light are generally biased. First, there is a strong selection bias, sometimes called the Malmquist bias or truncation bias, due to the brightness and proximity of celestial objects (Malmquist, 1920). It forces underestimation of the number of sources observed because it is more difficult to detect dim objects than bright ones, and it is more difficult to detect objects at large distances than at small distances away from the observer. Equivalently, the brighter sources are visible at greater distances. Thus, underrepresented samples are typically the outcome. Second, there is a bias that comes from the statistical fluctuations in measurement, called by some as the Eddington bias (Lynden-Bell, 1992). The same source may be measured to have various luminosities due to measurement error. Some sources will appear as other luminosity sources, which may cause misclassification of source objects. A worse issue is the fluctuation near the detection limit of the instrument which gives rise to missing data. Hence, every cosmological study must address the observational biases correctly, otherwise incorrect inferences may be drawn.

Complications with quality of observations are not limited to bias, but come from the measurement tools, also known as the limitations of the detector. For Earth based telescopes, weather, humidity and atmospheric particles prevent good observations and produce the blurring effect known as seeing. Imperfections in camera lens, charge-coupled devices (CCDs), or bandpass color filters produce greater systematic distortions in observations. These are calibrated for during every observational session. Additional distortions change light quality and its location. Madau reddening describes the occurrence that light from sources appears to be more red that the actual luminosity due to galactic dust (Madau et al., 1996). Gravitational lensing describes the occurrence when the sources appear to be stretched or have multiple reflections of themselves on an image due to strong gravitational pull of nearby invisible objects such as black holes or, even, dark matter. New algorithms for correct object detection and identification must be sought that incorporate these issues.

Resolution of aforementioned problems is a crucial component of cosmological inference in order to obtain valid estimates of uncertainty and biases associated with the observations and inferred quantities. However, it is a common practice in many cosmological studies that an existing method is used to estimate some unknown quantity; this estimate is used in another method for estimating a different quantity; and the plug-in layers of estimates continue. Substantial attention is devoted to calculation of the error estimates based on the final estimate; although, this is usually performed

under the assumptions that all intermediate plug-in estimates are known precisely and the errors in measurement are uncorrelated and random. When the true uncertainty is not handled properly across the layers of estimation, the resulting solution can be biased or incorrect. It is not surprising that astronomers turn to the statistical practice, where these issues are dealt with in a unified framework.

Indeed, majority of the current questions in astronomy are of statistical nature. Some example are as follows. How to model the points in an image representing the photon measurements of stars? How to classify the observed group of galaxies? How to account for measurement noise and stochastic signal of very dim sources? How to fit measurements of light spectra to non-linear astrophysical models? How to quantify uncertainty in the estimated parameters? Solutions to these problems make the field of astronomy very exciting, bringing together many disciplines including physics, mathematics, statistics, and computer science. In the review book, "Statistical Challenges in Astronomy", Eric Feigelson writes: "Powerful synergies thus emerge when astronomers and statisticians join in examining astrostatistical problems and approaches" (Feigelson and Babu, 2003).

One particular class of statistical methodology emerges to be useful in formulating problems in astronomy: Bayesian inference. It carries the notion that probability describes uncertainty. The aim of Bayesian approach is to update the existing scientific hypotheses (prior knowledge) with new evidence (data). The outcome is then expressed in terms of the degree of belief (probability), which allows for easy interpretation of the solution. In addition, with the help of computing, Bayesian methods can tackle very complex problems with relative ease, where the classical frequentist methods fail. Therefore, Bayesian approaches are naturally appealing to the astronomers.

0.1.2. Bayesian Statistics. Bayesian analysis is a statistical method for parameter estimation and prediction via the posterior distribution, which combines the observed distribution or likelihood with the prior distribution that summarizes the knowledge about the unknown parameters. The use of Bayes theorem allows us to quantify the uncertainty of the parameter in the posterior distribution made proportional to the product of the likelihood and prior distributions. The strength of the concept is unlimited. It gives rise to probabilistically coherent methodology for high-dimensional problems, small sample size problems, problems with incomplete observations, models with multiple layers of hierarchy in the parameters, or any other area where regular frequentist approaches may be difficult to apply, see, for example, Gelman et al. (2003).

To solidify our notation, define observed data as $y_i|\beta \overset{iid}{\sim} p(\beta), i = 1, \ldots, n$, the likelihood function based on data $\boldsymbol{y} = (y_1, \ldots, y_n)$ as $L(\beta) = p(\boldsymbol{y}|\beta) = \prod_{i=1}^{n} f_y(y_i|\beta)$, and encompass the prior information regarding the unknown parameter β into its probability density function $\pi(\beta)$. The Bayes theorem states that the posterior distribution of β given the data is achieved by:

$$p(\beta|\boldsymbol{y}) = \frac{L(\beta)\pi(\beta)}{\int L(\beta)\pi(\beta)d\beta} \propto L(\beta)\pi(\beta)$$

The posterior distribution, instead of a point estimate, provides a measure of uncertainty. All inference about the parameter is derived from the posterior distribution. The distributional summaries such as posterior mean (or median) can be used as parameter estimates; it is a good estimate in the sense that it minimizes expected posterior loss under the squared error loss (or absolute value error loss). Estimates of variance and modality are derived from posterior straightforwardly. The posterior predictive distribution of a new observation, $\tilde{y}$ (conditionally independent of the data given β), is determined by $p(\tilde{y}|y) = \int p(\tilde{y}, \beta|y)d\beta = \int p(\tilde{y}|\beta)p(\beta|y)d\beta$.

A central philosophy is molded based the intent to quantify and propagate the uncertainty characterized by probability. The unified intuitive framework, the ability to include prior information with observations, instinctive interpretation, and the effective empirical evidence makes Bayesian methods very attractive for physical data applications.

This dissertation presents additional research directions of Bayesian inference with use of computational methods of Markov Chain Monte Carlo (MCMC) applied in the area of astrostatistics. In the next section we describe important basic MCMC tools used for these purposes: the Gibbs sampler and Metropolis-Hastings sampler. The following chapters describe Bayesian methodology for missing data problems and spatial statistics problems for cosmological data.

0.2. Markov Chain Monte Carlo Methods

The field of Bayesian statistics bloomed with the development of Markov Chain Monte Carlo (MCMC) methods and increasingly powerful computing stations. Early inference relied on mathematical tractability of integrals and consisted of limited application that required analytic solutions. Starting with 1990's, however, major research directions generated multitudes of complex modeling strategies and Bayesian statistical approaches to broad range of problems relying on effective sampling algorithms of MCMC. For in-depth coverage of MCMC tools in statistics, we refer the reader to Robert and Casella (2004) and Gilks et al. (1996).

0.2.1. The Metropolis Algorithm. Metropolis Algorithm samples a sequence of random walk draws from a probability distribution. Direct sampling methods for sampling standard distributions, such as normal, uniform, gamma, or Poisson, are readily available in many statistical computer packages. The Metropolis algorithm becomes indispensable when the sampling is needed from nonstandard distributions or those that are known up to proportionality constant. The Metropolis algorithm is a rejection algorithm. It proceeds by sampling from a convenient candidate density $q(\beta^*|\beta^{(t-1)})$ of the parameter β^*, given the current state of the parameter, and decides to accept this proposal following a probabilistic rule. Most common candidate proposal density is a symmetric distribution $q(\beta^*|\beta^{(t-1)}) = N(\beta^*|\beta^{(t-1)}, v)$, where specification of the variance, v, is tuned during the burn in stage of the MCMC sampling period to allow for better mixing properties of the parameter (usually so that they acceptance rate is within 20% to 60%.). The algorithm is describes as follows:

ALGORITHM 1. The Metropolis algorithm repeats these steps, $t = 1, \ldots, T$:

Step 1: Draw β^* from $q(\beta|\beta^{(t-1)})$

Step 2: Compute the ratio $\alpha = h(\beta^*)/h(h(\beta^{(t-1)})$.

Step 3: If $\alpha \geq 1$, set $\beta^{(t)} = \beta^*$;

if $r < 1$, set

$$\beta^{(t)} = \begin{cases} \beta^* \text{ with probability } r \\ \beta^{(t-1)} \text{ with probability } 1 - r \end{cases}$$

The symmetric choice of the proposal density q is not always optimal. It is clearly evident if the parameter has a compact support. In this situation the Metropolis algorithm will result in high rate of rejections and waste computing time. The generalization is the Metropolis-Hasting (MH) algorithm, which allows non-symmetric candidate densities. MH algorithm replaces the acceptance ratio α in Step 2 of the Metropolis algorithm 1 with

$$\alpha = \frac{h(\beta^*)q(\beta^{(t-1)}|\beta^*)}{h(\beta^{(t-1)})q(\beta^*|\beta^{(t-1)})}.$$

Under mild regularity conditions, the draws β^t converge in distribution to the draws from the true posterior density $p(\beta, y)$ as $t \to \infty$ (Chib and Greenberg, 1995).

Many extensions and modifications to the Metropolis-Hastings algorithm have been proposed. Their goal is to speed up HM sampler convergence and improve the sampling coverage of the

parameter space. A very useful extension is called multiple-try Metropolis (MTM), which improves the sampling trajectory by increasing the acceptance rate and the step size of the sampler (Liu et al., 2000). In this algorithm, the next state of the MC chain is chosen from a set of samples at random with specially designed probability. It has the effects of reducing computation time and reducing autocorrelation within the chain. In the following chapters we use this approach in the implementation of our MCMC methods.

0.2.2. The Gibbs Sampler. The Gibbs sampler is a technique to sample from multidimensional posterior distribution assuming samples can be derived from each of full conditional distributions $p(\beta_j|\beta_{j\neq i,y}), j = 1,\ldots,p$. Under mild regularity conditions, the collection of full conditional distributions completely determine the joint posterior distribution $p(\beta|y)$ (Besag, 1974). This feature allows one to break up the high dimensional sampling problem into smaller manageable pieces or "easy" to sample full-conditional distributions. The hierarchical structure of many Bayesian problems naturally admits itself to this division. This feature also implies that one collects (albeit, correlated) samples from the joint posterior. The algorithm is described as follows:

ALGORITHM 2. For set of starting values $\{\beta_1^{(0)},\ldots,\beta_p^{(0)}\}$, Gibbs Sampler repeats these steps, $t = 1,\ldots,T$:

Step 1: Draw $\beta_1^{(t)}$ from $p(\beta_1|\beta_2^{(t-1)},\beta_3^{(t-1)},\ldots,\beta_p^{(t-1)},y)$
Step 2: Draw $\beta_2^{(t)}$ from $p(\beta_2|\beta_1^{(t)},\beta_3^{(t-1)},\ldots,\beta_p^{(t-1)},y)$
$\vdots$
Step 3: Draw $\beta_k^{(t)}$ from $p(\beta_k|\beta_1^{(t)},\beta_2^{(t)},\ldots,\beta_{p-1}^{(t)},y)$

Many extensions and modification to the Gibbs Sampler have been proposed, and some are very useful in practice. For more complex MCMC problems, for which the full conditional distributions are not in closed analytic form or for which the normalizing constant of the density in unknown, we can use Metropolis-within-Gibbs sampling method. Here, instead of directly sampling from the full conditional distribution $p(\beta_j|\beta_{j\neq i,y})$, one may use Metropolis-Hasting algorithm to generate β_j draws. Another very useful extension is the Blocked Gibbs Sampler, where the full conditional distributions are defined as joint conditionals in the event that some parameters can be naturally grouped and sampled from together given all other parameters and the data. In the following chapters we use these ideas in the implementation of our MCMC methods.

0.3. Outline of Road Ahead

This thesis concerns the following astrostatistical problems. Chapter 1 proposes a hierarchical Bayesian model for estimation of linear $\log(N) - \log(S)$ relationship between the log of the flux of sources and the log of the number of sources observed to that flux sensitivity. The challenge of this problem is survey incompleteness, which non-ignorable from the statistical point of view. The method attempts to correctly account for non-ignorable data and other uncertainty and biases resulting from the cosmological survey and the detector. Chapter 2 extends the hierarchical Bayesian model for estimation of piece-wise linear $\log(N) - \log(S)$ relationship. A Bayesian adaptive fence method for model selection in Bayesian settings is proposed.

CHAPTER 1

Complete Bayesian Analysis of the $\log(N) - \log(S)$ Problem Modeled via a Single Pareto Population

«We consider it a good principle to explain the phenomena by the simplest hypothesis possible.»

Claudius Ptolemaeus

— ancient Greek astronomer, mathematician

1.1. Introduction

Population characteristics of cosmological objects are often of central importance in cosmology and are a focus of astrostatistics. In this chapter we will focus on the study of the distribution of the flux. Flux is the observed brightness of astrophysical objects. Knowledge of the flux distribution is required to test and constrain theoretical assumptions about the Universe. The traditional approach is to estimate the number density distribution of the source flux empirically via a $\log(N) - \log(S)$ relationship. The $\log(N) - \log(S)$ relationship relates for the plot of the logarithm of the source flux, S, to the logarithm of the number of sources, N, observed to that flux sensitivity. In statistical terms, it gives the representation of the empirical survival function as a function of the log of the source flux. We refer to the next section for a detailed definition. We will present a method for estimating the $\log(N) - \log(S)$ relationship of the flux of X-ray sources, although the method is applicable to observations in other bands of electromagnetic spectrum.

The statistical nature of cosmological measurements and the observation process pose certain challenges in estimating the $\log(N) - \log(S)$ relationship. Main problems are the presence of noise and missing data. X-ray is a form of electromagnetic radiation. Its wavelength is much shorter (and frequency is higher) than that of visible light and radio waves. In terms of produced energy, X-rays expel much greater energy than radio waves. The rate at which the energy flows is known as the flux. For all cosmological data the measurements of the flux are not directly observed. Instead, the observations represent the cumulated flux received from the source, recorded as the

number of photon counts. The photon counts are subject to inherent Poisson-like variability. In addition, observations intended to measure the flux distribution are subject to both natural and detector induced uncertainties and biases. For example, source intensities can be contaminated by background luminosity or reduced due to their location away from the detector. One important consequence of these effects is that a subset of the source population of interest may be unobserved.

Early work focused on estimation of linear $\log(N) - \log(S)$ relationship. For reasons to be described later, linear relationship is associated with the Pareto distribution of the flux. Crawford et al. (1970) and Murdoch et al. (1973) derive maximum likelihood (ML) estimate of the slope parameter and apply the method to measurements of radio sources, for which the signal to noise ratio is generally high. To account of the difference between the flux and the photon counts, Murdoch et al. (1973) use normal approximation to measurement errors. Consequent series of papers on study of X-ray measurements apply the method of ML to estimate of the slope parameter under the assumption of Poisson error distribution (Maccacaro et al., 1982, 1987, Schmitt and Maccacaro, 1986). A recent work for estimating the parameters in linear and piece-wise linear $\log(N) - \log(S)$ relationships performed ML with application of expectation-maximization (EM) algorithm (Wong et al., 2014). Valid inference based on these methods is only possible for complete data surveys. X-ray measurements are particularly susceptible for incompleteness. Since fainter sources are more likely to be unobserved, the missing data mechanism is *non-ignorable* (Little and Rubin, 2002).

If the non-ignorable missing data mechanism is not accounted for, the estimation procedure may result in inferential bias. Bayesian methods are well-suited to these situations as they provide a unified and straightforward probabilistic framework. We develop a statistical method based on a Bayesian hierarchical model for estimating: (i) the number of sources unobserved due to the detector effects, (ii) the flux of observed sources, and, (iii) the parameters of the $\log(N) - \log(S)$ curve. By modeling the missing data mechanism (MDM, thereafter) we correct for detection biases and obtain posterior summaries for the bias-corrected source population.

Our Bayesian method for treatment of missing data in astrophysical surveys draws some parallels from the work presented by Loredo and Wasserman (1995). The authors develop a Bayesian approach for analyzing the distribution of gamma-ray burst peak photon fluxes and directions. Their method has a similar premise to account for all sources of uncertainty and biases attributed to the selection effects and non-detection. Our method differs in three respects. 1) Since our survey is very dim, we require to model the distribution of flux. 2) We derive the posterior distribution of

the model parameters marginalized across all missing data information instead of performing joint inference. 3) Our method produces a posterior distribution for the $\log(N) - \log(S)$ relationship instead of best-bit plug-in estimate.

This chapter is organized as follows. In section 1.2 we explain the base of the probabilistic framework: the power-law assumption and its connection to the Pareto distribution. In particular, section 1.2.1 introduces the missing data problem and how we can conceptualize missing data as a part of the model; section 1.2.2 describes the hierarchical Bayesian model. In section 1.3 we give the computational details of the method and insights to MCMC sampling procedures. In section 1.4 we describe the performance of our method, including validation of the method, assessment of the performance under model misspecification, and a tool for diagnosing model fitness. In section 1.5 we apply our method to a cosmological dataset of X-ray pulsar sources from *CHANDRA* Deep Field North and Deep Field South Surveys. We conclude the chapter with some discussions in section 1.6 where we offer possible extensions of the method.

1.2. Hierarchical Bayesian Modeling of Astrophysical Populations

The goal of our analysis is to study the distribution of fluxes for populations of astrophysical sources. Flux, also known as apparent brightness, is a measure of the amount of energy given off by an astronomical object, e.g., a star, over a fixed amount of time and area. Flux measurements make it easy for astronomers to compare the relative energy output of objects with very different sizes or ages. Population properties of the flux give insight about the stellar evolution and other astrophysical parameters.

Let S_i denote the flux ($ergs\ s^{-1}\ cm^{-2}$) of source i for $i = 1, \ldots, N$ and let $N(> S)$ be the number of sources in the population with flux above a threshold S. Historically the $N(> S)$ curve, or its log-scale counterpart, has provided a convenient way to summarize the distribution of fluxes in population. In this section we present a general framework and a hierarchical Bayesian modeling technique for estimating a source population distribution that accounts for detector induced biases, background noise and selection effects. We defer description of the detector effects until section 1.2.1, and begin with a discussion of basic population distribution modeling for astrophysical populations.

A standard assumption by cosmologists for the flux number density distribution is will follow a power-law. That is, the number of sources with flux exceeding a threshold S is assumed to obey a

relationship as a function of S:

$$N(>S) = \sum_{i=1}^{N} I_{\{S_i > S\}} \propto \alpha \cdot S^{-\theta}, \quad S > \tau > 0, \tag{1.1}$$

for some arbitrary positive constant τ, which we refer to as the minimum flux. Taking a logarithm transformation produces the linear $\log(N) - \log(S)$ curve:

$$\log_{10}(N(>S)) \propto \log_{10}(\alpha) - \theta \log_{10}(S). \tag{1.2}$$

By assuming a power-law for the survival function, there is an implied probability density function for a randomly selected source from the population. In particular, a straightforward derivation shows that, under independent sampling, the power-law form of the survival function in (1.1) and (1.2) uniquely corresponds to the Pareto probability density.

LEMMA 1. Let $S_i \overset{iid}{\sim} G$ where G is a probability distribution defined on (τ, ∞). If G has a power-law survival function of the form

$$\Pr(S_i > S) = \alpha \cdot S^{-\theta}, \quad S > \tau,$$

then G has a Pareto distribution.

This result is important in two respects. First, it shows that an assumption of linearity for the $\log(N) - \log(S)$ curve can be equivalently stated as an assumption that the source fluxes above a threshold τ follow a Pareto distribution. We denote this as $S_i \overset{iid}{\sim} \text{Pareto}\,(\theta, \tau), i = 1, \ldots, N$, for which the density is:

$$f(S_i|\theta, \tau) = \theta \tau^{\theta} S_i^{-(\theta+1)}, \quad S_i > \tau, \quad \tau, \theta > 0. \tag{1.3}$$

Second, by placing the linear $\log(N) - \log(S)$ assumption with its probabilistic equivalence we now have the basis for modeling the source fluxes under a hierarchical Bayesian framework. More importantly, the model that we develop can be explicitly checked via a simple goodness-of-fit procedure as detailed in section 1.4.4.

We motivate our approach in relation to the equivalence with a short discussion on direct estimation of survival curves. Survival curves play a crucial role in many biomedical applications, with the famous Kaplan-Meier estimator being the typical choice for estimating population survival curves

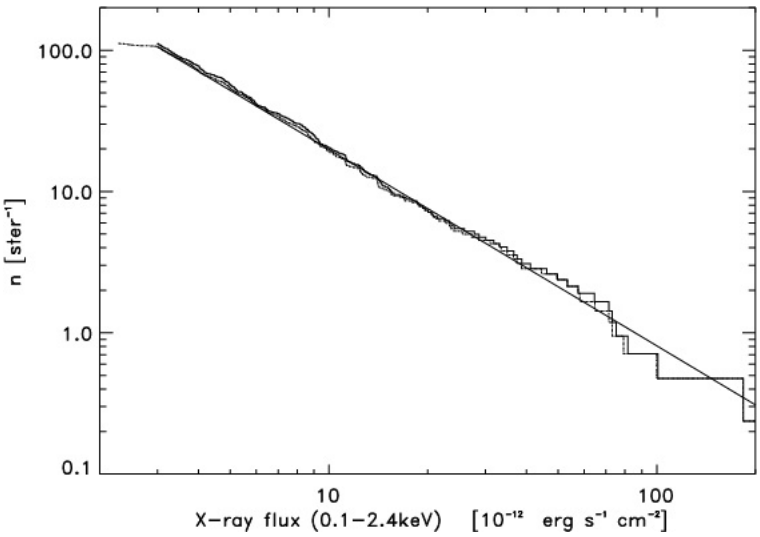

FIGURE 1.1. Example of $\log(N) - \log(S)$ relationship for X-ray sources. Linear fit and Kaplan-Meier estimates are drawn.

from sample data. The Kaplan-Meier estimator provides a consistent, non-parametric estimate of the population survival curve that allows for right-censoring in the observed data. In studies where little to no information about the underlying data generating process exists, non-parametric estimates are particularly desirable. Astrophysical studies have also applied survival analysis methods and Kaplan-Meier estimates (see Feigelson (1992) for overview). An example of a Kaplan-Meier estimate for linear $\log(N) - \log(S)$ relationship can be seen in Figure 1.1. Interestingly, when compared to biomedical studies, the problem of estimating the distribution of the flux possesses certain features inherently different, including that

(i) parametric models can be derived from first principles;

(ii) considerable additional information exists about the mechanism by which sources may not be observed; and

(iii) the fluxes (i.e., 'survival times') are not directly observed, and are subject to several sources of uncertainty (e.g, background contamination, inherent source variations, effective area uncertainties, image detection and pixelization effects).

Non-parametric methods make it challenging for features (ii) and (iii) to coherently include the full range of detector uncertainties. Feigelson (1992) discusses a restriction to special subsets (non-detections) of uncertainties to bypass the problem. Other astronomical studies suffer from the inability to handle missing data problems. Concurrent work on estimation of $\log(N) - \log(S)$

curvature parameters is done by Wong et al. (2014). The latter proposed to use the EM algorithm to estimate the sample parameters using maximum likelihood. However, a main drawback of their method is that the model cannot easily incorporate a MDM. Our fully Bayesian method is free of such obstructions. The goal of our methodology is to provide a flexible method (and accompanying software) to infer flux distributions from observed data that coherently accounts for the full spectrum of uncertainties introduced in the data collection process.

As outlined in (iii), the fluxes S_i are not observed directly in practice, so they must be inferred from other available measurements, typically photon counts. The photon counts are contaminated by detector effects in addition to their intrinsic uncertainties. We explicitly seek to capture three primary phenomena with our model: (a) background contamination, (b) intrinsic uncertainty, and, (c) missing data (i.e., incompleteness). By using a Bayesian hierarchical model we are able to account for all three phenomena simultaneously. As explained in detail in the next section, missing data is handled in a very natural way using Bayesian methods. In addition, Bayesian analysis allows for the introduction of external 'prior' information to facilitate estimation of unknown parameters.

1.2.1. Missing Data in $\log(N) - \log(S)$ Modeling. Incompleteness is an inherent challenge in cosmological surveys. The apparent brightness of stars is measured by using a detector, such as a CCD, that records how much energy strikes its light-sensitive surface each second. Bright sources tend to be observed more easily than dim ones, and very dim sources are not observed at all. Sources which physically exist but are not observed or not detected are henceforth referred to as missing sources. There are many conditions that can lead to missing sources – some of a statistical nature and others revolving around the location and calibration of the detector. For example, only a portion of X-ray emissions actually interacts with the detector resulting in Poisson-like noise in observed photon counts. Also, some dim sources are not detected because the background luminosity is brighter or just as bright as the source itself. Further problems with the observed image emerge due to detector specific effects, such as small exposure times (length of the observation period), large off-axis angle (large distance of source from the focal center of the detector), or small surface area (or detector aperture, which controls the angular resolution or blurring of the image). For ground-based telescopes, additional issue arrives in taking into account the absorption of light by Earth's atmosphere. However, it is not the focus of the study in the current data application for which observations are collected with the *CHANDRA* X-ray telescope which observes from space.

From a statistical perspective the key feature of this missing data is that the observed data now only corresponds to a, possibly biased, subset of the 'complete data'. For a scientist, inferential interest is in population parameters corresponding to the complete data, not the observed data. Hence, we separate two sources of error into the model conditional on the variables from the data collection process and the model that describes the data collection process (Gelman et al., 2003).

There is a rich statistical literature on the analysis of data with missing observations, see Little and Rubin (2002) for a review. The degree to which an analyst should be concerned about missing data is dependent on the mechanism by which the missingness arises. For example, it can be shown that randomly deleting observations from a dataset, while harmful for efficiency, typically does not affect the consistency of an estimator. Two restrictive but commonly used assumptions are that the data are either missing completely at random (MCAR) or missing at random (MAR). Under both of these assumptions the nature of the missing data does not depend on unobserved values and can be handled in a straightforward way.

To fix notation, we define $\boldsymbol{y} = (\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis})$ as the complete data vector separated into the observed, $\boldsymbol{y}_{obs}$, and missing, $\boldsymbol{y}_{mis}$, values, respectively, and I as the inclusion indicator, the random vector indicating whether each component of $\boldsymbol{y}$ is observed or missing. Let the model parameter of interest be β. The conditional distribution of the inclusion indicator given the complete data $\boldsymbol{y}$ is indexed by a parameter ϕ. We write the joint distribution of $(\boldsymbol{y}, I)$ given parameters (β, ϕ) as

$$p(\boldsymbol{y}, I|\beta, \phi) = p(\boldsymbol{y}|\beta)p(I|\boldsymbol{y}\phi)$$

and the observed information distribution as

$$p(\boldsymbol{y}_{obs}, I|\beta, \phi) = \int p(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}|\beta)p(I|\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, \phi)d\boldsymbol{y}_{mis}. \tag{1.4}$$

The MDM is defined by the conditional distribution of I given $\boldsymbol{y}$ and ϕ, which we refer to as *incompleteness function*. In particular, the MAR assumption implies that the MDM does not depend the missing values, so that the incompleteness function is

$$p(I|\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, \phi) = p(I|\boldsymbol{y}_{obs}, \phi),$$

and (1.4) can be simplified to

$$p(\boldsymbol{y}_{obs}, I|\beta, \phi) = p(\boldsymbol{y}_{obs}|\beta)p(I|\boldsymbol{y}_{obs}, \phi).$$

The stronger MCAR assumption implies that the data is observed at random and the MDM is completely independent of $\boldsymbol{y}$, which is written as

$$p(I|\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, \phi) = p(I|\phi),$$

so that (1.4) can be simplified as

$$p(\boldsymbol{y}_{obs}, I|\beta, \phi) = p(\boldsymbol{y}_{obs}|\beta)p(I|\phi).$$

The key property of the MDM is whether it can be considered to be 'ignorable' (Little and Rubin, 2002). The ignorable MDM occurs in the situation when the model parameters β and the missing-data distribution parameters ϕ are distinct and inferences for β can be made directly based on $p(\boldsymbol{y}_{obs}|\beta)$. In other words, the ignorable MDM can safely be ignored and inference about the parameters will remain valid. Even with ignorable missing data efficiency can sometimes be gained by incorporating knowledge of the MDM, such as in situations with MAR assumption.

In the $\log(N) - \log(S)$ problem, (holding other factors constant) lower flux sources have a lower probability of being observed than higher flux sources. Thus, the incompleteness function must depend on the flux of a source. Hence, the MDM is non-ignorable and must be accounted for in the analysis. By ignoring this fact and studying only observed sources, inference will typically be biased toward stochastically larger distributions. To bypass this issue, we may specify the flux threshold τ at a sufficiently large value and sources whose estimated flux falls below this limit are discarded, so that the probability of missingness is sufficiently small. This is a common approach, but has the drawback of discarding potentially useful data and not utilizing the knowledge contained in the incompleteness curves. We take another approach to solve this issue: by explicitly modeling the MDM.

Statistical applications with non-ignorable missing data are extremely challenging unless there is external knowledge about the MDM. Fortunately for $\log(N) - \log(S)$ analyses, such information is available. The incompleteness function directly provides information about the MDM that allows for full use of the data, and the ability to probe lower-flux ranges. From a Bayesian modeling standpoint, all aspects of the MDM need to be carefully translated into an incompleteness function that mathematically describes the probability of observing a source as a function of the source information and detector configuration. The incompleteness function is designed to incorporate all

knowledge about the underlying probabilistic MDM. That is, for any given flux, location, exposure and background intensity, it specifies the probability of observing a source. In practice, the incompleteness function can be derived from the detector sensitivity and properties of the observed image. Various approaches to obtaining the incompleteness function have been used in astrophysical applications, examples include Zezas et al. (2007) and Baloković et al. (2012). Since the incompleteness function encodes the MDM, parameter inference can be sensitive to the choice of the incompleteness function; thus, one needs to be careful in making such a choice.

In addition to missing data, some cosmological surveys also suffer from false detections. A false detection occurs when a 'source' listed in the observed data does not correspond to an actual source object. This can often be attributable to large background fluctuations. False detections can, in principle, be modeled. However, as this is not typically an issue with the X-ray data we examine here, we omit further discussion on this phenomena.

1.2.2. Probability Modeling of the $\log(N) - \log(S)$ Relationship. In this section we provide details of our probabilistic model of flux distributions in the presence of incompleteness, background contamination and other detector effects. In light of the discussion in section 1.2.1, we note that there are two different source populations: an 'observed source population' and a 'complete source population'. The observed source population corresponds to a typically biased subset of the complete source population. Inferentially, our goal is to estimate the $\log(N) - \log(S)$ relationship for the complete source population, not for the observed population. To do so we make explicit use of the missing data mechanism, which effectively describes the selection mechanism of the observed population. Our hierarchical Bayesian model therefore connects (i) a model for flux distributions in the complete source population via (1.3) below, (ii) the incompleteness function that describes the filtering from the complete source population to the observed sources, and, (iii) a model for the observable quantities incorporating all detector uncertainties.

Let N be the (unknown) total number of sources in the complete source population, n the number of observed sources and N_{mis} the number of missing sources so that $N = n + N_{mis}$. As discussed in section 1.2.1, sources may not be observed for a variety of reasons (e.g., 'weak' flux close to threshold τ, large off-axis angle, etc.). Since the parameter N is unknown, we specify a Negative-Binomial prior for the total number of sources in the population with flux above a given threshold, τ i.e., $N \sim$ Neg-Bin (a_N, b_N). The prior parameters a_N and b_N should be selected by

the analyst to reflect any prior information, or lack thereof, about the size of the source population. We will address the issue of selecting prior parameters in more detail in the context of specific data analysis in section 1.4.1 and specifically for the CDFN dataset in section 1.5.1.

Conditional on the total number of sources and model parameters, we assume that source fluxes for the complete source population follow a power-law or are distributed according to a Pareto distribution, as in (1.3). The model parameters are then the power-law slope, θ, and the flux population minimum threshold, τ. Note that in this context, the value of τ is not the same as a detector threshold which may be higher or lower than the flux population minimum threshold. See section 1.4.2 for further discussion of this distinction. For convenience we assume a conditionally conjugate Gamma prior distribution for θ and τ i.e., $\theta \sim \text{Gamma}\ (a_\theta, b_\theta)$ and $\tau \sim \text{Gamma}\ (a_m, b_m)$. In section 1.4.2 we also consider a conditional version of the model that fixes τ. For this conditional approach the best model can then be chosen using standard model selection techniques.

The model assumptions can be summarized as:

$$N \sim \text{Neg-Bin}\ (a_N, b_N)\,,\ \theta \sim \text{Gamma}\ (a, b)\,,\ \tau \sim \text{Gamma}\ (a_m, b_m) \tag{1.5}$$
$$S_i|\tau, \theta, N \overset{iid}{\sim} \text{Pareto}(\tau, \theta), \quad i = 1, \ldots, N$$

As noted, the power-law assumption in (1.1) is a theoretical relationship for the complete population of source fluxes and, depending on the incompleteness function, may not directly apply to the observed data. The data collected consist of photon counts for a subset of the population of sources, determined by the background noise, off-axis angle, and exposure map (and other detector effects). Define for source $i, i = 1, \ldots, N$:

$$Y_i^{tot} = Y_i^{src} + Y_i^{bkg}, \tag{1.6}$$
$$Y_i^{src}|S_i, B_i, L_i, E_i \overset{ind}{\sim} \text{Poisson}\ (\lambda(S_i, B_i, L_i, E_i))\,,$$
$$Y_i^{bkg}|B_i, L_i, E_i \overset{ind}{\sim} \text{Poisson}\ (k(B_i, L_i, E_i))\,,$$

with $\lambda_i = \lambda(S_i, B_i, L_i, E_i) = S_i E_i/\gamma, k_i = k(B_i, L_i, E_i, A_i) = B_i A_i$. In (1.6), γ denotes the energy per photon, B_i denotes the per-pixel photon background rate for the source, A_i denotes background area of the source, L_i denotes the off-axis angle, and E_i denotes the exposure map or effective area. The known functions λ_i and k_i represent the source and background photon count intensity for a given combination of intrinsic image effects. The quantities (B_i, L_i, E_i, A_i) are known for

all observed sources. For missing sources we assume a model for these quantities that reflects the properties of the detector; it is usually available together with the image information.

The last component of our hierarchical model specifies the probability of a source being detected. Let I_i be the indicator that source i is detected ($I_i = 1$ if source i is observed, $I_i = 0$ if missing). I_i is a stochastic indicator variable, depending on the intensity, location, background and effective area of the source.

$$I_i|S_i, B_i, L_i, E_i = \begin{cases} 1, & \Pr = g\left(S_i, B_i, L_i, E_i\right) \\ 0, & \Pr = 1 - g\left(S_i, B_i, L_i, E_i\right) \end{cases} \tag{1.7}$$

The incompleteness function g specifies the probability of detecting a source of a specified intensity under known background and observation settings. Define vector $S_{com} = (S_{obs}, S_{mis})^T$ as the flux vectors of observed and missing sources. For simplicity of notation similar partitions hold for all other source information.

The assumption (1.6) applies to the complete source population, so some of the Y_i^{tot} will not be observed. Even for the observed sources we do not observe the separate source and background counts Y_i^{src} and Y_i^{bkg}, and instead only observe the total count Y_i^{tot}. The observed and missing data for photon counts can be partitioned as

$$Y_{obs}^{tot} = \left(Y_{i_1}^{tot}, \ldots, Y_{i_n}^{tot}\right)^T,\ Y_{mis}^{tot} = \left(Y_{i_{n+1}}^{tot}, \ldots, Y_{i_N}^{tot}\right)^T,\ Y_{mis} = (Y_{obs}^{src}, Y_{mis}^{src}, Y_{mis}^{tot}) \tag{1.8}$$

where $\{i_1, \ldots, i_n\}$ correspond to the indices of the observed sources. In contrast to most statistical applications involving missing data, in this setting the number of missing data points, as well as their values, are both unknown.

1.2.3. Implementation. Having built our hierarchical model for flux distributions we can combine the model assumptions and prior distributions according to Bayes rule to obtain a posterior distribution for all unknown quantities. As with most hierarchical models, the posterior distribution cannot be summarized in a neat analytic form, and numerical techniques must instead be used to obtain samples from the distribution. Here we use Markov Chain Monte Carlo (MCMC) to produce samples from the posterior. The model described in equations (1.5)-(1.7) can be combined to yield a complete data posterior distribution

$$p(N, \theta, \tau, S_{com}, I_{com}, Y_{obs}^{src}, Y_{mis}^{src}, Y_{mis}^{tot}, B_{mis}, L_{mis}, E_{mis}|n, Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs}).$$

This complete-data posterior distribution includes a number of parameters related to the missing sources that are not of direct interest, such as B_{mis}, L_{mis} and E_{mis}. Therefore, we marginalize the posterior distribution across all parameters not of direct interest to obtain the desired posterior distribution for $(N, \theta, \tau, S_{com}, Y_{obs}^{src})$. Given the number of parameters we construct a blocked Gibbs sampler of this form

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

Additional unknown parameters can be optionally attained as conditional draws. For example, to produce a complete data $\log(N) - \log(S)$ curve, we require posterior samples of the missing fluxes. These can be obtained by sampling from $p(S_{mis}|n, N, \theta, \tau)$. Sampling of some blocks of parameters require Metropolis-Hastings and rejection sampling methods. See the next section for full details of the sampling process. Our analysis framework is also implemented in an `R` package called `logNlogS` that is expected to be publicly available from the `CRAN` archive.

1.3. Computational Details

Our interest is the joint posterior distribution of parameters and latent variables given the data. We now provide a summary of distributions necessary for implementing our model (see section 1.2.2) and a description to the computation required for sampling from the joint posterior. We refer to Appendix A for detailed derivations. The MCMC sampling scheme is based on the blocked Gibbs sampler that involve Metropolis, Metropolis-Hastings, rejection sampling, and numerical integration within the blocks. We derive the full conditional distributions necessary for the Gibbs sampler below.

We bring attention to one important strategy for sampling from our high-dimensional posterior distribution. Missing data provides a problem in that the dimension of the complete posterior distribution may change with every new iteration of MCMC. That is, since the number of sources in the population, N, is unknown, we need sample N and all the missing data components, $(S_{mis}, I_{mis}, Y_{mis}^{src}, Y_{mis}^{tot}, Y_{obs}^{src}, B_{mis}, L_{mis}, E_{mis})$. The dimension of the latter depends on N. There are methods treating varying-dimensional posterior problems, such as Reversible Jump MCMC (RJMCMC) (Green, 1995). However we found this method to be too time consuming computationally for the $\log(N) - \log(S)$ application. Also, it is a challenge to devise proper jumping rules and appropriate mappings between any pair of dimensions. Also, there is a high possibility for strong

autocorrelation of the parameters, which would require longer sampling time to reach stationarity and draw reasonably independent samples from the posterior. We instead use another strategy by marginalizing the full joint posterior distribution over missing sources. This approach allows us to keep the dimension of the parameters to be sampled fixed.

Our strategy is to take the complete data posterior
$p(N, \theta, \tau, S_{com}, I_{com}, Y^{src}_{com}, Y^{tot}_{mis}, B_{mis}, L_{mis}, E_{mis} | n, Y^{tot}_{obs}, B_{obs}, L_{obs}, E_{obs})$,
and integrate out the missing data components $(S_{mis},\ I_{mis}, Y^{src}_{mis}, Y^{tot}_{mis}, B_{mis}, L_{mis}, E_{mis})$, leaving parameters (N, θ, τ) and missing data components, flux and photon counts of the observed sources (S_{obs}, Y^{src}_{obs}), to be sampled from the marginalized joint-posterior. After collecting these samples via blocked Gibbs sampler, we can impute the flux of missing sources $(S_{mis}, B_{mis}, L_{mis}, E_{mis})$ conditional on the data and sampled parameters. We believe that this strategy avoids unnecessary dependence of information between the observed and missing sources. Also, as mentioned, the dimension of the sampled quantities $(N, \theta, \tau, S_{obs}, Y^{src}_{obs})$ stays fixed, which greatly simplifies the derivations of full conditionals, and sampling procedure as a result.

Following section 1.2.2, we use the following priors:

$$N \sim \text{Neg-Bin}\,(a_N, b_N),\ \theta \sim \text{Gamma}\,(a, b),\ \tau \sim \text{Gamma}\,(a_m, b_m)$$

In all cases, the hyper parameters of the priors are chosen and fixed according to prior solicitation from our collaborators. However, as we carefully examine in our simulation studies and sensitivity analysis (see section 1.4), any diffuse prior will have little to no effect on inference.
Let $\lambda_i = \lambda(S_{obs,i}, E_{obs,i}, L_{obs,i})$ and $k_i = k(E_{obs,i}, L_{obs,i})$. Define the structure for short-hand notation of the density of a random variable x with parameter β as Distr-Name$(x; \beta)$. For example, the conditional distribution of observed photon counts is:

$$p(Y^{tot}_{obs} | n, N, S_{obs}, B_{obs}, L_{obs}, E_{obs}) = \prod_{i=1}^{n} \text{Poisson}\,(Y^{tot}_{obs,i}; \lambda_i + k_i),$$

where $Y_{obs,i}^{tot} \sim \text{Poisson}\left(Y_{obs,i}^{tot}; \lambda_i + k_i\right)$ represents the Poisson density of the total photon count of source i with intensity $\lambda_i + k_i$. The latent data distributions are

$$S_{obs}|n, N, \theta, \tau, B_{obs}, L_{obs}, E_{obs} \sim \prod_{i=1}^{n} \text{Pareto}\left(S_i; \theta, \tau\right) g(S_i, B_i, L_i, E_i)$$

$$S_{mis}|n, N, \theta, \tau, B_{mis}, L_{mis}, E_{mis} \sim \prod_{i=1}^{N-n} \text{Pareto}\left(S_i; \theta, \tau\right)(1 - g(S_i, B_i, L_i, E_i))$$

$$Y_{obs}^{src}|n, N, Y_{obs}^{tot}, S_{obs} \sim \prod_{i=1}^{n} \text{Binomial}\left(Y_{obs,i}^{src}; Y_{obs,i}^{tot}, \frac{\lambda_i}{\lambda_i + k_i}\right)$$

$$Y_{mis}^{tot}|n, N, S_{mis} \sim \prod_{i=1}^{N-n} \text{Poisson}\left(Y_{mis,i}^{tot}; \lambda_i + k_i\right)$$

$$Y_{mis}^{src}|n, N, Y_{mis}^{tot}, S_{mis} \sim \prod_{i=1}^{N-n} \text{Binomial}\left(Y_{mis,i}^{src}; Y_{mis,i}^{tot}, \frac{\lambda_i}{\lambda_i + k_i}\right)$$

The distributions for B, L, E variables can be approximated directly from the observational process. Thus, the joint marginalized posterior is (see Appendix A for derivation):

$$\begin{aligned}
&p\left(N, \theta, \tau, S_{obs}, Y_{obs}^{src}|n, Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs}\right) \\
&\propto \binom{N}{n} \mathbb{I}_{\{n \leq N\}} \cdot \left(1 - \pi(\theta, \tau)\right)^{(N-n)} \\
&\quad \cdot \binom{N + a_N - 1}{a_N - 1} \left(\frac{1}{1 + b_N}\right)^{N} \left(\frac{b_N}{1 + b_N}\right)^{a_N} \mathbb{I}_{\{N \in \mathbb{Z}^+\}} \\
&\quad \cdot \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} \mathbb{I}_{\{\theta > 0\}} \\
&\quad \cdot \frac{b_m^{a_m}}{\Gamma(a_m)} \tau^{a_m - 1} e^{-b_m \tau} \mathbb{I}_{\{\tau > 0\}} \\
&\quad \cdot \left[\prod_{i=1}^{n} p\left(B_i, L_i, E_i\right) \cdot \theta \tau^{\theta} S_i^{-(\theta+1)} \mathbb{I}_{\{\tau < S_i\}} \cdot g(S_i, B_i, L_i, E_i)\right. \\
&\quad \cdot \frac{(\lambda_i + k_i)^{Y_i^{tot}}}{Y_i^{tot}!} e^{(\lambda_i + k_i)} \mathbb{I}_{\{Y_i^{tot} \in \mathbb{Z}^+\}} \\
&\quad \cdot \left.\binom{Y_i^{tot}}{Y_i^{src}} \left(\frac{\lambda_i}{\lambda_i + k_i}\right)^{Y_i^{src}} \left(1 - \frac{\lambda_i}{\lambda_i + k_i}\right)^{Y_i^{tot} - Y_i^{src}} \mathbb{I}_{\left\{Y_i^{src} \in \{0, 1, \ldots, Y_i^{tot}\}\right\}}\right]
\end{aligned}$$

In the posterior distribution above, we require to integrate out all missing source information, that includes the flux of the missing sources. The consequence of such integral is a definition of a marginal probability of observing a source, $\pi(\theta, \tau)$:

$$\begin{aligned}\pi(\theta,\tau) &= \int p(I|S,B,L,E)\cdot p(S,B,L,E|\theta,\tau)\ dS\ dB\ dE\ dL \\ &= \int g(S,B,L,E)\cdot p(S,B,L,E|\theta,\tau)\ dS\ dB\ dE\ dL.\end{aligned}$$

Care must be exercised to evaluate this multidimensional integral well.

The optimal sampling strategy for such a difficult posterior distribution is not obvious. We utilize the blocked Gibbs sampler to accomplish this task. For many blocks, there is no easy way of sampling, so we propose nested Metropolis-Hastings algorithms to collect a new sample within each block. For those parameters that have simpler forms for their full conditionals, other more direct sampling methods are implemented, such as numerical integration or rejection sampling. Next, we present the full conditional distributions for parameters used in the blocked Gibbs sampler, with a description about their sampling process. Blocks are separated into various types of parameters. We utilize independence and conditional independence between variables whenever possible.

1.3.1. Full-Conditional Distributions. In this section we state the full conditional distributions of parameters and describe their sampling methods.

Sampling Y_{obs}^{src}: Sample vector Y_{obs}^{src} component-wise: we have, for $i = 1, \ldots, n$,

$$\begin{aligned}p\left(Y_i^{src}|\cdot\right) &\propto p(Y_i^{src}|Y_i^{tot}, S_i, B_i, L_i, E_i) \\ &\propto \text{Binomial}\left(Y_i^{src}; Y_i^{tot}, \frac{\lambda(S_i,B_i,L_i,E_i)}{\lambda(S_i,B_i,L_i,E_i)+k(B_i,L_i,E_i)}\right)\end{aligned}$$

Sampling is done directly.

Sampling S_{obs}: Sample vector S_{obs} component-wise: we have, for $i = 1, \ldots, n$,

$$\begin{aligned} p\left(S_i|\cdot\right) \propto p\left(S_i|N,\theta,\tau\right) \cdot p\left(I_i = 1|S_i,B_i,L_i,E_i\right) \cdot p\left(Y_i^{tot}|S_i,B_i,L_i,E_i\right) \cdot \\ \cdot p\left(Y_i^{src}|Y_i^{tot},S_i,B_i,L_i,E_i\right) \\ \propto \text{Pareto}\left(S_i;\theta,\tau\right) \cdot g(S_i,B_i,L_i,E_i) \cdot \text{Poisson}\left(Y_i^{tot};\lambda(S_i,B_i,L_i,E_i)+k(B_i,L_i,E_i)\right) \cdot \\ \cdot \text{Binomial}\left(Y_i^{src};Y_i^{tot},\frac{\lambda(S_i,B_i,L_i,E_i)}{\lambda(S_i,B_i,L_i,E_i)+k(B_i,L_i,E_i)}\right) \end{aligned}$$

Sampling is done with the Metropolis-Hastings (MH) algorithm. We choose a truncated normal proposal distribution for S_i:

$$\text{Trunc-Normal}(S^{prop};S^{curr},v_S^2,lowBound=\tau) = \frac{\frac{1}{v_S}\phi(\frac{S^{prop}-S^{curr}}{v_S})}{1-\Phi(\frac{\tau-S^{curr}}{v_S})},$$

where ϕ and Φ are standard normal PDF and CDF, respectively, v_S^2 is a tuning parameter to maintain acceptance rate for MH somewhere between $20\% - 60\%$ and the lower truncation limit is τ.

Sampling θ: We have

$$\begin{aligned} p\left(\theta|\cdot\right) \propto p\left(\theta\right) \cdot p\left(S_{obs}|N,\theta,\tau\right) \cdot \left(1-\pi(\theta,\tau)\right)^{(N-n)} \\ \propto \left(1-\pi(\theta,\tau)\right)^{(N-n)} \cdot \text{Gamma}\left(\theta;a+n,b+\sum_{i=1}^{n}\log\left(\frac{S_i}{\tau}\right)\right) \end{aligned}$$

We propose to perform the sampling of θ with either the Metropolis-Hastings algorithm or rejection sampling. For MH algorithm, we considered two different proposals. Symmetric proposal distribution for θ is Normal $(\theta^{prop};\theta^{curr},v_{\theta,1}^2)$, where $v_{\theta,1}$ is a tuning parameter. An asymmetric proposal distribution for θ is Trunc-Normal$(\theta^{prop};\theta^{curr},v_{\theta,2}^2,lowBound=0)$, where $v_{\theta,2}$ is a tuning parameter and the lower truncation limit is zero. Based on our numerical studies, we found that the rejection sampler performs poorly, and MH performs equally well for the two proposal distributions. Thus we chose the sampling for θ to be performed via a Metropolis algorithm with a symmetric proposal.

Sampling τ: Sampling τ is the most difficult of all due to its deepest level in the model hierarchy. We have

$$\begin{aligned} p(\tau|\cdot) &\propto p(\tau,\theta,N)\cdot p\left(B_{obs},L_{obs},E_{obs}|\tau,\theta,N\right)\cdot p\left(n,S_{obs},I_{obs}|N,\theta,\tau,B_{obs},L_{obs},E_{obs}\right) \\ &\quad\cdot p\left(Y_{obs}^{tot},Y_{obs}^{src}|n,N,\theta,S_{obs},I_{obs},\tau,B_{obs},L_{obs},E_{obs}\right) \\ &\propto \mathbb{I}_{\{0<\tau<\min\{S_1,\ldots,S_n\}\}}\cdot \text{Gamma}\left(\tau;a_m+n\theta,b_m\right)\cdot\left(1-\pi(\theta,\tau)\right)^{N-n} \end{aligned}$$

The sampling of τ proceeds with the Metropolis-Hastings algorithm. Care must be exercised to make sure samples are drawn from the proper region. We considered multiple variations to the standard Metropolis to achieve this: Metropolis with normal proposal, MH with truncated-normal proposal distribution and Metropolis applied to a transformation of τ. The last method proved to have the best performance in reducing autocorrelation and is described as follows. We take a logarithm transformation of τ in order to preserve the positivity and to avoid numerical instability of taking samples very close to zero:

$$\begin{aligned} \eta &= \log(\tau) \\ p(\eta|\cdot) &\propto e^{\eta(n\theta+a_m+1)}\cdot e^{-b_m e^{\eta}}\cdot\left(1-\pi(\theta,\tau=e^{\eta})\right)^{N-n}\cdot\mathbb{I}_{\{\eta<\log(c_m)\}}, \end{aligned}$$

where $c_m=\min\{S_i\}_{i=1,\ldots,n}$. The upper bound for η, $\log(c_m)$, is reflected in the truncated normal distribution chosen as the asymmetric proporal distribution:

$$\text{Trunc-Normal}(\eta^{prop};\eta^{curr},v_\eta^2,upBound=\log(c_m))=\frac{\frac{1}{v_\eta}\phi\left(\frac{\eta^{prop}-\eta^{curr}}{v_\eta}\right)}{\Phi\left(\frac{\log(c_m)-\eta^{curr}}{v_\eta}\right)}.$$

where v_η^2 is a tuning parameter. The implementation of this MH algorithm turns out to be simple if we redefine the posterior distribution with the parameter of interest on the logarithmic scale. We also implemented an alternative to the Metropolis algorithm called multiple-try Metropolis (MTM) (Liu et al., 2000). Based on our simulations, both MH and MTM appear to work well.

Sampling N: We have

$$p\left(N|\cdot\right) \propto \left(\begin{array}{c} N \\ n \end{array}\right) \mathbb{I}_{\{n\leq N\}} \cdot \left(1-\pi(\theta,\tau)\right)^{(N-n)} \cdot p\left(N\right) \cdot p\left(S_{obs}|N,\theta,\tau\right)$$

$$\propto \frac{\Gamma(N+a_N)}{\Gamma(N-n+1)} \cdot \left(\frac{1}{b_N+1}\right)^N \cdot \left(1-\pi(\theta,\tau)\right)^{(N-n)} \mathbb{I}_{\{n\leq N\}}.$$

Sampling is done directly using the inverse-CDF method, with the CDF computed via numerical integration.

Sampling S_{mis}: The vector of parameters S_{mis} is not required as part of our joint posterior distribution, as it has been shown that we can average over these latent variables in the derivation of the posterior distribution. However, we typically would like to produce the $\log(N)-\log(S)$ plot, thus, we want to impute these latent variables. With model assumption, $S \sim \text{Pareto}\left(\theta,\tau\right)$, the probability of observing a missing source is $p(I=0|S,B,L,E) = 1-g(S,B,L,E)$. Note that the dimension of the S_{mis} vector is $N-n$, that is, it depends directly on the value of N and changes from iteration to iteration. We sample vector S_{mis} component-wise: for $i=1,\ldots,N-n$,

$$(B_i,L_i,E_i) \sim p(B_i,L_i,E_i)$$

$$S_{mis,i}|n,N,\theta,\tau,B_i,L_i,E_i,I_i=0 \sim \left(1-g(S_i,B_i,L_i,E_i)\right)\cdot \text{Pareto}\left(S_i;\theta,\tau\right).$$

Sampling is done via rejection sampling.

1.3.2. Additional details.

Computing $\pi(\theta,\tau)$: The marginal probability of observing a source as a function of θ and τ is a multidimensional integral. We must revert to numerical techniques to evaluate this value. Direct evaluation may proceed in two ways. It can be approximated numerically, such as by a combination of Riemann Sums and the Trapezoid Rule. Or it can be approximated via Monte Carlo sampling. Based on our numerical studies we have found that the latter method gives better performance in terms of speed and accuracy, especially if the parameter dimension is large.

We have

$$\begin{aligned}\pi(\theta,\tau) &= \int p(I|S,B,L,E)\cdot p(S,B,L,E|\theta,\tau)\ dS\ dB\ dE\ dL \\ &= \int g(S,B,L,E)\cdot p(S|\theta,\tau)\cdot p(B,L,E)\ dS\ dB\ dE\ dL.\end{aligned}$$

We collect Monte Carlo samples of B, L, E and S conditional on the parameters θ, τ and evaluate the incompleteness probability, $g(S, B, L, E)$. The empirical average of g is a good approximation to $\pi(\theta, \tau)$, provided that the Monte Carlo sample size is large enough.

Evaluating π at every instance of the blocked Gibbs sampler may hinder the speed of our algorithm. For this reason we have made substantial efforts to conduct numerical studies about computing $\pi(\theta, \tau)$. We found that to achieve the desired accuracy for parameter estimation, 100,000 Monte Carlo samples is sufficient. In this case, pre-computing values of π over a grid of parameters (vs. not pre-computing) can speed up the MCMC sampling tenfold. In addition, a pre-computed surface of π can be re-used for any images produced by the same survey, that is unchanged detector effects.

1.3.3. Validation. The complexity of our hierarchical Bayesian model and computation necessitates a validation method to make sure that the results are correct. Fortunately, the Bayesian methods lend themselves to automatic self-consistency checks and validation using simulated data (Cook et al., 2006). The main idea is to compare the results from the data-generating software and the model-fitting software. If the generating procedure used is correct and the software works properly, then the "true" generated parameter should resemble a sample from the empirical distribution of posterior parameter draws. Furthermore, the quantile of the "true" scalar parameter with respect to the posterior distribution should follow Uniform(0,1) distribution. This fact allows to construct powerful diagnostics of correctness of the software package. The procedure involves generating and analyzing the data according to the same model, followed by calculating posterior quantiles of each scalar parameter. We now describe the algorithm in more detail.

ALGORITHM 3. Step 1: Simulate "true" parameters from the prior $\beta^{(0)} \sim p(\beta)$, and data from the model, given $\boldsymbol{y}_{obs} \sim p(\boldsymbol{y}_{obs}|\beta^{(0)})$.

Step 2: Fit the model to obtain MCMC samples of parameters $\beta^{(l)} \sim p(\beta|\boldsymbol{y}_{obs}), l = 1, \ldots, L$ and compute posterior quantiles by the $\lfloor qL \rfloor$-th order statistic of the MCMC sample. That is, for $0 < q < 1$, compute $\hat{\beta}_q = \min\{\beta^{(l)} : \widehat{\Pr}(\beta < \beta_q) = \frac{1}{L}\sum_{l=1}^{L} \mathbb{I}_{\{\beta^{(l)} < \hat{\beta}_q\}} \geq q\} = \beta^{(l)}_{(\lfloor qL \rfloor)}$.

Step 3: Record whether or not the 'true' value of the parameter $\beta^{(0)}$ lies below the quantile $C = \mathbb{I}_{\{\beta^{(0)} \leq \hat{\beta}_q\}}$.

Step 4: Repeat Steps 1 & 2 a number of times, say, M, and calculate the average coverage $\frac{1}{M}\sum_{j=1}^{M} C_j$.

Step 5: For each parameter, plot average coverage vs. nominal coverage, q, to compare to 45 degree line.

The validation coverage plot produced visually checks agreement between the average and nominal coverage. Deviation from the nominal coverage could result due to incorrect programming of the algorithm and due to MCMC error; however, if the program is written correctly, MCMC error is expected to lie within the binomial confidence bounds at the nominal probability. That is, the coverage count based on M trials follows the binomial distribution with nominal coverage success probability, if the program is written correctly. A validation coverage plot for our single-Pareto $\log(N) - \log(S)$ model is shown in Figure 1.2. Actual and nominal coverage (colored lines) agree to within the binomial confidence bounds (dashed lines) for all levels of probability, hence our method samples from the target posterior distribution correctly. Each colored line corresponds to the average coverage of each parameter: $N, \theta, \tau = S_{min}$, and mean of fluxes, $S_i, i = 1, ..., n$. Taking average of S_i is needed because S's are latent variables, and not parameters, and we do not expect to achieve perfect coverage for each individual flux.

The potential of model failure can also be examined with posterior interval plots, in which Steps 1 & 2 in Algorithm 3 are repeated and extended to evaluate posterior credible sets $\mathcal{C}$ of level α: $\int_{\mathcal{C}} p(\beta|\boldsymbol{y}_{obs})d\beta = 1 - \alpha$. The estimate based on MCMC samples is $(\hat{\beta}_L, \hat{\beta}_U)$ such that $q(\hat{\beta}_L) = \alpha/2$ and $q(\hat{\beta}_U) = 1 - \alpha/2$. The posterior intervals of each of the M simulations are plotted as vertical strips vs. the corresponding true values of the parameter. The intersection of the intervals and the 45 degree line is expected to occur approximately $100(1-\alpha)\%$ of the time. Figure 1.3 displays 90%

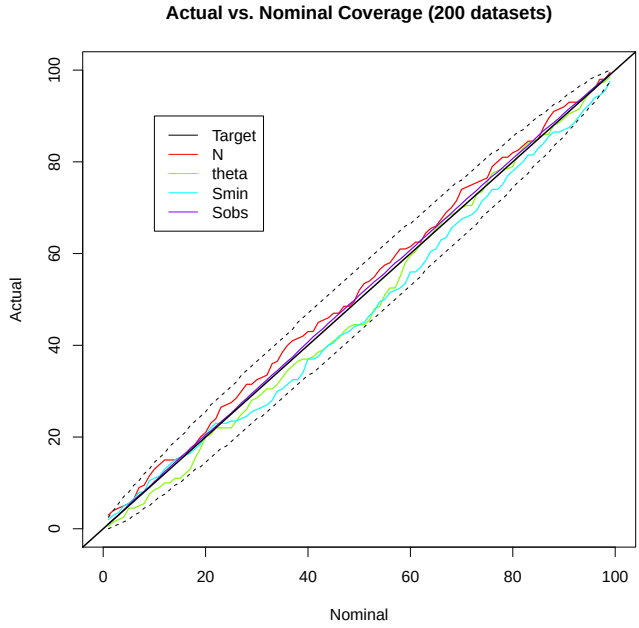

FIGURE 1.2. Coverage plot of main parameters of 200 dataset simulations during validation process.

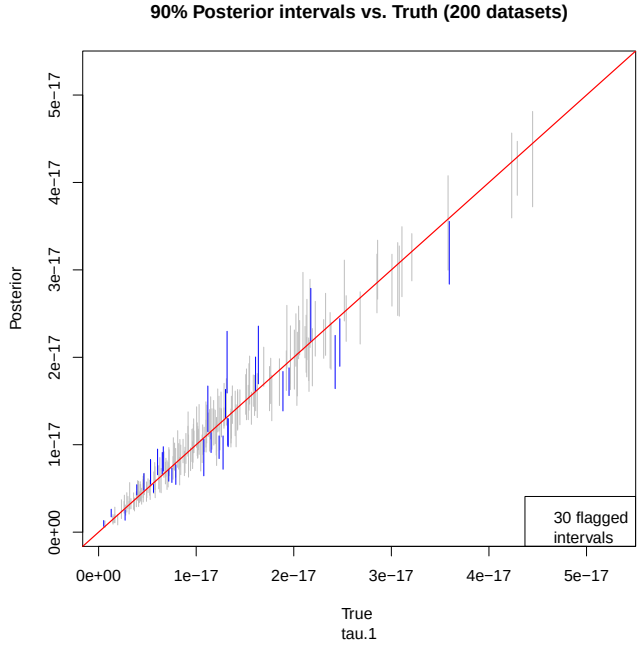

FIGURE 1.3. Posterior credible intervals of τ from 200 dataset simulations during validation process.

posterior intervals for parameter τ. There are 30 (or 15%) of the dark blue intervals which do not cover the truth, as expected.

Validation of the $\log(N) - \log(S)$ model is contingent on the correct specification of the cumulative incompleteness function, $\pi(\theta, \tau)$. As mentioned in section 1.3.2, no analytic solution exists for this integral and Monte Carlo simulation gives the fastest and most reliable approximation. We considered two version of MC approximation: pre-computing π over a dense grid of θ and τ before fitting the model and re-evaluating π during the model fitting stage. Our numerical experiments show that re-evaluation of π within every Gibbs step of the MCMC sampler incurs numerical error and induces biased coverage of key parameters of the model. The gain in the accuracy of the MC approximation by increasing the number of MC samples results in only marginal improvement to removing the bias in coverage. We experimented with various simulation scenarios up to 1,000,000 Monte Carlo samples for each integral estimate of π; all of them fail the validation. On the other hand, our simulations suggest that pre-computing π over a dense grid with 100,000 Monte Carlo samples is enough to achieve a validated result. This suggests that the Monte Carlo error has little effect on the parameter estimation when error on π does not change within MCMC stage; whereas the effect is amplified if Monte Carlo approximation for π is performed at each iteration within the MCMC. Hence, we suggest to pre-compute π if the analyst is interested in correctly estimating the tails of the distribution.

1.3.4. Statistical Inference. Once posterior samples have been obtained, those samples can be used in a number of different ways to summarize and visualize both marginal and joint posterior probability distributions. For all parameters, the samples from the posterior distribution are typically summarized into posterior estimates, such as the posterior mean, mode or median, and posterior credible intervals. In contrast to most non-Bayesian methods, having access to the full posterior distribution allows the analyst to observe asymmetry and multi-modality in the posterior distribution.

We now address the interpretation of several key parameters in the $\log(N) - \log(S)$ model of section 1.2.2, as well as what can be learned from our analysis. The most crucial inferential aspect for $\log(N) - \log(S)$ modeling is to realize that the characterization of the flux distribution can be done in two subtly different ways. The first characterization of the flux distribution comes from the slope θ in (1.2). This can be directly explored by plotting and summarizing the posterior samples for θ in the usual manner. In addition to θ, we can also examine the empirical $\log(N) - \log(S)$ relationship for the complete source population. This involves plotting the distribution of $\log(N) - \log(S)$

curves of the posterior samples of the flux for the observed and missing sources, as illustrated in Figure 1.11. The resulting posterior distribution of the $\log(N) - \log(S)$ curve can be used to visualize the level of uncertainty of our estimate of the curve and to investigate possible deviations from linearity in the curve. A formal method for the detection of non-linearity in the $\log(N) - \log(S)$ is more challenging and is addressed in section 1.4.4, where we introduce a 'goodness-of-fit' check for our model. Whereas all prior methods for estimation of $\log(N) - \log(S)$ relationship provide a point-estimate for $\log(N) - \log(S)$ (Crawford et al., 1970, Loredo and Wasserman, 1995, Maccacaro et al., 1982, Wong et al., 2014), our method is unique in expressing the uncertainty for the whole $\log(N) - \log(S)$ curve.

The values of other parameters also provide important information about the underlying astrophysical processes. The posterior distribution of N quantifies knowledge and uncertainty about the total number of sources in the complete source population. When the detection probability is very low, the complete source population can be much larger than the observed number of sources, thus the parameter estimates draw much of their information from the model assumptions and less from the data. Additionally, in situations of high-incompleteness, the posterior distributions tend to be wider to reflect the relative information in the observed data. This naturally reflects the inferential balance, that is, for settings with low incompleteness we must rely on external knowledge to understand the complete source population. In cases where such external information is available, there can be substantial gains in the accuracy and efficiency of the inference. In contrast, in situations of low-incompleteness, the total number of sources will be close to the number of observed sources and inference will typically be robust to the choice of incompleteness function. For these reasons, the sensitivity of inference about model parameters to the accuracy of the incompleteness function must be addressed. While our method requires accurate specification of the incompleteness function, particularly for high-incompleteness datasets, as illustrated in section 1.4.3, it drastically outperforms methods that ignore incompleteness altogether.

1.4. Simulation Studies of the Model Performance

We now investigate the ability of our model to estimate population model parameters, and the sensitivity of inference to model assumptions and choices of prior distributions. Most notably, we investigate the sensitivity of inference on θ, N, and τ to the choice of incompleteness function, and the impact of prior distributions on inference for θ and N.

We consider data simulated from our model with $\theta = 1.5$, $\tau = 10^{-16.5}$, and $N = 455$. The source-specific parameters (B_i, L_i, E_i) are sampled from the empirical joint distribution of those quantities for the CDFN survey to ensure compatibility with our analysis in section 2.7.1. The exposure time is fixed at $670,000$ *seconds* with a conversion factor between photon counts and photons of 430 cm^2ct/ph. The energy conversion factor is set at $\gamma = 1.6 \times 10^{-9}$ $ergs/ph$. The expected source and background photon counts are modeled as $\lambda_i = S_i E_i/\gamma$ and $k_i = B_i A_i$, where B_i is the background rate per pixel and A_i is the number of pixels covered by a source i. The proportion of observed simulated sources is 61%.

1.4.1. Sensitivity to Prior Distributions. In most settings the parameters of interest are likely to be θ, N and possibly τ, so we now investigate the sensitivity of inference to the choice of prior distribution for each of these parameters. We consider three different prior distributions for θ and obtain the posterior distribution for θ in each case, while N and τ are kept fixed at true values. The results are shown in Figure 1.4. In each case the prior distribution is shown as a dotted curve, the true value of θ is shown with a vertical line and the posterior distribution as a solid curve with the 95% central credible interval shaded. The left-most figure corresponds to a weak prior distribution for θ, which we recommend unless strong prior information is available about θ. The middle figure corresponds to a moderately informative prior distribution that is not centered at the true value but is nonetheless consistent with the true value. In both cases the posterior distribution effectively captures the truth. The right-most figure demonstrates what happens when a very strong, and incorrect prior distribution is used. In this case the data pushes the posterior distribution toward the true value, but the amount of data in this example is insufficient to overcome the misplaced certainty of the prior distribution. We therefore recommend moderately or weakly informative priors for θ in most settings.

Next we investigate the choice of prior distribution for N. We consider three prior distributions corresponding to a weak prior, a moderately informative prior that is consistent with the true value and a strongly informative but incorrectly specified prior while τ is kept fixed at the truth. Since inferential interest is primarily in θ, for each of the settings we also examine the corresponding impact on estimation of θ. The results for the three settings are shown in Figure 1.5. The left column displays the prior distributions for N, and the corresponding posterior distributions for N for each of the three priors. As expected the weak and moderately informative priors yield posterior

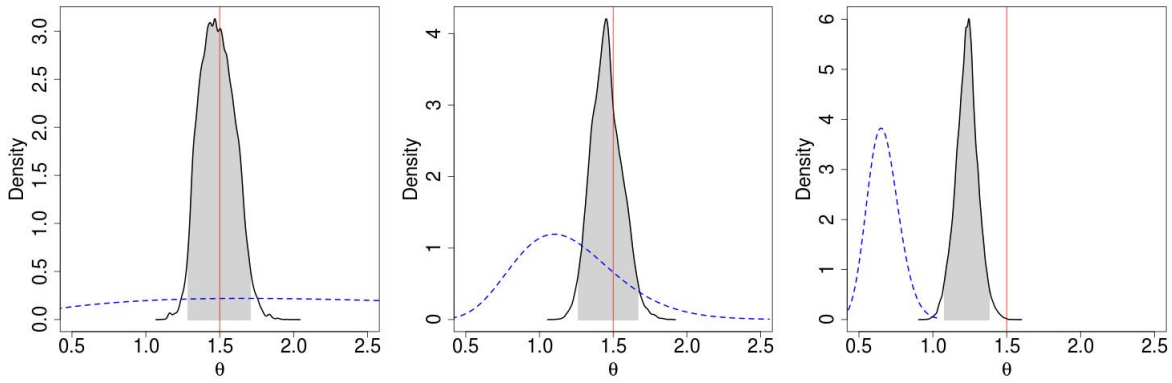

FIGURE 1.4. (L) Weak prior and corresponding posterior for θ, (C) a moderately informative prior for θ, and, (R) a strongly informative but incorrect prior for θ.

distributions that center around the true value of N, while the strongly informative and incorrect prior yields a posterior that is barely consistent with the true value. The right column displays the corresponding prior and posterior distributions for θ in each of the three settings. The results show that misspecification of the prior distribution of N has minimal impact on inference for θ; in all cases the central 95% posterior credible interval contains the true value. Hence, the estimates of θ are robust to the prior specifications of N. We note, however, that a stronger connection between θ and N is expected when τ is unknown and is to be estimated together with N and θ. We recommend weakly informative priors for all model parameters unless strong and reliable information is available to guide a more informative choice. A good "rule of thumb" for the selection of weak prior settings is to identify the plausible range for the parameters, and set the inner 50% interval of the prior to that range.

1.4.2. Sensitivity to Low-Threshold, τ. Estimates of $\log(N) - \log(S)$ relationship in other studies usually assume knowledge of the low flux threshold, chosen high enough to prevent missing sources in the survey. In this section we examine to what sensitivity can we reliably estimate the $\log(N) - \log(S)$ relationship, if we allow missing sources in the survey. Clearly, as $\tau \downarrow 0$, the detection probability decreases rapidly, thus inflating N_{mis} and imputing large numbers of missing sources. In this regime, we expect the posterior inference to reflect the prior and model structures. We remind the reader that the threshold of the detector is different from the flux population threshold, $\tau = S_{min}$. The detection threshold lowers the detection probability if it is greater than τ. For our applications, it is translated into a function of the detector effects described by $\pi_N(\theta, \tau)$. We would

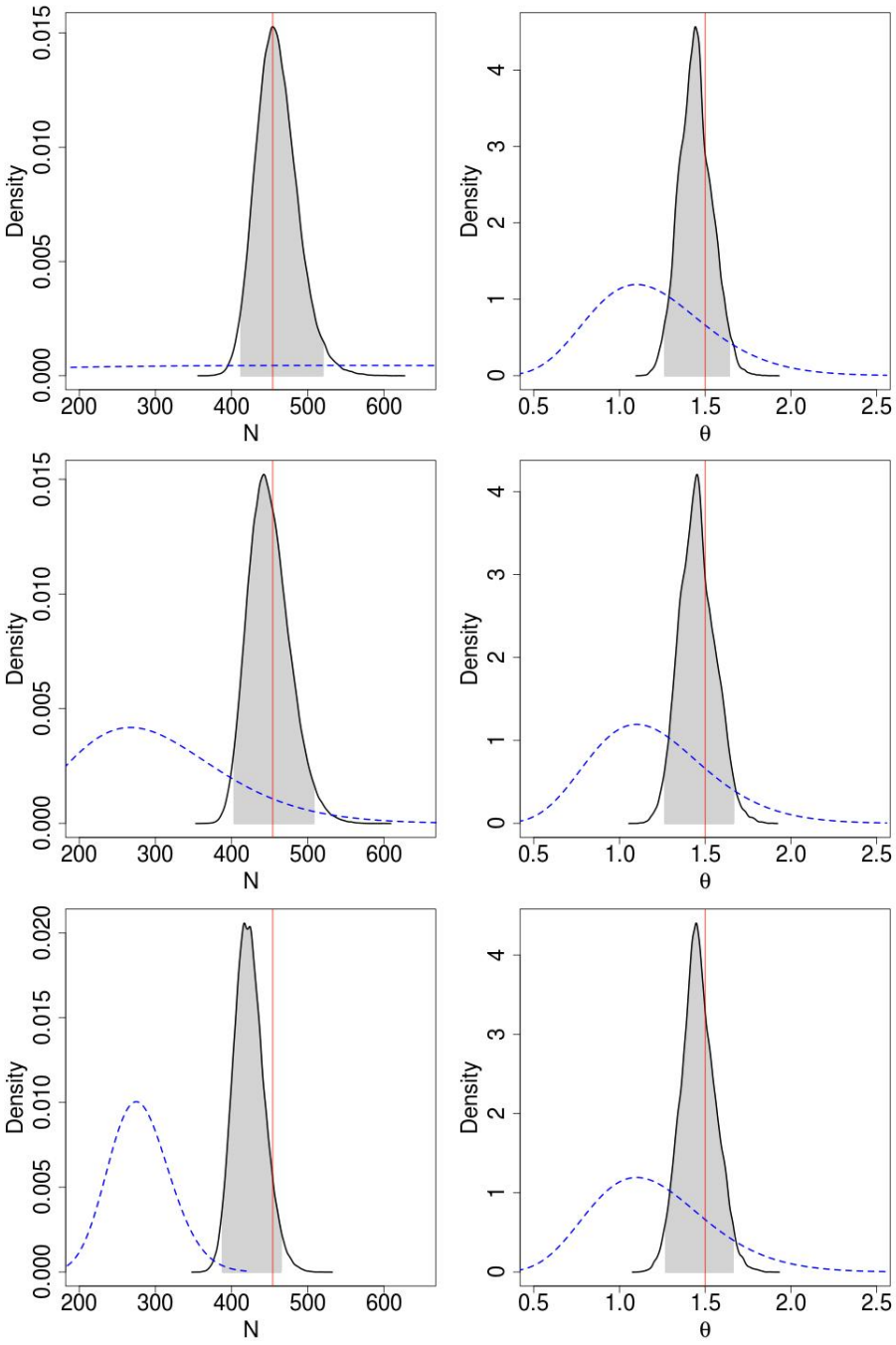

FIGURE 1.5. Prior and posterior distributions for N and θ. From top to bottom (Left column) these represent a weak prior for N, a moderately informative prior for N and a strongly informative but incorrect prior for N. The right column shows the corresponding prior and posterior distributions for θ.

like to investigate the reliability of our inferences as τ (or $\pi_N(\theta, \tau)$) decreases. We consider two aspects:

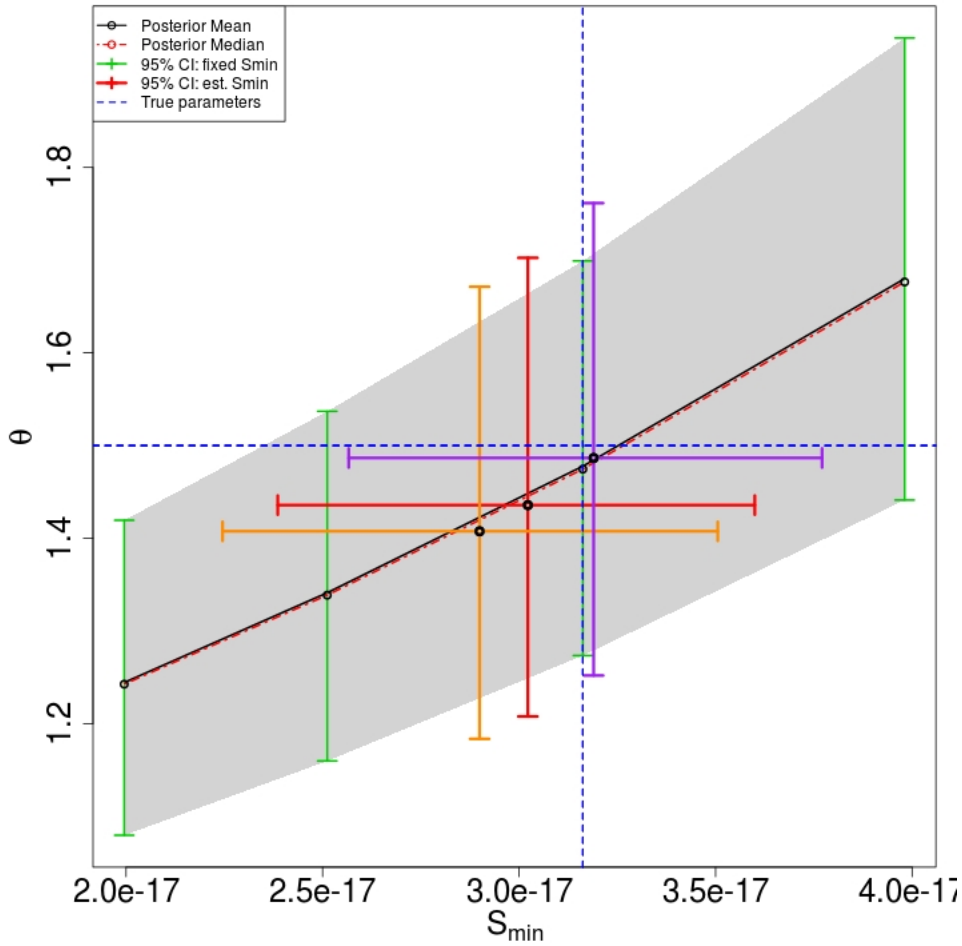

FIGURE 1.6. The grey regions provide the posterior 95% credible intervals for θ at the fixed τ case, with the θ estimate in the center line. The cross intervals show posterior dispersion in both θ and τ for varying priors on τ.

(1) Conditional on the specified model, how sensitive are our estimates (and scientific inferences) to the choice of threshold τ?

(2) If we had mis-specified the model (e.g., Schechter function or broken power-law instead of single power-law), how would our estimates be effected?

The solution to the first of these questions is a dataset-specific, standard sensitivity test. As for the second, it is primarily a method-specific evaluation of inferential robustness, and will be address in the next chapter.

There are two potential approaches to handling the minimum population flux τ: fixing it to a specific value, or, treating it as a regular parameter and obtaining a posterior distribution. Philosophically, the second approach is preferable as it is consistent with treatment of unknown quantities

in most Bayesian analyses. The argument for the former approach is that in small sample settings there can be minimal information in the data to appropriately constrain τ; as a result, estimation for all parameters can be impacted. In practice, the impact of small sample size can be addressed by the estimation of τ combined with a strongly informative prior distribution for τ.

To investigate the performance of our method for these two approaches we consider a larger simulation study with two scenarios. In the first scenario, we treat τ as known but not necessarily modeled with a correctly specified value. We simulate 200 datasets with $\log_{10}(\tau) = -16.5$, $\theta = 1.5$, and $N = 400$. For each of the 200 datasets we fit models with both τ fixed at values from $\log(\tau) = -18$ to $\log(\tau) = -15.5$ in increments of 0.1. For each dataset the posterior medians and 95% credible intervals were computed for θ. The second scenario assumes three gamma prior choices for τ. Broad informative prior uses $E[\tau] = 0.9 \times 10^{-16}, Var[\tau] = (0.8 * 10^{-16})^2$, mis-specified prior with true value in the upper tail uses $E[\tau] = 1.2 * 10^{-17}, Var[\tau] = (1.0 * 10^{-17})^2$, and mis-specified prior with true value in the upper tail uses $E[\tau] = 2.2 * 10^{-16}, Var[\tau] = (1.1 * 10^{-16})^2$. For each of the 200 datasets, we obtain credible intervals for both θ and τ. We then average the credible interval bounds and show the single cross region of coverage.

Figure 1.6 shows the 95% average credible intervals for θ for 200 datasets each fitted at a fixed set of τ values, and the average credible intervals for individual simulations in which both θ and τ were estimated. The horizontal portion of the red cross shows the average 95% posterior interval for τ, easily covering the true value. The vertical portion of the red crosses shows the average 95% credible interval for θ, also containing the true value of 1.5. The coverage of the true parameter coordinate $(10^{-16.5}, 1.5)$ is 96% out of 50 independently repeated crosses of varying weakly informative priors (the results for three of which are shown in the figure in red, orange and purple crosses). The credible intervals for θ for the fixed τ version of the model are shown by the green vertical bands. In most of the cases shown in Figure 1.6 the average posterior interval for θ covered the truth, except for one fixed scenario (green interval), when τ was fixed to $\tau = 10^{-16.7}$.

To obtain a more complete picture of the performance of the model when fixing τ, Figure 1.7 displays the bias, standard deviation and mean square error for estimating θ as well as the average posterior median and credible intervals for each fixed value of $\log(\tau)$. As expected, the bias is seen to be close to zero when τ is selected correctly. If τ is specified to be lower than the true value then estimates for θ have a slight negative bias induced by imputing too much missing data, thus flattening the slope of the $\log(N) - \log(S)$curve. Despite this slight bias the estimates for θ perform

reasonably well in contrast to estimates obtained when τ is fixed above the true value. In this case, a large bias and variance are induced, resulting in a poor estimate. This is straightforward to understand in that our model by definition forces all sources to have a flux above τ, and for those sources whose flux is below the artificially set threshold the model produces a poor fit. This illustrates how, in contrast to other methods for $\log(N) - \log(S)$ fitting, τ specifies the population minimum flux, not a threshold above which incompleteness is guaranteed to be minimal.

In light of the simulation results displayed in Figures 1.6 and 1.7, we recommend estimating the population minimum flux τ from the data unless strong and reliable prior information is available. We find that our proposed model is not especially sensitive to the prior of τ. If the user does prefer to fix τ then a conservative way to specify τ is to set it at lower values, however, one should be aware of the bias on the power-law slope estimate inherent with this specification: the model will often underestimate θ.

1.4.3. Sensitivity to the Incompleteness Function. As noted in section 1.2.1, it is anticipated that inference on some of the key model parameters may be sensitive to the choice of incompleteness function g, an issue we now address. To examine this we now consider fitting multiple, possibly incorrect, incompleteness functions to simulated data. For simplicity we consider four smooth incompleteness functions, shown in Figure 1.8. The true incompleteness function under which the data is generated is shown in the second row, the other rows show incompleteness functions that either systematically overestimate the detection probability (top) or systematically underestimate the detection probability (bottom). The middle and right columns show the prior and posterior distributions and true parameter values for N and θ for each of the four model fits. To isolate the impact of the incompleteness function the same prior distribution was used for each fit and the τ parameter was fixed at $10^{-16.5}$.

Figure 1.8 shows that θ can be estimated reasonably well when the incompleteness function is correct or overestimated (rows 1 and 2), yet it is overestimated when incompleteness is misspecified in the lower direction (rows 3 and 4). In all cases considered, the dispersion of θ stays reasonably constant. The estimate of N has a stronger connection to the specification of the detection probability. Only the correct specification results in the correct estimate of N. Overspecifying the incompleteness results in fewer missing sources and a lowered estimate of N. Underspecifying the incompleteness results in too many missing sources and a dramatic overestimation of N. We

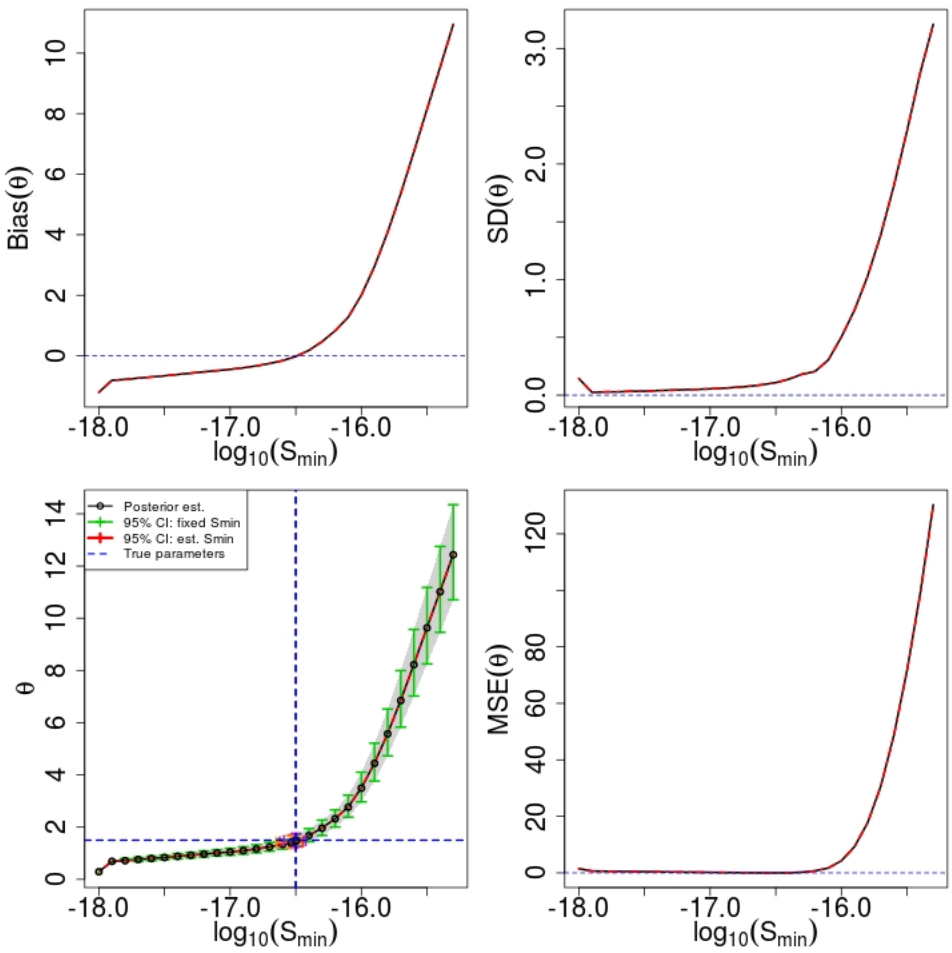

FIGURE 1.7. Sensitivity of τ on estimate of θ. Under the fixed τ scenarios, the plots show: (top-left) bias of θ, (top-right) standard deviation of θ, (mid-left) posterior regions and 95% credible intervals of θ, (mid-right) U-shape nature of MSE of θ.

conclude that it is safer to err on the larger specification of the detection probability, for which estimation of θ is reasonably stable.

1.4.4. Model Checking via Goodness-of-fit. For structured hierarchical models such as our $\log(N) - \log(S)$ model it is necessary to check whether the model assumptions are plausible. Luckily, Bayesian methods lend themselves to self-assessment via posterior predictive model checks (see, Rubin (1984)). The existence of draws from the posterior distribution make construction of posterior predictive model checks particularly easy which is not always possible for other methods of $\log(N) - \log(S)$ estimation (e.g., Schmitt and Maccacaro (1986), Wong et al. (2014)). Hence, the ability to check for the suitability of the model assumptions is an advantage of our proposed framework over the existing methods. We now describe our approach to checking the adequacy of the

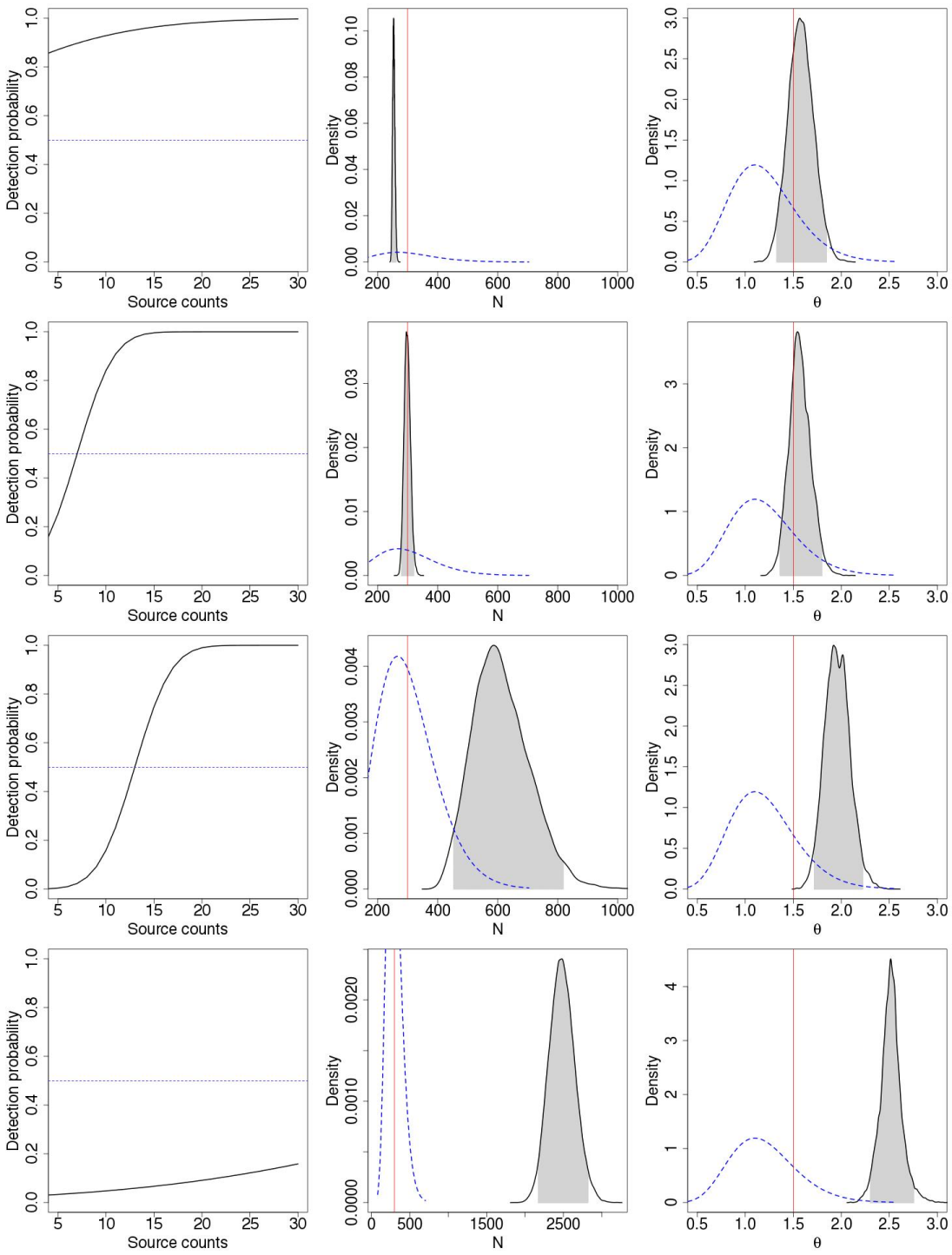

FIGURE 1.8. Left column: Four different incompleteness functions, Middle and Right columns: Corresponding prior and posterior distributions of N and θ. The 2nd row corresponds to the correct incompleteness function.

model. The predictive distribution is the conditional distribution of the new data $\boldsymbol{y}_{new}$ conditional on the observed $\boldsymbol{y}_{obs}$, integrating out all of the uncertainty in combined model parameters β, that

is,

$$p(\boldsymbol{y}_{new}|\boldsymbol{y}_{obs}) = \int p(\boldsymbol{y}_{new}, \beta|\boldsymbol{y}_{obs})d\beta = \int p(\boldsymbol{y}_{new}|\beta)p(\beta|\boldsymbol{y}_{obs})d\beta,$$

where the second identity follows only if the predictive distribution of Y_{new} is independent of $\boldsymbol{y}_{obs}$ given β. The posterior predictive check (PPC) was proposed by Rubin (1984), and expanded by Meng (1994) and Gelman et al. (1996). A thorough description and applications of PPC are described by Gelman et al. (2003). The main idea comes from the expectation that if the model specifications are appropriate, the predictive distribution of the new data would look 'similar' to the empirical distribution of the observed data, assuming that conditional independence holds. It follows that even functions of the data and the parameter derived from either distribution should be 'consistent' under the posterior predictive distribution (Meng, 1994). The degree of consistency would indicate the strength of model mis-fit, that is, failure of the applied model to describe the nature of the data.

We now describe the formal definition and procedure. Consider testing the hypothesis:

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

Based on the MCMC samples of model parameters we generate new datasets from the posterior predictive distribution. We take relevant summary statistics $T(Y)$ of the datasets to perform the test, and define the posterior predictive tail-area to evaluate the fit of a Bayesian model as the posterior predictive $p-$value (PPP) (Meng, 1994):

$$p_b = \Pr(T(\boldsymbol{y}_{new}) \geq T(\boldsymbol{y}_{obs})|\boldsymbol{y}_{obs}, \mathcal{H}_0) = \int \mathbb{I}_{\{T(\boldsymbol{y}_{new}) \geq T(\boldsymbol{y}_{obs})\}} p(\boldsymbol{y}_{new}|\boldsymbol{y}_{obs}, \mathcal{H}_0)d\boldsymbol{y}_{new}. \tag{1.9}$$

Since our interest is in the extrema of the distribution regardless of the direction, we use the corresponding two sided PPP:

$$p_b^* = 2 \cdot \min\{p_b, 1 - p_b\}.$$

Large p_b^* implies no obvious disagreement between the model and the observed data.

The choice of test statistic depends on the model assumptions we want to check, which allows some freedom of selection and dependency on the model parameters. PPC is flexible by avoiding a single global goodness-of-fit summary, which is, perhaps, a benefit. After all, posterior predictive checks have a long history of favorable empirical results and provide useful insights into model fitness for complex models Gelman (2007), Lynch and Western (2004).

Violations of model assumptions in our application are examined in a number of statistics based on a list of photon counts for the observed sources. The broad structure of the missing data mechanism is captured by the sample size of the replicates. The tail behavior of the observation process, the Poisson assumption, and the prior distribution of the detector properties $p(B, L, E)$ are addressed by the minimum and maximum photon count. General model ability to represent the observed data is addressed by the median or IQR of the photon counts. Finally, the power-law assumption of the model (1.1), i.e., the linearity of the $\log(N) - \log(S)$ plot of the flux, is addressed by a coefficient of determination, R^2, based on "crude" estimates of the flux, $S_i = (Y_i - k_i)\gamma/E_i$, (see (1.6) and definition of λ_i). We call it a crude estimate because it is based on the observed photon counts, not the missing. If the Pareto model is reasonable, then the crude $\log(N) - \log(S)$ plot is expected to appear linear and the crude R^2 values are expected to be close to 1.

The correspondence of the posterior predictive distribution of the future summary statistics with the value of observed summary statistic is summarized in p_b^*, but is best summarized graphically. We extend the notion of univariate summary statistics described above to bivariate summaries in order to get additional insight to the correlations between model properties. We plot posterior predictive replicates of two univariate statistics in a scatter plot and examine the relative standing of the coordinate of two observed statistics against the bivariate density of the resulting plot. We define the bivariate posterior predictive p-value as the extreme-tail probability of the resulting bivariate density (also see (1.9)); it is the proportion points below the slice of the bivariate density at the observed coordinate. As before, it represents a measure of surprise when some aspects of the model under $\mathcal{H}_0$ are not represented by the data. To demonstrate, consider bivariate posterior predictive distributions for two statistics: (i) the number of observed sources, and (ii) the median photon count for the observed sources. In this case, let $T(Y_{new}) = (\text{length}_{\{Y_{new}\}}, \text{median}_{\{Y_{new}\}})^T$.

Figure 1.9 shows an example of the consistent fit (left) and the poor fit (right). In the left plot, the correct model is fit to the simulated data and the posterior predictive samples give the number of observed sources consistent with the observed source number indicated by a vertical line. In the right plot, an incorrect model is flagged as a poor fit; since the sizes of the posterior predictive replicates are too small compared to the size of observed data. The figure also shows how small $p-$value indicates that the model cannot capture this aspect of the data based on the value of sample size (length) and median statistics. In the left plot, where the correct model is fit to the simulated data, the posterior predictive samples are consistent with the values of the statistics for

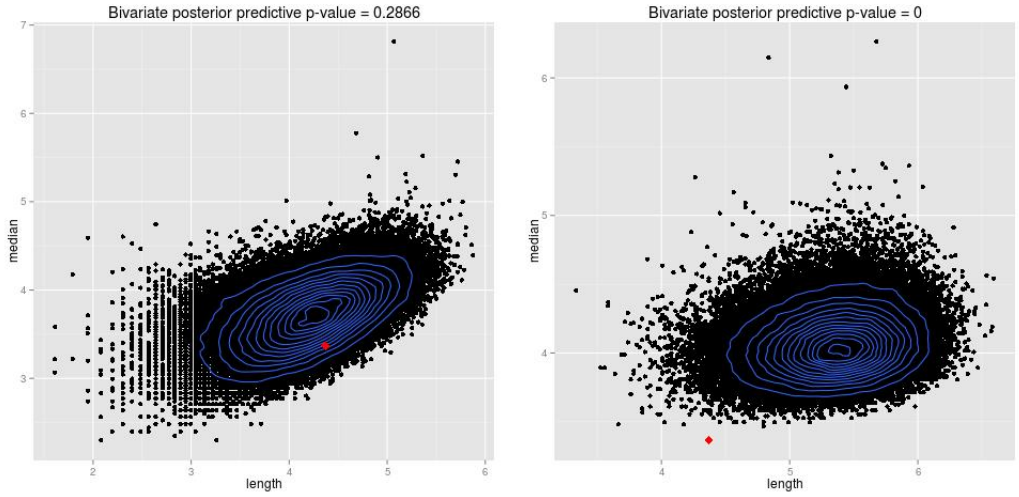

FIGURE 1.9. Bivariate posterior predictive scatter plot for the conditional model: (left) fitted τ equal to truth, (right) fitted τ larger to truth.

the observed data, as displayed by the red dot. In contrast, the right plot shows an example where an incorrect model was fit to the data, in this case τ was fixed to an incorrect value above the true value, and the posterior predictive distribution is not consistent with the values of the statistics for the observed data. The bivariate posterior predictive $p-$value is approximated as the tail area of the extreme region outside the contour of the observed statistic. For the consistent model the bivariate posterior predictive $p-$value is 0.287, for the incorrect model it is very close to zero.

Like all model checking procedures and goodness-of-fit tests, posterior predictive checking has limitations in detecting violations of model assumptions (Bayarri and Berger, 1998). The posterior predictive $p-$value is not a pivotal quantity by construction like the classical $p-$value and it cannot have the same interpretation. Instead it should be considered as an informational summary of the evidence of discrepancies between the model in question and the data. Inconsistency between the observed data and the posterior predictive replicates indicates a lack of fit of the model. However, the absence of evidence for a violation does not guarantee that the model fits the data well. Note that we are not concerned with the Type I error rate or probability of rejecting the null hypothesis when it is actually true. Also, since posterior predictive $p-$values are not calibrated and since this procedure in a sense makes more than one use of the observed data, especially when many checks are performed, the presence of borderline posterior predictive $p-$values is expected and it is not an indictment. We strongly encourage the use of posterior predictive checks and $p-$values to diagnose potential violations of the model assumptions.

1.5. Application to Astronomical Data

1.5.1. Application: *CHANDRA* Deep Field North. We apply our methodology to a sample of 225 X-ray sources from the *CHANDRA* Deep Field North (CDFN) survey. It is the deepest 0.5-8.0 keV survey ever made, and nearly 600 X-ray sources are detected. This survey is comprised of 2 Ms of *CHANDRA* ACIS-I exposure covering 448 sq. arcmin. To preserve the signal in the sources and avoid possible issues with false detections our sample is restricted to sources with an off-axis angle below 8 arcmins. The combined color image of CDFN is shown in figure 1.10.

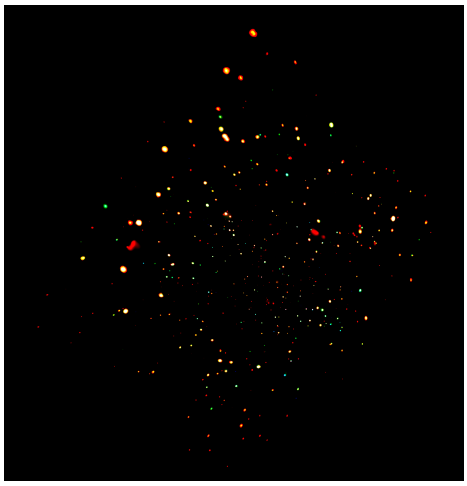

FIGURE 1.10. "True-color" *CHANDRA* image of the whole CDFN

We assume moderately informative priors, elicited from collaborators. $\theta \sim \Gamma(a = 12, b = 10)$, $\tau \sim \Gamma(a_m = 1.78, b_m = 1.48 \cdot 10^{17})$, $N \sim$ Neg-Bin $(a_N = 9.278, b_N = 0.03)$. Incompleteness probability table and detector effects frequency table were directly provided by our collaborators. The detection probability for a given source intensity, background, and off-axis angle is estimated from simulations (Zezas and Fabbiano, 2002).

Using these priors, our Bayesian method yields a posterior median for the power-law slope $\hat{\theta}$ of 0.667, with a 95% credible interval (CI) of $(0.563, 0.780)$. The estimate agrees with the other studies completed (Wong et al., 2014). The population minimum flux is estimated at the posterior median $\hat{\tau}$ of $10^{-16.28}$ with 95% CI $(10^{-16.37}, 10^{-16.20})$. The corresponding population size is estimated with a posterior median $\hat{N}$ of 293 with 95% CI (275, 313), estimating on average 77% completeness of the survey. Table 1.1 summarizes the posterior estimates and central 95% credible intervals for the other key parameters of our model. The posterior draws of the flux for the complete source population give rise to the $\log(N) - \log(S)$ plot shown in Figure 1.11. Each curve corresponds to a

	Mean	SD	2.5%	97.5%
N	293.2	9.79	275.0	313.0
θ	0.6666	0.0559	0.5627	0.780
τ	$10^{-16.2782}$	$10^{-17.2640}$	$10^{-16.3745}$	$10^{-16.1989}$

TABLE 1.1. Posterior estimates of major parameters for the CDFN dataset of section 1.5.1

posterior sample for the fluxes of the complete source population, with missing source flux shown in red and observed source flux in gray. The plot appears to be approximately linear, with no obvious breaks or changes in slope. The width of the $\log(N) - \log(S)$ curve reflects the posterior uncertainty in the flux estimates.

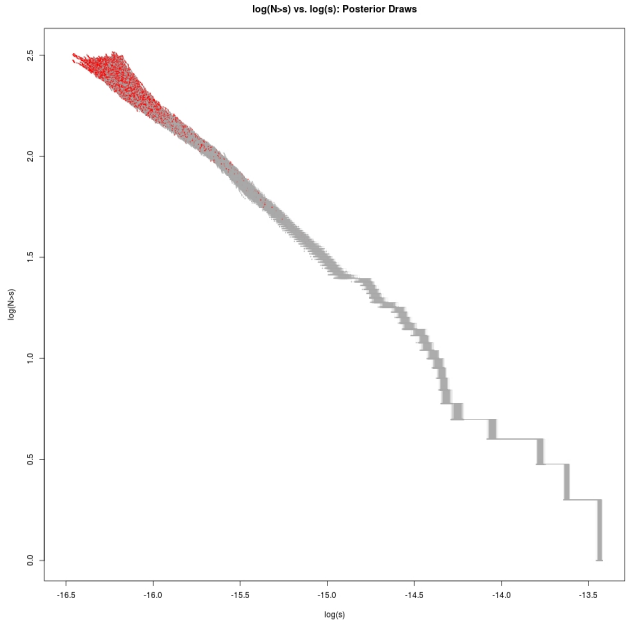

FIGURE 1.11. The $\log(N) - \log(S)$ plot for the CDFN data. Each line in the plot corresponds to a set of fluxes for the complete source population sampled from a single iteration of MCMC scheme with observed sources shown in grey and missing sources in red. The plot shows a sample of 1000 posterior draws.

To examine the adequacy of the model assumptions we use the posterior predictive checks described in section 1.4.4. Posterior predictive $p-$values for a selection of summary statistics are presented in table 1.2. Both the univariate and bivariate posterior predictive p-values are large (> 0.078) for all features we considered, hence no features are flagged as extreme, indicating there is no lack of fit in all aspects of the predictive distribution. Note that this does not rule out the possible presence of breaks or slope variation in the $\log(N) - \log(S)$ curve, since the comparison is performed to the distribution of the photon counts, not fluxes. Also, most values of the crude

R^2 are above 90%, hence linearity of the $\log(N) - \log(S)$ is justified. Selected bivariate plots are given in Figure 1.12 (where we log-transformed all statistics of photon counts for readability). We emphasize that the posterior predictive check has only the potential to indicate problems with the fit and it does not differentiate the models that fit adequately. Our fitted model assumes a linear structure for the $\log(N) - \log(S)$ curve at the population level, while curvilinear counterparts are not considered. An appropriate procedure for selection between candidate Bayesian models is possible; however, we will defer this discussion to be the main topic of the next chapter.

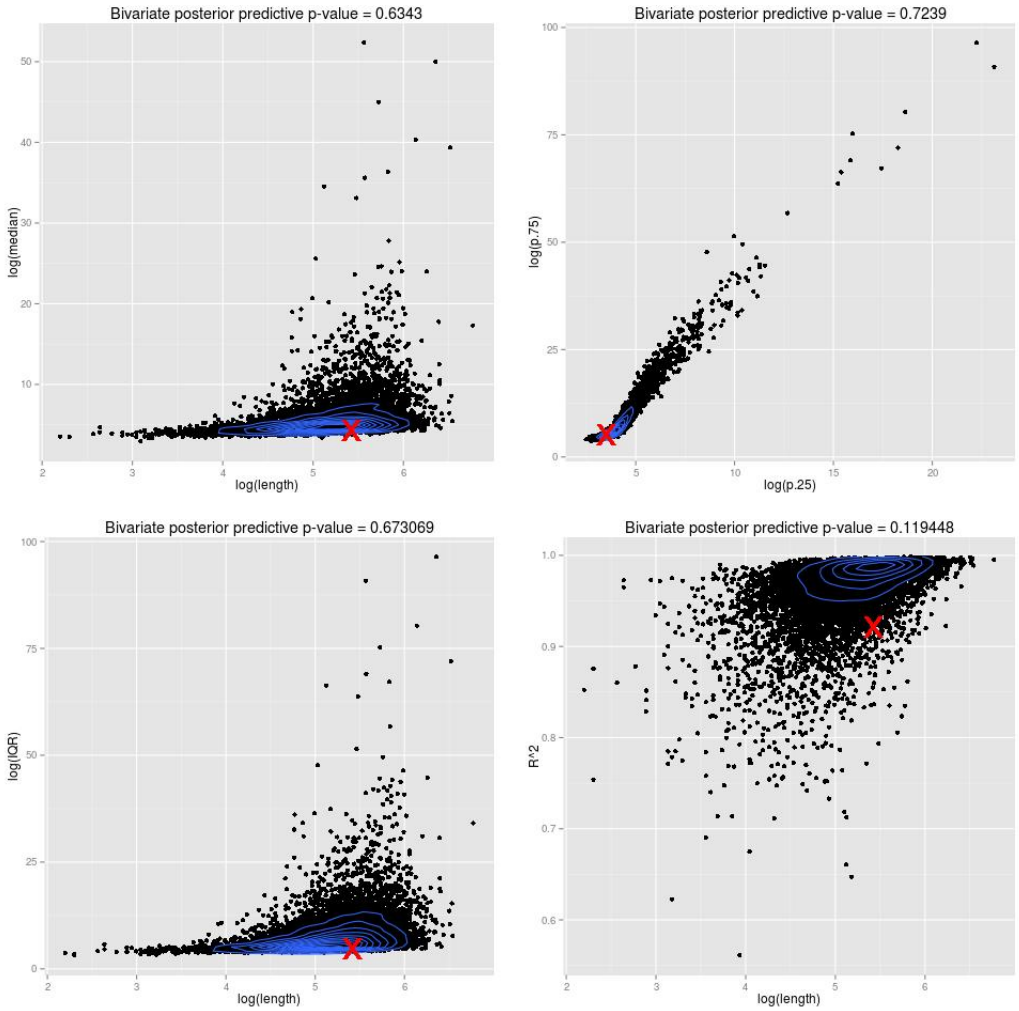

FIGURE 1.12. The bivariate posterior predictive plots show that the single Pareto model fit is fairly adequate. The only unusual feature is the bivariate posterior predictive plot of log(length) vs. crude estimate of R^2 with p-value much below 0.05 level.

We also perform a comparative analysis of the misspecification of incompleteness by ignoring the missing data. The estimate of θ has a posterior median for the power-law slope of 0.626 with 95%

CI $(0.539, 0.720)$. The posterior median estimate of the population minimum flux is $10^{-16.32}$ with 95% CI $(10^{-16.37}, 10^{-16.26})$. Both estimates are not much higher than those in the analysis with correct incompleteness specification. However, ignoring missing information does underestimate the uncertainty in the estimates. This implies that if one was to ignore the effect of missing data, one would be misguided to think that there are fewer sources, and the $\log(N) - \log(S)$ is more steep than it actually is. The $\log(N) - \log(S)$ plot is shown in Figure 1.13. It seems very similar by eye to the original analysis.

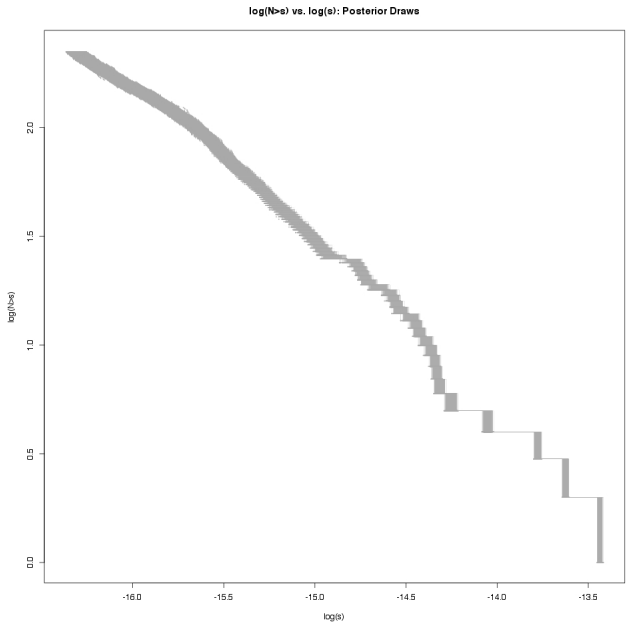

FIGURE 1.13. The $\log(N) - \log(S)$plot for the CDFN data while ignoring missing data.

1.5.2. Application: *CHANDRA* **Deep Field South.** A more recent survey was done on another part of the southern sky: *CHANDRA* Deep Field South (CDFS). This survey was selected for analysis as a more conclusive example of the existence of non-linearity in the $\log(N) - \log(S)$ relationship. CDFS is another deep 0.5-7.0 keV survey covering 0.11 square degrees with over 2000 detected X-ray sources. This survey is comprised of 11 days of *CHANDRA* ACIS-I exposure. We consider a sample of 358 sources. The combined color image of CDFS is shown in figure 1.14.

Our Bayesian method yields the posterior median of power-law slope $\hat{\theta}$ of 0.3367, with a 95% credible interval (CI) of $(0.2863, 0.3946)$. The estimate does not agree with the expectations. The population minimum flux is estimated at the posterior median $\hat{\tau}$ of $10^{-17.30}$ with 95% CI $(10^{-17.81}, 10^{-16.98})$. The corresponding population size is estimated at the posterior median $\hat{N}$ of

Statistic/feature	Posterior predictive p-value
Number of observed sources	0.3389
Minimum photon count	0.2488
Maximum photon count	0.0786
Median photon count	0.1658
Lower quartile of photon counts	0.1015
Upper quantile of photon counts	0.1319
Photon count IQR	0.1404
Crude estimate of R^2	0.0974
Number of observed sources vs. med photon count	0.6343
Lower quartile vs. upper quartile of photon counts	0.7239
Number of observed sources vs. photon count IQR	0.6730
Number of observed sources vs. crude estimate of R^2	0.1194

TABLE 1.2. Univariate and bivariate posterior predictive p-values for assessing the adequacy of the model assumptions for the CDFN dataset.

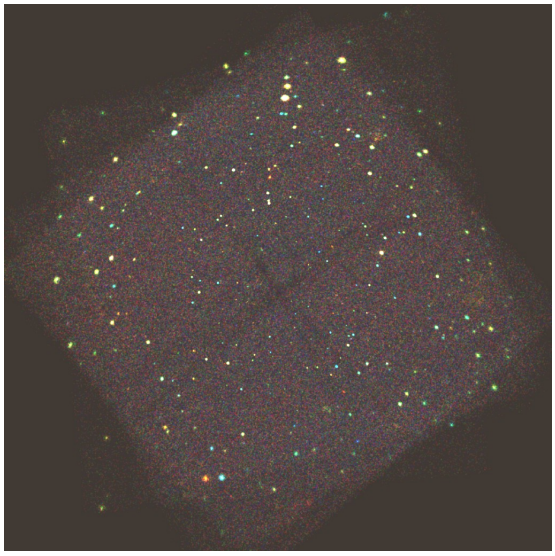

FIGURE 1.14. "True-color" *CHANDRA* image of the whole CDFS

629 with 95% CI (497,907), estimating on average 57% completeness of the survey. Table 1.3 summarizes the posterior estimates and central 95% credible intervals for the other key parameters of our model. The posterior draws of the flux for the complete source population give rise to the $\log(N) - \log(S)$ plot shown in Figure 1.15. The plot clearly does not appear to be linear. The $\log(N) - \log(S)$ curve reflects large posterior uncertainty in the flux estimates.

We examine the adequacy of the model assumptions with posterior predictive checks. Posterior predictive $p-$values for a selection of summary statistics are presented in table 1.4. The posterior predictive check reinforces the lack of fit of the model. The univariate posterior predictive p-values

	Mean	SD	2.5%	97.5%
N	648.0	104.1	497.0	907.0
θ	0.3378	0.0278	0.2863	0.39460
τ	$10^{-17.2832}$	$10^{-17.6531}$	$10^{-17.8108}$	$10^{-16.9817}$

TABLE 1.3. Posterior estimates of major parameters for the CDFS dataset of section 2.7.2

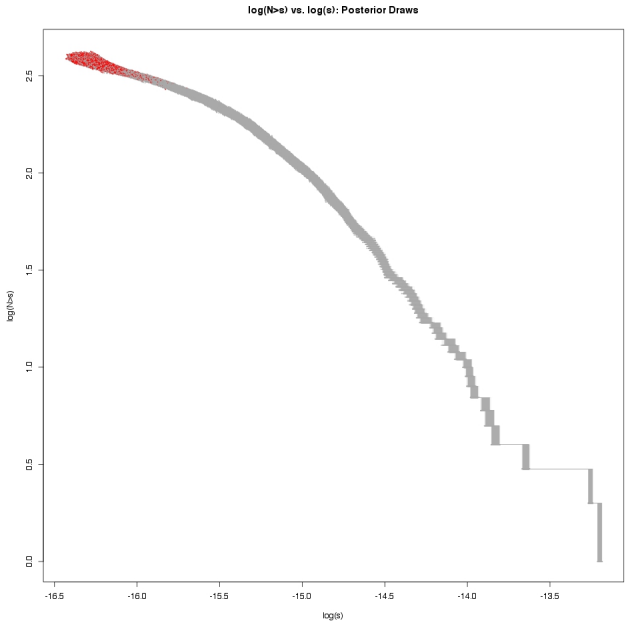

FIGURE 1.15. The $\log(N) - \log(S)$plot for the CDFS data.

are small (> 0.051) for the maximum photon count and the crude estimate of R^2, indicating there is a lack of fit. The bivariate p-value is 0.0015 for length of photon sample vs crude estimate of R^2 of the flux. Also, most values of the crude R^2 are above 95% for predicted datasets and only 87% for the observed dataset, hence the linearity of the $\log(N) - \log(S)$ is only due to the model assumption but the model fit is poor. Selected PPC plots given in Figure 1.16 demonstrate lack of model fit. We conclude that the simple Pareto model does not provide an appropriate fit to these data. We will examine other models which allow for non-linearity in the $\log(N) - \log(S)$ in the next chapter.

1.6. Discussion and Concluding Remarks

We have presented a comprehensive method for estimation of the $\log(N) - \log(S)$ relationship using a hierarchical Bayesian model. The strengths of the model are many. First, it allows a comprehensive study of the incompleteness of surveys by correctly accounting for missing data and

Statistic	Posterior predictive p-value
Number of observed sources	0.1700
Minimum photon count	0.2375
Maximum photon count	0.0500
Median photon count	0.3088
Lower quartile of photon counts	0.1806
Upper quantile of photon counts	0.2306
Photon count IQR	0.2371
Crude estimate of R^2	0.0069
Number of observed sources vs. max photon count	0.3551
Lower quartile vs. upper quartile of photon counts	0.9590
Number of observed sources vs. photon count IQR	0.6867
Number of observed sources vs. crude estimate of R^2	0.0015

TABLE 1.4. Univariate and bivariate posterior predictive p-values for assessing the adequacy of the model assumptions for the CDFS dataset.

bias from detector effects. Second, as a by-product, it provides an easy way of imputing missing information, such as estimates of the flux for observed and missing sources. Third, the method is built on a strong probabilistic foundation that has a support from physical observations. Fourth, it allows goodness-of-fit diagnostic checks. Fifth, our method works reasonably fast.

One must keep in mind that our model depends heavily on the specification of incompleteness curve. The sensitivity studies have shown that misspecification of the priors of unknown parameters is not as crucial for biased inference as misspecification of the incompleteness function. Hence, most efforts must be put to obtain valid incompleteness curves. On the other hand, this is to be expected, just as calibrating the detector is a requirement for unbiased inference in astrophysical surveys.

Other potential limitations of the current model are that it allows only straight-line relationships for the $\log(N) - \log(S)$ curve. The real question of interest is whether the power-law is sufficient or perhaps curvature in the $\log(N) - \log(S)$ plot really exists and is not due to the incompleteness of the survey. The next chapter deals with possible extensions to the model to allow flexible forms of the curve yet still preserving the correspondence between standard probabilistic assumptions and physically motivated models.

By modeling the $\log(N) - \log(S)$ relationship within a hierarchical Bayesian framework we achieve flexibility in describing the properties of both the source population and the detector induced uncertainties. Our method explicitly corrects for the non-ignorable missing data mechanism that is often ignored by competing methods.

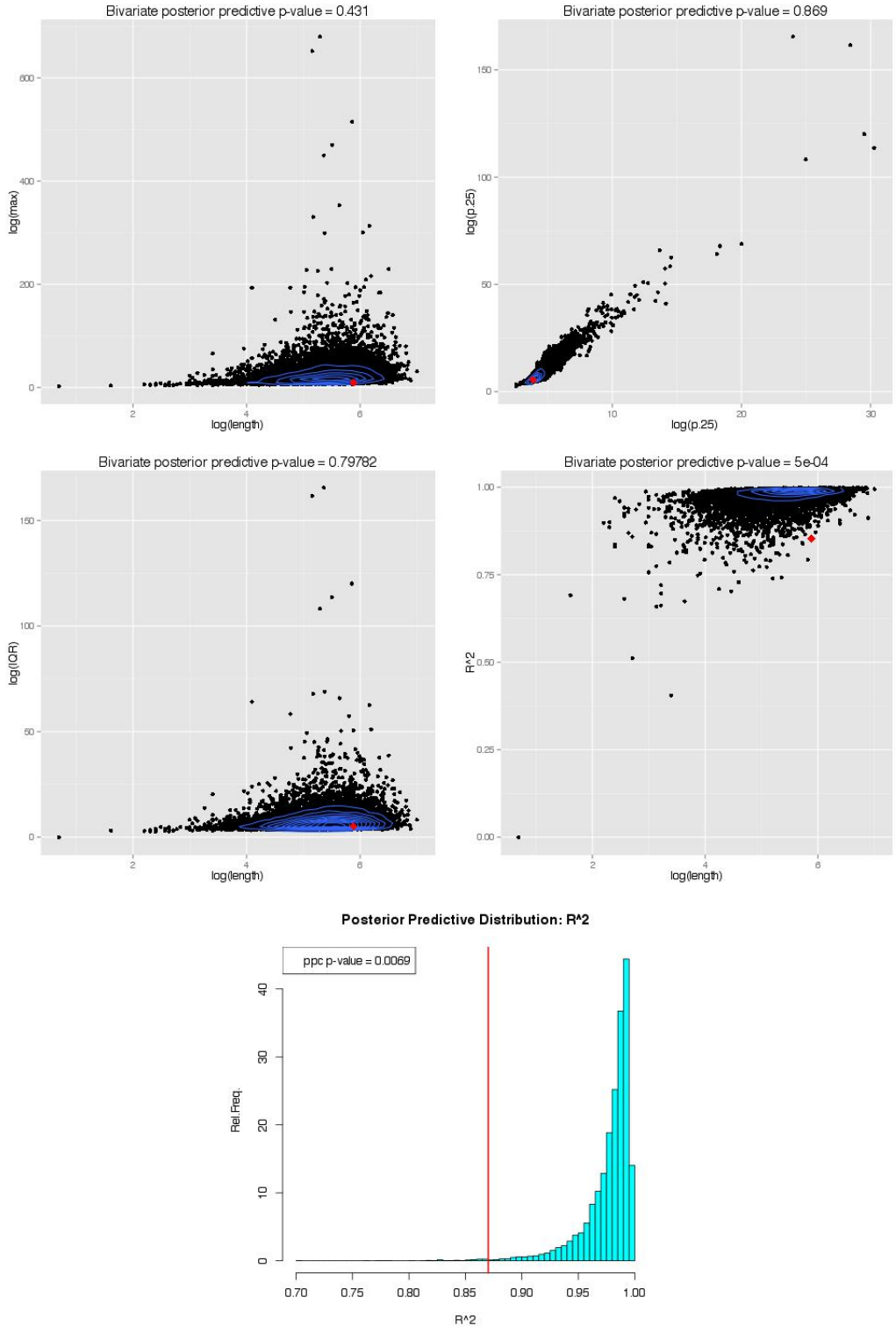

FIGURE 1.16. The posterior predictive plots show that the single Pareto model fit of CDFS dataset is not adequate. The posterior predictive p-value is around 0 for the univariate PPC (histogram plot).

CHAPTER 2

Analysis of the $\log(N) - \log(S)$ Problem Modeled via a Broken-Pareto Population –

«We all agree that your theory is crazy,
but is it crazy enough?»

NIELS BOHR
— Jewish Danish physicist

2.1. Introduction

When observing very faint sources, it is common for $\log(N) - \log(S)$ relationship to appear either curved or piece-wise linear, rather than linear. It is of interest to discern whether the curvature is due to the missing data near the detection boundary, or whether the curvature and/or non-linearity is a real property of $\log(N) - \log(S)$ for faint source populations. The single power-law model, described in the previous chapter, does not allow sufficient flexibility to answer these questions. In extending the single power-law model we must be careful to retain the astrophysical interpretation of the model, as well as the requirement that the $\log(N) - \log(S)$ shape corresponds to a valid population distribution of the flux. Note that this requirement restricts many available curvature models (e.g., a general polynomial). We choose to model the $\log(N) - \log(S)$ relationship as a combination of multiple power-laws, connected at the knots. This model is typically referred to as a "broken" power-law in the astrophysical literature, for example, see Zezas and Fabbiano (2002), Jóhannesson et al. (2006), Kim et al. (2007), and a recent paper by Wong et al. (2014). In section 2.2.1 we show that this model is equivalent to a mixture of truncated Pareto distributions for the flux. Other model options are available as general mixtures of Pareto distributions, however, these models may not provide desired shapes of $\log(N) - \log(S)$ plot. In section 2.3 we provide the details of parameter estimation in the broken-Pareto model. In order to select between various competing models, we review existing methods for model selection in section 2.4 and implement a novel Bayesian model selection procedure to decide between a single- and broken-Pareto model in section 2.6.

2.2. Broken Power-Law Models

The estimation of piece-wise linear $\log(N) - \log(S)$ relationship of X-Ray sources is common in the astrophysical literature. For example, Mateos et al. (2008) estimate two and three piece linear $\log(N) - \log(S)$ relationship of X-ray sources collected over many bands of energy. The source fluxes here assume to come from a mixture of flux populations. Mateos et al. (2008) use the maximum likelihood to estimate parameters of the flux populations and perform chi-square goodness-of-fit to approve of their model. Wong et al. (2014) also considers estimation of the piece-wise linear $\log(N) - \log(S)$ relationship for X-ray sources. Their method uses interwoven EM algorithm to estimate the power-law slopes, the breakpoints, and the number of breakpoints in the resulting mixture population of the flux. Both of these methods, however, cannot account for missing data and require to limit their survey at a minimum flux. Ignoring the missing data may potentially result in a biased estimation of parameters and narrow confidence intervals. In this section we propose a coherent approach to estimating the piece-wise linear $\log(N) - \log(S)$ relationship, while accounting for the non-ignorable missing data process and detector-induced effects. We build the Bayesian hierarchical model from the first principle assumptions to the flux distribution and follow up with the algorithm for parameter inference. Our method, similar to Mateos et al. (2008), assumes a known number of pieces in the $\log(N) - \log(S)$ relationship. In order to select appropriate number of pieces we examine methods for model selection in sections 2.4 through 2.6. We apply our method to the CDFN and CDFS *CHANDRA* surveys in section 2.7. The results of the CDFN analysis is compared to those provided by Wong et al. (2014).

2.2.1. Single Broken Power-Law Model. To generalize the basic power-law model for the $\log(N) - \log(S)$ distribution we have several options. First, we note that, under independent sampling the linear $\log(N) - \log(S)$ plot corresponds to a Pareto distribution for the complete-data fluxes. More general shapes for the population $\log(N) - \log(S)$ curve will correspond to different complete-data flux distributions. Recall the duality of the power-law model: $\log(N) - \log(S)$ is linear if and only if the flux distribution is a Pareto distribution. Formally, let the complete-data flux distribution G have c.d.f. F_G, and suppose $S_i \overset{iid}{\sim} G$. Define

$$\log_{10}\left(1 - F_G(s)\right) := H\left(\log_{10}(s)\right).$$

Then the function H is linear if and only if G is the Pareto distribution. Since the linearity of $\log(N) - \log(S)$ has both theoretical and empirical support, the simplest and most commonly used generalization is a broken power-law:

$$\log_{10}\left(1 - F_G(s)\right) = \begin{cases} \alpha_1 - \theta_1 \log_{10}(s), & \tau_1 \leq s < \tau_2 \\ \alpha_2 - \theta_2 \log_{10}(s), & s \geq \tau_2 \end{cases}, \tag{2.1}$$

subject to the continuity constraint that $(\alpha_1 - \alpha_2) = (\theta_1 - \theta_2)\log(\tau_2)$. Here, we define the Pareto minimum τ_1 and the break point as τ_2. It is natural to ask if the broken power-law in (2.1) corresponds to a known distribution. The answer, as may be expected, is 'yes': a mixture of (truncated) Pareto distributions. Similarly to the power-law setting, the result is also 'if and only if' result.

LEMMA 2. Any distribution whose $\log(N) - \log(S)$ plot is a broken power-law can be represented as a mixture of a truncated Pareto distribution and an (untruncated) Pareto distribution.

That is, we have:

$$Y \sim \left[1 - \left(\frac{\tau_2}{\tau_1}\right)^{-\theta_1}\right] X_1 + \left(\frac{\tau_2}{\tau_1}\right)^{-\theta_1} X_2, \tag{2.2}$$

where: $X_1 \sim$ Truncated-Pareto $(\tau_1, \theta_1, \tau_2)$ with CDF given by

$$F_1(s) = \frac{1 - \left(\frac{s}{\tau_1}\right)^{-\theta_1}}{1 - \left(\frac{\tau_2}{\tau_1}\right)^{-\theta_1}}, \qquad \tau_1 \leq s < \tau_2 \tag{2.3}$$

and $X_2 \sim$ Pareto (τ_2, θ_2). The proof of Lemma 2 is found in Appendix B. It is important to note that the continuity constraint restricts the distribution of Y to contain only 4 free parameters instead of 5 (two for each straight line and the break-point location).

The broken-Pareto CDF can be explicitly shown to be:

$$F_G(s) = \begin{cases} 1 - \left(\frac{s}{\tau_1}\right)^{-\theta_1}, & \tau_1 \leq s < \tau_2 \\ 1 - \left(\frac{\tau_2}{\tau_1}\right)^{-\theta_1}\left(\frac{s}{\tau_2}\right)^{-\theta_2}, & s \geq \tau_2 \end{cases}$$

with the density:

$$f_G(s) = \begin{cases} \frac{\theta_1}{\tau_1}\left(\frac{s}{\tau_1}\right)^{-(\theta_1+1)}, & \tau_1 \le s < \tau_2 \\ \frac{\theta_2}{\tau_2}\left(\frac{\tau_2}{\tau_1}\right)^{-\theta_2}\left(\frac{s}{\tau_2}\right)^{-(\theta_2+1)}, & s \ge \tau_2. \end{cases} \tag{2.4}$$

We can generalize the broken power-law idea further to any number of truncated Pareto mixtures. However, first we must mention another consequence of such model: it is impossible for an unconstrained mixture of two (untruncated) Pareto distributions to have a broken power-law in the plot of $\log(N) - \log(S)$. Suppose that $X_1 \sim \text{Pareto}(\theta_1, \tau_1)$, $X_2 \sim \text{Pareto}(\theta_2, \tau_1)$, $Y = pX_1 + (1-p)X_2$, for some $p \in [0,1]$, and let the CDF of X_1, X_2, Y be $F_1(s) = 1 - e^{\alpha_1}s^{\theta_1}, F_2(s) = 1 - e^{\alpha_2}s^{\theta_2}$, and $F_Y(s) = pF_1(s) + (1-p)F_2(s)$, respectively.

Assume that there exists a broken power-law relationship such that the $\log(N) - \log(S)$ plot is made up of two connected straight lines with slopes θ_1 and θ_2 at some connection point $s = B$. Then the $\log(N) - \log(S)$ plot will be described by this curve:

$$Q(s) = \alpha_1 - \theta_1 \log_{10}(s) + \mathbb{I}\{s \ge B\}\left[(\alpha_2 - \alpha_1) - (\theta_2 - \theta_1)\log_{10}(s)\right]$$

We expect this curve to be derived from logarithm of the CDF of the Pareto mixture:

$$\begin{aligned} \log_{10}(1 - F_Y(s)) &= \log_{10}\{1 - [pF_1(s) + (1-p)F_2(s)]\} \\ &= \log_{10}\left[1 - p\left(1 - e^{\alpha_1}s^{-\theta_1}\right) - (1-p)\left(1 - e^{\alpha_2}s^{-\theta_2}\right)\right] \\ &= \log_{10}\left(pe^{\alpha_1}s^{-\theta_1} + (1-p)e^{\alpha_2}s^{-\theta_2}\right) \end{aligned}$$

It is obvious that $Q(s) \neq \log_{10}(1 - F_Y(s))$ for any $p \in (0,1)$, since it is impossible to distribute the logarithm inside the parentheses. Hence, the two Pareto populations in the mixture do not overlap if the power-law is piece-wise linear.

2.2.2. Multiple Broken Power-Law Model. The broken power-law model of section 2.2.1 can be further generalized to a piece-wise linear $\log(N) - \log(S)$ relationship. In other words, we can allow for an arbitrary number m mixture pieces or, equivalently, $m-1$ break-points. The setting is similar to the single broken power-law model, with the analogous probabilistic model being a mixture of truncated Pareto distributions and a single untruncated Pareto distribution. Let $\tau_2, \ldots, \tau_m$ denote the locations of the breakpoints, τ_1 denote the minimum flux, and $\tau_{m+1} = \infty$. If

we assume a piece-wise linear theoretical $\log(N) - \log(S)$ relationship then:

$$F_G(s) = \begin{cases} 1 - \alpha_1^* s^{-\theta_1}, & \tau_1 \leq s < \tau_2 \\ 1 - \alpha_2^* s^{-\theta_2}, & \tau_2 \leq s < \tau_3 \\ \vdots & \vdots \\ 1 - \alpha_{m-1}^* s^{-\theta_{m-1}}, & \tau_{m-1} \leq s < \tau_m \\ 1 - \alpha_m^* s^{-\theta_m}, & s \geq \tau_m \end{cases}$$

In a similar manner to the single broken power-law setting we can show that:

$$F_j(s) = \frac{1}{p_j} \left\{ 1 - \sum_{i=1}^{j-1} p_i \right\} \left[1 - \left(\frac{s}{\tau_j} \right)^{-\theta_j} \right], \qquad j = 1, \ldots, m. \tag{2.5}$$

Constraints on the CDF lead to a recursive relationship among the mixture probabilities:

$$p_j = \left[1 - \left(\frac{\tau_{j+1}}{\tau_j} \right)^{-\theta_j} \right] \left(1 - \sum_{i=1}^{j-1} p_i \right), \qquad j = 1, \ldots, m, \tag{2.6}$$

where $\left(\frac{\tau_{j+1}}{\tau_j} \right)^{-\theta_j}$ is understood as 0 when $j = m$. Plugging into (2.5) we obtain:

$$F_j(s) = \frac{1 - \left(\frac{s}{\tau_j} \right)^{-\theta_j}}{1 - \left(\frac{\tau_{j+1}}{\tau_j} \right)^{-\theta_j}} \mathbb{I}_{\{\tau_j \leq s < \tau_{j+1}\}}, \qquad j = 1, \ldots, m.$$

In other words, the multiple broken power-law assumption corresponds to the following probabilistic model:

$$Y \sim I_1 X_1 + I_2 X_2 + \cdots + I_m X_m$$

where:

$$I_j \sim \text{Multinomial}\left(1, p_1, p_2, \ldots, p_m\right),$$

$$X_j \sim \text{Truncated-Pareto}\left(\tau_j, \theta_j, \tau_{j+1}\right), \qquad j = 1, \ldots, m.$$

and $p_1, \ldots, p_m$ are defined by th following:

$$p_j = \left[1 - \left(\frac{\tau_{j+1}}{\tau_j} \right)^{-\theta_j} \right] \prod_{i=1}^{j-1} \left(\frac{\tau_{i+1}}{\tau_i} \right)^{-\theta_i}, \tag{2.7}$$

which gives rise to the following identity:

$$1 - \sum_{i=1}^{j} p_i = \prod_{i=1}^{j} \left(\frac{\tau_{i+1}}{\tau_i} \right)^{-\theta_i} \tag{2.8}$$

(see Appendix B). Note that the final Pareto distribution is actually untruncated since $\tau_{m+1} = \infty$. As expected, the continuity constraints leave only $2m$ free parameters: m slopes for the line segments and $m - 1$ breakpoints, plus the minimum point τ_1.

2.3. Posterior Inference for Multiple Broken Power-Law Model

The previous section described the distribution of the fluxes only. We now describe the full Bayesian hierarchical model including modeling the photon counts and the missing data mechanism for general multiple broken power-law.

2.3.1. Building the Posterior Distribution. Construction of the model for $\log(N) - \log(S)$ based on multiple broken power-law follows a similar structure to that of the single power-law model. Again, let N be the (unknown) total number of sources in the complete source population, with n and N_{mis} the number of observed and missing sources respectively, so that $N = n + N_{mis}$. We assume a Negative-Binomial prior distribution for the total number of sources in the population with flux above a given threshold, τ_1 i.e., $N \sim$ Neg-Bin (a_N, b_N). Conditional on the total number of sources and model parameters, we assume that source fluxes for the complete source population follow a broken power-law, i.e., a mixture of truncated Pareto distributions. We assume that the number of mixture populations m is known. The model parameters are the m power-law slopes, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$, the flux population minimum threshold τ_1, and consequent breakpoints $\tau_2, \ldots, \tau_m$. We note again that the value of τ_1 is the flux population minimum threshold, which is not the same as a flux detector threshold. We assume a conditionally conjugate Gamma prior distribution for θ and τ_1 i.e., $\theta_j \sim$ Gamma $(a_j, b_j)\,, j = 1, ..., m$ and $\tau_1 \sim$ Gamma (a_τ, b_τ).

In order to build the broken power-law model assumption into our Bayesian hierarchical model, we only need to modify two distributional assumptions from the single Pareto model scenario. First, the model for the flux $S = (S_1, \ldots, S_N)$ is changed to a broken-Pareto. Next, we need additional prior distributions for the breakpoints $\tau = (\tau_1, \ldots, \tau_m)$. It can be shown (see Appendix B) that the

general m-component broken power-law density of the flux is expressed as:

$$f_Y(s) = \sum_{j=1}^{m} \left\{ \prod_{i=1}^{j-1} \left(\frac{\tau_{i+1}}{\tau_i} \right)^{-\theta_i} \right\} \left(\frac{\theta_j}{\tau_j} \right) \left(\frac{s}{\tau_j} \right)^{-(\theta_j+1)} \mathbb{I}_{\{\tau_j \le s < \tau_{j+1}\}}, \tag{2.9}$$

where we define $\prod_{i=1}^{0} \left(\frac{\tau_{i+1}}{\tau_i} \right)^{-\theta_i} = 1$. The choice of prior for $\tau_1, \ldots, \tau_m$ is tricky because of the complicated support of this multivariate distribution: $0 < \tau_1 < \tau_2 < \cdots < \tau_m$. The dependence among variables τ_j and further dependence of the latent flux variables S_i: $\tau_j \le S_i < \tau_{j+1}$ can greatly reduce the efficiency of the Gibbs sampler of these parameters. We examined a number of various sampling strategies of τ_j and found that the following procedure works well. We propose to split the Gibbs sampler into two blocks: $[\tau_1]$ and $[\tau_2, \ldots, \tau_m]$ and utilize a transformation of variables in the latter block to another space, where the transformed variables are unconstrained and independent. We let $\tau_1 \sim \text{Gamma}(a_\tau, b_\tau)$. We further define a log transformation on τ_j as $\eta_j = h_j(\tau_j | \tau_{j-1}) = \log(\tau_j - \tau_{j-1})$ for $j = 2, \ldots, m$. In this situation $\tilde{\tau} = (\tau_2, \ldots, \tau_m)^T$ can be expressed as:

$$\tilde{\tau} = h^{-1}(\eta | \tau_1) = \begin{pmatrix} \tau_1 + e^{\eta_2} \\ \tau_1 + e^{\eta_2} + e^{\eta_3} \\ \vdots \\ \tau_1 + \sum_{j=2}^{m} e^{\eta_j} \end{pmatrix} \tag{2.10}$$

where we assume $\eta = (\eta_2, \ldots, \eta_m)^T \overset{indep}{\sim} \text{Multivariate-Normal}(\mu, C)$ with $\mu = (\mu_2, \ldots, \mu_m)^T$ and $C = \text{diag}\{c_2^{-1}, \ldots, c_m^{-1}\}$. Let $\tilde{\tau} = (\tau_2, \ldots, \tau_m)^T$. This transformation preserves non-negativity and increasing order of τ_j's, hence the sampling of the breakpoints $\tilde{\tau}$ can be efficiently performed on the space of η.

Using (2.9) we can now derive the posterior distribution and describe the sampling strategies of all unknown parameters, where we marginalize the complete data posterior across all the missing source information. Assume the total number of broken-Pareto pieces, m is known in advance. Let

$\theta = (\theta_1, \ldots, \theta_m)^T$ and $\tau = (\tau_1, \ldots, \tau_m)^T$. The posterior distribution of the break-point model is:

$$
\begin{aligned}
(2.11)\quad & p(N, \theta, \tau, S_{obs}, Y_{obs}^{src} | n, Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs}) \\
& \propto \left[\binom{N}{n} \mathbb{I}_{\{n \le N\}} \cdot (1 - \pi(\theta, \tau))^{(N-n)} \right] \cdot \left[\prod_{j=1}^{m} \frac{b_j^{a_j}}{\Gamma(a_j)} \theta_j^{a_j - 1} e^{-b_j \theta_j} \mathbb{I}_{\{\theta_j > 0\}} \right] \\
& \quad \cdot \left[\binom{N + a_N - 1}{a_N - 1} \left(\frac{1}{1 + b_N} \right)^N \left(\frac{b_N}{1 + b_N} \right)^{a_N} \mathbb{I}_{\{N \in \mathbb{Z}^+\}} \right] \\
& \quad \cdot p(\tau_1, \ldots, \tau_m) \, \mathbb{I}_{\{0 < \tau_1 < \tau_1 < \cdots < \tau_m\}} \cdot \left[\prod_{i=1}^{n} p(B_i, L_i, E_i) \cdot g(S_i, B_i, L_i, E_i) \right. \\
& \quad \cdot \sum_{j=1}^{m} \left\{ \prod_{l=1}^{j-1} \left(\frac{\tau_{l+1}}{\tau_l} \right)^{-\theta_l} \right\} \left(\frac{\theta_j}{\tau_j} \right) \left(\frac{S_i}{\tau_j} \right)^{-(\theta_j + 1)} \mathbb{I}_{\{\tau_j \le S_i < \tau_{j+1}\}} \cdot \frac{(\lambda_i + k_i)^{Y_i^{tot}}}{Y_i^{tot}!} e^{(\lambda_i + k_i)} \mathbb{I}_{\{Y_i^{tot} \in \mathbb{Z}^+\}} \\
& \quad \left. \cdot \binom{Y_i^{tot}}{Y_i^{src}} \left(\frac{\lambda_i}{\lambda_i + k_i} \right)^{Y_i^{src}} \left(1 - \frac{\lambda_i}{\lambda_i + k_i} \right)^{Y_i^{tot} - Y_i^{src}} \mathbb{I}_{\left\{Y_i^{src} \in \{0, 1, \ldots, Y_i^{tot}\}\right\}} \right],
\end{aligned}
$$

with $\tau_{m+1} = +\infty$, $\lambda_i \equiv \lambda(S_i, B_i, L_i, E_i)$, $k_i \equiv k(B_i, L_i, E_i)$, $\prod_{i=1}^{0} \left(\frac{\tau_{i+1}}{\tau_i} \right)^{-\theta_i} \equiv 1$, and $\pi(\theta, \tau) = \int g(S, B, L, E) \cdot p(S, B, L, E | \theta, \tau) \; dS \; dB \; dE \; dL$, the marginal probability of detecting a source. Full derivation of this posterior is given in Appendix B.

We note that the form of this distribution is not much more complex than the single-Pareto posterior distribution of previous chapter. The main differences are in the prior distribution of the break-points and the broken-Pareto likelihood function (2.9). Sampling from the posterior distribution via MCMC is described in the next section.

2.3.2. Sampling of Parameters. The posterior distribution (2.11) allows us to compute full conditional distributions of all unknown parameters and describe their sampling techniques. In this section we describe the sampling methods for the parameters $\theta = (\theta_1, \ldots, \theta_m)^T$, $\tilde{\tau} = (\tau_2, \ldots, \tau_m)^T$, and S_{obs}. Note that the sampling methods for parameters $N, Y_{obs}^{tot}, Y_{obs}^{src}$ for this hierarchical broken-Pareto model are very similar to that of the single-Pareto model, so we omit further discussion of this topic. The details of derivation of the full conditional distributions can be found in the Appendix B.

Sampling $\theta = (\theta_1, \ldots, \theta_m)^T$*:* We have

$$p(\theta|\,\cdot\,) \propto \left[(1-\pi(\theta,\tau))^{N-n}\right] \cdot \cdot \prod_{j=1}^{m} \text{Gamma}\left(\theta_j;\ a_j + n(j) - 1,\ b_j + \mathbb{I}_{\{j\neq m\}} \log\left(\frac{\tau_{j+1}}{\tau_j}\right) \sum_{i=1}^{m} \left[n(i)\mathbb{I}_{\{i\geq j+1\}}\right] + \sum_{i\in\mathcal{I}(j)} \log\left(\frac{s_i}{\tau_j}\right)\right),$$

where $\mathcal{I}(j) = \{i : \tau_j \leq s_i < \tau_{j+1}\}$ and $n(j)$ is the cardinality of $\mathcal{I}(j)$ i.e., $\mathcal{I}(j)$ $(n(j))$ denotes the set (number) of source indices whose flux is contained in the interval corresponding to the j-th mixture component.

All terms apart from those involving $\pi(\theta,\tau)$ factorize in terms of $\theta_1, \ldots, \theta_m$. This partial factorization allows for the exact (conditional) posterior draws to be obtained by rejection sampling (see Appendix B). The rejection sampling procedure is beneficial because it is guaranteed to produce conditionally independent posterior draws. However, in practice the acceptance rate of this rejection sampler can be very low. For this reason we choose to use the Metropolis-Hastings algorithm to obtain approximate (conditional) posterior draws of θ. We use normal distribution proposals with the variance tuned to insure acceptance for MH between 20%-60%. The sampling via MH procedure is selected at random with success probability 0.9, otherwise, the rejection sampling method is used.

Sampling $\tilde{\tau} = (\tau_2, \ldots, \tau_m)^T$ *via* $\eta = (\eta_2, \ldots, \eta_m)^T$*:* Recall that $\tilde{\tau} = h^{-1}(\eta|\tau_1)$ and consider components $\tau_j = h_j^{-1}(\eta|\tau_1) = \tau_1 + \sum_{k=2}^{j} e^{\eta_k}$, as in (2.10). We have

$$p(\eta|\,\cdot\,) = p(h(\tilde{\tau}|\tau_1)|\,\cdot\,) \propto \left[(1-\pi(\theta,\tau))^{(N-n)}\right] \cdot \exp\left[-\frac{1}{2}\sum_{j=2}^{m}\{c_j(\eta_j - \mu_j)\}^2\right] \cdot \mathbb{I}_{\{\tau_1<\tau_2<\cdots<\tau_m\}} \cdot \left[\prod_{j=1}^{m}\left\{\prod_{l=1}^{j-1}\left(\frac{\tau_{l+1}}{\tau_l}\right)^{-\theta_l}\right\}^{n(j)} \prod_{i\in\mathcal{I}(j)}\left(\frac{\theta_j}{\tau_j}\right)\left(\frac{s_i}{\tau_j}\right)^{-(\theta_j+1)} \mathbb{I}_{\{\tau_1<\min(s_1,\ldots,s_n)\}}\right],$$

where $\mathcal{I}(j) = \{i : \tau_j \leq s_i < \tau_{j+1}\}$ and $n(j)$ is the cardinality of $\mathcal{I}(j)$. Sampling of whole η vector is done via the Metropolis-Hastings algorithm. We use normal proposal distribution for η with a variance tuning parameter. This transformation satisfies all constraints on τ: $0 < \tau_1 < \tau_2 < \cdots < \tau_m$, but performs sampling in η space.

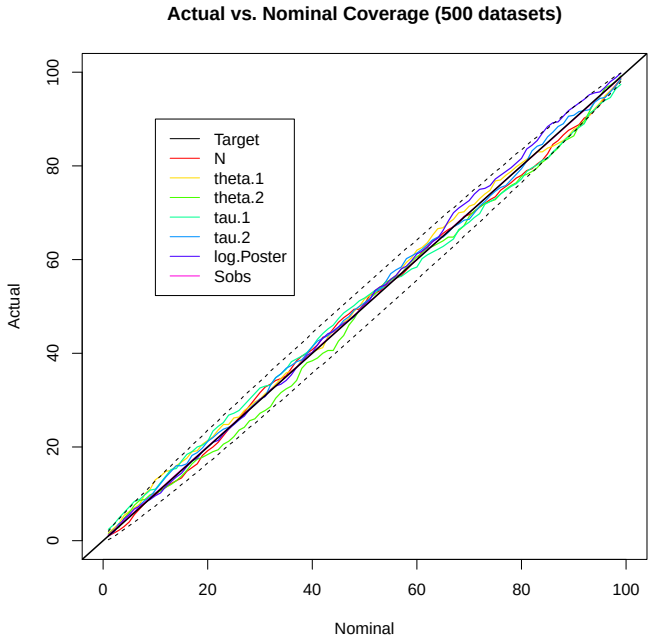

FIGURE 2.1. Coverage line plot for the broken-Pareto model. Validation is implied since all parameter quantiles (colored lines) correspond to the nominal levels and fall close to the 45 degree line.

Sampling $S_{obs} = (S_1, \ldots, S_n)^T$ *:* For $i = 1, \ldots, n$, we have

$$p\left(S_i|\cdot\right) \propto g(S_i, B_i, L_i, E_i) \cdot \text{Pareto}\left(S_i; \theta, \tau\right)$$
$$\cdot \text{Poisson}\left(Y_i^{tot}; \lambda(S_i, B_i, L_i, E_i) + k(B_i, L_i, E_i)\right) \cdot$$
$$\cdot \text{Binomial}\left(Y_i^{src}; Y_i^{tot}, \frac{\lambda(S_i, B_i, L_i, E_i)}{\lambda(S_i, B_i, L_i, E_i) + k(B_i, L_i, E_i)}\right)$$

Sampling is done with the Metropolis-Hastings algorithm. We use a normal distribution proposal with the tuning variance parameter.

2.3.3. Validation. As described in previous chapter, the MCMC samples from the posterior distribution admit themselves to a self-consistency check. We perform the validation check of the broken-Pareto model to estimate parameters N, θ, τ, S_{obs}. Results in Figure 2.3.3 shows that the convergence to the stationary distribution is reached because all coverage proportions (colored lines) lie within the approximate binomial error bounds (dashed black lines).

We note that the validation is much harder to achieve for missing data models because it requires the accurate computation of the marginal probability of observing a source, π. Our results show

that numerical approximation to π performed within the Gibbs sampler makes it much harder to attain a validated result. We suggest pre-computing and smoothing π prior to application of the Gibbs. We find that relaxing the stringent requirements on knowledge of probability of observing a source to within 0.01 of the truth results in reasonable parameter estimates. In practice, once the algorithm has been validated, convergence of the Monte Carlo chain should be verified through the trace plots of the parameter draws and numerical summaries, such as the effective sample size.

2.3.4. Parameter Inference. The inferences for the $\log(N) - \log(S)$ parameters are based on the posterior MCMC draws. The main interest falls on the estimates of broken power-law slopes, $\theta_1, \ldots, \theta_m$, and location of the flux breakpoints $\tau_2, \ldots, \tau_m$. Of secondary interest are estimates of the minimum threshold, τ_1, and the population size N. We use posterior mean, median, or mode to represent the estimates and construct 95% posterior credible intervals to represent the uncertainty in these estimates. The $\log(N) - \log(S)$ plot can be constructed after the sampling of S_{mis} given MCMC draws of other parameters. This plot helps to visually examine the potential curvature of $\log(N) - \log(S)$ and helps to determine if the parameter estimates are reasonable from the astrophysical point of view. We also examine the posterior predictive checks of this model to verify goodness-of-fit.

We emphasize that the current $\log(N) - \log(S)$ model is a conditional model assuming the knowledge of the number of broken-Pareto components, m. Formal methodology for the estimation of the number of breakpoints is an obvious extension of the method. We leave this problem for future work, but supply ideas of the associated challenges in the Discussion section below. On the other hand, in the subsequent sections we describe how to select the number of broken-Pareto components utilizing model selection criteria in Bayesian settings and a novel model selection approach called Bayesian adaptive fence method.

2.4. Model Selection

When one is presented with multiple plausible models, it is of interest to have an automated procedure for model selection. In $\log(N) - \log(S)$ setting it is crucial to be able to choose among the candidate models, e.g., single Pareto model vs. broken-Pareto model with one break point. Model selection procedures have been adapted for Bayesian methodology and are typically designed based on evaluation of model performance given the data. Even though it is a widely researched topic, it is often a difficult problem that does not have a unique best solution. We present some popular

methods for Bayesian model selection, including the Bayes Factor and Information Criteria. In section 2.5 we present simulation results examining their performance in various classical scenarios for model selection and identify their weaknesses. In section 2.6 we introduce a new method for model selection called the Bayesian adaptive fence method and show its improvement over the present methods in simulation.

2.4.1. Bayes Factor. In comparing two available models M_1 and M_2, a popular method for evaluating model performance is the Bayes Factor, which is the ratio of the marginal likelihoods of the data under model 1 and model 2. Suppose that the data is in the form of a vector y, and a continuous parameter vector $\beta \in \Omega_{M_k}$ under model M_k where $k = 1, 2$. When M_k is the true model, the marginal likelihood is given by:

$$p(y|M_k) = \int_{\Omega_{M_k}} p(y|\beta, M_k)p(\beta|M_k)d\beta. \tag{2.12}$$

This quantity is sometimes referred to as the Bayesian evidence of model M_k and represents the average of the likelihood $p(y|\beta, M_k)$ under the prior $p(\beta|M_k)$. Simple application of Bayes rule gives the posterior probability

$$p(M_k|y) = \frac{p(y|M_k)p(M_k)}{p(y|M_1)p(M_1) + p(y|M_2)p(M_2)}, \quad (k = 1, 2), \tag{2.13}$$

where $p(M_k)$ is the model prior with $p(M_1) = 1 - p(M_2)$. Taking the ratio of the posterior probabilities, we have:

$$\frac{p(M_1|y)}{p(M_2|y)} = \frac{p(y|M_1)}{p(y|M_2)}\frac{p(M_1)}{p(M_2)} = BF_{12}\frac{p(M_1)}{p(M_2)}, \tag{2.14}$$

where the Bayes Factor (BF) is defined as

$$BF_{12} = \frac{p(y|M_1)}{p(y|M_2)}. \tag{2.15}$$

The Bayes Factor in (2.14) is the ratio of the posterior odds and prior odds when M_1, M_2 are the only model choices. It is the factor by which the relative odds between two models improve after accounting for the data. Hence, the value BF_{12} represents the change in strength of evidence provided by the data in favor of one scientific theory (M_1) as opposed to another (M_2). Jeffreys (1961) and Kass and Raftery (1995) give reference scales for interpretation of strength of evidence (see Table 2.1).

$\log(B_{12})$	Evidence against M_2
0 – 0.5	Not worth mentioning
0.5 – 1	Substantial
1 – 2	Strong
>2	Decisive

TABLE 2.1. Interpretation of the evidence against model M_2 compared to model M_1 based on logarithm of the Bayes Factor.

When multiple candidate models are involved, enumerable by a set $\mathcal{M} = \{M_1, M_2, ...\}$, the same idea generalizes to Bayes factor BF_{jk}. It shows the change in strength of evidence for M_j against M_k.

The marginal probability $p(y|M_k)$ is central for evaluating the BF. However, in the majority of hierarchical Bayesian models, the unknown parameter has multiple dimensions and makes the evaluation of (2.12) intractable. In practice, therefore the marginal probabilities must be approximated based on the Monte Carlo draws of the parameter. The difficulty with this approach stems from the fact that the use of prior distribution as the importance density is extremely inefficient in practice because the prior is too broad to provide good samples that maximize the likelihood. If the parameter has multiple dimensions, then the likelihood peak is very narrow, so it becomes virtually impossible to efficiently sample good parameter configurations from the prior. For this reason, samples from the posterior distribution may be more useful in providing a better approximation to the integral. A comprehensive review of standard methods for approximating marginal probabilities (2.12) is given in Kass and Raftery (1995), Chen et al. (2000), and Ardia et al. (2009). In the following we will be focusing on model M_k for fixed k. Thus, for simplicity of notation, we drop dependence on M_k, so that

$$p(y|M_k) \stackrel{\text{def}}{=} p(y) = \int_\Omega p(y|\beta)p(\beta)d\beta$$

is the familiar marginal density of the data, a normalizing constant of the joint posterior density.

Newton and Raftery (1994) show that, due to the identity,

$$E\left[\frac{1}{p(y|\beta)}\middle| y\right] = \int \frac{1}{p(y|\beta)}p(\beta|y)d\beta = \int \frac{p(y|\beta)p(\beta)}{p(y|\beta)p(y)}d\beta = \frac{1}{p(y)}\int p(\beta)d\beta = \frac{1}{p(y)}, \tag{2.16}$$

a natural estimator of (2.12) is the harmonic mean of the likelihood, evaluated at the posterior samples of parameters $\beta^{(1)}, ..., \beta^{(L)}$:

$$p(y) \approx \left(\frac{1}{L} \sum_{i=1}^{L} \frac{1}{p(y|\beta^{(i)})} \right)^{-1}. \tag{2.17}$$

This estimator has been a part of large debate. $p(y)$ is known to be very sensitive to changes to the prior, however, the estimator (2.17) is typically not sensitive to the prior choice. Also, the variance of the harmonic mean estimator is often not finite (Kass and Raftery, 1995), hence this estimator is not expected to perform well. For example, Robert et al. (2009) or Neal (2008) show that this estimator performs poorly. We present the BF result based on (2.17) for simplicity of computation and for comparison.

Many approximations to the Bayesian evidence or directly to the Bayes Factor have been proposed in the literature. Unfortunately, all methods we considered are not computationally viable when applied to the $\log(N) - \log(S)$ problem. Laplace approximation to the integral is not appropriate due to the high dimensionality of the parameter space and highly skewed posterior distributions. Newton and Raftery (1994) suggest modifications to (2.17) using simulated annealing and an iterative approximation to the marginal likelihood. Due to the extreme small order of magnitude of the flux parameters with skewed posteriors, numerical error is usually incurred. Meng and Wong (1996) describe a bridge sampling estimate of the BF, which is another twist on the importance sampling idea. We implemented an approximation based on the harmonic mean and geometric mean (Meng and Wong, 1996), but found that neither of these methods were precise enough to give useful results. Skilling (2004) describes a nested sampling algorithm, in which the multidimensional integral is recast into a one-dimensional integral for ease of numerical evaluation. The $\log(N) - \log(S)$ problem has many types of parameters with highly skewed and peaked distributions, and we have not yet found an appropriate recasting function.

If the BF does not have an analytic solution and priors have similar volume, it is not always clear if the approximated BF value can be trusted, and whether improvement in quality of fit of the model is actually visible. Besides, the BF does not measure goodness of fit of the model and does not penalize for overfitting. Penalization type class of model selection methods have been proposed that set a default penalty for complex models. Given the computational and theoretical

issues with the Bayes Factor for the $\log(N) - \log(S)$ problem, we now consider other methods for model selection.

2.4.2. Information Criteria. The most widely used information criterion is the Akaike Information Criterion (AIC) (Akaike, 1974), defined as:

$$AIC = -2 \log p(y|\hat{\beta}_{MLE}) + 2k, \tag{2.18}$$

where the conditional likelihood, $p(y|\beta)$, is maximized at the maximum likelihood estimate $\hat{\beta}_{MLE}$ of the unknown parameter β, and k is the number of parameters in the model. AIC consists of two terms: a measure of goodness of fit, which decreases with more complex models, and a penalty for model complexity, which increases for more complex models. The model is selected based on the minimum AIC. In practice for Bayesian problems, AIC has a good performance when flat priors are used and when a maximum likelihood estimate is easily available; however, in other situations AIC is not guaranteed to work well. From a Bayesian point of view, the AIC may be considered as -2 times the estimate of out-of-sample predictive accuracy, the expected log predictive distribution $E[\log p(y^*|\hat{\beta}(y))] = \int \log p(y^*|\hat{\beta}(y))p(y^*)dy^*$, where the posterior distribution is summarized by a maximum likelihood point estimate of the parameter. This expression cannot be estimated directly. The standard approach is to use the log posterior density of the observed data given $\hat{\beta}(y)$ and correct for bias due to overfitting. The value k is the bias correction for the amount of increase in predictive accuracy given the maximum likelihood estimate by fitting k parameters. Informative prior distributions and hierarchical structures, such as in $\log(N) - \log(S)$ problem, tend to reduce the amount of overfitting, so k may be too large for needed bias correction. For our $\log(N) - \log(S)$ model, AIC is also not applicable because the MLE is not available due to nature of the missing data.

A modified Bayesian version of AIC is defined by Spiegelhalter et al. (2002) and is called the Deviance Information Criterion (DIC). It is computed by replacing the MLE by its posterior estimate $\tilde{\beta}_{Bayes}$, for example $\tilde{\beta}_{PostMean} = E[\beta|y]$, and by replacing k with a data-driven bias correction:

$$DIC = -2 \log p(y|\tilde{\beta}_{Bayes}) + 2p_{DIC}. \tag{2.19}$$

The term $\log p(y|\tilde{\beta}_{Bayes})$ is the estimate of the expected log predictive density, which we consider as a semi-Bayesian version of the measure of predictive accuracy. The bias correction term, p_{DIC},

is the estimated effective number of parameters and is defined by:

$$p_{DIC} = 2\left\{\log p(y|\tilde{\beta}_{Bayes}) - E[\log p(y|\beta)|y]\right\},$$

where the expectation is the average of β over the posterior distribution. In practice, the posterior expectation is approximated from simulations by replacing the expectation $E[\log p(y|\beta)|y]$ by average over the L posterior draws $\beta^{(l)}_{MCMC}$:

$$\frac{1}{L}\sum_{l=1}^{L}\left\{\sum_{i=1}^{N}\log p(y_i|\beta^{(l)}_{MCMC})\right\}.$$

Similarly, the plug-in estimate of β, the posterior mean, can be evaluated as $\frac{1}{L}\sum_{l=1}^{L}\beta^{(l)}_{MCMC}$. $\tilde{\beta}_{Bayes}$ estimate may not be unique. Using posterior mean, median, or mode, we arrive at DIC_{Mean}, DIC_{Median}, and DIC_{Mode}, respectively.

The advantage of DIC is in the ease computation, because it can be evaluated using already available MCMC draws of the parameter vector. The DIC approach gained popularity from its implementation in `WinBugs` package for Bayesian data analysis. Among many criticisms of the DIC, two issues stand out. For missing data problems, the definition of a parameter, and hence, of the DIC, becomes somewhat arbitrary. For example, Celeux et al. (2003) gives eight modifications to the DIC. Another issue is that p_{DIC} is not guaranteed to be positive for models outside of log-concave densities. An alternative definition of the penalty that guarantees positivity is:

$$p_{DIC,V} = 2Var[\log p(y|\beta)|y],$$

with corresponding value for DIC: DIC_V. Gelman et al. (2013) show that both penalties give the correct limit for fixed model and increasing sample size, while p_{DIC} is more numerically stable. In the next subsections we present the performance of the DIC method for model selection for hierarchical models.

A "fully" Bayesian version of AIC is introduced by Watanabe (2010) and is called the Watanabe-Akaike Information Criterion, or a "widely applicable information criterion" (WAIC). It assumes that all observations are independent. It uses the expected log pointwise predictive density as a new dataset as measure of predictive accuracy instead of plug-in predictive density, as in DIC. The

correction for the effective number of parameters to adjust for overfitting is defined as:

$$p_{WAIC1} = 2\sum_{i=1}^{N} \{\log E[p(y_i|\beta)|y] - E[\log p(y_i|\beta)|y]\},$$

where first term inside the summation is the log pointwise predictive density of the i-th data point:

$$\log E[p(y_i|\beta)|y] = \log \int p(y_i|\beta)p(\beta|y)d\beta$$

and is estimated in practice with the use of parameter draws form the posterior distribution:

$$\log \left\{ \frac{1}{L} \sum_{l=1}^{L} p(y_i|\beta_{MCMC}^{(l)}) \right\}.$$

The second term is evaluated similarly:

$$\bar{p} = \left\{ \frac{1}{L} \sum_{l=1}^{L} \log p(y_i|\beta_{MCMC}^{(l)}) \right\}.$$

An alternative definition is to use the variance of the individual terms in log predictive density combined across all data points:

$$p_{WAIC2} = \sum_{i=1}^{N} Var[\log p(y_i|\beta)|y],$$

where the variance is estimated by

$$\frac{1}{L-1} \sum_{l=1}^{L} \left\{ \log p(y_i|\beta_{MCMC}^{(l)}) - \bar{p} \right\}^2.$$

Note that the p_{WAIC2} expression omits the factor of 2. It shows the fluctuation of the posterior distribution. With respect to the posterior distribution, $\log p(y_i|\beta), 1 \leq i \leq N$ are not independent even if $y_i, 1 \leq i \leq N$ are, indeed, independent. The WAIC is defined similarly to AIC and DIC in estimating predictive accuracy with bias correction:

$$WAICj = -2\sum_{i=1}^{N} \log E[p(y_i|\beta)|y] + 2p_{WAICj}.$$

Finally, another well known information criterion is due to Schwarz (1978), who defines the Bayesian Information Criterion (BIC). BIC has a similar form to AIC (2.18), however, the penalty

is replaced by $k\log(n)$, that is,

$$BIC = -2\log p(y|\hat{\beta}_{MLE}) + k\log(n),$$

which for large datasets gives a larger penalty per parameter when compared to AIC. BIC is derived from approximating marginal probability of the data, which is different from the goal of other information criteria to approximate the predictive accuracy of the model. Nevertheless, we present the BIC results whenever possible for comparison.

2.5. Simulation Studies for Bayesian Model Selection

In this section we perform multiple simulation studies for model selection performance. We consider the classical setting, where the goal is to select the true model that is one of the models in the candidate set. Admittedly, it is rare for the true model to be part of the candidate set of models, and the usual approach in practice is to focus on the selection of a model with the goal of optimal prediction. In this situation, the chosen model can be incorrect but useful. From here arises the common Bayesian point of view to compare model selection rules in hopes of understanding the fitted models, instead of selecting one best model. On the other hand, for problems like $\log(N) - \log(S)$, where the selection of a useful model is desirable, the Bayesian point of view may not acceptable. One is then to search for the best available Bayesian model selection rule (among AIC, DIC, WAIC and BIC). Hence, the classical settings we consider now are useful in comparing performance of model selection methods to know which method has a potential to perform poorly even in reasonably simple scenarios.

2.5.1. Performance of Information Criteria in Bayesian Multiple Linear Regression.

We now examine the performance of DIC and WAIC in the classical setting of Bayesian multiple linear regression. We judge the performance in terms of correctly selecting a correct model out of 3 candidate nested models. It is known that DIC tends to overfit the data and is not consistent as $n \to \infty$. Little is known about WAIC. Usually, these statistics are used in comparison when applied to the single dataset. However little information about reliability is known in this case.

Selection Criteria	N20	N50	N100	N200	N500	N1000	N10000
DIC_{mean}	0.94	0.96	0.92	0.84	0.88	0.89	0.82
DIC_{median}	0.94	0.96	0.92	0.84	0.88	0.89	0.82
DIC_{mode}	0.88	0.92	0.89	0.84	0.82	0.88	0.82
DIC_V	1.00	1.00	0.99	0.97	0.98	0.94	0.84
WAIC_1	0.90	0.94	0.92	0.80	0.86	0.88	0.82
WAIC_2	0.91	0.95	0.92	0.82	0.86	0.89	0.82
AIC	0.76	0.85	0.86	0.74	0.82	0.84	0.80
BIC	0.84	0.94	0.96	0.94	1.00	0.98	1.00

TABLE 2.2. (Flat prior) Proportion that Model M_2 with 2 predictors was selected against Models M_1 and M_3. Model M_1 was never selected.

Selection Criteria	N20	N50	N100	N200	N500	N1000	N10000
DIC_{mean}	0.94	0.96	0.92	0.84	0.88	0.89	0.82
DIC_{median}	0.94	0.96	0.92	0.84	0.88	0.89	0.82
DIC_{mode}	0.88	0.92	0.89	0.84	0.82	0.88	0.82
DIC_V	1.00	1.00	1.00	0.97	0.98	0.94	0.84
WAIC_1	0.90	0.94	0.92	0.80	0.86	0.88	0.82
WAIC_2	0.91	0.95	0.92	0.82	0.86	0.89	0.82
AIC	0.76	0.85	0.86	0.74	0.82	0.84	0.80
BIC	0.84	0.94	0.96	0.94	1.00	0.98	1.00

TABLE 2.3. (Informative prior) Proportion that Model M_2 with 2 predictors was selected against Models M_1 and M_3. Model M_1 was never selected.

Three nested models are considered as follows:

$$
\begin{aligned}
M_1 : & y_i = \beta_1 + \beta_2 x_{1i} + \epsilon_i \\
M_2 : & y_i = \beta_1 + \beta_2 x_{1i} + \beta_3 x_{2i} + \epsilon_i \\
M_3 : & y_i = \beta_1 + \beta_2 x_{1i} + \beta_3 x_{2i} + \beta_4 x_{3i} + \epsilon_i \\
& \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)
\end{aligned}
$$

We simulate data under M_2, with true $\beta_1 = 2, \beta_2 = 1, \beta_3 = 1$, $\sigma = 0.5$, and independent set of predictors $X_j \sim \text{Normal}\,(0, 1)$, $j = 1, 2, 3, 4$, which corresponds to a fairly low signal-to-noise. To probe consistency, we simulate the data of various scales for $y_1, \ldots, y_N$, where $N = 20, 50, 100, 200, 500, 1000, 10000$. To evaluate under- and overfitting, we fit all 3 nested models to each dataset. To do this, we assume conjugate priors: $\sigma^2 \sim \text{Scaled-Inverse-Chi-Squared}\,(\nu_0, s_0^2)$ and $\beta|\sigma^2 \sim \text{Normal}\,(\beta_0, \sigma^2\Lambda_0^{-1})$. Informative priors were set with $\nu_0 = 50, s_0^2 = 0.5$, $\beta_0 = (2, 1, \ldots, 1)^T, \Lambda_0^{-1} = 2\mathbb{I}$. Flat priors can be achieved by setting ensured with $s_0^2 = 0$ and $\Lambda_0^{-1} = \mathbf{0}$.

We simulate 10000 Monte Carlo draws of the parameters from the posterior distribution. The simulation was repeated 200 times for each dataset. We select the model based on the minimum DIC, WAIC, AIC, and BIC, and report the proportions of correctly selected models out of 200.

Results in Tables 2.2 and 2.3 show that the DIC and WAIC perform well for both prior scenarios. Both tables give very similar results. Methods never select the under-fitted model M_1. Some methods select the overfitted model M_3 that includes all important predictors in addition to a third predictor. The proportion of overfitting ranges between 0% and 18% at various sample size datasets. This means that all methods perform well with at least 82% of probability of correctly selecting true model out of three candidates. In both prior setting scenarios, DIC_V outperforms all other methods for small to moderate sample sizes, achieving perfect selection in a few cases. Our results demonstrate that, with increasing sample sizes, the rate of correctly selecting the true model does not increase to 1 for any method. Hence, neither DIC nor WAIC are consistent for model selection. The performance of AIC and BIC is as expected. BIC has a perfect selection for large sample sizes, it is consistent. AIC is not consistent and performs slightly worse than DIC and WAIC. These results suggest the model selection based on DIC and WAIC in linear models can perform well, with DIC_V being the best among measures based on posterior parameter draws.

2.5.2. Performance of Information Criteria in Simulation Study of Flux Data. Model selection criteria of AIC, DIC, and WAIC are derived under the assumption of approximate Gaussian estimation. We now consider a hierarchical Bayesian model for which the MCMC draws of parameters produce highly skewed posteriors. We demonstrate that in these situations, performance of the information criteria for model selection is not adequate. The model is defined as follows:

$$
\begin{aligned}
&Y_i|\tau,\theta,m \sim \text{Pareto}\,(\tau,\theta)^T, \tau=(\tau_1,...,\tau_m)^T, \theta=(\theta_1,...,\theta_m)\\
&\theta \overset{iid}{\sim} \text{Gamma}\,(a,b)\\
&\tau_1 \sim \text{Gamma}\,(\alpha,\beta)\\
&\eta_j \overset{iid}{\sim} \text{Normal}(\mu,c)\\
&\tau_j = \tau_1 + \sum_{k=2}^{j} e^{\eta_k}, j=2,...,m
\end{aligned}
$$

The above model is the Flux model part of the hierarchy within $\log(N)-\log(S)$ for complete data scenario (no missing data). We assume the number of Pareto mixture components, m, is known

in advance. Hence, our model can be viewed as a conditional model given m. Nevertheless, we do not treat m as a parameter. When $m = 1$, the model is called bp0, simple Pareto model without breakpoints. When $m = 2$, the model is called bp1, the broken-Pareto model with 1 breakpoint. When $m = 3$, the model is called bp2, the broken-Pareto model with 2 breakpoints. Even though these models are not nested theoretically, they can be considered approximately nested as the breakpoints move closer to the minimum flux, τ_1. In results that follow, the true model of the data generating process is bp1, the "underfitted" model is bp0, and the "overfitted" model is bp2. The goal is to select the correct model using information criteria DIC, WAIC, AIC, and BIC.

The hyperparameters of the priors of unknown parameters θ, τ, and η are assumed to be completely known. We consider two settings: vaguely informative priors (Weak Prior) use $\theta_j \sim \text{Gamma}\,(2.5, 1), \eta \sim \text{Normal}(-37.7, 0.7)$, and $\tau_1 \sim \text{Gamma}\,(1.02, 1.02 \times 10^{16})$ to match $E[\tau_1] = 10^{-16}$ and $SD[\tau_1] = 9.9 \times 10^{-17}$; informative priors (Informative Prior) use $\theta_j \sim \text{Gamma}\,(25, 16), \eta \sim \text{Normal}(-38, 0.6)$, and $\tau_1 \sim \text{Gamma}\,(9, 6 \times 10^{17})$ to match $E[\tau_1] = 1.5 \times 10^{-17}$ and $SD[\tau_1] = 5 \times 10^{-18}$.

Every dataset was generated under the true model with the following parameters: $\theta = (1.1, 1.5), \tau = (10^{-17}, 2 \times 10^{-17})$, with $m = 2$ (True model bp1). We generated the data at various sample sizes: $N = 15$, 50, 100, 200, and 500. Each scenario was repeated 200 times.

MCMC was used to estimate the parameters under bp0, bp1, and bp2 models. We omit the derivation here, noting that θ can be sampled directly from gamma distribution, and the flux breakpoints require Metropolis. For Metropolis updates of τ_1 and η_j, the proposal variances were tuned every 100 iterations out of first 5,000, resulting in an approximate acceptance rate of 35%. We used 10,000 burn-in and additional 50,000 kept iterations for our MCMC sampler. All posterior evaluations were based on these kept draws. Model selection criteria DIC and WAIC are based directly on the posterior draws of parameters. In order to produce AIC and BIC, we also computed MLE of the parameters assuming a traditional frequentist approach via log likelihood (derivation omitted).

Table 2.4 shows that all methods perform rather poorly in selecting the correct bp1 model. The best performance is seen only for larger sample size data $N = 500$, where the best model selection method is WAIC_2 followed by WAIC_1 and BIC. Only at $N = 500$ the DIC at median and mode, WAIC_1, WAIC_2, and BIC correctly select the true model with probability just above 50%. At smaller sample sizes all methods demonstrate their failure to select the correct model with

N	bp	DIC_{mean}	DIC_{med}	DIC_{mode}	DIC_V	$WAIC_1$	$WAIC_2$	AIC	BIC
15	0	0.20	0.24	0.38	0.89	0.77	0.67	0.54	0.72
	1	0.11	0.14	0.17	0.08	0.08	0.13	0.34	0.26
	2	0.69	0.62	0.45	0.04	0.15	0.20	0.12	0.03
50	0	0.14	0.25	0.39	0.93	0.67	0.61	0.18	0.79
	1	0.27	0.20	0.20	0.07	0.18	0.18	0.36	0.18
	2	0.59	0.55	0.41	0.01	0.15	0.20	0.46	0.03
100	0	0.12	0.17	0.38	0.89	0.62	0.58	0.07	0.72
	1	0.24	0.33	0.21	0.09	0.18	0.20	0.23	0.25
	2	0.65	0.49	0.41	0.03	0.19	0.21	0.70	0.03
200	0	0.07	0.12	0.33	0.88	0.48	0.47	0.01	0.68
	1	0.40	0.34	0.29	0.09	0.36	0.39	0.10	0.30
	2	0.54	0.53	0.39	0.04	0.15	0.14	0.90	0.01
500	0	0.01	0.04	0.10	0.65	0.14	0.14	0.00	0.34
	1	0.44	0.52	0.56	0.26	0.71	0.73	0.01	0.62
	2	0.55	0.45	0.34	0.09	0.14	0.13	0.98	0.04

TABLE 2.4. (Weak prior) Proportion of selecting Model bp0, bp1, and bp2 when True Model is bp1. For example, for sample size N=15, the BIC procedure selected model bp0 72%, model bp1 26%, and model bp0 3% out of 200 datasets.

1 breakpoint. For example, the DIC at mean, median, mode tend to select a larger model with 2 breakpoints instead of 1. DIC_V, $WAIC_1$, $WAIC_2$, BIC methods tend to over-penalize, whereas DIC at mean, median, mode methods tend to under-penalize. AIC under-penalizes for large sample sizes, but over-penalizes for small sample sizes. BIC vaguely exhibits consistency behavior when the sample size is increasing. AIC and DIC are not consistent with increasing sample size.

Overall, no good model selection procedure exists from the ones we examined. It is possible that the selection criteria are not capable in picking up the difference between models. For a given dataset, the values of the criterion function are very similar between different models. It appears that for the smaller sample size the selection based on minimum value of the criteria among the 200 datasets is almost due to chance. Evidently, there is no difference in predictive power among the 3 models when judged by DIC or WAIC.

In this simulation, the DIC procedure has a very serious problem: negative penalty is a common occurrence. The penalty term is the measure of complexity and is associated with the effective number of parameters, so it is unreasonable for its measure to be negative. A negative penalty will further decrease the value of the criterion and hence can mistakenly point to select a model. We found that 62% of the datasets produced a negative penalty in DIC at mean, 10% of the datasets produced that in DIC at median, and 6% of the datasets produced that in DIC at mode. Hence, DIC method is not recommended for model selection in application to the flux data.

N	bp	DIC_{mean}	DIC_{med}	DIC_{mode}	DIC_V	$WAIC_1$	$WAIC_2$	AIC	BIC
15	0	0.26	0.23	0.28	0.68	0.46	0.40	0.54	0.72
	1	0.24	0.29	0.29	0.20	0.28	0.30	0.34	0.26
	2	0.50	0.47	0.42	0.13	0.27	0.30	0.12	0.03
50	0	0.16	0.23	0.35	0.84	0.47	0.44	0.18	0.79
	1	0.26	0.26	0.27	0.12	0.27	0.28	0.36	0.18
	2	0.58	0.51	0.38	0.04	0.26	0.28	0.46	0.03
100	0	0.14	0.14	0.34	0.82	0.36	0.34	0.07	0.72
	1	0.24	0.25	0.20	0.12	0.30	0.32	0.23	0.25
	2	0.61	0.61	0.46	0.06	0.34	0.34	0.70	0.03
200	0	0.07	0.10	0.28	0.82	0.32	0.30	0.01	0.68
	1	0.35	0.31	0.33	0.14	0.39	0.41	0.10	0.30
	2	0.58	0.59	0.39	0.04	0.29	0.28	0.90	0.01
500	0	0.03	0.04	0.12	0.55	0.10	0.10	0.00	0.34
	1	0.50	0.50	0.56	0.33	0.68	0.69	0.01	0.62
	2	0.47	0.47	0.32	0.12	0.22	0.21	0.98	0.04

TABLE 2.5. (Informative pior) Proportion of selecting Model bp0, bp1, and bp2 when True Model is bp1. The structure is similar to Table 2.4.

Result in Table 2.5 is similar to the weak prior simulation result we just discussed. Only at $N = 500$ most methods attain the correct model selection proportion of 50% or more. $WAIC_1$ and $WAIC_2$ methods show the best performance, followed by BIC. DIC at mean and median, and AIC tend to overfit and select the larger model with 2 breakpoints instead of 1. For smaller sample sizes all methods fail to select the correct model. Overpenalizing methods are DIC_V, $WAIC_1$, $WAIC_2$. Underpenalizing methods are DIC at mean, median, mode. AIC and BIC result did not change from before.

To summarize, all methods demonstrate their inability to select the correct model with 1 break-point. This result supports the notion that DIC nor WAIC should not be used for classical model selection in hierarchical model with non-Gaussian errors. The negative penalty of the DIC at mean, median and mode render these procedures unusable.

2.5.3. Performance of Information Criteria in Simulation for $\log(N) - \log(S)$. We examine the performance of model selection methods for choosing the number of Pareto mixture components in the $\log(N) - \log(S)$ problem. We perform a simulation study with 23 experimental settings summarized in Table 2.6.

Each experimental setting is fitted with a single Pareto model and a broken-Pareto model with one and two break-points. These parameter values are chosen to mimic Wong et al. (2014). We set the background noise, off-axis angle, effective areas, source area, and expected background counts

to $B_i = 0.1$, $L_i = 4.294$, $E_i = 10^{19}$, $Area_i = 100$ and $k_i = B_i Area_i = 10$, respectively for all $i = 1, \ldots, N$. The exposure time is set at $670,000$ *seconds*. Recall that the expected source counts are $\lambda_i = S_i E_i / \gamma$, where the energy conversion factor is set at $\gamma = 1.6 \times 10^{-9}$ $ergs/ph$. Conditional on the dimension m, the number of free parameters in the model is $1+2m+n$. This number includes n unobserved fluxes of observed sources, which usually changes for different datasets according to the probability of observing a source $g = \Phi((\lambda_i + 700)/1400)$, where Φ denotes the standard normal CDF.

In all settings, we assume a uniform distribution for $p(B_i), p(L_i), p(E_i)$ of volume 1, so that priors for (B_i, L_i, E_i) has no effect. We summarize prior distribution assumptions for the parameters as follows.

For single Pareto model, $m = 1$,

- $N \sim$ Neg-Bin $(\alpha = 9.278, \beta = 0.0309)$ so that $E[N] = 300$ and $Var[N] = 100^2$;
- $\theta_1 \sim$ Gamma $(a = 3.5, b = 2.5)$;
- $\tau_1 \sim$ Gamma $(a_m = 1.494, b_m = 2.716 \times 10^{16})$ so that $E[\tau_1] = 5.5 \times 10^{-17}$ and $Var[\tau_1] = (4.5 \times 10^{-17})^2$.

For broken-Pareto model with one break point, $m = 2$,

- $N \sim$ Neg-Bin $(\alpha = 9.278, \beta = 0.0309)$ so that $E[N] = 300$ and $Var[N] = 100^2$;
- $\theta_1 \sim$ Gamma $(a_1 = 4, b_1 = 3), \theta_2 \sim$ Gamma $(a_2 = 5, b_2 = 3)$;
- $\tau_1 \sim$ Gamma $(a_m = 2.1511, b_m = 9.7778 \times 10^{16})$ so that $E[\tau_1] = 2.2 \times 10^{-17}$ and $Var[\tau_1] = (1.5 \times 10^{-17})^2$;
- $\tau_2 | \tau_1 = e^{\eta_2} + \tau_1$ where $\eta_2 \sim$ Log-Normal$(\mu = -38, \sigma = 0.7)$.

For broken-Pareto model with two break points, $m = 3$,

- $N \sim$ Neg-Bin $(\alpha = 9.278, \beta = 0.0309)$ so that $E[N] = 300$ and $Var[N] = 100^2$;
- $\theta_1 \sim$ Gamma $(a_1 = 3.5, b_1 = 2.4), \theta_2 \sim$ Gamma $(a_2 = 5, b_2 = 3), \theta_3 \sim$ Gamma $(a_3 = 7, b_2 = 3.6)$;
- $\tau_1 \sim$ Gamma $(a_m = 2.1511, b_m = 9.7778 \times 10^{16})$ so that $E[\tau_1] = 2.2 \times 10^{-17}$ and $Var[\tau_1] = (1.5 \times 10^{-17})^2$;
- $\tau_2 | \tau_1 = e^{\eta_2} + \tau_1$ where $\eta_2 \sim$ Log-Normal$(\mu = -38, \sigma = 0.7)$, and
- $\tau_3 | \tau_1, \tau_2 = e^{\eta_3} + \tau_1 + \tau_2$ where $\eta_3 \sim$ Log-Normal$(\mu = -37.7, \sigma = 0.8)$.

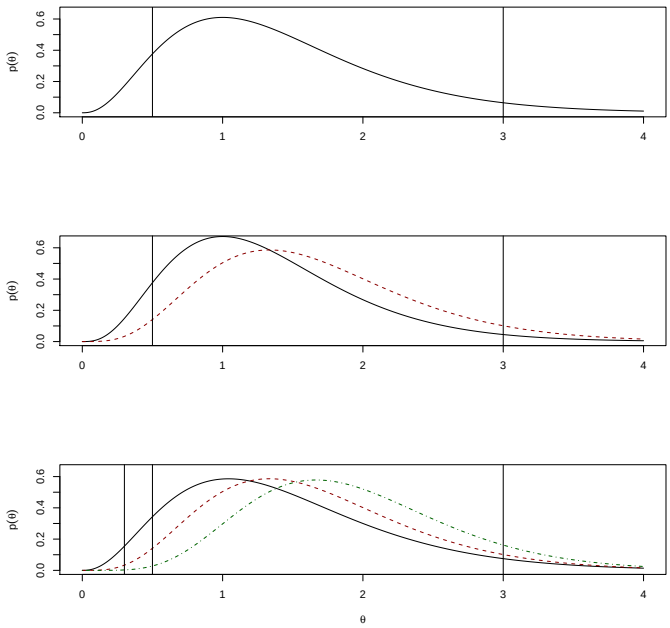

FIGURE 2.2. Prior densities for θ parameters. Top: $m = 1$, middle: $m = 2$, bottom $m = 3$.

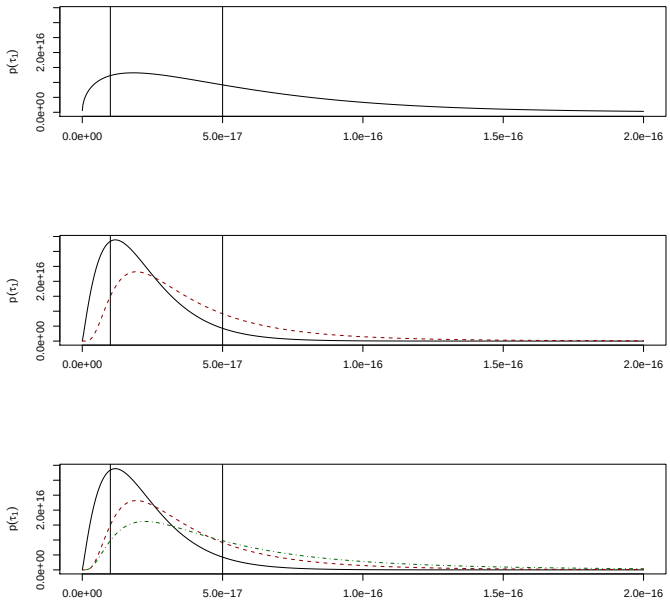

FIGURE 2.3. Prior densities for τ parameters. Top: $m = 1$, middle: $m = 2$, bottom $m = 3$.

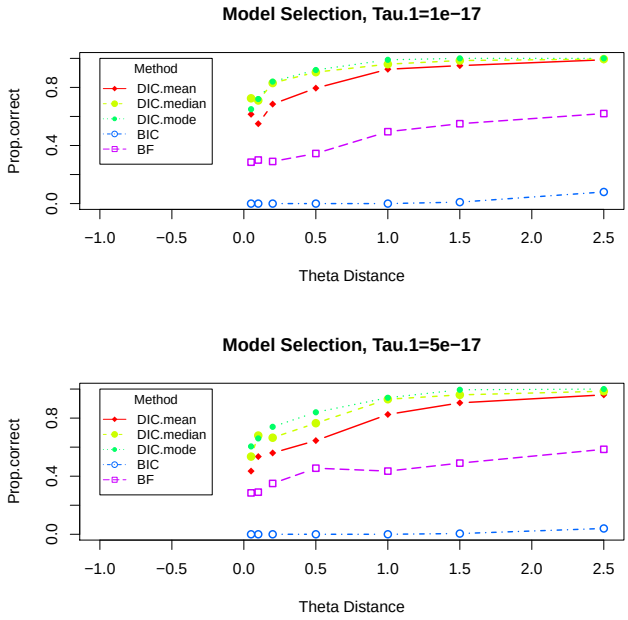

FIGURE 2.4. Model selection performance compared across gap of θ corresponding to $\theta_2 - \theta_1$.

The coverage of these priors can be visualized in Figure 2.2 and Figure 2.3. The vertical lines in the plot represent locations of the true parameters.

For each experimental setting, we generate and analyze 200 datasets attempting to fit either a single Pareto model or a broken Pareto model with one breakpoint. Experimental settings 1-20 have a unique true model, whereas for settings 21-23, both models are incorrect, but the broken Pareto model with 1 breakpoint is slightly better than no break point model. The analysis is done with our MCMC procedure based on 110,000 iterations and 10,000 burn-in samples. The model selection results based on DIC, BIC, and Bayes Factor are provided below in Table 2.7. We report three DIC statistics based on plug-in estimates of parameters using posterior mean, median, and mode. (Crude estimate of) BIC was evaluated from the average of log-likelihoods. Bayes Factor was approximated via harmonic mean estimates of normalizing constants.

The favored measure of model selection is DIC based on posterior mode, followed by DIC based on posterior median. The DIC selects the true model more than 80% of the time when the separation between two slopes of the power-laws is reasonably large, i.e., $\theta_2 - \theta_1 > 0.5$. Selection of the correct model is much more difficult when two slopes are nearly identical, thus showing a nearly linear $\log(N) - \log(S)$. Still, in these situations, the DIC at the mode has at least 60% rate of correct

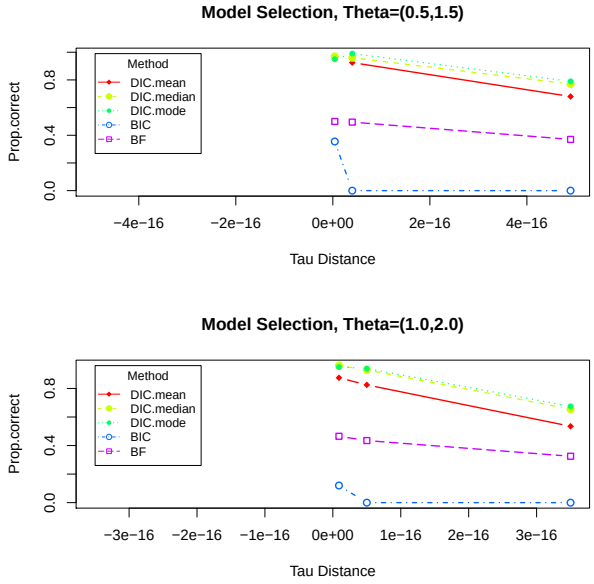

FIGURE 2.5. Model selection performance compared across gap of τ corresponding to $\tau_2 - \tau_1$ in broken Pareto models with a single breakpoint.

model selection. We consider two sets of settings 6-10 and 11-16 for which the distances range between 0.05 to 2.5. These results are summarized in two panels of Figure 2.4.

The separation between the breakpoint and the minimum flux, $\tau_2 - \tau_1$ is also very important for good performance of model selection. The large enough sample size of each Pareto component drives good estimation of parameters. In the event of a broken Pareto model with one breakpoint, the ideal is to have roughly 50% of all sources in each of the two populations in the mixture. A very small distance between τ_1 and τ_2 forces the majority of sources (observed and missing) to be in the second population of the mixture. In this situation, the first population is small in size and the overall number of missing sources in this population is small. Simulating these missing sources would not provide enough flexibility to $\log(N) - \log(S)$ to appear linear, and hence, the proportion to correct model selection is high, around 95%. On the other hand, a very large distance forces the majority of sources to be in the first population of the mixture, which also includes the majority of all missing sources. The proportion of correctly selected models is reduced to around 66% in some cases. We consider settings 17,6,18 and 19,13,20 for which the distances were selected in such a way as to force approx 15%, 50%, 85% to the first mixture, respectively. These results are summarized in the two panels of Figure 2.5.

Setting	True Model	$\log_{10}(\tau_1)$	$\log_{10}(\tau_2)$	$\log_{10}(\tau_3)$	N	θ_1	θ_2	θ_3
1	bp0	-17.0000			300	0.50		
2	bp0	-16.3010			300	1.00		
3	bp1	-17.0000	-16.3010		300	0.50	0.55	
4	bp1	-17.0000	-16.3010		300	0.50	0.60	
5	bp1	-17.0000	-16.3010		300	0.50	0.70	
6	bp1	-17.0000	-16.3010		300	0.50	1.00	
7	bp1	-17.0000	-16.3010		300	0.50	1.50	
8	bp1	-17.0000	-16.3010		300	0.50	2.00	
9	bp1	-17.0000	-16.3010		300	0.50	3.00	
10	bp1	-16.3010	-16.0000		300	1.00	1.05	
11	bp1	-16.3010	-16.0000		300	1.00	1.10	
12	bp1	-16.3010	-16.0000		300	1.00	1.20	
13	bp1	-16.3010	-16.0000		300	1.00	1.50	
14	bp1	-16.3010	-16.0000		300	1.00	2.00	
15	bp1	-16.3010	-16.0000		300	1.00	2.50	
16	bp1	-16.3010	-16.0000		300	1.00	3.50	
17	bp1	-17.0000	-16.8539		300	0.50	1.50	
18	bp1	-17.0000	-15.3010		300	0.50	1.50	
19	bp1	-16.3010	-16.2291		300	1.00	2.00	
20	bp1	-16.3010	-15.3979		300	1.00	2.00	
21	bp2	-17.0000	-16.0969	-15.7447	300	0.30	1.00	3.00
22	bp2	-17.0000	-16.0969	-15.7447	300	0.50	0.70	0.90
23	bp2	-17.0000	-16.0969	-15.7447	300	1.50	1.70	1.90

TABLE 2.6. Simulation Settings for Model Selection

Performance of model selection based on Bayes Factor is not favorable to DIC. The criterion BIC penalizes larger parameter models too much and always selects a single power-law. We will no longer consider BF and BIC for future model selection.

	DIC.mean	DIC.median	DIC.mode	BIC	BF
Set.1	0.51	0.43	0.41	1.00	0.23
Set.2	0.52	0.43	0.40	1.00	0.20
Set.3	0.61	0.72	0.65	0.00	0.28
Set.4	0.55	0.71	0.72	0.00	0.30
Set.5	0.69	0.83	0.84	0.00	0.29
Set.6	0.80	0.91	0.92	0.00	0.34
Set.7	0.93	0.96	0.99	0.00	0.49
Set.8	0.95	0.98	1.00	0.01	0.55
Set.9	0.99	0.99	1.00	0.08	0.62
Set.10	0.43	0.54	0.60	0.00	0.28
Set.11	0.54	0.68	0.66	0.00	0.29
Set.12	0.56	0.67	0.74	0.00	0.35
Set.13	0.65	0.77	0.84	0.00	0.46
Set.14	0.82	0.93	0.94	0.00	0.43
Set.15	0.91	0.96	0.99	0.01	0.49
Set.16	0.96	0.98	1.00	0.04	0.58
Set.17	0.95	0.97	0.95	0.35	0.50
Set.18	0.68	0.77	0.79	0.00	0.37
Set.19	0.88	0.96	0.95	0.12	0.47
Set.20	0.54	0.66	0.68	0.00	0.33
Set.21	0.90	0.94	0.99	0.03	0.53
Set.22	0.67	0.78	0.85	0.00	0.34
Set.23	0.51	0.61	0.58	0.00	0.24

TABLE 2.7. Model Selection Results: DIC at mean, DIC at median, DIC at mode, BIC based on average log-likelihood, and BF based on harmonic mean.

2.6. Bayesian Adaptive Fence Method

Jiang et al. (2008) introduced a different approach to model selection called the adaptive fence method. The idea of the fence method is to construct a fence to isolate a set of "correct" models that are suitable to reasonably describe the data and then choose an "optimal" model according to a criterion of optimality from the models within the fence. The fence is constructed using the inequality:

$$Q(M) - Q(\tilde{M}) \leq c, \tag{2.20}$$

where Q is a measure of lack of fit, M is a candidate model, $\tilde{M}$ is the baseline model whose Q is minimum, and c is a constant. For a given c, a model is labeled M_c if it is in the fence and satisfies the optimality criterion. Let M_{opt} be the actual optimal model. The standard optimality criterion is, but is not limited to, the minimal dimension criterion. In such a case, M_{opt} is a true model with minimal dimension. Since selection of a model with high parsimony is desirable, the minimal dimension criterion is a good choice. Also, it works well when the candidate models are submodels of a full model, $\tilde{M}$, which necessarily has the minimum Q.

Adaptive Fence (AF) selects the cut-off c by maximizing the empirical probability of selection. Ideally, this means to maximize

$$p = P(M_c = M_{opt}) \tag{2.21}$$

over c. The probability P in (2.21) is approximated under model $\tilde{M}$ via the use of a parametric bootstrap as follows. First estimate the unknown parameters under the full model $\tilde{M}$. Treating the estimated parameters as the true parameters, draw B bootstrap samples under $\tilde{M}$. Fit all candidate models in the set $\mathcal{M}$, including the full model, to each bootstrap sample. For a given c, let $p^*(M) = P^*(M_c = M)$ be the relative frequency among all bootstrap samples that model M satisfies (2.20) and is optimal. Finally, let $p^* = \max_{M \in \mathcal{M}} p^*(M)$ be the maximal probability of selection under $\tilde{M}$ for each c. The plot of p^* vs c usually resembles a "W-shape". $p^* = 1$ at $c = 0$, for which the full model is always selected because it is the only model in the fence. $p^* = 1$ at a very large c, say c_{max}, for which the minimum model is always selected because all candidates are in the fence and criteria is minimum dimension. However, p^* also peaks somewhere in between this range of c. AF method aims to maximize p^* over c restricted to $(0, c_{max})$. In other words, it aims

to select the cut-off c corresponding to a "significant peak in the middle" in p^*. Consistency of the selected model under the chosen c is guaranteed when the sample size increases to infinity and the true model is among the candidate model set $\mathcal{M}$, see Jiang et al. (2008).

The fence method is versatile in that it does not restrict the choice of goodness-of-fit measure, parameter estimation, and optimality criterion. Goodness of fit measure Q can incorporate aims for estimation or prediction. In particular, choice of Q can be related to methods of estimation. For example, estimation by maximum likelihood goes naturally with negative log-likelihood evaluated at MLE as the Q. One important requirement is that the chosen goodness of fit measure should separate the candidate models. In the event that models perform equally well by a choice of Q, it will be difficult to detect a significant peak in the p^* vs. c plot.

There are situations when the fence method does not perform well because p^* vs. c looks "V-shaped" and no significant peak in the middle is detected. It can occur if Q is the same for all models, or in special situations where the true model is on the boundary of the candidate set, for example, either the full model or the minimum model. Jiang examined possible strategies to overcome the latter problem (e.g., Jiang et al., 2008). On the other hand, the fence method shows very good performance if the true model is in the middle of the candidate set.

We extend the AF method for model selection in application to Bayesian data analysis. We call it Bayesian Adaptive Fence (BAF) method. So far, the fence method has not been considered in Bayesian inference. In many Bayesian problems including $\log(N) - \log(S)$, MLE is difficult to evaluate, but MCMC draws from the posterior distribution are easily available. It seems desirable and natural to define Q based on the posterior draws of parameters. We can use numerous definitions of the negative log-likelihood following DIC and WAIC derivations. For example, one may consider the negative log likelihood evaluated at the posterior mean: $Q = -\log p(y|\tilde{\beta}_{PostMean})$. We keep the optimality criterion as minimum-dimension for the model within the fence. Performance of these choices of Q is not as good as the negative log-likelihood evaluated at the MLE because in some cases the Q values are very similar for various models.

We present performance of the AF method applied to the Flux data simulation. In this simulation, the MLE is available and the data is complete. We consider the following measures of goodness

of fit:

$$
\begin{aligned}
Q1 &= -\log p(y|\hat{\beta}_{MLE}),\\
Q2 &= -\log p(y|\tilde{\beta}_{PostMean}),\\
Q3 &= -\log p(y|\tilde{\beta}_{PostMedian}),\\
Q4 &= -\log p(y|\tilde{\beta}_{PostMode}),\\
Q5 &= -\frac{1}{L}\sum_{l=1}^{L}\log p(y|\beta^{(l)}_{MCMC}),\\
Q6 &= -\sum_{i=1}^{N}\log\left[\frac{1}{L}\sum_{l=1}^{L}p(y_i|\beta^{(l)}_{MCMC})\right].
\end{aligned}
$$

The first measure Q1 is the same as in the original fence method procedure proposed by Jiang et al. (2008) using the negative log-likelihood maximized at MLE. Measures Q2-Q4 are modeled after the predictive measure based on DIC. Measure Q6 is modeled after the predictive measure based on WAIC.

Table 2.8 gives the result of applying BAF method for model selection for the flux data, according to weak prior assumption. We average the proportion of selecting the correct model over 50 datasets, each evaluated by using 100 bootstrap samples. It shows that the fence method works best when using the $Q1$ measure. However, it performs reasonably well for $Q6$, the approximated log of the average pointwise log-densities based on the posterior draws of parameters. This suggests that the WAIC-type approximation to measure of lack of fit is much more reasonable and stable estimate compared with DIC-type. It is important to design a good measure of lack of fit for the fence method.

In $\log(N) - \log(S)$ problem, the selection is needed regarding the presence of the breakpoint leading to the candidate set of models. Here we consider three cases: M_1 model with no breakpoint, M_2 model with one breakpoint, M_3 model with two breakpoints. These three models are not nested theoretically. However, model M_2 will appear to have the same $\log(N) - \log(S)$ plot as M_1 if the breakpoint is very close to τ_1 and missing data exists. Hence, the models can be considered nested in the practical sense.

Recall that the fence works well when the true model is "in the middle". Thus, for convenience, we consider introducing two additional fake models. Let model M_0 represent a model with a known,

N	bp	Q1	Q2	Q3	Q4	Q5	Q6
15	0	0.00	0.85	0.60	0.30	0.20	0.00
	1	0.90	0.15	0.40	0.65	0.55	0.65
	2	0.10	0.00	0.00	0.05	0.25	0.35
50	0	0.00	0.40	0.15	0.05	0.00	0.00
	1	1.00	0.50	0.60	0.60	0.40	0.65
	2	0.00	0.10	0.25	0.35	0.60	0.35
100	0	0.00	0.70	0.05	0.00	0.05	0.00
	1	1.00	0.30	0.60	0.55	0.55	0.75
	2	0.00	0.05	0.35	0.45	0.40	0.25
200	0	0.00	0.35	0.05	0.00	0.00	0.00
	1	1.00	0.50	0.55	0.55	0.60	0.80
	2	0.00	0.15	0.40	0.45	0.40	0.20
500	0	0.00	0.20	0.00	0.00	0.00	0.00
	1	1.00	0.70	0.70	0.70	0.70	0.75
	2	0.00	0.10	0.30	0.30	0.30	0.25

TABLE 2.8. (Weak prior) Proportion of selecting Model bp0, bp1, and bp2 when True Model is bp1 for Flux simulation using BAF method with automated detection of a peak in p* plot.

incorrect value for parameter θ, thus leaving only 1 unknown parameter τ to estimate. Note that M_0 is the model of minimum dimension. Now the single Pareto model M_1 is no longer considered on the boundary. Let model M_4 be a full phantom model that theoretically overfits the data. The specific parametrization of M_4 is not important. Instead, it is important to know that its value Q is the lowest among Q of other candidates. We suggest to use $Q(M_4) = \min_{M\in\{M_0,M_1,M_2,M_3\}} Q(M) - \epsilon$, for a positive constant ϵ. The actual value of epsilon does not matter as long as it helps to keep $Q(M_4)$ below other Q but on relatively the same scale as other Q. The use of M_4 makes sure that none of M_1, M_2, M_3 are full models. With the addition of M_0 and M_4, the candidate set $\mathcal{M}$ consists of 5 approximately nested models. So that the AF method can be applied. Our method draws some similarities with the invisible fence method described in Jiang et al. (2011).

We apply the BAF method to the $\log(N) - \log(S)$ simulation, for which MLE are not available, some data are missing (non-ignorable), and the model contains an additional level of hierarchy compared to the Flux data simulation. It is important to design a good measure of lack of fit for the fence method. We design the following goodness-of-fit measures Q as functions of posterior

Case	bp	Q1	Q2	Q3	Q4	Q5	Q6
1	0*	1.00	0.97	0.97	1.00	1.00	1.00
1	1	0.00	0.03	0.03	0.00	0.00	0.00
1	2	0.00	0.00	0.00	0.00	0.00	0.00
5	0	0.03	0.00	0.00	0.00	0.00	0.03
5	1*	0.34	0.47	0.57	0.70	0.33	0.30
5	2	0.63	0.53	0.43	0.30	0.67	0.67
7	0	0.03	0.00	0.00	0.00	0.00	0.03
7	1*	0.54	0.47	0.57	0.70	0.33	0.30
7	2	0.43	0.53	0.43	0.30	0.67	0.67
21	0	0.00	0.00	0.00	0.00	0.00	0.00
21	1	0.20	0.10	0.07	0.00	0.20	0.20
21	2*	0.80	0.90	0.93	1.00	0.80	0.80
22	0	0.10	0.00	0.00	0.00	0.00	0.05
22	1	0.20	0.20	0.25	0.35	0.25	0.25
22	2*	0.70	0.80	0.75	0.65	0.75	0.70

TABLE 2.9. Model selection proportions for $\log(N) - \log(S)$ simulation using Bayesian adaptive Fence Method with automated detection of a peak in p* plot.

parameter draws:

$$
\begin{aligned}
Q1 &= -\sum_{i=1}^{N} \log\left[\frac{1}{L}\sum_{l=1}^{L} p(y_i|\beta^{(l)}_{MCMC})\right], \\
Q2 &= -\log p(y|\tilde{\beta}_{PostMean}), \\
Q3 &= -\log p(y|\tilde{\beta}_{PostMedian}), \\
Q4 &= -\log p(y|\tilde{\beta}_{PostMode}), \\
Q5 &= -\frac{1}{L}\sum_{l=1}^{L} \log p(y|\beta^{(l)}_{MCMC}), \\
Q6 &= -Median\left\{\log p(y|\beta^{(l)}_{MCMC}), l = 1, \ldots, L\right\}.
\end{aligned}
$$

An example of the plot p* vs c is shown in Figure 2.6. The method does not produce a "W-shape" because we have introduced the best-fitting model, the baseline model "4" (in red), against which all other comparisons are made. By default, the choice of the baseline model only occurs at minimum c. Since this model does not exist and can never be selected, we assign it to have probability of 0. The automated procedure proceeds to select the first significant peak and its corresponding model. In this case, model "2" (in red) is selected.

Table 2.9 reports the model selection results based on the BAF method for the $\log(N) - \log(S)$ simulation. The models listed with an asterisk are the true models to be selected. The fence method

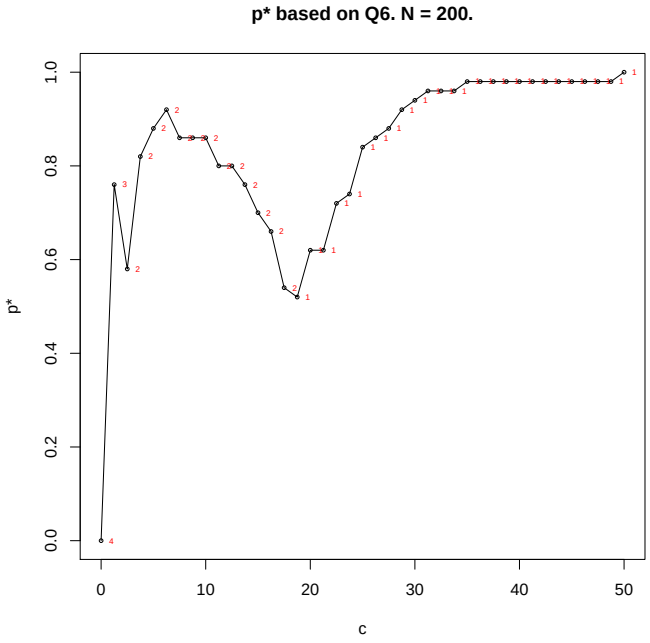

FIGURE 2.6. p* vs c plot for the Bayesian adaptive fence method. The model associated with each c is shown as a number in red.

result is based on the 100 bootstrap resamples of 30 datasets. The simulation Case 1, 21, and 22, perform better in selecting the correct model than do DIC and WAIC methods, as shown in previous section. The Cases 5 and 7 do not give a better performance, and tend to overestimate in some cases. Overestimation problem with the DIC and WAIC for $\log(N) - \log(S)$ analysis has been seen noted before. These results suggest that the BAF method outperforms in some cases the usual Bayesian methods for model selection. The BAF method outperforms all other methods of DIC, WAIC, AIC and BIC in model selection. We conclude that the BAF method has a potential to improve model selection in Bayesian data applications.

The original paper by Jiang et al. (2008) uses a parametric bootstrap in order to approximate the probability of selection under the full model. The parametric bootstrap gives tighter confidence bounds for parameter estimation than the standard normal theory, and also automatically inherits any preexisting dependence structure in the parameters. Hence any functional, such as p*, formed from the estimate parameters produced via bootstrapping also inherits such benefits. Thus, the parametric bootstrap is expected to give better performance than the non-parametric bootstrap. The BAF method does not share all of the benefits of the FM.

In our simulation studies we find that the choice of Q is very important when the BAF method is used. When Q is chosen primarily on the consideration for the prediction, BAF method can sometimes select an overfitted model. Care must be put into such a choice. The selection of the significant peak in the plot of p* vs c is also not very straightforward. When the automated procedures are used to evaluate the significant peak, the final choice depends on the smoothing in the p* vs. c plot. We recommend to consider various smoothing options for comparison. A highly significant peak is expected to show up for most smoothing settings.

From a computational viewpoint, the fence method takes time or abundant computing resources. However, we believe that with increasingly vast computing capabilities, this drawback of the fence method is a small price to pay for a reasonable model selection for tool Bayesian inference. We conclude that the Bayesian adaptive fence method is a powerful tool for model selection and should be considered for model selection in the hierarchical Bayesian inference with missing data.

2.7. Data Analysis

We revisit the analysis of the two datasets CDFN and CDFS to estimate the $\log(N) - \log(S)$ relationship of potential mixtures of the flux distributions. Ultimately, our goal is to select the most useful model among the postulated candidate models. We compare BAF, DIC, and WAIC methods when applied to the data. To assess that the chosen model has a good fit, we also perform posterior predictive check, described in section 1.4.4 of previous chapter.

2.7.1. Application: Chandra Deep Field North. We apply our model and the model selection to the real data. We begin with the CDFN dataset. Recall that this dataset consists of 225 X-Ray sources of *CHANDRA* observation of the Northern sky. The sample of sources is based on off-axis angle threshold of 8 arcmin. Previous $\log(N) - \log(S)$ analysis had shown that the presence of a break in $\log(N) - \log(S)$ is possible, yet it is not clear. We apply the BAF method to select an appropriate model for this data choosing between three candidate models: the single Pareto, bp0; the broken Pareto with 1 breakpoint bp1; and the broken Pareto with 2 breakpoints, bp2. In addition to the BAF, we report the model selection results of DIC and WAIC and examine posterior predictive p-values.

We assume moderately informative priors, elicited from collaborators. For $j = 1, 2, 3$, $\theta_j \sim$ Gamma $(15, 30)$, $\tau_1 \sim$ Gamma $(2.78, 1.85 \times 10^{17})$, $N \sim$ Neg-Bin $(9.278, 0.03)$, and for $k = 2, 3$,

BAF criterion	Sel. model	Method	bp0	bp1	bp2
Q_1	bp2	DIC_{Mean}	1839.3	**1833.1**	1839.4
Q_2	bp2	DIC_{Median}	1842.4	**1835.1**	1839.6
Q_3	bp2	DIC_{Mode}	1835.0	**1817.8**	1830.2
Q_4	bp2	DIC_V	1985.2	1862.8	**1856.7**
Q_5	bp2	WAIC_1	1716.5	1701.1	**1698.9**
Q_6	bp2	WAIC_2	1797.4	1780.5	**1778.1**

TABLE 2.10. BAF, DIC, and WAIC for CDFN analysis. BAF always selects bp2 model. Lowest DIC and WAIC of the selected model is marked in bold.

Posterior Predictive Statistic	bp0	bp1	bp2
Number of observed sources	0.21	0.21	0.21
Minimum photon count	0.31	0.31	0.31
Maximum photon count	0.10	0.08	0.07
Median photon count	0.28	0.25	0.24
Lower quartile of photon counts	0.14	0.12	0.11
Upper quantile of photon counts	0.19	0.17	0.14
Photon count IQR	0.21	0.20	0.17
Crude estimate of R^2	0.13	0.12	0.14
Number of observed sources vs. med photon count	0.69	0.69	0.60
Lower quartile vs. upper quartile of photon counts	0.73	0.64	0.56
Number of observed sources vs. photon count IQR	0.70	0.71	0.64
Number of observed sources vs. crude estimate of R^2	0.17	0.14	0.21

TABLE 2.11. Posterior predictive p-values for CDFN analysis.

$\eta_k \sim \text{Normal}\,(-38, 0.7^2)$. Incompleteness probability table and detector effects frequency table were directly provided by our collaborators.

Table 2.10 reports the model selection results for BAF, DIC, and WAIC for selecting between candidate models bp0, bp1, and bp2. BAF method consistently selects bp2 model under every criterion function Q we considered. (For the list of criterion functions Q_1 through Q_6, please refer to the previous section.) This result implies that the data strongly suggests that bp2 model is appropriate. The DIC and WAIC are not in agreement in their selected model. The favored model is bp1 for DIC at mean, median, and mode measures, but it is model bp2 based on DIC_V and WAIC measures. Noting that the actual criterion measures for the DIC at mean are fairly similar across candidate models, we do not recommend its use for model selection in CDFN analysis. DIC_V and WAIC results agree with the BAF method model selection. We therefore conclude that the model bp2, the broken-Pareto model with 2 breakpoints, provides a good fit to these data. The chosen model by our model selection criteria implies that 3 populations of the source flux are present in the CDFN survey.

The posterior predictive p-values (PPP) are given in Table 2.11. The table shows that all candidate models bp0, bp1, and bp2 provide a reasonable fit these data because PPP are large. The maximum photon count of the posterior predictive datasets shows the lowest PPP (ranging from 0.1 to 0.07); however, it is still within the reasonable bounds. We conclude that the chosen model by model selection, bp2, has no significant departures in posterior predictive datasets and it should perform well in prediction.

Table 2.12 reports the parameter estimates of the candidate models bp0, bp1, bp2 of the Bayesian analysis of CDFN data after accounting for incompleteness. The same table also reports the estimated parameters of the competing method by Wong et al. (2014), which estimated the breakpoints of other $\log(N) - \log(S)$ parameters via interwoven EM algorithm, but ignored the missing data structure. The final chosen model based on BAF is bp2 by BAF, while the estimated model by interwoven EM is bp1. The parameter estimates of the second and third flux population from our method agrees very well with the parameter estimates of the model fitted by Wong. Our method for bp2 fit shows that the first two flux populations have two power-law slopes estimated by posterior mean as 0.52 (with 95% central credible of $(0.31, 0.77)$) and 0.48 $(0.30, 0.67)$. Wong's method for bp1 fit shows that the first flux population has a power-law slope of 0.48 (standard error 0.06). Similarly, the first two flux populations of our method have two minimum flux thresholds estimated as $10^{-16.46}(10^{-16.62}, 10^{-16.35})$ and $10^{-16.11}(10^{-16.40}, 10^{-15.80})$. Wong's method fits the minimum flux threshold as $10^{-16.344}(10^{0.03})$. The last untruncated Pareto flux population has the estimated power-law slope of 0.78 $(0.62, 0.91)$ by our method, and 0.854 (0.224) by Wong's method, and the estimated breakpoint of $10^{-15.81}(10^{-16.08}, 10^{-15.59})$ by our method, and $10^{-15.57}(10^{0.271})$ by Wong's method. The main difference between these models is that accounting for missing data reflects a third possible population of the flux in the lower flux boundary with very similar population characteristics as the next truncated Pareto flux population.

Our analysis suggests the possibility of a slightly overfitted model. The model with one breakpoint exhibits a bimodal posterior distribution in the power-law slope. One of the modes is at 0.5, while the other mode is at 1.10. There are two potential reasons for this jump. First, it is conceivable that the MCMC sampler needs to be ran for longer than 50000 iterations. Second, it is very likely that the incompleteness function has not been perfectly specified. Observing that the model with two breakpoints estimates has two nearly identical power-law slopes (around 0.5) and the distance between the minimum flux and the first breakpoint is fairly negligible, we consider the possibility

Model	bp0	bp1	bp2	bp1 (Wong et.al.,2014)
Parameter	Mean [Median] (95% central interval)			Estimate (SE)
θ_1	0.66 [0.66] $(0.55, 0.77)$	1.08 [0.59] $(0.47, 1.46)$	0.52 [0.52] $(0.31, 0.77)$	0.483 (0.060)
θ_2		0.63 [0.62] $(0.53, 0.76)$	0.48 [0.48] $(0.30, 0.67)$	0.854 (0.224)
θ_3			0.78 [0.74] $(0.62, 0.91)$	
$\log_{10}(\tau_1)$	-16.28 [-16.27] $(-16.40, -16.21)$	-16.34 [-16.34] $(-16.44, -16.29)$	-16.46 [-16.46] $(-16.62, -16.35)$	-16.344 (0.030)
$\log_{10}(\tau_2)$		-16.23 [-16.27] $(-16.35, -15.86)$	-16.11 [-16.15] $(-16.40, -15.80)$	-15.657 (0.271)
$\log_{10}(\tau_3)$			-15.81 [-15.84] $(-16.08, -15.59)$	
N	294 [293] $(275, 314)$	296 [296] $(272, 322)$	280 [278] $(257, 312)$	

TABLE 2.12. Parameter estimates of major parameters for the CDFN data based on bp0, bp1, and bp2 models. Compared to the competing method for estimation of the number of breakpoints and parameters via interwoven EM (Wong et al., 2014).

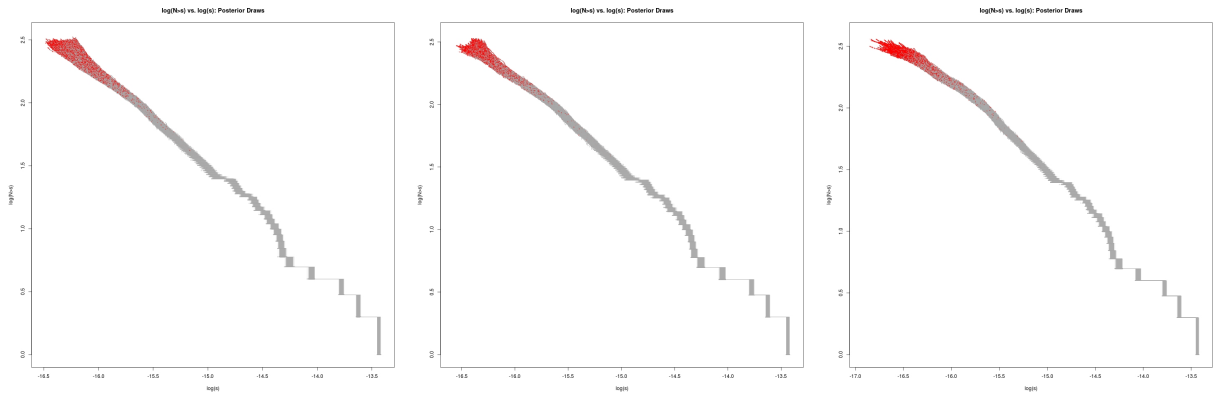

FIGURE 2.7. $\log(N) - \log(S)$ plots for models bp0 (left), bp1 (middle), and bp2 (right) for CDFN data.

that CDFN survey exhibits two flux populations. Moreover, the location of the breakpoint and the $\log(N) - \log(S)$parameters are consistent to the result by Wong et al. (2014).

Figure 2.7 shows the $\log(N) - \log(S)$ fits based on our method for models bp0, bp1, and bp2. Even though all models produce very similar figures, inferential properties of each model are different. Uncertainty in the $\log(N) - \log(S)$ relationship due to missing data is clearly demonstrated in the lower tail of the curve. The method estimates approximately 80% completeness of the survey.

BAF criterion	Sel. model	Method	bp0	bp1	bp2
Q_1	bp1	DIC_{Mean}	2979.5	**2968.9**	2988.9
Q_2	bp1	DIC_{Median}	2987.4	**2975.3**	2993.6
Q_3	bp1	DIC_{Mode}	2989.5	**2974.2**	2990.2
Q_4	bp1	DIC_V	3126.2	**3091.0**	3634.7
Q_5	bp0	WAIC_1	2785.2	**2770.9**	2795.1
Q_6	bp1	WAIC_2	2909.1	**2894.3**	2913.7

TABLE 2.13. BAF, DIC, and WAIC for CDFS analysis. BAF tends to select bp1 model. Lowest DIC and WAIC of the selected model is marked in bold. Both DIC and WAIC select bp1 model.

Posterior Predictive Statistic	bp0	bp1	bp2
Number of observed sources	0.03	0.10	0.23
Minimum photon count	0.01	0.02	0.00
Maximum photon count	0.38	0.17	0.03
Median photon count	0.03	0.09	0.00
Lower quartile of photon counts	0.00	0.00	0.00
Upper quantile of photon counts	0.06	0.24	0.02
Photon count IQR	0.03	0.28	0.04
Crude estimate of R^2	0.12	0.03	0.01
Number of observed sources vs. med photon count	0.06	0.04	0.00
Lower quartile vs. upper quartile of photon counts	0.00	0.00	0.00
Number of observed sources vs. photon count IQR	0.15	0.45	0.11
Number of observed sources vs. crude estimate of R^2	0.03	0.02	0.01

TABLE 2.14. Posterior predictive p-values for CDFS analysis.

2.7.2. Application: Chandra Deep Field South. We return to our data analysis of CDFS. Recall that this data consists of 358 X-Ray sources of Chandra observation of the Southern sky in 0.5-7.0 keV. The analysis from previous chapter suggested evidence of non-linear $\log(N) - \log(S)$. The CDFS survey is a much fainter survey than the CDFN. Our method currently does not handle false positive sources. For that reason we consider a subset of these data based on the following cuts. First, we limit the off-axis angle to 10.5 *arcmin*, and then we cut all potentially false sources with the negative significance of measure from the `wavdetect` software and a crude estimate of the source photon counts of -10. The resulting subsample contains 341 sources. We use the BAF, DIC, and WAIC to select a model for this data from the familiar three candidate models: bp0, bp1, and bp2. In addition to the model selection, we report the PPP to make sure the chosen model is appropriate.

Model	bp0	bp1	bp2
Parameter	Mean [Median] (95% central interval)		
θ_1	0.57 [0.57] $(0.52, 0.64)$	0.49 [0.49] $(0.43, 0.56)$	0.14 [0.13] $(0.07, 0.27)$
θ_2		0.95 [0.95] $(0.74, 1.19)$	0.37 [0.36] $(0.21, 0.52)$
θ_3			0.85 [0.84] $(0.70, 1.00)$
$\log_{10}(\tau_1)$	-16.00 $[-16.00]$ $(-16.04, -15.96)$	-16.04 $[-16.03]$ $(-16.08, -15.99)$	-16.27 $[-16.27]$ $(-16.44, -16.15)$
$\log_{10}(\tau_2)$		-14.72 $[-14.73]$ $(-14.78, -14.63)$	-15.85 $[-15.89]$ $(-16.05, -15.55)$
$\log_{10}(\tau_3)$			-15.40 $[-15.39]$ $(-15.60, -15.27)$
N	501 [501] $(472, 532)$	419 [419] $(400, 440)$	377 [376] $(364, 390)$

TABLE 2.15. Parameter estimates of major parameters for the CDFS data based on bp0, bp1, and bp2 models.

We assume moderately informative priors, elicited from collaborators. For $j = 1, 2, 3$, $\theta_j \sim$ Gamma $(12, 16)$, $\tau_1 \sim$ Gamma $(1.38, 3.46 \times 10^{16})$, $N \sim$ Neg-Bin $(4.05, 0.014)$, and for $k = 2, 3$, $\eta_k \sim$ Normal $(-38, 0.7^2)$.

Table 2.13 summarizes model selection results based on the BAF, DIC, and WAIC. Majority of the methods favor the model bp1, the broken Pareto with 1 breakpoint. The posterior predictive p-values are given in Table 2.14. All models show a poor fit with many PPP around 0, especially for bp2 model. We believe that this result can be due to the fact that it is challenging to specify the marginal detection probability, $\pi(\theta, \tau)$, correctly. Our pre-computed π over a 6-dimensional grid of the parameters (θ, τ) seems to have been insufficient. The $\log(N) - \log(S)$ plots for fitting all models ti CDFS are shown in Figures2.8. The long thin tail of low flux in bp2 model fit indicate potential misspecification of the detection probability. Hence, further investigation of analysis this dataset is needed.

The parameter estimates of all candidate models are provided in Table 2.15. The chosen model, bp1, implies that the CDFS dataset contains two flux populations. The power-law slope of the first flux population is estimated a posterior mean of 0.49 with 95% credible intervals (0.43, 0.56). The low flux threshold is estimated at $10^{-16.04}$ $(10^{-16.08}, 10^{15.99})$. The second flux population estimates

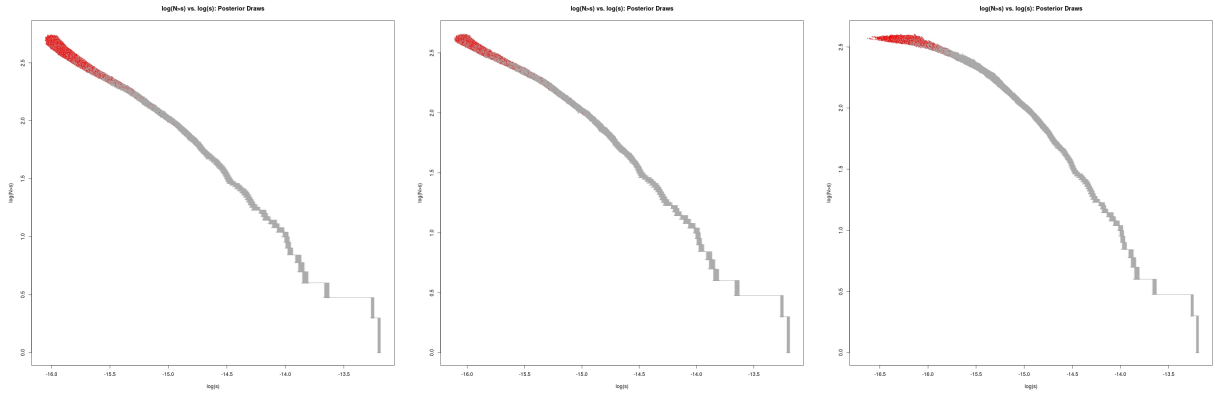

FIGURE 2.8. $\log(N) - \log(S)$ plots for models bp0 (left), bp1 (middle), and bp2 (right) for CDFS data.

the power-law slope of 0.95 (0.74, 1.19) and the flux breakpoint of $10^{-14.72}(10^{14.78}, 10^{14.63})$. The method estimates approximately 81% completeness of the survey.

2.8. Discussion and Concluding Remarks

Estimation of $\log(N) - \log(S)$ is a challenging problem from the methodological point of view. Astrophysical assumptions about the flux distribution yields probabilistic foundation of the model. It requires to correctly account for missing data and bias from detector effects. It necessitates the proper treatment of the missing data. It calls for a unified estimation approach in order to estimate the power-law slope and the flux to build the $\log(N) - \log(S)$ curve. We have presented a comprehensive method for estimation of the $\log(N) - \log(S)$ relationship for broken power-laws using a hierarchical Bayesian model. Our method explicitly corrects bias associated with the non-ignorable missing data mechanism that is often ignored by competing methods. We also presented a Bayesian adaptive fence (BAF) method for discrimination and selection of the models of optimal number of breakpoints. While BAF is computationally intensive, it has the potential to fill the model selection methodology gap in Bayesian data analysis.

2.8.1. Limitations and Alternative Approaches. One limitation of our method is that it cannot account for false positive sources, that is, sources that are recorded as such, but are not actually there. However, it would be straightforward to extend our method to overcome this problem. Consider a probability of a source to be a false positive as a function, g, of the flux and the detector effects. Define FP as the event that the source is incorrectly identified as a source, and define Obs to be the event that the source is observed. Then the probability of a true observed source

in a survey is $\Pr(FP^c \cap Obs) = (1 - \Pr(FP)) \Pr(Obs|FP^c) = (1 - g)\pi$. Hence, the probabilistic paradigm, upon which this method is built, allows the investigator to include all possible observation uncertainties in similar manner.

Another obvious drawback of our approach for the $\log(N) - \log(S)$ problem is the inability to estimate the number breakpoints, or, equivalently, the number of broken-Pareto components, m. Our hierarchical model is derived conditionally on the number of breakpoints. However, it is conceivable to design a model of deeper hierarchy, in which the number of breakpoints is also a parameter. In this case, the dimension parameters N and m form the outer layer of the hierarchy, and all other parameters θ, τ, S_{obs}, etc., form the inner layers. The main challenge of this approach is the change in dimension of parameter vectors θ and τ for different MCMC draws of m.

The standard MCMC applications rely on the fixed parameter dimension to simulate from posterior distribution. A special extension to MCMC methodology for trans-dimensional problems is called Reversible-Jump Markov chain Monte Carlo (RJMCMC) (Green, 1995). It assumes that the stationary distribution is the joint posterior distribution of a set of states of a model indicator and model dimension. For the $\log(N) - \log(S)$ problem, for example, the model can be the single- or broken-Pareto with number of components $m = 1$ or $m = 2$, respectively. The number of models in the set does not have to be finite. To allow valid proposals of the model/dimension state, one must design a one to one and differentiable mapping function with non-zero support. The differential in the model dimension is to be absorbed with a random component, so that the total dimension of the proposed state vs current state is matched. Dimension matching can be achieved with many possible mapping functions. However, some give much better convergence performance than others.

Designing appropriate mappings for RJMCMC is not an obvious task. For $\log(N) - \log(S)$ problem, the MCMC chain of the sampler must be able to move to and from any model with 0, 1, 2, or more breakpoints. The easiest mapping to consider is the identity mapping. However, it is not guaranteed to produce efficient sampling. It is difficult to come up with effective proposal distribution for trans-dimensional jumps. Nevertheless, we believe that the application of RJMCMC can provide a useful extension to our method.

The specification of single and broken power-law forms for the $\log(N) - \log(S)$ relationship are physically motivated, but we may wish to consider more flexible forms to allow for possible contamination of populations. To achieve this, one may consider a general mixture of Pareto distributions for the flux. Ideally, we would be able to estimate the mixture probabilities and the

minimum flux of each Pareto mixture, but practically this may be infeasible. We have developed the mixture Pareto model for fixed minimum flux vector, however, the estimation of this parameter requires more work. Our current simulations have not attained convergence. We believe that alternate MCMC sampling procedures, such as ancillary-sufficiency interweaving strategy (Yu and Meng, 2011) or Hamiltonian MCMC (Neil, 2011) may be used to boost MCMC efficiency.

Accounting for contamination of flux populations via a mixture Pareto flux distribution is reasonable, however, it does not provide a flexible enough model to account all variations of slopes in $\log(N) - \log(S)$ relationship. Various configurations of mixture proportions and slopes result in extremely smoothed "elbows" in $\log(N) - \log(S)$. At the same time, it is impossible to achieve a sharp "elbows" using this model. Hence, other models are needed. One may specify a more flexible functional form for $p(S_{com}|\theta, \tau)$, such as higher order polynomial or other smoother. These models can be embedded within the broader Bayesian computational framework, however, executing the integration based on these models will require much work.

2.8.2. Model Selection Performance. The ultimate goal of model selection is two-fold; it is desirable to select the correct model and for the selected model to predict well. In practice, a single model selection rule is not always capable of achieving both sides of the goal. Thus, selection consistency can be interpreted by focusing on either the consistent model selection or prediction. In the former sense, the consistency of a model selection criterion means that the probability of the selected model equal to the true model converges to 1. This definition is obvious assuming that the true model is approximately one of the candidate models. In the latter sense, assuming that none of the candidate models can possibly match the complexity of the true model, consistency means that prediction of the future observations is the best for the chosen model among all other models.

Model selection rules AIC, DIC, and WAIC are designed to predict optimally. It is not surprising that their performance in selecting the correct model is poor, as we have shown in section 2.5. However, optimal prediction is not the main goal for $\log(N) - \log(S)$ problem. In many situations, the estimates of the low-flux component of the broken-Pareto model are fall below the reasonable detection limits. As a result, the missing data can considerably influence the lower flux tail of $\log(N) - \log(S)$ curve. The prediction criterion is usually in favor of such "overfitted" models. In addition, our simulations show that the DIC and WAIC can have very similar measures across the candidate models, which implies that they may be not sensitive enough to capture differences

between models in some cases. Thus, we believe that DIC and WAIC are not appropriate for the selection of the number of breakpoints.

Conversely, the BAF method does not discriminate between the goals of model selection because the selected model is closest to the true model implying good predictive performance. The original fence method (Jiang et al., 2008) is designed to select the optimal model in two steps. The first is to identify the set of true or best approximating models by utilizing a measure of goodness-of-fit or best prediction. The second is to select a model based on some optimality criterion, which need not be unique. For $\log(N) - \log(S)$ problem the approximate subsets nature of candidate models gives natural optimality criterion of minimum dimension. The final selected model is optimal in the sense that it cannot be further reduced or simplified, and consistent if $n \to \infty$ (Jiang et al., 2008). In Bayesian application, consistency is not assured, but may be approximate under some regularity conditions. We defer the proof of this result to future work. Our simulations show an improvement in model selection over DIC and WAIC. We therefore promote the use of BAF method for model selection in Bayesian inference.

From a practical point of view, the Bayesian adaptive fence method has a few obvious disadvantages. First, the adaptive nature of the fence method requires high computational cost because each candidate model needs to be fitted to each bootstrap dataset via MCMC. The $\log(N) - \log(S)$ method is already computationally intensive. Serial evaluation of the posterior all bootstrap resamples on a single computing machine is out of question. However, if one has parallel computing capabilities, the computation time can be reduced to the computing time of a single MCMC fit.

Second, the device of introducing the most parsimonious but incorrect model, M_0, to serve as a boundary of candidate model space can be considered as ad hoc. It can be argued that in reality the true model and the true parameters are unknown, therefore it would seem strange to label model M_0 as incorrect. Also, even if the true model was a minimal dimension model (single Pareto model M_1), for which the parameter estimates are readily available, the definition of an "incorrect" model is ambiguous. Say, if the true model is a single Pareto model with true $\theta = 0.5$, then an incorrect model would be any model for which $\theta \neq 0.5$. The choices $\theta = 0.3$ and $\theta = 0.7$ both mis-specify the models to be incorrect, yet the model selection result via BAF can be very different for these models. Then which of the two mis-specified models is appropriate to be used as M_0? We argue that both of them are appropriate. From this point of view, it is most useful to fit model M_1 to get an estimate of θ, and then try a number of M_0 model candidates with θ specified in the neighborhood

of M_1 estimate and look for consistency, i.e., the middle ground. The values of θ that are specified too high or too low will produce the measure of goodness-of-fit, Q, that is too large compared to Q of other candidate models. Since our goal is to select a model inside the candidate set but not the boundary model M_0, the middle ground will necessarily produce a model on the inside.

Third, the standard non-parametric bootstrap (resampling of size n with replacement) has a potential to produce a sub-optimal result in model selection. Bootstrapping is a particularly useful tool for estimating the distribution of any statistic when the original data can be considered an independent identically distributed sample. In Bayesian inference, the estimators are produced based on posterior MCMC draws instead of the data alone. Furthermore, in $\log(N)-\log(S)$ problem, the photon count observations are independent but are not identically distributed. Therefore, the distributional properties of the bootstrap samples are a slightly different from that of the original data. Efron observed that the total new information carried from the original data to the bootstrap resample with replacement is roughly $(1 - e^{-1})n = 0.632n$ (Efron, 1983). That is, only 63.2% of the observations in the resample are truly independent, which limits the finite sample performance of resulting statistics (estimating the probability that the optimal model is in the fence). Possible extension to improve the sampling properties of the bootstrap is to use the sequential resampling scheme due to Rao et al. (1997). The sequential resampling proceeds with drawing observations sequentially one at a time, randomly and with replacement, until at least $0.632n$ distinct observations are collected. Asymptotic consistency of this sampling scheme has been established. We consider examination of this and other bootstrap schemes to achieve the same goal as future work.

The ideas of this chapter are important to the astrophysical community because it provides a unified Bayesian approach for parameter estimation and its uncertainty. The method is unique in application to estimation of $\log(N) - \log(S)$ by appropriate treatment of missing data. Similar approaches can be applied in other astronomical surveys with missing data.

APPENDIX A

Appendix A: Single-Pareto Model for $\log(N) - \log(S)$

In this appendix we provide proofs and sketches of derivations necessary for chapter 1.

A.1. Proof of Lemma 1

PROOF OF LEMMA 1. Given the survival function of G, we can easily derive the distribution function G and the density function g of S_i.

$$G(s) = \Pr(S_i < s) = 1 - \Pr(S_i > s) = \begin{cases} 1 - \alpha \cdot s^{-\theta}, & \text{if } s > \tau \\ 0, & \text{if } s \leq \tau \end{cases}$$

$$g(s) = \frac{d}{ds} G(s) = \theta \cdot \alpha \cdot s^{-\theta-1}, \text{ if } s > \tau, \text{ and 0 elsewhere.}$$

With total probability rule, we obtain the required proportionality constant.

$$\int_{-\infty}^{\infty} g(s)ds = \int_{\tau}^{\infty} \theta \cdot \alpha \cdot s^{-\theta-1} ds = \alpha \cdot \tau^{-\theta} = 1$$

Which implies that $\alpha = \tau^{\theta}$. Hence, $G(s) = 1 - \left(\frac{s}{\tau}\right)^{-\theta}$. By uniqueness of the distribution function, G is a Pareto distribution. □

A.2. Model Assumptions for Single Pareto Model

The distributional assumptions are listed below. These components will be useful for derivation of the posterior distribution. We fill-in the gaps to derivations of section 1.3.

$$p(N) = \begin{pmatrix} N + a_N - 1 \\ a_N - 1 \end{pmatrix} \left(\frac{1}{1+b_N}\right)^N \left(\frac{b_N}{1+b_N}\right)^{a_N} \quad \sim \quad \text{Neg-Bin}(N; a_N, b_N)$$

$$p(n|N,\theta) = \begin{pmatrix} N \\ n \end{pmatrix} (\pi(\theta))^n (1 - \pi(\theta))^{N-n} \quad \sim \quad \text{Binomial}(n; N, \pi(\theta))$$

$$p(\theta) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta} \quad \sim \quad \text{Gamma}(\theta; a, b)$$
$$p(\tau) = \frac{b_m^{a_m}}{\Gamma(a_m)}\tau^{a_m-1}e^{-b_m\tau} \quad \sim \quad \text{Gamma}(\tau; a_m, b_m)$$

$$g(S, B, L, E) = \Pr(I = 1|S, B, L, E)$$
$$\pi(\theta, \tau) = \int p(I = 1|S, B, L, E) \cdot p(S, B, L, E|\theta, \tau)\ dS\ dB\ dE\ dL$$
$$= \int g(S, B, L, E) \cdot p(S, B, L, E|\theta, \tau)\ dS\ dB\ dE\ dL$$

$$p(S_{obs}|n, N, \theta, \tau, B_{obs}, L_{obs}, E_{obs}) = \prod_{i=1}^{n} p(S_{obs,i}|n, N, \theta, \tau, B_{obs,i}, L_{obs,i}, E_{obs,i})$$
$$= \prod_{i=1}^{n} p(S_i, I_i = 1|n, \theta, \tau, B_i, L_i, E_i)$$
$$= \prod_{i=1}^{n} [p(S_i|n, \theta, \tau)\, p(I_i = 1|S_i, B_i, L_i, E_i)]$$
$$= \prod_{i=1}^{n} \left[\theta\tau^{-1}\left(\frac{S_i}{\tau}\right)^{-(\theta+1)} g(S_i, B_i, L_i, E_i)\right]$$
$$\sim \prod_{i=1}^{n} \text{Pareto}(S_i; \theta, \tau) g(S_i, B_i, L_i, E_i)$$

$$p(S_{mis}|n, N, \theta, \tau, B_{obs}, L_{obs}, E_{obs}) = \prod_{i=1}^{n} p(S_{mis,i}|n, N, \theta, \tau, B_{obs,i}, L_{obs,i}, E_{obs,i})$$
$$= \prod_{i=1}^{N-n} p(S_i, I_i = 0|n, N, \theta, \tau, B_i, L_i, E_i)$$
$$= \prod_{i=1}^{N-n} [p(S_i|n, N, \theta, \tau)\, p(I_i = 0|S_i, B_i, L_i, E_i)]$$
$$= \prod_{i=1}^{N-n} \left[\theta\tau^{-1}\left(\frac{S_i}{\tau}\right)^{-(\theta+1)} (1 - g(S_i, B_i, L_i, E_i))\right]$$
$$\sim \prod_{i=1}^{N-n} \text{Pareto}(S_i; \theta, \tau)(1 - g(S_i, B_i, L_i, E_i))$$

Observe that, in the case of "step"-function g:

$$
\begin{aligned}
&p\left(S_{obs}|n,N,\theta,\tau,B_{obs},L_{obs},E_{obs}\right)=\prod_{i=1}^{n}p\left(S_{obs,i}|n,N,\theta\right)=\prod_{i=1}^{n}p\left(S_i,I_i|n,\theta\right)\\
&=\prod_{i=1}^{n}p\left(S_i|n,\theta\right)\mathbb{I}\{S_i>C_i(E_i)\}\\
&=\prod_{i=1}^{n}p\left(S_i|n,\theta,\tau\right)\mathbb{I}\{S_i>C_i(E_i),S_i>\tau\}\\
&=\prod_{i=1}^{n}p\left(S_i|n,\theta,\tau,C_i(E_i)\right)\mathbb{I}\{S_i>\max\{\tau,C_i(E_i)\}\}\\
&\sim\text{Pareto}\left(\theta,\max\{\tau,C_i(E_i)\}\right),\ \text{for each } i=1,\ldots,n.
\end{aligned}
$$

Let $\lambda_i=\lambda(S_{obs,i},B_{obs,i},E_{obs,i},L_{obs,i})$ and $k_i=k(B_{obs,i},E_{obs,i},L_{obs,i})$.

$$
\begin{aligned}
&p\left(Y_{obs}^{tot}|n,N,S_{obs}\right)=\prod_{i=1}^{n}p\left(Y_{obs,i}^{tot}|n,N,S_{obs,i}\right)\\
&=\prod_{i=1}^{n}\frac{\left(\lambda_i+k_i\right)^{Y_{obs,i}^{tot}}e^{-(\lambda_i+k_i)}}{Y_{obs,i}^{tot}!}\\
&\sim\prod_{i=1}^{n}\text{Poisson}\left(Y_{obs,i}^{tot};\lambda_i+k_i\right)\\
&p\left(Y_{obs}^{src}|n,N,Y_{obs}^{tot},S_{obs}\right)=\prod_{i=1}^{n}p\left(Y_{obs,i}^{src}|n,N,Y_{obs,i}^{tot},S_{obs,i}\right)\\
&=\prod_{i=1}^{n}p(Y_{obs,i}^{src},Y_{obs,i}^{tot}|n,N,S_{obs,i})\frac{1}{p(Y_{obs,i}^{tot})}
\end{aligned}
$$

Let $X=Y_{obs,i}^{src}\sim\text{Poisson}\left(\lambda_i\right),\ V=Y_{obs,i}^{bkg}\sim\text{Poisson}\left(k_i\right),\ W=Y_{obs,i}^{tot}=X+V,\ X\perp V.$

Then we have,

$$
\begin{aligned}
&p_{X|W=w}(x)=\frac{\Pr(X=x,W=w)}{\Pr(W=w)}=\frac{\Pr(X=x)\Pr(V=w-x)}{\Pr(W=w)}\\
&=\frac{\frac{\lambda_i^x e^{-\lambda_i}}{x!}\cdot\frac{k_i^{w-x}e^{-k_i}}{(w-x)!}}{\frac{(\lambda_i^u+k_i)^w e^{-(\lambda_i+k_i)}}{w!}}=\frac{w!}{x!(w-x)!}\left(\frac{\lambda_i}{\lambda_i+k_i}\right)^x\left(\frac{k_i}{\lambda_i+k_i}\right)^{w-x}
\end{aligned}
$$

Hence, $X|W=w\sim\text{Binomial}\left(X;w,\frac{\lambda_i}{\lambda_i+k_i}\right).$

$$
\begin{aligned}
p\left(Y_{obs}^{src}|n,N,Y_{obs}^{tot},S_{obs}\right) &= \prod_{i=1}^{n} p(Y_{obs,i}^{src},Y_{obs,i}^{tot}|n,N,S_{obs,i})\frac{1}{p(Y_{obs,i}^{tot})} \\
&= \prod_{i=1}^{n} \frac{\lambda_i^{Y_{obs,i}^{src}} k^{(Y_{obs,i}^{tot}-Y_{obs,i}^{src})} e^{-(\lambda_i+k_i)}}{Y_{obs,i}^{src}!(Y_{obs,i}^{tot}-Y_{obs,i}^{src})!} \frac{Y_{obs,i}^{tot}!}{(\lambda_i+k_i)^{Y_{obs,i}^{tot}} e^{-(\lambda_i+k_i)}} \\
&= \prod_{i=1}^{n} \frac{Y_{obs,i}^{tot}!}{Y_{obs,i}^{src}!(Y_{obs,i}^{tot}-Y_{obs,i}^{src})!} \left(\frac{\lambda_i}{\lambda_i+k_i}\right)^{Y_{obs,i}^{src}} \left(\frac{k_i}{\lambda_i+k_i}\right)^{Y_{obs,i}^{tot}-Y_{obs,i}^{src}} \\
&= \prod_{i=1}^{n} \begin{pmatrix} Y_{obs,i}^{tot} \\ Y_{obs,i}^{src} \end{pmatrix} \left(\frac{\lambda_i}{\lambda_i+k_i}\right)^{Y_{obs,i}^{src}} \left(1-\frac{\lambda_i}{\lambda_i+k_i}\right)^{Y_{obs,i}^{tot}-Y_{obs,i}^{src}} \\
&\sim \prod_{i=1}^{n} \text{Binomial}\left(Y_{obs,i}^{src}; Y_{obs,i}^{tot}, \frac{\lambda_i}{\lambda_i+k_i}\right)
\end{aligned}
$$

$$
\begin{aligned}
p\left(Y_{mis}^{tot}|n,N,S_{mis}\right) &= \prod_{i=1}^{N-n} p\left(Y_{mis,i}^{tot}|n,N,S_{mis,i}\right) \\
&= \prod_{i=1}^{N-n} \frac{(\lambda_i+k_i)^{Y_{mis,i}^{tot}} e^{-(\lambda_i+k_i)}}{Y_{mis,i}^{tot}!} \\
&\sim \prod_{i=1}^{N-n} \text{Poisson}\left(Y_{mis,i}^{tot}; \lambda_i+k_i\right)
\end{aligned}
$$

$$
\begin{aligned}
p\left(Y_{mis}^{src}|n,N,Y_{mis}^{tot},S_{mis}\right) &= \prod_{i=1}^{N-n} p\left(Y_{mis,i}^{src}|n,N,Y_{mis,i}^{tot},S_{mis,i}\right) \\
&= \prod_{i=1}^{N-n} p(Y_{mis,i}^{src},Y_{mis,i}^{tot}|n,N,S_{mis,i})\frac{1}{p(Y_{mis,i}^{tot})} \\
&= \prod_{i=1}^{N-n} \frac{\lambda_i^{Y_{mis,i}^{src}} k^{(Y_{mis,i}^{tot}-Y_{mis,i}^{src})} e^{-(\lambda_i+k_i)}}{Y_{mis,i}^{src}!(Y_{mis,i}^{tot}-Y_{mis,i}^{src})!} \frac{Y_{mis,i}^{tot}!}{(\lambda_i+k_i)^{Y_{mis,i}^{tot}} e^{-(\lambda_i+k_i)}} \\
&= \prod_{i=1}^{N-n} \begin{pmatrix} Y_{mis,i}^{tot} \\ Y_{mis,i}^{src} \end{pmatrix} \left(\frac{\lambda_i}{\lambda_i+k_i}\right)^{Y_{mis,i}^{src}} \left(1-\frac{\lambda_i}{\lambda_i+k_i}\right)^{Y_{mis,i}^{tot}-Y_{mis,i}^{src}} \\
&\sim \prod_{i=1}^{N-n} \text{Binomial}\left(Y_{mis,i}^{src}; Y_{mis,i}^{tot}, \frac{\lambda_i}{\lambda_i+k_i}\right)
\end{aligned}
$$

A.3. Derivation of Posterior Distribution for Single Pareto Model

Using above calculations, we now derive the joint posterior distribution of the parameters of interest.

$$
\begin{aligned}
&p\left(N, \theta, \tau, S_{obs}, Y_{obs}^{src} | n, Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs}\right) \\
&= \frac{1}{p(n, Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs})} \cdot p\left(N, \theta, \tau, S_{obs}, Y_{obs}^{src}, Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs}, n\right) \\
&= \frac{1}{p(n, Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs})} \cdot \\
&\quad \cdot \int p\left(N, \theta, \tau, I_{com}, S_{com}, Y_{com}^{src}, Y_{com}^{tot}, B_{com}, L_{com}, E_{com}, n\right) \\
&\qquad dI_{mis}\ dS_{mis}\ dB_{mis}\ dL_{mis}\ dE_{mis}\ dY_{mis}^{src}\ dY_{mis}^{tot} \\
&= \frac{1}{p(n, Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs})} \cdot p\left(N, \theta, \tau\right) \\
&\quad \cdot \int p\left(n, I_{com}, S_{com}, Y_{com}^{src}, Y_{com}^{tot}, B_{com}, L_{com}, E_{com} | N, \theta, \tau\right) \\
&\qquad dI_{mis}\ dS_{mis}\ dB_{mis}\ dL_{mis}\ dE_{mis}\ dY_{mis}^{src}\ dY_{mis}^{tot} \\
&= \frac{1}{p(n, Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs})} \cdot p\left(N, \theta, \tau\right) \\
&\quad \cdot \int p\left(I_{com}, S_{com}, Y_{com}^{src}, Y_{com}^{tot}, B_{com}, L_{com}, E_{com} | N, \theta, \tau\right) \cdot \mathbb{I}_{\{\sum_{i=1}^{N} I_i = n\}} \\
&\qquad dI_{mis}\ dS_{mis}\ dB_{mis}\ dL_{mis}\ dE_{mis}\ dY_{mis}^{src}\ dY_{mis}^{tot} \\
&= \frac{1}{p(n, Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs})} \cdot p\left(N, \theta, \tau\right) \\
&\quad \cdot \int_{\mathcal{A}} p\left(I_{com}, S_{com}, Y_{com}^{src}, Y_{com}^{tot}, B_{com}, L_{com}, E_{com} | N, \theta, \tau\right) \\
&\qquad dI_{mis}\ dS_{mis}\ dB_{mis}\ dL_{mis}\ dE_{mis}\ dY_{mis}^{src}\ dY_{mis}^{tot}
\end{aligned}
$$

where $\mathcal{A} = \{$all permutations of vector I of length N with entries $I_j \in \{0, 1\}$, s.t. $\sum_j I_j = n\}$

Note: how do we know which components S_i of S_{com} are the missing sources? i.e. how to perform such integral w.r.t. I_{mis}? We integrate all $N - n$ components S_i for which $I_i = 0$. That is, integrate across all permutations of I_{com} vector of length N in which there are exactly n 1's. Consider integrating out the first missing source, supposedly if $N - n = 1$; there are N possibilities that there

is a source with $I_i = 0$ and all other components are 1. In this special case,

$$\begin{aligned}
&\int_{\mathcal{A}} p\left(I_{com}, S_{com}, Y_{com}^{src}, Y_{com}^{tot}, B_{com}, L_{com}, E_{com} | N, \theta, \tau\right) \\
&\qquad dI_{mis,1}\ dS_{mis,1}\ dB_{mis,1}\ dL_{mis,1}\ dE_{mis,1}\ dY_{mis,1}^{src}\ dY_{mis,1}^{tot} \\
&= \int_{\mathcal{A}} \prod_{i=1}^{N} p(I_i, S_i, B_i, L_i, E_i, Y_i^{tot}, Y_i^{src} | N, \theta, \tau) \\
&\qquad dI_{mis,1}\ dS_{mis,1}\ dB_{mis,1}\ dL_{mis,1}\ dE_{mis,1}\ dY_{mis,1}^{src}\ dY_{mis,1}^{tot} \\
&= N \cdot \int \prod_{i=2}^{N} p(I_i = 1, S_i, B_i, L_i, E_i, Y_i^{tot}, Y_i^{src} | N, \theta, \tau) \\
&\quad \cdot p(I_1 = 0, S_1, B_1, L_1, E_1, Y_1^{tot}, Y_1^{src} | N, \theta, \tau)\ dS_1\ dB_1\ dL_1\ dE_1\ dY_1^{src}\ dY_1^{tot} \\
&= N \cdot p(I_{obs}, S_{obs}, B_{obs}, L_{obs}, E_{obs}, Y_{obs}^{src}, Y_{obs}^{tot} | N, \theta, \tau) \\
&\quad \cdot \int p(B_1, L_1, E_1) \cdot \Bigg[\int p(S_1 | \theta, \tau) \cdot p(I_1 = 0 | S_1, B_1, L_1, E_1)\ dS_1 \\
&\qquad \cdot \int p(Y_1^{tot} | N, \theta, S_1, B_1, L_1, E_1)\ dY_1^{tot} \\
&\qquad \cdot \int p(Y_1^{src} | N, \theta, S_1, B_1, L_1, E_1, Y_1^{tot})\ dY_1^{src} \Bigg]\ dB_1\ dL_1\ dE_1 \\
&= N \cdot p(I_{obs}, S_{obs}, B_{obs}, L_{obs}, E_{obs}, Y_{obs}^{src}, Y_{obs}^{tot} | N, \theta, \tau) \\
&\quad \cdot \int p(B_1, L_1, E_1) \cdot \Bigg\{ \int (1 - g(S_1, B_1, L_1, E_1))\, p(S_1 | \theta, \tau)\ dS_1 \cdot \\
&\qquad \cdot \left[\sum_{y_1^{tot}=0}^{\infty} \frac{(\lambda_1 + k_1)^{y_1^{tot}}}{y_1^{tot}!} e^{(\lambda_1 + k_1)} \right] \\
&\qquad \cdot \left[\sum_{y_1^{src}=0}^{y_1^{tot}} \begin{pmatrix} y_1^{tot} \\ y_1^{src} \end{pmatrix} \left(\frac{\lambda_1}{\lambda_1 + k_1} \right)^{y_1^{src}} \left(1 - \frac{\lambda_1}{\lambda_1 + k_1} \right)^{y_1^{tot} - y_1^{src}} \right] \Bigg\}\ dB_1\ dL_1\ dE_1 \\
&\text{(with } \lambda_1 \equiv \lambda(S_1, B_1, L_1, E_1) \text{ and } k_1 \equiv k(B_1, L_1, E_1)) \\
&= N \cdot p(I_{obs}, S_{obs}, B_{obs}, L_{obs}, E_{obs}, Y_{obs}^{src}, Y_{obs}^{tot} | N, \theta, \tau) \\
&\quad \cdot \int (1 - g(S_1, B_1, L_1, E_1)) \cdot p(S_1 | \theta, \tau) \cdot p(B_1, L_1, E_1)\ dS_1\ dB_1\ dL_1\ dE_1 \\
&= N \cdot p(n, S_{obs}, B_{obs}, L_{obs}, E_{obs}, Y_{obs}^{src}, Y_{obs}^{tot} | N, \theta, \tau) \cdot (1 - \pi(\theta, \tau))
\end{aligned}$$

Now consider that $N-n=2$ and we integrate out both of these missing sources. There are $\begin{pmatrix} N \\ 2 \end{pmatrix}$ arrangements of I containing exactly two 0's. Also, by independence of sources, the integral over all other missing variables will simplify to a product of two $1-\pi$ values: $(1-\pi(\theta,\tau))^2$. Continuing in this manner we get the following result for $N-n$ missing sources.

$$
\begin{aligned}
& p\left(N,\theta,\tau,S_{obs},Y_{obs}^{src}|n,Y_{obs}^{tot},B_{obs},L_{obs},E_{obs}\right) \\
&= \frac{1}{p(n,Y_{obs}^{tot},B_{obs},L_{obs},E_{obs})}\cdot p\left(N,\theta,\tau\right) \\
&\quad \cdot \int_{\mathcal{A}} p\left(I_{com},S_{com},Y_{com}^{src},Y_{com}^{tot},B_{com},L_{com},E_{com}|N,\theta,\tau\right) \\
&\qquad dI_{mis}\ dS_{mis}\ dB_{mis}\ dL_{mis}\ dE_{mis}\ dY_{mis}^{src}\ dY_{mis}^{tot} \\
&= \frac{1}{p(n,Y_{obs}^{tot},B_{obs},L_{obs},E_{obs})}\cdot p\left(N,\theta,\tau\right) \\
&\quad \cdot \begin{pmatrix} N \\ n \end{pmatrix} \mathbb{I}_{\{n\leq N\}}\cdot p(I_{obs},S_{obs},B_{obs},L_{obs},E_{obs},Y_{obs}^{src},Y_{obs}^{tot}|N,\theta,\tau)\cdot\ (1-\pi(\theta,\tau))^{(N-n)} \\
&= \frac{1}{p(n,Y_{obs}^{tot},B_{obs},L_{obs},E_{obs})}\cdot \begin{pmatrix} N \\ n \end{pmatrix} \mathbb{I}_{\{n\leq N\}}\cdot\ (1-\pi(\theta,\tau))^{(N-n)} \\
&\quad \cdot p\left(N\right)\cdot p\left(\theta|N\right)\cdot p\left(\tau|N,\theta\right)\cdot p\left(B_{obs},L_{obs},E_{obs}|N,\theta,\tau\right) \\
&\quad \cdot p\left(I_{obs},S_{obs}|N,\theta,\tau,B_{obs},L_{obs},E_{obs}\right) \\
&\quad \cdot p\left(Y_{obs}^{tot}|N,\theta,\tau,B_{obs},L_{obs},E_{obs},I_{obs},S_{obs}\right) \\
&\quad \cdot p\left(Y_{obs}^{src}|Y_{obs}^{tot},N,\theta,\tau,B_{obs},L_{obs},E_{obs},I_{obs},S_{obs}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{p(n, Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs})} \cdot \begin{pmatrix} N \\ n \end{pmatrix} \mathbb{I}_{\{n \leq N\}} \cdot \left(1 - \pi(\theta, \tau)\right)^{(N-n)} \\
&\quad \cdot p\left(N\right) \cdot p\left(\theta\right) \cdot p\left(\tau\right) \cdot p\left(B_{obs}, L_{obs}, E_{obs}\right) \cdot p\left(S_{obs} | N, \theta, \tau\right) \cdot p\left(I_{obs} | S_{obs}, B_{obs}, L_{obs}, E_{obs}\right) \\
&\quad \cdot p\left(Y_{obs}^{tot} | B_{obs}, L_{obs}, E_{obs}, I_{obs}, S_{obs}\right) \cdot p\left(Y_{obs}^{src} | Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs}, I_{obs}, S_{obs}\right) \\
&= \frac{1}{p(n, Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs})} \cdot \begin{pmatrix} N \\ n \end{pmatrix} \mathbb{I}_{\{n \leq N\}} \cdot \left(1 - \pi(\theta, \tau)\right)^{(N-n)} \\
&\quad \cdot \begin{pmatrix} N + a_N - 1 \\ a_N - 1 \end{pmatrix} \left(\frac{1}{1 + b_N}\right)^{N} \left(\frac{b_N}{1 + b_N}\right)^{a_N} \mathbb{I}_{\{N \in \mathbb{Z}^+\}} \\
&\quad \cdot \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} \mathbb{I}_{\{\theta > 0\}} \\
&\quad \cdot \frac{b_m^{a_m}}{\Gamma(a_m)} \tau^{a_m - 1} e^{-b_m \tau} \mathbb{I}_{\{\tau > 0\}} \\
&\quad \cdot \left[\prod_{i=1}^{n} p\left(B_i, L_i, E_i\right) \cdot \theta \tau^{\theta} S_i^{-(\theta+1)} \mathbb{I}_{\{\tau < S_i\}} \cdot g(S_i, B_i, L_i, E_i) \right. \\
&\quad \cdot \frac{(\lambda_i + k_i)^{Y_i^{tot}}}{Y_i^{tot}!} e^{(\lambda_i + k_i)} \mathbb{I}_{\{Y_i^{tot} \in \mathbb{Z}^+\}} \\
&\quad \left. \cdot \begin{pmatrix} Y_i^{tot} \\ Y_i^{src} \end{pmatrix} \left(\frac{\lambda_i}{\lambda_i + k_i}\right)^{Y_i^{src}} \left(1 - \frac{\lambda_i}{\lambda_i + k_i}\right)^{Y_i^{tot} - Y_i^{src}} \mathbb{I}_{\left\{Y_i^{src} \in \{0, 1, \ldots, Y_i^{tot}\}\right\}} \right]
\end{aligned}
$$

with $\lambda_i \equiv \lambda(S_i, B_i, L_i, E_i)$ and $k_i \equiv k(B_i, L_i, E_i)$.

A.4. Full-Conditional Distributions for Single Pareto Model

A.4.1. Sampling Y_{obs}^{src}: Deriving the full conditional distribution for Y_{obs}^{src}, we arrive at:

$$p\left(Y_{obs}^{src}|\cdot\right) \propto p(Y_{obs}^{src}|Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs}, I_{obs}, S_{obs})$$
$$= \prod_{i=1}^{n} \left(\begin{array}{c} Y_i^{tot} \\ Y_i^{src} \end{array} \right) \left(\frac{\lambda_i}{\lambda_i + k_i}\right)^{Y_i^{src}} \left(1 - \frac{\lambda_i}{\lambda_i + k_i}\right)^{Y_i^{tot} - Y_i^{src}} \mathbb{I}_{\left\{Y_i^{src} \in \{0,1,\ldots,Y_i^{tot}\}\right\}}$$

Thus, by independence of observed sources, we can sample vector Y_{obs}^{src} component-wise:
for $i = 1, \ldots, n$,

$$p\left(Y_i^{src}|\cdot\right) \propto p(Y_i^{src}|Y_i^{tot}, S_i, B_i, L_i, E_i)$$
$$\sim \text{Binomial}\left(Y_i^{src}; Y_i^{tot}, \frac{\lambda(S_i, B_i, L_i, E_i)}{\lambda(S_i, B_i, L_i, E_i) + k(B_i, L_i, E_i)}\right)$$

A.4.2. Sampling S_{obs}: Deriving the full conditional distribution for S_{obs}, we arrive at:

$$p\left(S_{obs}|\cdot\right) \propto p\left(S_{obs}|N, \theta, \tau\right) \cdot p\left(I_{obs}|S_{obs}, B_{obs}, L_{obs}, E_{obs}\right) \cdot$$
$$\cdot p\left(Y_{obs}^{tot}|B_{obs}, L_{obs}, E_{obs}, I_{obs}, S_{obs}\right) \cdot p\left(Y_{obs}^{src}|Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs}, I_{obs}, S_{obs}\right)$$
$$= \left[\prod_{i=1}^{n} \theta\tau^{\theta} S_i^{-(\theta+1)} \mathbb{I}_{\{\tau < S_i\}} \cdot g(S_i, B_i, L_i, E_i) \cdot \frac{(\lambda_i + k_i)^{Y_i^{tot}}}{Y_i^{tot}!} e^{(\lambda_i + k_i)} \mathbb{I}_{\{Y_i^{tot} \in \mathbb{Z}^+\}} \cdot \right.$$
$$\left. \cdot \left(\begin{array}{c} Y_i^{tot} \\ Y_i^{src} \end{array} \right) \left(\frac{\lambda_i}{\lambda_i + k_i}\right)^{Y_i^{src}} \left(1 - \frac{\lambda_i}{\lambda_i + k_i}\right)^{Y_i^{tot} - Y_i^{src}} \mathbb{I}_{\left\{Y_i^{src} \in \{0,1,\ldots,Y_i^{tot}\}\right\}} \right]$$

Thus, by independence of observed sources, we can sample vector S_{obs} component-wise:
for $i = 1, \ldots, n$,

$$p\left(S_i|\cdot\right) \propto p\left(S_i|N, \theta, \tau\right) \cdot p\left(I_i = 1|S_i, B_i, L_i, E_i\right) \cdot p\left(Y_i^{tot}|S_i, B_i, L_i, E_i\right) \cdot$$
$$\cdot p\left(Y_i^{src}|Y_i^{tot}, S_i, B_i, L_i, E_i\right)$$
$$\sim \text{Pareto}\left(S_i; \theta, \tau\right) \cdot g(S_i, B_i, L_i, E_i) \cdot \text{Poisson}\left(Y_i^{tot}; \lambda(S_i, B_i, L_i, E_i) + k(B_i, L_i, E_i)\right) \cdot$$
$$\cdot \text{Binomial}\left(Y_i^{src}; Y_i^{tot}, \frac{\lambda(S_i, B_i, L_i, E_i)}{\lambda(S_i, B_i, L_i, E_i) + k(B_i, L_i, E_i)}\right)$$

A.4.3. Sampling θ: We now derive the full conditional distribution for θ.

$$
\begin{aligned}
p\left(\theta|\cdot\right) &\propto p\left(\theta\right)\cdot p\left(S_{obs}|N,\theta,\tau\right)\cdot\left(1-\pi(\theta,\tau)\right)^{(N-n)} \\
&= \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}\mathbb{I}_{\{\theta>0\}}\cdot\left[\prod_{i=1}^{n}\theta\tau^{-1}\left(\frac{S_i}{\tau}\right)^{-(\theta+1)}\mathbb{I}_{\{\tau<S_i\}}\right]\cdot\left(1-\pi(\theta,\tau)\right)^{(N-n)} \\
&\propto \theta^{a-1}e^{-\theta b}\cdot\theta^n e^{\sum_{i=1}^{n}\log\left(\frac{S_i}{\tau}\right)^{-\theta}}\mathbb{I}_{\{\theta>0\}}\cdot\left(1-\pi(\theta,\tau)\right)^{(N-n)} \\
&= \theta^{(a+n-1)}e^{-\theta\left[b+\sum_{i=1}^{n}\log\left(\frac{S_i}{\tau}\right)\right]}\mathbb{I}_{\{\theta>0\}}\cdot\left(1-\pi(\theta,\tau)\right)^{(N-n)} \\
&\propto \left(1-\pi(\theta,\tau)\right)^{(N-n)}\cdot\text{Gamma}\left(\theta;a+n,b+\sum_{i=1}^{n}\log\left(\frac{S_i}{\tau}\right)\right)
\end{aligned}
$$

A.4.4. Sampling N: We now derive the full conditional distribution for N.

$$
\begin{aligned}
p\left(N|\cdot\right) &\propto \binom{N}{n}\mathbb{I}_{\{n\leq N\}}\cdot\left(1-\pi(\theta,\tau)\right)^{(N-n)}\cdot p\left(N\right)\cdot p\left(S_{obs}|N,\theta,\tau\right) \\
&= \binom{N}{n}\mathbb{I}_{\{n\leq N\}}\cdot\left(1-\pi(\theta,\tau)\right)^{(N-n)}\cdot\binom{N+a_N-1}{a_N-1}\left(\frac{1}{1+b_N}\right)^N\left(\frac{b_N}{1+b_N}\right)^{a_N}\mathbb{I}_{\{N\in\mathbb{Z}^+\}} \\
&\propto \frac{\Gamma(N+1)}{\Gamma(n+1)\Gamma(N-n+1)}\cdot\left(1-\pi(\theta,\tau)\right)^{(N-n)}\cdot\frac{\Gamma(N+a_N)}{\Gamma(a_N)\Gamma(N+1)}\left(\frac{1}{b_N+1}\right)^N\mathbb{I}_{\{n\leq N\}} \\
&\propto \frac{\Gamma(N+a_N)}{\Gamma(N-n+1)}\cdot\left(\frac{1}{b_N+1}\right)^N\cdot\left(1-\pi(\theta,\tau)\right)^{(N-n)}\mathbb{I}_{\{n\leq N\}}
\end{aligned}
$$

A.4.5. Sampling τ: We now derive the full conditional distribution of τ.

$$\tau\sim\text{Gamma}\left(a_m,b_m\right),\quad a_m,b_m>0,s>0.$$

Recall that for the single power-law model, the flux density is a Pareto:

$$p(s|\theta,\tau)=\theta\tau^{\theta}s^{-(\theta+1)},\quad s>\tau.$$

A gamma prior on τ gives the following full conditional distribution of τ:

$$\begin{aligned}
p(\tau|\cdot) &\propto p(\tau,\theta,N)\cdot p\left(B_{obs},L_{obs},E_{obs}|\tau,\theta,N\right)\cdot p\left(n,S_{obs},I_{obs}|N,\theta,\tau,B_{obs},L_{obs},E_{obs}\right)\\
&\quad\cdot p\left(Y_{obs}^{tot},Y_{obs}^{src}|n,N,\theta,S_{obs},I_{obs},\tau,B_{obs},L_{obs},E_{obs}\right)\\
&= p(\tau)\cdot p(\theta)\cdot p(N)\cdot p\left(B_{obs},L_{obs},E_{obs}\right)\cdot p\left(n,S_{obs},I_{obs}|N,\theta,\tau,B_{obs},L_{obs},E_{obs}\right)\\
&\quad\cdot p\left(Y_{obs}^{tot},Y_{obs}^{src}|n,N,\theta,S_{obs},I_{obs},\tau,B_{obs},L_{obs},E_{obs}\right)\\
&\propto p(\tau)\cdot p\left(n,S_{obs},I_{obs}|N,\theta,\tau,B_{obs},L_{obs},E_{obs}\right)\\
&= \frac{b_m^{a_m}}{\Gamma(a_m)}\tau^{a_m-1}e^{-b_m\tau}\mathbb{I}_{\{\tau>0\}}\\
&\quad\cdot\begin{pmatrix}N\\ n\end{pmatrix}(1-\pi(\theta,\tau))^{N-n}\,\mathbb{I}_{\{n\le N\}}\cdot\prod_{i=1}^{n}\theta\tau^{\theta}S_i^{-(\theta+1)}g(S_i,B_i,L_i,E_i)\mathbb{I}_{\{\tau<S_i\}}\\
&\propto \tau^{n\theta+a_m-1}\cdot e^{-b_m\tau}\cdot(1-\pi(\theta,\tau))^{N-n}\cdot\mathbb{I}_{\{0<\tau<c_m\}},\text{where } c_m=\min\{S_1,\ldots,S_n\},\\
&\propto \text{Gamma}\left(\tau;a_m+n\theta,b_m\right)\cdot(1-\pi(\theta,\tau))^{N-n}\cdot\mathbb{I}_{\{0<\tau<c_m\}}
\end{aligned}$$

The sampling of τ proceeds with the Metropolis-Hastings algorithm. Care must be exercised to make sure samples are drawn in the proper region. In order to preserve the positivity of the parameter and to avoid numerical instability of making samples very close to zero, we first take a logarithm transformation:

$$\begin{aligned}
\eta &= \log(\tau)\\
p(\eta|\cdot) &= e^{\eta}\cdot p(\tau=e^{\eta}|\cdot)\propto e^{\eta(n\theta+a_m+1)}\cdot e^{-b_m e^{\eta}}\cdot(1-\pi(\theta,\tau=e^{\eta}))^{N-n}\cdot\mathbb{I}_{\{\eta<\log(c_m)\}}.
\end{aligned}$$

The upper bound for η is reflected in the truncated normal distribution chosen as the asymmetric jumping distribution.

$$J(\eta^{prop}|\eta^{curr})=\frac{\frac{1}{\sigma}\phi\left(\frac{\eta^{prop}-\eta^{curr}}{\sigma}\right)}{\Phi\left(\frac{\log(c_m)-\eta^{curr}}{\sigma}\right)}.$$

(For standard normal PDF ϕ and CDF Φ.)

The algorithm proceeds as follows. Consider a starting point: $\eta^{curr}=\log(\tau^{curr})$, where τ^{curr} is the current state value of the minimum flux.

ALGORITHM 4. Step 1: Sample a proposal $\eta^{prop} \sim J(\eta^{prop}|\eta^{curr})$.

Step 2: Compute the ratio of densities for MH algorithm:

$$\alpha = \frac{p(\eta^{prop}|data) \cdot J(\eta^{curr}|\eta^{prop})}{p(\eta^{curr}|data) \cdot J(\eta^{prop}|\eta^{curr})} = \frac{e^{\eta^{prop}} \cdot p(\tau^{prop} = e^{\eta^{prop}}| \cdot) \cdot \Phi(\frac{\log(c_m)-\eta^{curr}}{\sigma})}{e^{\eta^{curr}} \cdot p(\tau^{curr} = e^{\eta^{curr}}| \cdot) \cdot \Phi(\frac{\log(c_m)-\eta^{prop}}{\sigma})}$$

Step 3: Draw $U \sim \text{Uniform}\,(0, 1)$.

Step 4: If $U < \alpha$, accept $\eta^{new} = \eta^{prop}$, otherwise keep $\eta^{new} = \eta^{curr}$.

Step 5: Transform back $\tau^{new} = e^{\eta^{new}}$ for the new draw of the minimum flux.

A.4.6. Sampling S_{mis}: The vector of parameters S_{mis} is not required as part of our Gibbs sampler, and it has been shown that we can average over these latent variables in the derivation of the posterior distribution. However, we require to produce the $\log(N) - \log(S)$ plot, thus, we want impute these latent variables. At the same time, we do not want to introduce the dependence of S_{mis} on all parameters in our Gibbs sampler. Instead of re-deriving the posterior distribution with S_{mis} included in the parameter set, we proceed with the unconditional approach. Through model assumption, $S \sim \text{Pareto}\,(\theta, \tau)$, and the probability of observing a missing source is $p(I = 0|S, B, L, E) = 1 - g(S, B, L, E)$. Note that the dimension of S_{mis} vector is $N - n$, that is, it depends directly on the value of N and changes from iteration to iteration.

$$p\left(S_{mis}|n, N, \theta, \tau, B_{mis}, L_{mis}, E_{mis}, I_{mis}\right) = \prod_{i=1}^{N-n} p(S_i, I_i = 0|n, N, \theta, \tau, B_i, L_i, E_i)$$

$$= \prod_{i=1}^{N-n} p(S_i|n, N, \theta, \tau) \cdot p(I_i = 0|S_i, B_i, L_i, E_i) = \prod_{i=1}^{N-n} \theta\tau^{-1} \left(\frac{S_i}{\tau}\right)^{-(\theta+1)} (1 - g(S_i, B_i, L_i, E_i))$$

Hence, for $i = 1, \ldots, N - n$,

$$(B_i, L_i, E_i) \sim p(B_i, L_i, E_i)$$

$$S_i|n, N, \theta, \tau, B_i, L_i, E_i, I_i = 0 \sim (1 - g(S_i, B_i, L_i, E_i)) \cdot \text{Pareto}\,(S_i; \theta, \tau).$$

Sampling is done via rejection sampling method.

APPENDIX B

Appendix B: Broken-Pareto Model for $\log(N) - \log(S)$

In this Appendix we provide a sketch of distributional derivations and whenever necessary description to the computation required for the model in chapter 2.

B.1. Summary of Distributions

(1) Pareto distribution. $X \sim \text{Pareto}\,(\tau_1, \theta)$

$$\text{pdf:} \qquad f_X(x) = \theta \frac{\tau_1^\theta}{x^{\theta+1}}, \qquad x \geq \tau_1.$$

$$\text{cdf:} \qquad F_X(x) = 1 - \frac{\tau_1^\theta}{x^\theta}, \qquad x \geq \tau_1.$$

$$\text{quantile:} \qquad x_q = \frac{\tau_1}{(1-q)^{1/\theta}}.$$

(2) Truncated-Pareto distribution. $X \sim \text{Truncated-Pareto}(\tau_1, \theta, \tau)$

$$\text{pdf:} \qquad f_X(x) = \theta \frac{x^{-(\theta+1)}}{\tau_1^{-\theta} - \tau^{-\theta}}, \qquad \tau_1 \leq x < \tau.$$

$$\text{cdf:} \qquad F_X(x) = \frac{\tau_1^{-\theta} - x^{-\theta}}{\tau_1^{-\theta} - \tau^{-\theta}}, \qquad \tau_1 \leq x < \tau.$$

$$\text{quantile:} \qquad x_q = \left(\tau_1^{-\theta} - q(\tau_1^{-\theta} - \tau^{-\theta})\right)^{-1/\theta}.$$

(3) Mixture of Truncated-Pareto distributions. $Y \sim p_1 X_1 + p_2 X_2 + \cdots + p_m X_m$, where

$$X_j \sim \text{Truncated-Pareto}\,(\tau_j, \theta_j, \tau_{j+1}),$$

$$0 < \tau_1 \leq \tau_2 \leq \cdots \leq \tau_m,$$

$$0 < p_j < 1, \ \sum_{j=1}^{m} p_j = 1,$$

$$p_j = \left[1 - \left(\frac{\tau_{j+1}}{\tau_j}\right)^{-\theta_j}\right] \prod_{i=1}^{j-1} \left(\frac{\tau_{i+1}}{\tau_i}\right)^{-\theta_i}, \ j = 1, \ldots, m.$$

$$\text{pdf:} \qquad f_Y(x) = \sum_{j=1}^{m} \left\{ \prod_{i=1}^{j-1} \left(\frac{\tau_{i+1}}{\tau_i} \right)^{-\theta_i} \right\} \left(\frac{\theta_j}{\tau_j} \right) \left(\frac{x}{\tau_j} \right)^{-(\theta_j+1)} \mathbb{I}_{\{\tau_j \le x < \tau_{j+1}\}}.$$

$$\begin{aligned} \text{cdf:} \qquad F_Y(x) &= \sum_{j=1}^{m} p_j \mathbb{I}_{\{x > \tau_{j+1}\}} + \sum_{j=1}^{m} p_j F_{X_j}(x) \mathbb{I}_{\{\tau_j \le x < \tau_{j+1}\}} \\ &= \sum_{j=1}^{m} \left\{ \prod_{i=1}^{j-1} \left(\frac{\tau_{i+1}}{\tau_i} \right)^{-\theta_i} \right\} \left[1 - \left(\frac{s}{\tau_j} \right)^{-\theta_j} \right] \mathbb{I}_{\{\tau_j \le x < \tau_{j+1}\}}. \end{aligned}$$

(4) Mixture of Pareto distributions. $Y \sim p_1 X_1 + p_2 X_2 + \cdots + p_m X_m$,

where

$$\begin{aligned} &X_j \sim \text{Pareto}\,(\tau_j, \theta_j)\,, \\ &0 < \tau_1 \le \tau_2 \le \cdots \le \tau_m, \\ &0 < p_j < 1,\ \sum_{j=1}^{m} p_j = 1,\ j = 1, \ldots, m. \end{aligned}$$

$$\text{pdf:} \qquad f_Y(x) = \sum_{j=1}^{m} p_j \theta_j \frac{x^{-(\theta+1)}}{\tau_j^{-\theta}} \mathbb{I}_{\{x \ge \tau_j\}}.$$

$$\text{cdf:} \qquad F_Y(x) = \sum_{j=1}^{m} p_j \left(1 - \frac{x^{-\theta_j}}{\tau_j^{-\theta}} \right) \mathbb{I}_{\{x \ge \tau_j\}}.$$

B.2. Proofs of Lemma 2 and Identity for p_j

Proof of Lemma 2. Given the CDF defined by the broken power-law in (2.1), we can represent G in the form:

$$Y = pX_1 + (1-p)X_2 \sim G,$$

where $p \in [0,1]$, $X_1 \in [\tau_1, \tau_2]$ and $X_2 \in (\tau_2, \infty)$. Denote the CDF of X_1 and X_2 by F_1 and F_2, respectively, then we have the following four required conditions on F_1 and F_2:

$$(1) \quad F_1(\tau_1) = 0, \qquad (2) \quad F_1(\tau_2) = 1, \qquad (3) \quad F_2(\tau_2) = 0, \qquad (4) \quad \lim_{s \to \infty} F_2(s) = 1.$$

For simplicity we denote $\alpha_j^* = 10^{\alpha_j}$ for $j = 1, 2$. The above conditions and piecewise linearity yield the following constraints:

$$F_G(s) = \begin{cases} pF_1(s) = 1 - \alpha_1^* s^{-\theta_1} & s \leq \tau_2 \\ p + (1-p)F_2(s) = 1 - \alpha_2^* s^{-\theta_2} & s > \tau_2. \end{cases}.$$

The lower limits (1) and (3) imply $\alpha_1^* = \tau_1^{\theta_1}$ and $\alpha_2^* = (1-p)\tau_2^{\theta_2}$. So far we have:

$$F_1(s) = \frac{1 - \left(\frac{s}{\tau_1}\right)^{-\theta_1}}{p} \tag{B.1}$$

$$F_2(s) = \frac{(1-p) - (1-p)\left(\frac{s}{\tau_2}\right)^{-\theta_2}}{1-p} = 1 - \left(\frac{s}{\tau_2}\right)^{-\theta_2}$$

On the other hand, the upper limit (2) and (B.1) allow us to solve for p:

$$p = 1 - \left(\frac{\tau_2}{\tau_1}\right)^{-\theta_1}.$$

So, the first mixture component distribution is:

$$F_1(s) = \frac{1 - \left(\frac{s}{\tau_1}\right)^{-\theta_1}}{1 - \left(\frac{\tau_2}{\tau_1}\right)^{-\theta_1}}, \qquad \tau_1 < s < \tau_2. \tag{B.2}$$

Thus, we have:

$$Y \sim \left[1 - \left(\frac{\tau_2}{\tau_1}\right)^{-\theta_1}\right] X_1 + \left(\frac{\tau_2}{\tau_1}\right)^{-\theta_1} X_2,$$

where: $X_1 \sim$ Truncated-Pareto $(\tau_1, \theta_1, \tau_2)$ with CDF given by (B.2), and $X_2 \sim$ Pareto (τ_2, θ_2). It is important to note that the continuity constraint restricts the distribution of Y to contain only 4 free parameters instead of 5 (two for each straight line and the break-point location). □

Proof of Identity (2.8). Consider the general m-component broken power-law density of the flux. We begin with the flux mixture $Y \sim p_1 X_1 + p_2 X_2 + \cdots + p_m X_m$, where $X_j \sim \text{Pareto}\,(\tau_j, \theta_j)$, $\tau_1 \leq \tau_2 \leq \cdots \leq \tau_m$, $0 < p_j < 1$, $\sum_{j=1}^m p_j = 1$, $j = 1, \ldots, m$. The CDF of Y is:

$$F_Y(s) = \sum_{j=1}^{m} p_j F_j(s) = \sum_{j=1}^{m} \left\{ 1 - \sum_{i=1}^{j-1} p_i \right\} \left[1 - \left(\frac{s}{\tau_j} \right)^{-\theta_j} \right] \mathbb{I}_{\{\tau_j \leq s < \tau_{j+1}\}}.$$

The density is given by differentiating:

$$f_Y(s) = \sum_{j=1}^{m} \left\{ 1 - \sum_{i=1}^{j-1} p_i \right\} \left(\frac{\theta_j}{\tau_j} \right) \left(\frac{s}{\tau_j} \right)^{-(\theta_j+1)} \mathbb{I}_{\{\tau_j \leq s < \tau_{j+1}\}}.$$

Constraints on F_j give rise to the recursive relationship of p_j, $j = 1, \ldots, m$:

$$p_j = \left[1 - \left(\frac{\tau_{j+1}}{\tau_j} \right)^{-\theta_j} \right] \left(1 - \sum_{i=1}^{j-1} p_i \right). \tag{B.3}$$

Based on this recursive relationship, we can show that:

$$1 - \sum_{i=1}^{j} p_i = p_j \left[\frac{\left(\frac{\tau_{j+1}}{\tau_j} \right)^{-\theta_j}}{1 - \left(\frac{\tau_{j+1}}{\tau_j} \right)^{-\theta_j}} \right]. \tag{B.4}$$

Combining (B.3) and (B.4) we see that:

$$p_{j+1} = p_j \left[1 - \left(\frac{\tau_{j+2}}{\tau_{j+1}} \right)^{-\theta_{j+1}} \right] \left[\frac{\left(\frac{\tau_{j+1}}{\tau_j} \right)^{-\theta_j}}{1 - \left(\frac{\tau_{j+1}}{\tau_j} \right)^{-\theta_j}} \right].$$

Noting the cancellation of successive terms, we obtain:

$$p_j = \left[1 - \left(\frac{\tau_{j+1}}{\tau_j} \right)^{-\theta_j} \right] \prod_{i=1}^{j-1} \left(\frac{\tau_{i+1}}{\tau_i} \right)^{-\theta_i}. \tag{B.5}$$

Lastly, we use (B.5) to show that for $q = 1, \ldots, m$:

$$\begin{aligned} 1 - \sum_{j=1}^{q} p_j &= 1 - \sum_{j=1}^{q} \left\{ \left[1 - \left(\frac{\tau_{j+1}}{\tau_j} \right)^{-\theta_j} \right] \prod_{i=1}^{j-1} \left(\frac{\tau_{i+1}}{\tau_i} \right)^{-\theta_i} \right\} \\ &= 1 - \sum_{j=1}^{q} \left\{ \prod_{i=1}^{j-1} \left(\frac{\tau_{i+1}}{\tau_i} \right)^{-\theta_i} - \prod_{i=1}^{j} \left(\frac{\tau_{i+1}}{\tau_i} \right)^{-\theta_i} \right\}. \end{aligned}$$

Let $c(j) = \prod_{i=1}^{j} \left(\frac{\tau_{i+1}}{\tau_i}\right)^{-\theta_i}$ with restriction that $c(0) = 1$. Then we have:

$$\begin{aligned} 1 - \sum_{j=1}^{q} p_j &= 1 - \sum_{j=1}^{q} \{c(j-1) - c(j)\} \\ &= 1 - (c(0) - c(q)) = c(q) \\ &= \prod_{i=1}^{q} \left(\frac{\tau_{i+1}}{\tau_i}\right)^{-\theta_i}. \end{aligned}$$

□

B.3. Derivation of Posterior Distribution for Broken-Pareto Model

The posterior derivation under the broken-Pareto setting follows the same strategy as that for single Pareto setting. That is, we marginalize across the missing source information to use only the observed data. The only change in the posterior distribution is the p.d.f. model for the flux. It follows that the posterior distribution of the break-point model is:

$$\begin{aligned} &p(N, \theta, \tau, S_{obs}, Y_{obs}^{src} | n, Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs}) \\ &\qquad = \frac{1}{p(n, Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs})} \cdot p\left(N, \theta, \tau\right) \\ &\qquad \cdot \int_{\mathcal{A}} p\left(I_{com}, S_{com}, Y_{com}^{src}, Y_{com}^{tot}, B_{com}, L_{com}, E_{com} | N, \theta, \tau\right) \\ &\qquad\qquad dI_{mis}\ dS_{mis}\ dB_{mis}\ dL_{mis}\ dE_{mis}\ dY_{mis}^{src}\ dY_{mis}^{tot} \end{aligned}$$

where

$$\mathcal{A} \quad = \quad \{\text{all permutations of vector } I \text{ of length } N \text{ with entries } I_j \in \{0,1\}, \text{ s.t.} \sum_j I_j \quad = \quad n.\}$$

Thus

$$\begin{aligned} &p(N, \theta, \tau, S_{obs}, Y_{obs}^{src} | n, Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs}) \\ &\qquad = \frac{1}{p(n, Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs})} \cdot p\left(N, \theta, \tau\right) \cdot \left[\begin{pmatrix} N \\ n \end{pmatrix} \mathbb{I}_{\{n \leq N\}} \cdot \left(1 - \pi(\theta, \tau)\right)^{(N-n)} \right] \\ &\qquad\qquad \cdot p(I_{obs}, S_{obs}, B_{obs}, L_{obs}, E_{obs}, Y_{obs}^{src}, Y_{obs}^{tot} | N, \theta, \tau) \end{aligned}$$

$$= \frac{1}{p(n, Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs})} \cdot \left[\begin{pmatrix} N \\ n \end{pmatrix} \mathbb{I}_{\{n \leq N\}} \cdot (1 - \pi(\theta, \tau))^{(N-n)} \right]$$

$$\cdot p(N) \cdot p(\theta) \cdot p(\tau) \cdot p(B_{obs}, L_{obs}, E_{obs})$$

$$\cdot p(S_{obs}|N, \theta, \tau) \cdot p(I_{obs}|S_{obs}, B_{obs}, L_{obs}, E_{obs})$$

$$\cdot p\left(Y_{obs}^{tot}|B_{obs}, L_{obs}, E_{obs}, I_{obs}, S_{obs}\right) \cdot p\left(Y_{obs}^{src}|Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs}, I_{obs}, S_{obs}\right)$$

$$\text{(B.6)} \quad = \frac{1}{p(n, Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs})} \cdot \left[\begin{pmatrix} N \\ n \end{pmatrix} \mathbb{I}_{\{n \leq N\}} \cdot (1 - \pi(\theta, \tau))^{(N-n)} \right]$$

$$\cdot \left[\begin{pmatrix} N + a_N - 1 \\ a_N - 1 \end{pmatrix} \left(\frac{1}{1 + b_N} \right)^N \left(\frac{b_N}{1 + b_N} \right)^{a_N} \mathbb{I}_{\{N \in \mathbb{Z}^+\}} \right] \cdot \left[\prod_{j=1}^{m} \frac{b^a}{\Gamma(a)} \theta_j^{a-1} e^{-b\theta_j} \mathbb{I}_{\{\theta_j > 0\}} \right]$$

$$\cdot p(\tau_1, \ldots, \tau_m) \, \mathbb{I}_{\{0 < \tau_1 < \tau_2 < \cdots < \tau_m\}} \cdot \left[\prod_{i=1}^{n} p(B_i, L_i, E_i) \cdot g(S_i, B_i, L_i, E_i) \right.$$

$$\cdot \sum_{j=1}^{m} \left\{ \prod_{l=1}^{j-1} \left(\frac{\tau_{l+1}}{\tau_l} \right)^{-\theta_l} \right\} \left(\frac{\theta_j}{\tau_j} \right) \left(\frac{S_i}{\tau_j} \right)^{-(\theta_j + 1)} \cdot \mathbb{I}_{\{\tau_j \leq S_i < \tau_{j+1}\}}$$

$$\cdot \frac{(\lambda_i + k_i)^{Y_i^{tot}}}{Y_i^{tot}!} e^{(\lambda_i + k_i)} \mathbb{I}_{\{Y_i^{tot} \in \mathbb{Z}^+\}}$$

$$\left. \cdot \begin{pmatrix} Y_i^{tot} \\ Y_i^{src} \end{pmatrix} \left(\frac{\lambda_i}{\lambda_i + k_i} \right)^{Y_i^{src}} \left(1 - \frac{\lambda_i}{\lambda_i + k_i} \right)^{Y_i^{tot} - Y_i^{src}} \mathbb{I}_{\left\{Y_i^{src} \in \{0, 1, \ldots, Y_i^{tot}\}\right\}} \right]$$

with $\tau_{m+1} = +\infty$, $\lambda_i \equiv \lambda(S_i, B_i, L_i, E_i)$, $k_i \equiv k(B_i, L_i, E_i)$, and $\prod_{l=1}^{0} \left(\frac{\tau_{l+1}}{\tau_l} \right)^{-\theta_l} \equiv 1$.

B.4. Full-Conditional Distributions for Broken-Pareto Model

B.4.1. Sampling $\theta = (\theta_1, \ldots, \theta_m)^T$: Using (B.6) we can derive the conditional posterior distribution of $\theta = (\theta_1, \ldots, \theta_m)^T$:

$$p(\theta|N, S_{obs}, \tau) \propto \left[\prod_{j=1}^{m} \theta_j^{a_j - 1} e^{-\theta_j b_j} \mathbb{I}_{\{\theta_j > 0\}} \right] \cdot \left[\binom{n}{m} \pi(\theta, \tau)^n (1 - \pi(\theta, \tau))^{N-n} \right]$$

$$\cdot \left[\prod_{i=1}^{n} p(B_i, L_i, E_i) \, g(S_i, B_i, L_i, E_i) \sum_{j=1}^{m} \left\{ \prod_{l=1}^{j-1} \left(\frac{\tau_{l+1}}{\tau_l} \right)^{-\theta_l} \right\} \left(\frac{\theta_j}{\tau_j} \right) \left(\frac{s_i}{\tau_j} \right)^{-(\theta_j + 1)} \mathbb{I}_{\{\tau_j \leq s_i < \tau_{j+1}\}} \right]$$

Distributing the successive sum terms of the product, we note that the indicators $\mathbb{I}_{\{\tau_j \leq s_i < \tau_{j+1}\}}$ will eliminate all but single j-th terms of the sum. Define $\mathcal{I}(j) = \{i : \tau_j \leq s_i < \tau_{j+1}\}$ and $n(j)$ is the cardinality of $\mathcal{I}(j)$ i.e., $\mathcal{I}(j)$ $(n(j))$ denotes the set (number) of source indices whose flux is contained in the interval corresponding to the j-th mixture component. Collecting all product terms with common powers, we have:

$$\begin{aligned}
&\propto \left[\prod_{j=1}^{m} \theta_j^{a_j-1} e^{-\theta_j b_j} \mathbb{I}_{\{\theta_j>0\}}\right] \cdot \left[(1-\pi(\theta,\tau))^{N-n}\right] \\
&\qquad \cdot \left[\prod_{j=1}^{m} \left\{\prod_{l=1}^{j-1} \left(\frac{\tau_{l+1}}{\tau_l}\right)^{-\theta_l}\right\}^{n(j)} \prod_{i\in\mathcal{I}(j)} \left(\frac{\theta_j}{\tau_j}\right)\left(\frac{s_i}{\tau_j}\right)^{-(\theta_j+1)}\right] \\
&\qquad\qquad \propto \left[(1-\pi(\theta,\tau))^{N-n}\right] \cdot \left[\prod_{j=1}^{m} \theta_j^{a_j+n(j)-1} \mathbb{I}_{\{\theta_j>0\}}\right] \\
&\qquad\qquad\qquad \cdot \left[\prod_{j=1}^{m} e^{-\theta_j b_j}\right] \cdot \left[\prod_{j=1}^{m} \left\{\prod_{l=1}^{j-1} \left(\frac{\tau_{l+1}}{\tau_l}\right)^{-\theta_l}\right\}^{n(j)} \cdot \prod_{j=1}^{m} e^{-\theta_j \sum_{i\in\mathcal{I}(j)} \log\left(\frac{s_i}{\tau_j}\right)}\right].
\end{aligned}$$

All terms apart from those involving $\pi(\theta,\tau)$ factorize in terms of $\theta_1,\ldots,\theta_m$:

$$\begin{aligned}
p(\theta|N,S_{obs},\tau) &\propto \left[(1-\pi(\theta,\tau))^{N-n}\right] \prod_{j=1}^{m} \theta_j^{a_j+n(j)-1} \mathbb{I}_{\{\theta_j>0\}} \\
&\qquad \cdot \exp\left\{-\left[\theta_j b_j + n(j)\mathbb{I}_{\{j\neq 0\}} \sum_{l=1}^{j-1} \theta_l \log\left(\frac{\tau_{l+1}}{\tau_l}\right) + \theta_j \sum_{i\in\mathcal{I}(j)} \log\left(\frac{s_i}{\tau_j}\right)\right]\right\} \\
&= \left[(1-\pi(\theta,\tau))^{N-n}\right] \prod_{j=1}^{m} \theta_j^{a_j+n(j)-1} \mathbb{I}_{\{\theta_j>0\}} \\
&\qquad \cdot \exp\left\{-\theta_j \left[b_j + \mathbb{I}_{\{j\neq m\}} \log\left(\frac{\tau_{j+1}}{\tau_j}\right) \sum_{i=1}^{m} \left[n(i)\mathbb{I}_{\{i\geq j+1\}}\right] + \sum_{i\in\mathcal{I}(j)} \log\left(\frac{s_i}{\tau_j}\right)\right]\right\}.
\end{aligned}$$

This partial factorization allows for the exact (conditional) posterior draws to be obtained by rejection sampling as follows:

ALGORITHM 5. Step 1: For $j = 1, \ldots, m$, draw:

$$\theta_j^* \sim \text{Gamma}\left(a_j + n(j), b_j + \mathbb{I}_{\{j \neq m\}} \log\left(\frac{\tau_{j+1}}{\tau_j}\right) \sum_{i=1}^{m} \left[n(i)\mathbb{I}_{\{i \geq j+1\}}\right] + \sum_{i \in \mathcal{I}(j)} \log\left(\frac{s_i}{\tau_j}\right)\right),$$

and denote $\theta^* = (\theta_1^*, \ldots, \theta_m^*)$.

Step 2: Generate $U \sim U[0,1]$. If $U \leq (1 - \pi(\theta^*, \tau))^{N-n}$, then accept θ^*, else return to step 5.

Alternatively, this partial factorization also allows for the approximate (conditional) posterior draws to be obtained by Metropolis-Hastings algorithm. The sampling via MH procedure will be selected at random with success probability 0.9.

B.4.2. Sampling $\tilde{\tau} = (\tau_2, \ldots, \tau_m)^T$ via $\tilde{\eta} = (\eta_2, \ldots, \eta_m)^T$: Recall that $\tau = h^{-1}(\eta|\tau_1)$ and define components $\tau_j = h_j^{-1}(\eta|\tau_1) = \tau_1 + \sum_{k=2}^{j} e^{\eta_k}$, as in (2.10). Using (B.6) we can derive the conditional posterior distribution of $\tilde{\eta} = (\eta_2, \ldots, \eta_m)^T$:

$$p(\tilde{\eta}|N, S_{obs}, \theta, \tau_1) \propto \left[\begin{pmatrix} N \\ n \end{pmatrix} \mathbb{I}_{\{n \leq N\}} \left(1 - \pi(\theta, \tau)\right)^{(N-n)}\right] p(\eta_2, \ldots, \eta_m|\tau_1) \mathbb{I}_{\{\tau_1 < \tau_2 < \cdots < \tau_m\}}$$

$$\cdot \left[\prod_{i=1}^{n} p\left(B_i, L_i, E_i\right) g(S_i, B_i, L_i, E_i) \sum_{j=1}^{m} \left\{\prod_{l=1}^{j-1} \left(\frac{\tau_{l+1}}{\tau_l}\right)^{-\theta_l}\right\} \left(\frac{\theta_j}{\tau_j}\right) \left(\frac{S_i}{\tau_j}\right)^{-(\theta_j+1)} \mathbb{I}_{\{\tau_j \leq S_i < \tau_{j+1}\}}\right]$$

$$\propto \left[\left(1 - \pi(\theta, \tau)\right)^{(N-n)}\right] \cdot \exp\left[-\frac{1}{2} \sum_{j=2}^{m} \left\{c_j(\eta_j - \mu_j)\right\}^2\right] \mathbb{I}_{\{\tau_1 < \tau_2 < \cdots < \tau_m\}}$$

$$\cdot \left[\prod_{j=1}^{m} \left\{\prod_{l=1}^{j-1} \left(\frac{\tau_{l+1}}{\tau_l}\right)^{-\theta_l}\right\}^{n(j)} \prod_{i \in \mathcal{I}(j)} \left(\frac{\theta_j}{\tau_j}\right) \left(\frac{s_i}{\tau_j}\right)^{-(\theta_j+1)} \mathbb{I}_{\{\tau_1 < \min(s_1, \ldots, s_n)\}}\right],$$

where $\mathcal{I}(j) = \{i : \tau_j \leq s_i < \tau_{j+1}\}$ and $n(j)$ is the cardinality of $\mathcal{I}(j)$.

It is impossible to factor out $\tau_2, \ldots, \tau_m$ and further simplify the expression. Sampling of whole η vector is done via Metropolis-Hastings algorithm. The proposal distribution must satisfy all constraints on τ. As described earlier during the derivation of the joint posterior distribution, the constraint $0 < \tau_1 < \tau_2 < \cdots < \tau_m$ can be implemented via a variable transformation (2.10), where $h(\tau_2, \ldots, \tau_m|\tau_1) \in \mathtt{R}^{m-1}$.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Ardia, D., Hoogerheide, L. F., and Van Dijk, H. K. (2009). To Bridge, to Warp or to Wrap? A Comperative Study of Monte Carlo Methods for Efficient Evaluation of Marginal Likelihood. Technical report, Tinbergen Institute.

Baloković, M., Smolčić, V., Ivezić, v., Zamorani, G., Schinnerer, E., and Kelly, B. C. (2012). Disclosing the Radio Loudness Distribution Dichotomy in Quasars: An Unbiased Monte Carlo Approach Applied to the SDSS-FIRST Quasar Sample. *The Astrophysical Journal*, 759(1):30.

Bayarri, M. J. and Berger, J. O. (1998). Quantifying Surprise in the Data and Model Verification. In Bernardo, J., Berger, J., Dawid, A., and Smith, A., editors, *Bayesian Statistics 6. Proceedings of the Sixth Valencia International Meeting*, pages 53–82, Oxford, England. Oxford University Press.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B*, 36(2):192–236.

Celeux, G., Forbes, F., Robert, C. P., Titterington, D. M., and de recherche en économie et statistique (Paris, F. (2003). *Deviance Information Criteria for Missing Data Models*. Documents de travail du CREST. INSEE.

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer, NY, New York.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *Journal of the American Statistical Association*, 49(4):327–335.

Cook, S. R., Gelman, A., and Rubin, D. B. (2006). Validation of Software for Bayesian Models Using Posterior Quantiles. *Journal of Computational and Graphical Statistics*, 15(3):675–692.

Crawford, D. F., Jauncey, D. L., and Murdoch, H. S. (1970). Maximum-Likelihood Estimation of the Slope from Number-Flux Counts of Radio Sources. *The Astrophysical Journal*, 162:405.

Efron, B. (1983). Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *Journal of the American Statistical Association*, 78(382):316–331.

Feigelson, E. D. (1992). Censoring in Astronomical Data Due to Nondetections. In Feigelson, E. D. and Babu, G. J., editors, *Statistical Challenges in Modern Astronomy*, pages 221–237. Springer New York, New York, NY.

Feigelson, E. D. and Babu, G. J. (2003). *Statistical Challenges in Astronomy*. Springer New York, New York, NY.

Gelman, A. (2007). Comment: Bayesian Checking of the Second Levels of Hierarchical Models. *Statistical Science*, 22(3):349–352.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. Chapman and Hall/CRC Press, 2nd edition.

Gelman, A., Hwang, J., and Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. Vol.6, No.4. *Statistica Sinica*, 6(4):733–807.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*, volume 39. Chapman & Hall/CRC, New York.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Jeffreys, S. H. (1961). *Theory of Probability*. Oxford University Press, Oxford, England, 3rd edition.

Jiang, J., Nguyen, T., and Rao, J. S. (2011). Invisible fence methods and the identification of differentially expressed gene sets. *Statistics and Its Interface*, 4(3):403–415.

Jiang, J., Rao, J. S., Gu, Z., and Nguyen, T. (2008). Fence methods for mixed model selection. *The Annals of Statistics*, 36(4):1669–1692.

Jóhannesson, G., Björnsson, G., and Gudmundsson, E. H. (2006). Afterglow Light Curves and Broken Power Laws: A Statistical Study. *The Astrophysical Journal*, 640(1):L5–L8.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kim, M., Wilkes, B. J., Kim, D., Green, P. J., Barkhouse, W. A., Lee, M. G., Silverman, J. D., and Tananbaum, H. D. (2007). Chandra Multiwavelength Project XRay Point Source Number Counts and the Cosmic XRay Background. *The Astrophysical Journal*, 659(1):29–51.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley, 2nd edition.

Liu, J. S., Liang, F., and Wong, W. H. (2000). The Multiple-Try Method and Local Optimization in Metropolis Sampling. *Journal of the American Statistical Association*, 95(449):121–134.

Loredo, T. J. and Wasserman, I. M. (1995). Inferring the spatial and energy distribution of gamma-ray burst sources. 1: Methodology. *The Astrophysical Journal Supplement Series*, 96:261.

Lynch, S. M. and Western, B. (2004). Bayesian Posterior Predictive Checks for Complex Models. *Sociological Methods & Research*, 32(3):301–335.

Lynden-Bell, D. (1992). Eddington-Malmquist Bias, Streaming Motions, and the Distribution of Galaxies. In Feigelson, E. and Babu, G. J., editors, *Statistical Challenges in Modern Astronomy*, pages 201–16. New York: Springer-Verlag.

Maccacaro, T., Gioia, I. M., Zamorani, G., Feigelson, E. D., Fener, M., Giacconi, R., Griffiths, R. E., Murray, S. S., Stocke, J., and Liebert, J. (1982). A medium sensitivity X-ray survey using the Einstein Observatory - The log N-log S relation for extragalactic X-ray sources. *The Astrophysical Journal*, 253:504.

Maccacaro, T., Romaine, S., and Schmitt, H. M. M. (1987). LogN-logS slope determination in imaging X-ray astronomy. *IN: Observational cosmology; Proceedings of the IAU Symposium*, pages 597–599.

Madau, P., Ferguson, H. C., Dickinson, M. E., Giavalisco, M., Steidel, C. C., and Fruchter, A. (1996). High-redshift galaxies in the Hubble Deep Field: colour selection and star formation history to z˜4. *Monthly Notices of the Royal Astronomical Society*, 283:1388–1404.

Malmquist, K. (1920). A study of the stars of spectral type A. *Medd. Lund. Astr. Obs.*, 22.

Mateos, S., Warwick, R. S., Carrera, F. J., Stewart, G. C., Ebrero, J., Della Ceca, R., Caccianiga, A., Gilli, R., Page, M. J., Treister, E., Tedds, J. A., Watson, M. G., Lamer, G., Saxton, R. D., Brunner, H., and Page, C. G. (2008). High precision X-ray log N log S distributions: implications for the obscured AGN population. *Astronomy and Astrophysics*, 492(1):51–69.

Meng, X.-L. (1994). Posterior Predictive p-Values. *The Annals of Statistics*, 22(3):1142–1160.

Meng, X.-l. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. In *Statistica Sinica*, pages 831–860.

Murdoch, H. S., Crawford, D. F., and Jauncey, D. L. (1973). Maximum-Likelihood Estimation of the Number-Flux Distribution of Radio Sources in the Presence of Noise and Confusion. *The Astrophysical Journal*, 183:1.

Neal, R. (2008). The Harmonic Mean of the Likelihood: Worst Monte Carlo Method Ever | Radford Neal's blog on WordPress.com.

Neil, R. (2011). MCMC for Using Hamiltonian Dynamics. In Brooks, S., Gelman, A., Jones, G., and Xiao-Li, M., editors, *Handbook of Markov Chain Monte Carlo*, pages 113–162. Chapman & Hall, Boca Raton, FL.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1):3–48.

Rao, C., Pathak, P., and Koltchinskii, V. (1997). Bootstrap by sequential resampling. *Journal of Statistical Planning and Inference*, 64(2):257–281.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods. 2nd edition.* Springer, NY.

Robert, C. P., Wraith, D., Goggans, P. M., and Chan, C.-Y. (2009). Computational methods for Bayesian model choice. In *I Can*, number 2009, pages 251–262.

Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applies Statistician. *The Annals of Statistics*, 12(4):1151–1172.

Schmitt, J. H. M. M. and Maccacaro, T. (1986). Number-counts slope estimation in the presence of Poisson noise. *The Astrophysical Journal*, 310:334.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

Skilling, J. (2004). Nested Sampling. In *Bayesian Inference and Maximum Enthropy Methods in Science and Engineering: 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 735, pages 395–405. AIP Publishing.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594.

Wong, R. K. W., Baines, P., Aue, A., Lee, T. C. M., and Kashyap, V. L. (2014). Automatic estimation of flux distributions of astrophysical source populations. *The Annals of Applied Statistics*, 8(3):1690–1712.

Yu, Y. and Meng, X.-L. (2011). To Center or Not to Center: That Is Not the QuestionAn AncillaritySufficiency Interweaving Strategy (ASIS) for Boosting MCMC Efficiency. *Journal of Computational and Graphical Statistics*, 20(3):531–570.

Zezas, A. and Fabbiano, G. (2002). Chandra Observations of The Antennae Galaxies (NGC 4038/4039). IV. The XRay Source Luminosity Function and the Nature of Ultraluminous XRay Sources. *The Astrophysical Journal*, 577(2):726–737.

Zezas, A., Fabbiano, G., Baldi, A., Schweizer, F., King, A. R., Rots, A. H., and Ponman, T. J. (2007). Chandra Monitoring Observations of the Antennae Galaxies. II. XRay Luminosity Functions. *The Astrophysical Journal*, 661(1):135–148.