# Bayesian Estimation of Luminosity Distributions and Model Based Classification of Astrophysical Sources

a thesis presented for the degree of

Doctor of Philosophy of Imperial College London

and the

Diploma of Imperial College

by

Vasileios Stampoulis

Department of Mathematics

Imperial College

180 Queen's Gate, London SW7 2BZ

September 2017

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Signed: Vasileios Stampoulis

Thesis advisor: Professor David van Dyk Vasileios Stampoulis

# Bayesian Estimation of Luminosity Distributions and Model Based Classification of Astrophysical Sources


## Abstract

The distribution of the flux (observed luminosity) of astrophysical objects is of great interest as a measure of the evolution of various types of astronomical source populations and for testing theoretical assumptions about the Universe. This distribution is examined using the cumulative distribution of the number of sources ($N$) detected at a given flux ($S$), known as the $\log(N) - \log(S)$ curve to astronomers. Estimating the $\log(N) - \log(S)$ curve from observational data can be quite challenging though, since statistical fluctuations in the measurements and detector biases often lead to measurement uncertainties. Moreover, the location of the source with respect to the centre of observation and the background contamination can lead to non-detection of sources (missing data). This phenomenon becomes more apparent for low flux objects, thus indicating that the missing data mechanism is non-ignorable. In order to avoid inferential biases, it is vital that the different sources of uncertainties, potential bias and missing data mechanism be properly accounted for. However, the majority of the methods in the relevant literature for estimating the $\log(N) - \log(S)$ curve are based on the assumption of complete surveys with non missing data.

In this thesis, we present a Bayesian hierarchical model that properly accounts for the missing data mechanism and the other sources of uncertainty. More specifically, we model the joint distribution of the complete data and model parameters and then derive the posterior distribution of the model parameters marginalised across all missing data information. We utilise a Blocked Gibbs sampler in order to extract samples from the joint posterior distribution of the parameters of interest. By using a Bayesian approach, we produce a posterior distribution for the $\log(N) - \log(S)$ curve instead of a best-fit estimate. We apply this method to the Chandra Deep Field South (CDFS) dataset.

Furthermore, approaching this complicated problem from a fully Bayesian angle enables us to appropriately model the uncertainty about the conversion factor between

observed source photon counts and observed luminosity. Using relevant spectral data for the observed sources, the uncertainty about the flux-to-count conversion factor $\gamma$ for each observed source is expressed through MCMC draws from the posterior distribution of $\gamma$ for each source. In order to account for this uncertainty in the non-detected sources, we develop a novel statistical approach for fitting a hierarchical prior on the flux-to-count conversion factor based on the MCMC samples from the observed sources (a statistical approach that can be used in many modelling problems of similar nature). We derive in a similar manner the posterior distribution of the model parameters, marginalised across the missing data, and we explore the impact in our posterior estimates of the parameters of interest in the CDFS dataset.

Studying the $\log(N) - \log(S)$ relationship for different source populations can give us further insight into the differences between the various types of astronomical populations. Hence, we propose a new soft-clustering scheme for classifying galaxies in different activity classes (Star Forming Galaxies, LINERs, Seyferts and Composites) using simultaneously 4 optical emission-line ratios ($[\mathrm{N_{II}}]/\mathrm{H}\alpha$, $[\mathrm{S_{II}}]/\mathrm{H}\alpha$, $[\mathrm{O_I}]/\mathrm{H}\alpha$ and $[\mathrm{O_{III}}]/\mathrm{H}\beta$). The most widely used classification approach is based on 3 diagnostic diagrams, which are 2-dimensional projections of those emission line ratios. Those diagnostics assume fixed classification boundaries, which are developed through theoretical models. However, the use of multiple diagnostic diagrams independently of one another often gives contradicting classifications for the same galaxy, and the fact that those diagrams are 2-dimensional projections of a complex multi-dimensional space is limiting the power of those diagnostics. In contrast, we present a data-driven soft clustering scheme that estimates the posterior probability of each galaxy belonging to each activity class. More specifically, we fit a large number of multivariate Gaussian distributions to the Sloan Digital Sky Survey (SDSS) dataset in order to capture local structures and subsequently group the multivariate Gaussian distributions to represent the complex multi-dimensional structure of the joint distribution of the 4 galaxy activity classes. Finally, we discuss how this soft-clustering can lead to estimates of population-specific $\log(N) - \log(S)$ relationships.

# List of Publications

1) *McKeough, K., Siemiginowska, A., Cheung, C.C., Stawarz, L., Kashyap, V.L., Stein, N., Stampoulis, V., van Dyk, D.A., Wardle, J.F.C., Lee, N.P. and Harris, D.E., 2016. Detecting Relativistic X-ray Jets in High-Redshift Quasars. The Astrophysical Journal, 833(1), p.123.*

This paper describes a novel statistical procedure used for fitting a hierarchical prior to posterior distributions that were computed using another statistical procedure. This methodology is used in the estimating a prior for the flux-to-count conversion factor in Chapter 3.

2) *Stampoulis, V., van Dyk, D. A., Kashyap, V. L., and Zezas, A. (2017). Multidimensional Data Driven Classification of Active Galaxies. Monthly Notices of the Royal Astronomical Society, revision requested and submitted.*

This paper contains the clustering methodology derived in Chapter 4.

3) *Stampoulis, V., van Dyk, D. A., Kashyap, V. L., and Zezas, A. (2017). Estimating the* $\log(N) - \log(S)$ *with Count-to-Flux Conversion Factor Uncertainty, in progress.*

This paper includes the hierarchical Bayesian model for the estimation of the $\log(N) - \log(S)$ relationship developed in Chapters 2 and 3.

*To my parents, Georgios and Pinelopi.*

# Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor Prof. David van Dyk for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance was crucial during all the time of my research and writing of this thesis.

Besides my advisor, I would like to thank my astrophysicists collaborators, Dr. Andreas Zezas and Dr. Vinay Kashyap. They guided my path through the magnificent world of astrophysics and helped me work on a series of very interesting problems. This thesis is the result of this fruitful collaboration. Moreover, I extend my deepest appreciation to the members of the CHASC International Astrostatistics Center for their insightful comments and discussions. I am looking forward to continue working with them.

I would also like to thank my parents Georgios and Pinelopi and my brother Konstantinos. They have supported me throughout my studies - and my life in general - in every possible way, and I am immensely grateful to them.

I would also like to acknowledge the support from the Roth scholarship scheme, funded by the Department of Mathematics of Imperial College London, as well as the RISE EU programme which funded my extended stay to the Harvard-Smithsonian Center for Astrophysics.

# List of figures

# List of tables

# Contents

# 1 Introduction

## 1.1 Astrostatistics

In recent years, there has been a constantly increasing use of advanced statistical techniques in solving challenging problems in different fields of science. The easy access to a vast amount of data has given rise to a series of questions such as how we can interpret the observations, what dependencies exist between variables as well as how can we update theoretical models in the presence of a great number of new observations.

One of the scientific fields that finds itself with serious challenges in statistical treatments of data is undoubtedly Astrophysics. The sheer amount of data that is available for analysis to astronomers has increased dramatically over the last decades as a result of the increased number and capabilities of both earth-based and space telescopes. On the one hand, the astronomical community has ready access to exciting new data, but on the other hand it is inevitably faced with the challenging task of enabling efficient and objective scientific exploitation of those enormous multi-faceted datasets. Consequently, statistics has become an essential part of the process leading to the correct interpretation of the physical phenomena.

The challenges that astronomers face in analysing the astronomical data come both as a result of the nature of the data and of the type of questions the scientific

community puts forward. The astronomical data are typically multidimensional, subject to heteroscedastic errors in measurement and have very complex structures. Moreover, the astronomical observations cannot be repeated since the universe is constantly evolving, thus the replication of experiments is not possible. At the same time astronomers are interested in drawing conclusions for the physical mechanisms and laws that govern the universe represented through complex models. The highly complicated nature of the astronomical data and the subsequent scientific questions lead to many subtle inference problems that require sophisticated statistical tools. As a result, the establishment of whole new statistical frameworks, modelling techniques and innovative computational methods is required in order to answer the questions posed by the astrophysical problems.

This interweaving of statistics and astrophysics, known as Astrostatistics, should not be viewed as only devoted to the development of new methods for dealing with astrophysical problems, but also as an opportunity to establish new general statistical techniques and expand the boundaries of statistical thinking.

## 1.2 The Scientific Problem

The inspiration for this research has its origins in the long-standing astrophysical problem of estimating the distribution of the flux. The flux* is the power per unit area radiated from an astronomical source, whether this source could be a galaxy or a star or any other type of source. This density is estimated traditionally by using a $\log(N) - \log(S)$ relationship, where $S$ is the source flux and $N$ is the number of sources observed to that flux sensitivity. This relationship has traditionally been assumed to be either linear, or piece-wise linear.

The distribution of the flux of astrophysical objects is of great interest as a measure of the evolution of various types of astronomical source populations and for testing theoretical assumptions about the Universe. The $\log(N) - \log(S)$ relationship provides an overall picture of source populations and facilitates the comparison with models for populations and their evolution. The applications extend to analysing populations of black-holes and neutron stars in galaxies, populations of stars in star-

*The related term luminosity is used for the total energy radiated from an astronomical source.

clusters or distribution of dark matter in the universe.

More specifically, astronomical objects that are at small redshifts (where the geometry of the Universe is well approximated by the Euclidean geometry) which do not evolve with cosmic time and are uniformly distributed in the Universe, are characterised by a $\log(N) - \log(S)$ curve with slope equal to 1.5 (Maccacaro et al. 1987). Under these hypotheses, any class of objects has the same $\log(N) - \log(S)$ relationship regardless of its luminosity function. However, when we look at astronomical objects at high redshifts, we expect that the $\log(N) - \log(S)$ curve will be less steep. This flattening of the curve provides us with information regarding the geometry of the Universe and on the luminosity function of the sources. Furthermore, if the astronomical objects exhibit some form of evolution with cosmic time, then this can affect the aforementioned slope of 1.5. In other words, any departure from the slope of 1.5 is indicative of some effect.

There are many factors and complications that make the estimation of the $\log(N) - \log(S)$ relationship quite challenging. In astronomical observations, the flux is not directly measured. The collected data represent the cumulated flux received from the source which is stored as photon counts. The recorded photon counts are subject to Poisson like variability. Those statistical fluctuations in measurements introduce a bias, namely the Eddington bias (Teerikorpi 2004). More specifically, a set of sources with a single luminosity will, upon observation, be spread out due to measurement error. If the statistical scatter is close to some kind of detection threshold, then the inferred luminosity based on only the detected sources will end up being a biased estimate. Furthermore, astronomical objects that are either at smaller distances relative to the measurement instrument, or are brighter, are more likely to be detected. This phenomenon leads to a strong selection bias known as Malmquist bias (Malmquist 1920). The statistical issues associated with the detection of astronomical sources also include issues such as estimating the probability that the observed source is just a background fluctuation or the uncertainty in estimating the flux given the background and the instrumental inefficiencies.

### 1.2.1 Existing Approaches and their Inefficiencies

Early work on the estimation of the $\log(N)-\log(S)$ relationship focused on assuming a power law distribution for the flux, which corresponds to a linear $\log(N)-\log(S)$ relationship. Murdoch et al. (1973) use maximum likelihood (ML) estimation in order to estimate the parameters of the distribution of the flux. They model the existence of experimental errors by employing normal approximations, arguing that the measurement errors have a significant effect on the number counts. A linear relationship, but with Poisson errors, is utilised in Maccacaro et al. (1982) and Schmitt & Maccacaro (1986). Georgakakis et al. (2008) present a ML method with Poisson errors for estimating a piece-wise linear $\log(N)-\log(S)$ relationship that takes into account the non-negligible spatial variations in sensitivity across the astronomical survey. A recent approach that attempts to estimate both the number of different linear pieces and their location is proposed by Wong et al. (2014). The authors present a statically interesting refinement to the more standard ML methods by using an interwoven Expectation-Maximisation algorithm. Zezas et al. (2007) explore the idea of incorporating the uncertainty of the flux-to-count conversion parameter in the ML framework; they discuss how the uncertainty of the flux-to-count conversion parameter for each source can be included in a log-likelihood estimation of a $\log(N)-\log(S)$ curve by assuming that the distribution of the flux-to-count conversion parameter can be approximated by a multivariate normal distribution with variance taken from the covariance matrix of the spectral fit.

A very important issue with the majority of the existing literature regarding the estimation of the $\log(N)-\log(S)$ relationship is the inherent assumption in the proposed methods of having a complete dataset with no missing sources. However, as we described previously, incompleteness in astronomical surveys is an unavoidable phenomenon, which -if not properly accounted for- can lead to inferential biases. Thus, since the data we observe is a biased subset of the complete population, it is of outmost importance to take into consideration this issue in order to draw proper inferential conclusions.

### 1.2.2 Research Direction

Our work is based on the Bayesian approach introduced by Udaltsova (2014) for estimating the $\log(N)-\log(S)$ relationship. The author proposes a hierarchical

Bayesian model in order to account for the measurement and detector biases as well as the missing data mechanism. More specifically, the suggested approach models the joint distribution of the complete data and model parameters, and then derives the posterior distribution of the model parameters marginalised across all missing data information.

There are many examples in the astro-statistical literature that a Bayesian approach has been used in order to account for problems with missing data (non-detection) and selection effects. Kelly (2007) derives a multi-level Gibbs sampler for estimating the parameters of linear regression when measurement errors and intrinsic scatter (variations in the physical properties of astronomical sources that are not completely captured by the variables included in the model) exist. The method is generalised for cases with multiple independent variables, non-detections, and selection effects, exhibiting excellent performance. Buchner et al. (2015) develop a non-parametric method to incorporate uncertainties from measurements, incompleteness of the data, and limited sample size. They employ a Hamiltonian Markov chain Monte Carlo to obtain estimates of the parameters.

Our research extends the work of Udaltsova (2014) in two directions. Initially, we propose a series of extensions to the model regarding the estimation and the sampling from the joint distribution of the background noise $B$, the off axis angle $L$ and the exposure map $E$, as well as the selection of the incompleteness function. More specifically, we incorporate a more detailed incompleteness function which depends on the background, distance from the centre and total time of exposure. The proper selection of the incompleteness function is of paramount importance since it plays a crucial role in the posterior inference. We also utilise survey specific background and exposure time distributions when sampling for the missing data and, finally, we re-derive the conditional posterior distributions for the parameters of interest while correcting errors in the mathematical derivations of the conditional posterior distributions.

Furthermore, in our endeavour to fully address any sources of uncertainty that might affect considerably and in systematic manner the posterior inference, a new framework was developed. Within its remit, we were able to appropriately model the uncertainty about the conversion factor between observed source photon counts and observed flux. In the vast majority of the relevant literature that tries to estimate

the $\log(N) - \log(S)$ relationship, this conversion factor, $\gamma$, is assumed to be constant for all the sources. However, the actual value of this parameter depends on the spectrum of each astronomical source. Using relevant spectral data for the observed sources, the uncertainty about the flux-to-count conversion factor $\gamma$ for each observed source is expressed through Markov Chain Monte Carlo (MCMC) draws from the posterior distribution of $\gamma$ for each source. In order to account for this uncertainty in the non-detected sources, we develop a novel statistical approach for fitting a hierarchical prior on the flux-to-count conversion factor based on the MCMC samples from the observed sources (a statistical approach that can be used in many modelling problems of similar nature). In order to draw posterior inference for the $\log(N) - \log(S)$ relationship, we derive in a similar manner the posterior distribution of the model parameters, marginalised across the missing data, and we explore the impact on our posterior estimates of the parameters of interest in the Chandra Deep Field South dataset.

Studying the $\log(N) - \log(S)$ relationship for different source populations can give us further insight into the differences between the various types of astronomical populations. Based on that, we delved into a long and heavily researched classification problem in Astronomy, which is the classification of galaxies to different activity classes (Star Forming Galaxies, LINERs, Seyferts and Composites). Despite the extensive literature on this problem though, the existing classification schemes are mostly purely theoretical in nature and do not offer a robust classification methodology. Therefore, we propose a new soft-clustering data driven classification scheme (SoDDA), for classifying galaxies in different activity classes using 4 optical emission-line ratios ($[\mathrm{N_{II}}]/\mathrm{H}\alpha$, $[\mathrm{S_{II}}]/\mathrm{H}\alpha$, $[\mathrm{O_{I}}]/\mathrm{H}\alpha$ and $[\mathrm{O_{III}}]/\mathrm{H}\beta$). The most widely used classification approach is based on 3 diagnostic diagrams, which are 2-dimensional projections of those emission line ratios. Those diagnostics assume fixed classification boundaries, which are developed through theoretical models. However, the use of multiple diagnostic diagrams independently of one another often gives contradicting classifications for the same galaxy, and the fact that those diagrams are 2-dimensional projections of a complex multi-dimensional space is limiting the power of those diagnostics. To tackle this issue, we present a data-driven soft clustering scheme that estimates the posterior probability of each galaxy belonging to each activity class. More specifically, we fit a large number of multivariate Gaussian distributions to the Sloan Digital Sky Survey (SDSS) dataset in order to capture local structures and subsequently group the multivariate Gaussian distributions to represent the complex multi-dimensional structure of the joint

distribution of the 4 galaxy activity classes. Moreover, since the use of linear hard boundaries for this classification problem is widespread in the astronomical community, we extract 4-dimensional linear boundaries for our classification scheme using Support Vector Machines (SVM). The classification accuracy of those hard boundaries is very close to that of SoDDA, even after removing one of the dimensions ($[O_I]/H\alpha$ which is difficult to measure for many objects). This indicates the power of our classification and the importance of considering all the dimensions of the problem jointly. Finally, we discuss how this soft-clustering can lead to estimates of population-specific $\log(N) - \log(S)$ relationships.

## 1.3 Statistical Background

Bayesian analysis is an extremely versatile statistical tool that allows us to make parameter estimation through the posterior distribution, which is a combination of the observed data and our prior knowledge about the unknown parameters. As general notation, let $\theta$ denote the parameters of interest and $y$ denote the observed data. The joint probability distribution of $\theta$ and $y$ can be written as the product of the prior distribution $p(\theta)$ and the sampling distribution $p(y|\theta)$, i.e.

$$p(\theta, y) = p(\theta) \cdot p(y|\theta)$$

Using Bayes' theorem, we get the posterior distribution as:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta) \cdot p(y|\theta)}{p(y)},$$

where $p(y)$ is the marginal distribution of the data $y$.

### 1.3.1 Missing data

In the Bayesian way of thinking, there is no distinction between missing data and parameters since both of them are uncertain and are characterised by a joint posterior distribution (Rubin 1996). In the missing data literature, two notions are used in the context of unintentional missing data, the missing at random (MAR) notion and the missing completely at random (MCAR) notion (see Gelman et al. 2014).

If we define as $y = (y_{obs}, y_{mis})$ the complete data and $I$ an indicator variable which is 1 if the component of $y$ is observed, then the joint distribution of $(y, I)$ given parameters $(\theta, \phi)$ is:

$$p(y, I|\theta, \phi) = p(y|\theta) \cdot p(I|y, \phi)$$

The conditional distribution of $I$ given the complete data and the unknown parameter $\phi$, i.e. $p(I|y, \phi)$, describes the missing data mechanism and represents the incompleteness function, while the distribution of the observed data can be written:

$$p(y_{obs}, I|\theta, \phi) = \int p(y|\theta) \cdot p(I|y, \phi) \, \mathrm{d}y_{mis} \tag{1.1}$$

Under the MAR hypothesis, the distribution of incompleteness function doesn't depend on the missing values, i.e. $p(I|y, \phi) = p(I|y_{obs}, \phi)$ and thus 1.1 simplifies to $p(y_{obs}, I|\theta, \phi) = p(y_{obs}|\theta) \cdot p(I|y_{obs}, \phi)$.

If we also add the assumption that the distribution of the missing data mechanism is independent of $y$, i.e. $p(I|y, \phi) = p(I|\phi)$, then we speak of data that are observed at random and we have the notion of MCAR in which the missing data mechanism can be ignored. Through Bayesian analysis, we can deal with the non-ignorable missing data mechanism we have in the $\log(N) - \log(S)$ problem.

### 1.3.2 Markov Chain Monte Carlo Methods

In order to draw posterior inference about the parameters of interest, we draw samples from the posterior distribution in order to summarise it by estimating posterior statistics such as the mean, variance and percentiles. In simple Bayesian models, it is often easy to draw samples directly from the posterior distribution of the parameters $p(\theta|y)$ where $\theta$ denotes the parameters vector and $y$ the observed data. Nevertheless, in more complicated contexts where we have multiple parameters and the posterior distribution is not a standard distribution (as in the Bayesian model we develop in the next Chapters for $\log(N) - \log(S)$ ), the perhaps most common approach is to dive into the realm of Markov Chain Monte Carlo (MCMC) simulation methods.

The rise of MCMC methods in the late 1980s has opened new horizons in the field of Bayesian Statistics, allowing statisticians to explore complex models and don't limit their modelling in problems that are analytically tractable. MCMC simulation is basically a general method for simulation based on drawing values of the parameters $\theta$ from approximate distributions and then correct the draws to better approximate the posterior distribution. Those values are drawn sequentially from a transition probability distribution based only on the value of the last draw, thus forming a Markov Chain [†]. The key part of the MCMC methods is to construct the transition probability distributions so that the Markov chain converges to a unique stationary distribution [‡] - the posterior distribution $p(\theta|y)$.

A very important notion regarding the proof of the convergence of the MCMC is the reversibility of a Markov chain. A transition probability distribution is said to be reversible with respect to an initial distribution if for the Markov chain they define, the distribution of $\theta^1, \theta^2, ..., \theta^n$ has the same distribution as $\theta^{k-1}, \theta^{k-2}, ..., \theta^{k-n}$ for all $n, k$. This definition has an immediate consequence -reversibility implies that the Markov chain has a unique stationary distribution, but not vice-versa. Thus, in the MCMC framework, we are interested in updates that are reversible. For the scope of this research, we are interested into two algorithms, the Metropolis-Hastings Algorithm and the Gibbs Sampler.

**The Metropolis-Hastings Algorithm**. The Metropolis-Hastings Algorithm is an adaptation of a random walk which takes advantage of an acceptance/rejection rule in order to converge to the posterior distribution. Metropolis & Ulam (1949) and Metropolis et al. (1953) were the first to present Markov chain simulation of probability distributions, effectively putting forward what is known today as the Metropolis algorithm. Hastings (1970) extended the Metropolis algorithm by using non symmetric transition probability distributions. Green (1995) summarises and

[†]A sequence $\theta^1, \theta^2, ....$ of random elements of some set is a Markov chain, if the conditional distribution of $\theta^{n+1}$ given $\theta^1, ..., \theta^n$ depends only on $\theta^n$. We call state space the set in which $\theta^i$ takes values. The marginal distribution of $\theta^1$ is called the initial distribution. If there exists a set of numbers $P_{ij}$ such as when the chain is in state $i$, the probability that the next state is $j$ is $P_{ij}$, then we say that the collection $\{\theta^n,\ n > 0\}$ is a Markov chain with transition probabilities $P_{ij}$.

[‡]A Markov chain is said to be stationary or invariant or equilibrium if for every positive integer $k$, the conditional distribution of $(\theta^{n+2}, ..., \theta^{n+k})$ given $\theta^{n+1}$ does not depend on $n$. In other words, a stationary distribution of a Markov chain is a probability distribution that remains unchanged in the Markov chain as time progresses. An initial distribution is said to be stationary for some transition probability distribution if the Markov chain specified by this initial distribution and the transition probability distribution is stationary.

generalises the algorithm.

The Metropolis-Hastings is an adaptation of a random walk that uses an acceptance/rejection update to converge to the stationary distribution. More specifically, suppose that the posterior distribution we want to sample from has unnormalised density $h$ - this is the stationary distribution of the MCMC sampler we want to construct. The Metropolis-Hastings Algorithm iterates between the following steps:

Step 1: If the current state is $\theta^{t-1}$, propose a move to state $\theta^*$, having as conditional probability given $\theta^{t-1}$ the $q(\theta^*|\theta^{t-1})$.

Step 2: Compute the Hastings ratio $r(\theta^{t-1}, \theta^*) = \frac{h(\theta^*) \cdot q(\theta^{t-1}|\theta^*)}{h(\theta^{t-1}) \cdot q(\theta^*|\theta^{t-1})}$.

Step 3 (known as Metropolis rejection): Accept the proposed move $\theta^*$ with probability $a(\theta^{t-1}, \theta^*) = \min(1, r(\theta^{t-1}, \theta^*))$.

In order to prove that the stationary distribution of the Markov chain generated by the Metropolis-Hastings algorithm is the target posterior distribution, we assume that the chain is started at step $t-1$ with a draw $\theta^{t-1}$ from the target posterior distribution $h(\theta)$. Following Brooks et al. (2011), if we assume two points $\theta_a$ and $\theta_b$ such as $h(\theta_b)q(\theta_a \mid \theta_b) > h(\theta_a)q(\theta_b \mid \theta_a)$, then the unconditional probability of moving from $\theta_a$ to $\theta_b$ is

$$\begin{aligned} P(\theta^* = \theta_b, \theta^{t-1} = \theta_a) &= h(\theta_a) \cdot q(\theta_b \mid \theta_a) \cdot \min\left(\frac{h(\theta_b) \cdot q(\theta_a|\theta_b)}{h(\theta_a) \cdot q(\theta_b|\theta_a)}, 1\right) \\ &= h(\theta_a) \cdot q(\theta_b \mid \theta_a) \end{aligned}$$

since, because of our assumption, the probability of acceptance is 1. The unconditional probability of moving from $\theta_b$ to $\theta_a$ is

$$P(\theta^* = \theta_a, \theta^{t-1} = \theta_b) = h(\theta_b) \cdot q(\theta_a \mid \theta_b) \cdot \frac{h(\theta_a) \cdot q(\theta_b|\theta_a)}{h(\theta_b) \cdot q(\theta_a|\theta_b)}$$
$$= h(\theta_a) \cdot q(\theta_b \mid \theta_a)$$

which is the same as the probability of transition from $\theta_a$ to $\theta_b$. Thus, the joint density of $(\theta^*, \theta^{t-1})$ is symmetric and $h(\theta)$ is the stationary distribution of the Markov Chain produced using the Metropolis-Hastings update.

The candidate probability density $q$ can be either symmetric, in which case we have the special case of Metropolis Algorithm, or asymmetric in the more general case when the choice of a symmetric density is not optimal. The power of the Metropolis-Hastings Algorithm lies on its ability to draw samples from any distribution, both one-dimensional and high-dimensional. An interesting discussion about the rates of convergence of MCMC can be found in Rosenthal (1995).

**The Gibbs Sampler**. The Gibbs Sampler is a MCMC method for drawing samples from multi-dimensional posterior distributions. Besag (1974) proves that under mild regularity conditions, the joint posterior distribution $p(\theta|y)$ is completely determined by the conditional distributions $p(\theta_j|\theta_{j\neq i}, y)$. It was first used by Geman & Geman (1984) in an application to image processing.

In essence, the Gibbs sampler updates the posterior conditional distribution of one component of the state vector given the rest of the components, i.e. given the current state $(\theta_1^{t-1}, ..., \theta_p^{t-1})$, the algorithm performs the following steps:

Step 1: Draw $\theta_1^t$ from $p(\theta_1|\theta_2^{t-1}, ..., \theta_p^{t-1}, y)$.

Step 2: Draw $\theta_2^t$ from $p(\theta_2|\theta_1^t, \theta_3^{t-1}..., \theta_p^{t-1}, y)$.

Step 3: ...

Step $p$: Draw $\theta_p^t$ from $p(\theta_p|\theta_1^t, ..., \theta_{p-1}^t, y)$.

In a Gibbs update, the proposal is from a conditional distribution of the desired equilibrium distribution and it is always accepted. It is easy to prove that Gibbs is a special case of Metropolis-Hastings. More specifically, following Brooks et al. (2011), we split the state vector $\theta = (u, v)$ and allow the proposal to alter only $u$, but not $v$. We rewrite the unnormalised density $h(u, v) = g(v)q(v, u)$, where $g(v)$ is an unnormalised marginal of $v$ and $q(v, u)$ is the normalised conditional distribution of $u$ given $v$. If we assume a Metropolis-Hastings update with $q$ as the proposal distribution and the update is $\theta^* = (u^*, v)$, then the Hastings ratio is

$$\begin{aligned} r(\theta^{t-1}, \theta^*) &= \frac{h(u^*, v) \cdot q(u, v)}{h(u, v) \cdot q(v, u^*)} \\ &= \frac{g(v)q(v, u^*)q(u, v)}{g(v)q(v, u)q(v, u^*)} \\ &= 1, \end{aligned}$$

hence the proposal is always accepted.

**Convergence of MCMC.** In order to draw posterior inference from MCMC simulation, we use the collection of the simulated draws from our iterative simulation scheme. However, in order to have correct inference from the MCMC method, it is required that the iterative simulation scheme has converged to the target posterior distribution. If the iterations have not proceeded long enough, the simulations may have not converged to the stationary distribution. Monitoring convergence is a notoriously difficult problem in the MCMC practise and there is no single remedy.

A very important concern with MCMC is the within-sequence serial correlation, i.e. the samples we draw are correlated. Although in the long run the iterative simulation draws converge to the target posterior distribution, nearby samples are correlated between them. If this correlation is high, then the convergence to the equilibrium could be slow.

Moreover, when a Markov chain appears to have converged to the equilibrium

distribution when it has not, we are faced with the phenomenon called pseudo-convergence. This can happen when the state space is poorly connected by the Markov chain dynamics and it takes a lot of time to move from one part of the state space to the others.

In order to handle those issues, it is often considered good practise in the MCMC literature (Gelman et al. 2014) to run multiple different chains with starting values dispersed in the state space (after disregarding an initial sample known as burn-in). The convergence is examined by comparing the variation between and within simulated sequences. This is achieved by estimating the $\hat{R}$ statistic, suggested in Gelman & Rubin (1992), for monitoring the convergence of scalar estimands. More specifically, if we assume that we have $m$ parallel chains, each of length $n$, and $B$ is the between sequence variance and $W$ the within sequence variance, then this statistic is defined as:

$$\hat{R} = \sqrt{\frac{\hat{\text{var}}^+(\psi|y)}{W}}, \quad \text{where } \hat{\text{var}}^+(\psi|y) = \frac{n-1}{n}W + \frac{1}{n}B$$

with $\hat{\text{var}}^+(\psi|y)$ representing the marginal posterior variance of the estimand. If the value of $\hat{R}$ is close to 1, this is an indication that the chains have converged to the target posterior distribution.

Furthermore, if the efficiency of the our MCMC simulation scheme is too low -in other words we need a lot of time to obtain convergence to the equilibrium-, we should consider using a more efficient approach. Many extensions have been proposed in the literature with the aim to increase the efficiency (see Brooks et al. 2011). The paper by van Dyk & Meng (2001) contains an extended discussion on techniques and extensions for improving the convergence and reducing the auto-correlation in complex MCMC problems. The Blocked Gibbs Sampler (Roberts & Sahu 1997) groups some of the parameters together and samples them from their joint distribution given the state of the other parameters. This is particularly useful in cases which some parameters could be physically grouped together or in the case of latent variables. In the case where the posterior conditional distribution cannot be sampled directly, we can employ the Metropolis- Hastings algorithm giving rise to the Metropolis-within-Gibbs sampling scheme (using the terminology by Gilks et al. 1995). A rather interesting approach that we explore in our research is the Partially Collapsed Gibbs (PCG) sampler.

**Partially Collapsed Gibbs (PCG)**: The PCG, as described by van Dyk & Park (2008) and Park & van Dyk (2009), is based on the notion that by reducing conditioning, we increase the variance of the complete conditional distributions of a Gibbs sampler. In practise, this translates to replacing a subset of the complete conditional distributions by distributions that condition on fewer of the unknown quantities.

In the Gibbs sampler terminology, collapsing refers to the process of integrating a joint posterior distribution over a subset of unknown quantities. This process results to a "collapsed" posterior distribution which offer an improved rate of convergence (Liu et al. 1994). The collapsing though can be challenging to implement under some scenarios, such as when the complete conditional distributions of the collapsed posterior distributions might be harder to work with than the conditional distributions of the original posterior distribution.

The partially collapsed Gibbs sampler aims to combine conditional distributions from the original posterior distribution with conditional distributions from one or more collapsed posterior distributions. In other words, in PCG sampling, we collapse only those conditional distributions that offer computational advantage without complicating the parameters updating. This strategy should be used carefully, since the resulting conditional distributions may not be functionally compatible and changing the order of the draws in the Gibbs scheme might alter the equilibrium distribution (see van Dyk & Jiao 2015).

### 1.3.3 Hierarchical Models

The term hierarchical (or multi-level) models is used to describe statistical models that involve parameters that can be regarded as connected in some manner by the structure of the problem. This dependence should be reflected in the joint probability model. Hierarchical models are appropriate for many statistical applications in which the parameters have a hierarchical structure. For instance, in astronomy the measurement of the population of a type of astronomical sources can be regarded as a hierarchical model since the population is described by some population parameters and the properties of the individual objects are described by the object-level

parameters.

Hierarchical models are especially useful, since for many statistical applications, we are mostly interested in a subset of the parameters that are considered to refer to the entire population. Thus, we are interested in the marginalised posterior of those parameters. Furthermore, another motivation about the hierarchical formulation is that the data are directly linked only to the object- level parameters. For instance, in astronomy a population itself cannot be measured explicitly but rather only through individual astronomical objects.

In order to formulate the structure of hierarchical models, we explore a simple two-level model in which we split the parameter space $\theta$ into two parts, i.e. $\theta = (\phi, \psi)$, where $\phi$ are the population parameters and $\psi$ the object-level parameters. Suppose we observe $n$ objects from a population, where the data for object $i$ can be modelled as $p(\psi_i|y_i)$ and the object level parameter as $p(\psi_i|\phi)$. Then, the joint prior distribution of all the parameters can be written as:

$$p(\theta) = p(\phi, \psi) = p(\psi \mid \phi) \cdot p(\phi)$$

Thus, the joint posterior distribution of the population and object-level parameters is:

$$p(\phi, \psi \mid y) \propto p(\phi) \cdot p(\psi \mid \phi) \cdot p(y \mid \psi, \phi)$$

It is often the case that we are not interested on the properties of individual objects, but rather on the population parameters. The power of hierarchical models lies on the fact that we can integrate over the object-level parameters and obtain a marginal posterior distribution of the population parameters, i.e.

$$p(\phi \mid y) = p(\phi) \int p(\psi|\phi) p(y|\psi) d\psi$$

Furthermore, if we are interested in the properties of a specific object $i$, the posterior can be marginalised as

$$p(\psi_i|y_i) \propto \int p(\phi) \prod_{j=1, j\neq i}^{N} \Big[ \int p(\psi_j|\phi) p(y_j|\psi_j) d\psi_j \Big] d\phi,$$

which is the Bayesian equivalent of shrinkage[§] estimation (see Efron & Morris 1973, 1975). In other words, the posterior knowledge of the properties of an object is increased not just by the measurement associated with that object, but also by the constraints on the population from which it has come through.

The structure of the hierarchical models is such that Gibbs sampling could easily be applied in order to draw samples from the posterior distribution. More specifically, the posterior distribution of object $i$ conditioned on all other parameters can be written as:

$$p(\psi_i|\phi, y, \psi_{-i}) = p(\psi_i|\phi) p(y_i|\psi_i),$$

and the conditional probability of the population level parameters can be written as

$$p(\phi|\psi, y) = p(\phi) \prod_{i=1}^{N} p(\psi_i|\phi).$$

[§]The term shrinkage refers to the idea that an estimator is improved by combining it with other information.

Thus, drawing samples from the posterior distribution can be achieved through Gibbs sampling, although the exact sampling strategy for each individual conditional distribution depends on the form of that distribution (might have to utilise Metropolis within Gibbs).

### 1.3.4 Choice of Priors

The choice of priors distributions in a Bayesian context is a key part of Bayesian inference. The combination of the prior distribution with the probability of the data results to the posterior distribution. There are two main issues with choosing a prior distribution; (a) what information about the unknown parameter does the prior distribution contains, and (b) what are the properties of the resulting posterior distribution.

For the first issue, in a statistical application with well-identified parameters and large sample sizes, reasonable choices of prior distributions will have minor effects on posterior inferences. The dependence of the posterior inference on the choice of the prior distribution should be checked by a sensitivity analysis: comparing posterior inferences under different reasonable choices of prior distribution. However, if the sample size is small, or if the available data provide only indirect information about the parameters of interest, the choice of the prior distribution has a bigger impact on the posterior inference. In case of hierarchical models, the priors can be set up hierarchically, so that clusters of parameters have shared prior distributions, which can themselves be estimated from data.

The properties of the resulting posterior distribution should be also taken into consideration in the choice of a prior distribution. For instance, if the posterior distribution is in the same family as the prior distribution, then the prior is called a conjugate prior for the likelihood distribution. This offers many computational advantages in the posterior inference. A very important issue with the choice of prior is to ensure that the resulting posterior distribution is proper probability distribution. Kass & Wasserman (1996) provide a more in-depth discussion about the theoretical principles regarding the selection of priors.

## 1.4 Outline

In Chapter 2, we describe the interesting approach of Udaltsova (2014), which models the estimation of $\log(N) - \log(S)$ as a hierarchical Bayesian problem. Subsequently, we present our proposed extensions of this model and we perform an extended model validation in 2 different ways. The application of our method on the Chandra Deep Filed South dataset is presented in the last part of the Chapter.

In Chapter 3 we develop a methodology for incorporating the uncertainty of the flux-to-count rate conversion factor $\gamma$ to the Bayesian hierarchical model we developed for estimating the $\log(N) - \log(S)$ relationship. We discuss how we can construct a prior distribution for $\gamma$ for the sources of a specific astronomic survey, both for the observed and the unobserved sources. Afterwards, we extend the Bayesian hierarchical model by including the uncertainty of $\gamma$ and extract the joint posterior distribution of all the parameters of interest. The sampling algorithm for drawing samples from this posterior distribution is thoroughly analysed, and we also present model validation results. Finally, we apply our methodology to the Chandra Deep Field South survey, and we compare the results with those from the model that assumes constant $\gamma$ for all the sources.

Chapter 4 revolves around the problem of classifying galaxies to different activity classes based on emission line ratios. Initially, the scientific problem is presented, including the existing approaches. Our innovative, data-driven approach is described in detail along with the actual implementation. The performance of our multidimensional data driven classification scheme is compared with the most commonly used scheme. We also introduce multidimensional linear decision boundaries that we compare in terms of their prediction accuracy with both our new method and the most commonly used scheme. Finally, we discuss how this classification scheme can be combined with the hierarchical Bayesian method we developed in the first two Chapters in order to produce population specific $\log(N) - \log(S)$ curves.

Finally, Chapter 5 contains a discussion of this research and propose further avenues of research.

# 2

# Bayesian Analysis of the $\log(N) - \log(S)$ Problem

The study of the population properties of the flux is of high importance as we saw in the previous chapter. Modelling the distribution of the fluxes for populations of astronomical sources provides information about the stellar evolution and the geometry of the universe.

The first section of this chapter describes the interesting approach of Udaltsova (2014), which models the estimation of $\log(N) - \log(S)$ as a hierarchical Bayesian problem. This section summarises the main points provided from Udaltsova (2014), although some formulas have beer re-derived in the current work to properly correct mathematical errors found in Udaltsova (2014).

The following sections of the chapter include the original work developed in our research. More specifically, Section 2 describes our proposed extension of this model, while Section 3 contains the validation of our algorithm performed with 2 different ways and Section 4 discusses posterior inference and model validation. Section 5 presents the application of our method on the Chandra Deep Filed South dataset, followed by a small discussion in Section 6.

## 2.1 Probability Modelling of the $\log(N) - \log(S)$ Relationship

As we discussed in the previous chapter, the early work of cosmologists was assuming a linear $\log(N) - \log(S)$ relationship, which means that the flux density distribution follows a power law distribution. Following the notation from Udaltsova (2014) and assuming that we have an ideal, complete survey with no missing data, if we define as $S_i$ $(ergs^{-1}cm^{-2})$ the flux of each astronomical source $i$ , with $i = 0, \ldots, N$, and $N(> S)$ the number of sources with flux greater than $S$, then the power law assumption gives rise to a relationship of the form:

$$N(> S) = \sum_{i=1}^{N} I_{S_i > S} \propto aS^{-\theta}, S > \tau > 0 \tag{2.1}$$

where $\tau$ is a positive constant indicating the minimum flux. The log transformation of this relationship leads to a linear relationship between $\log(N)$ and $\log(S)$, i.e.

$$\log_{10} N(> S) = \log_{10} a - \theta \log_{10} S \tag{2.2}$$

However, it is common for many astronomical datasets for the $\log(N) - \log(S)$ relationship to appear piece-wise linear or curved. In this case, the distribution of the flux is represented by multiple power laws, connected at the knots. This model is known as broken power law in the astronomical literature (Zezas & Fabbiano 2002, Jóhannesson et al. 2006, Wong et al. 2014). For example, the piece wise linear relationship of $\log(N) - \log(S)$ with 1 break is of the form:

$$\log_{10} N(> S) = \begin{cases} a_1 - \theta_1 \log_{10}(S), & \tau_1 \leq S < \tau_2 \\ a_2 - \theta_2 \log_{10}(S), & S \geq \tau_2 \end{cases} \tag{2.3}$$

where $\tau_1$ is the minimum population threshold and $\tau_2$ is the breakpoint.

Section 2.1.1 describes the probabilistic modelling of the single power law model (linear $\log(N) - \log(S)$ relationship), while section 2.1.3 extends the modelling to the broken power law case.

### 2.1.1 Single Power Law Model

The single power law model for the distribution of the flux corresponds to a linear $\log(N) - \log(S)$ relationship. In the introductory chapter we discussed about the incompleteness of astronomical surveys, which as a result provide us only with a subset of the 'complete source population', the 'observed source population'. The $\log(N) - \log(S)$ relationship we are interested in is about the complete source population and not the observed source population which is a biased subset of the complete.

As it was noted in the previous Chapter, incompleteness is an inherent challenge in astronomical surveys. Missing data arise as a result of many reasons; fainter sources are less likely to be observed due to detector sensitivity or Poisson like noise in observed photon counts. The statistical question that arises is whether the subset of the data that is observed is a biased subset of the population.

In the $\log(N) - \log(S)$ problem, the probability of observing a source depends on the source flux. Thus, the missing data mechanism is non-ignorable since the observed data is a biased subset of the complete population. In order to accurately model the missing data mechanism, we should use all the external knowledge. In the $\log(N) - \log(S)$ framework, this knowledge is available and we will describe in detail the incompleteness function in the following subsections.

Udaltsova (2014) proposes a hierarchical Bayesian model which consists of a model about the distribution of the flux in the complete source population, an incompleteness function representing the missing data mechanism that describes the selection mechanism of the observed source population as well as a model for all the observable quantities and detector uncertainties. More specifically, Udaltsova (2014) proves that the linear relationship between $\log(N)$ and $\log(S)$ is equivalent to the statement that the flux follows a Pareto distribution, $S_i \sim \text{Pareto}\,(\theta, \tau)$, for which the probability density function is:

$$f(S_i|\theta,\tau) = \theta\tau^{\theta} S_i^{-(\theta+1)}, \quad S_i > \tau, \quad \tau, \theta > 0 \tag{2.4}$$

where $\theta$ denotes the power-law slope and $\tau$ is the flux population minimum threshold. It is important to note that $\tau$ does not represent the lowest detection threshold of the detector, rather the theoretical limit that we could observe a source. We assume for convenience conjugate prior distributions for $\theta$ and $\tau$, namely Gamma distributions; $\theta \sim \Gamma(a_\theta, b_\theta)$ and $\tau \sim \Gamma(a_m, b_m)$. As a reminder, if a random variable $x$ follows a Gamma distribution with shape parameter $a$ and rate parameter $b$, then the density is:

$$f(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} \times 10^{-bx}, \quad a > 0, b > 0. \tag{2.5}$$

Due to the missing data mechanism we only observe a subset of the complete source population. If we define as $n$ the number of observed sources, $N$ the unknown total number of sources and $N_{mis}$ the number of missing sources we have that $N = n + N_{mis}$. We assume a Negative-Binomial prior for the total number of sources, i.e. $N \sim$ Neg-Bin $(a_N, b_N)$, with probability density function given by:

$$f(N \mid a_N, b_N) = \binom{N + a_N - 1}{a_N - 1} \Big(\frac{1}{1+b_N}\Big)^N \Big(\frac{b_N}{1+b_N}\Big)^{a_N}. \tag{2.6}$$

Another important aspect that should be taken into consideration is the fact that we do not directly observe the flux. The data collection consists of photon counts for the observed source population. The photon counts that we measure on the detector for each source $i$ include both the counts from the actual astronomical source, $Y_i^{src}$, and the background contamination, $Y_i^{bkg}$, i.e. $Y_i^{tot} = Y_i^{src} + Y_i^{bkg}$. Moreover, the number of detected counts is affected by:

- The off-axis angle $L$, the distance from the centre of the field of view. The detector sensitivity decreases as we move away from the centre.

- The per-pixel photon background rate for the source $B$ (counts/pixel).

- The exposure map $E$. The exposure map for a given observation combines the effective area of the telescope and detector with a map of the dwell time versus

sky position, accumulated by following the telescope pointing motion during the observation. The effective area varies with position on the detector and is also energy dependent. At a given energy, the effective area as a function of position on the detector is called the instrument map. The map of dwell time versus pointing direction, built up by the telescope pointing motion, is called the aspect histogram. Combining the instrument map with the aspect histogram, the exposure map is a map of the total exposure as a function of position on the sky.

- There is also a dependence on the flux to count rate conversion factor $\gamma$, which depends on the spectral model that is assumed and the energy band of the source. In the majority of the relevant literature so far, the factor $\gamma$ is assumed to be constant for all the sources. For this chapter we will assume that it is constant for all the sources as in the relevant literature. In the next chapter we will develop a methodology for incorporating the uncertainty of this factor to the model.

Taken all the above into consideration, we define for each source i:

$$Y_i^{tot} = Y_i^{src} + Y_i^{bkg}, \tag{2.7}$$

$$Y_i^{src} | S_i, E_i, \gamma_i \overset{\text{ind}}{\sim} Poisson\ (\ \lambda(S_i, E_i, \gamma_i)\ ), \tag{2.8}$$

$$Y_i^{bkg} | B_i, A_i \overset{\text{ind}}{\sim} Poisson\ (\ k(B_i, A_i)\ ), \tag{2.9}$$

where $A_i$ denotes the background area of the source, $\lambda(S_i, E_i, \gamma_i) = S_i E_i / \gamma_i$ and $k(B_i, A_i) = B_i A_i$. The quantities $(E_i, B_i, L_i, A_i)$ are known for all the observed sources. Udaltsova (2014) assumes that the distributions of those parameters for the unobserved sources are the same as those of the observed sources. We will suggest a different assumption for that joined distribution in the next section, which will be more survey specific.

In the previous subsections, we elaborated that the probability of an astronomical source being detected depends on parameters such as the background, the off-axis angle etc. Udaltsova (2014) proposes a detection probability curve $g = g(S, B, L, E, \gamma)$,

which gives us the probability of observing an astronomical source for give flux, background, off-axis angle and exposure value. We will also refer to the function $g$ as the incompleteness function. We define an indicator variable $I_i$ which indicates whether a source is detected ($I_i = 1$ for detection or $I_i = 0$ otherwise). As expected, $I_i = 1$ with probability $g(S_i, E_i, B_i, L_i, \gamma)$.

### 2.1.2 Computational Details of the Single Power Law Model

In the previous section we define the probability modelling for the flux distributions. In order to obtain the posterior distribution, we combine all the model assumptions with the prior distributions. Define as $S_{\text{com}} = (S_{\text{obs}}, S_{\text{mis}})$ the flux vector of observed and missing sources. Similarly $Y^{\text{tot}}_{\text{obs}} = (Y^{\text{tot}}_{i=1}, ..., Y^{\text{tot}}_{i=n})$, $Y^{\text{tot}}_{\text{mis}} = (Y^{\text{tot}}_{i=n+1}, ..., Y^{\text{tot}}_{i=N})$ where $i = 1, .., n$ corresponds to the observed sources. The complete data posterior distribution can be summarised as:

$$p(N, \theta, \tau, S_{\text{com}}, I_{\text{com}}, Y^{\text{src}}_{\text{obs}}, Y^{\text{src}}_{\text{mis}}, Y^{\text{tot}}_{\text{mis}}, B_{\text{mis}}, L_{\text{mis}}, E_{\text{mis}} | n, Y^{\text{tot}}_{\text{obs}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}}) \tag{2.10}$$

The missing data in this posterior distribution gives rise to significant computational difficulties. The total number of sources $N$ in the population is unknown, so the dimension of the complete data posterior distribution will change through every iteration of the MCMC. So, instead of sampling a great number of missing parameters in each iteration we implement another sampling strategy. More specifically, we marginalise the full joint posterior distribution over the missing sources, i.e. we integrate out the missing data parameters $(S_{\text{mis}}, I_{\text{mis}}, Y^{\text{src}}_{\text{mis}}, Y^{\text{tot}}_{\text{mis}}, B_{\text{mis}}, L_{\text{mis}}, E_{\text{mis}})$. This leaves the main parameters of interest $(N, \theta, \tau)$ and the missing parameters flux and photon counts $(S_{\text{obs}}, Y^{\text{src}}_{\text{obs}})$ of the observed sources in the marginalised joint posterior. By using this sampling scheme, the dimension of the sampled quantities is kept constant.

If we define as $\lambda_i = \lambda(S_i, E_i, B_i, L_i)$ and $\kappa_i = \kappa(B_i, A_i)$, then the marginalised joint-posterior is (see Udaltsova (2014) for details of the derivation):

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

where the $\pi(\theta, \tau)$ denotes the marginal probability of observing a source as a function of $\theta$ and $\tau$, i.e.

$$
\begin{aligned}
\pi(\theta, \tau) &= \int g(I | S, B, L, E) \cdot p(S, B, L, E | \theta, \tau) \, \mathrm{d}S \, \mathrm{d}B \, \mathrm{d}E \, \mathrm{d}L \\
&= \int g(I | S, B, L, E) \cdot p(S | \theta, \tau) \cdot p(B, L, E) \, \mathrm{d}S \, \mathrm{d}B \, \mathrm{d}E \, \mathrm{d}L
\end{aligned} \tag{2.11}
$$

The sampling of such a complicated posterior distribution is not trivial. Udaltsova (2014) suggests utilising a Blocked Gibbs sampler in which each conditional posterior distribution requires a different sampling strategy. The details can be found at Udaltsova (2014). The Gibbs sampler algorithm can be described as follows:

**Blocked Gibbs sampler**: For number of iterations $i = 1, ..., T$:

- Sample $\boldsymbol{Y}_{\text{obs}}^{\text{src}}$ component-wise for the observed sources $i = 1, ..., n$ as:

$$p(Y_i^{\text{src}}|\cdot) \sim Binomial\Big(Y_i^{\text{src}}; Y_i^{\text{tot}}, \frac{\lambda_i}{\lambda_1 + \kappa_i}\Big).$$

- Sample $\boldsymbol{S_{obs}}$ component-wise for $i = 1, ..., n$ as:

$$p(S_i|\cdot) \sim Pareto(S_i|N, \theta, \tau)\mathbb{I}_{\{\tau<S_i\}} \cdot g(S_i, B_i, L_i, E_i) \cdot Poisson(Y_i^{tot}; \lambda_i + \kappa_i)$$
$$\cdot Binomial\Big(Y_i^{src}; Y_i^{tot}, \frac{\lambda_i}{\lambda_1 + \kappa_i}\Big).$$

  where $\lambda(S_i, E_i, B_i, L_i) = S_i E_i/\gamma$ and $\kappa(B_i, A_i) = B_i A_i$.

- Sample $\boldsymbol{\theta}$ as:

$$p(\theta|\cdot) \propto (1 - \pi(\theta, \tau))^{N-n} \cdot Gamma\Big(\theta; a + n, b + \sum_{i=1}^{n} log(\frac{S_i}{\tau})\Big)$$

- Sample $\boldsymbol{N}$ as:

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

- Sample $\boldsymbol{\tau}$ as:

$$p(\tau|\cdot) \propto \tau^{n\theta + a_m - 1} \cdot e^{-b_m \tau} \cdot (1 - \pi(\theta, \tau))^{N-n} \cdot \mathbb{I}_{\{\tau<c_m\}}$$

  where $c_m = min(S_1, ..., S_n)$.

### 2.1.3 Broken Power-Law Models

We discussed that using a piece-wise linear $\log(N) - \log(S)$ approach is common in the relevant literature and more appropriate for some astrophysical populations.

The Chandra Deep Field South dataset, that we examine at the last section of this Chapter, shows strong evidence of a broken power law model.

For the cases where the $\log(N) - \log(S)$ curve appears to be curved or piece-wise linear, the single power law model fails to capture this non-linearity. Assuming a known number $m$ of pieces in the $\log(N) - \log(S)$ relationship, Udaltsova (2014) proves that a mixture of $m-1$ truncated Pareto distributions and an untruncated Pareto distribution corresponds to a piece-wise linear $\log(N) - \log(S)$ relationship with $m$ linear sections. This $m$-component broken power-law distribution for the flux $S$ has density:

$$f(S) = p(S|\theta_1, ..., \theta_m, \tau_1, ..., \tau_m) = \sum_{j=1}^{m} \Big[ \prod_{i=1}^{j-1} \Big( \frac{\tau_{i+1}}{\tau_i} \Big)^{-\theta_i} \Big] \Big( \frac{\theta_j}{\tau_j} \Big) \Big( \frac{S}{\tau_j} \Big)^{-(\theta_j+1)} \mathbb{I}_{\{\tau_j \leq S \leq \tau_{j+1}\}}, \tag{2.12}$$

where $\theta = (\theta_1, ..., \theta_m)$ are the $m$ power-law slopes, $\tau_1$ is the flux population minimum threshold, $(\tau_2, ..., \tau_m)$ are the consequent breakpoints and $\prod_{i=1}^{0} \left( \frac{\tau_{i+1}}{\tau_i} \right)^{-\theta_i} = 1$.

The construction of the posterior distribution for the multiple power-law model follows the same structure to that of the single power law model. The only differences are the density for the flux complete source population, which is no longer a Pareto distribution, and the additional prior distributions for the parameters of the $m$-component broken power-law distribution. More specifically, conditional on the total unknown number of sources, the source fluxes for the complete source population follow a $m$-component broken power-law distribution. The priors are chosen as in Udaltsova (2014). i.e. Gamma prior distributions for the $m$ power slopes $(\theta_1, ..., \theta_m)$ and $\tau_1$, i.e. $\theta_j \sim \text{Gamma}(a_j, b_j)$, $j = 1, ..., m$ and $\tau_1 \sim \text{Gamma}(a_\tau, b_\tau)$.

In order to preserve the non-negativity and increasing order of $\tau_2, ..., \tau_m$, the transformation $\eta_j = h_j(\tau_j \mid \tau_{j-1}) = \log(\tau_j - \tau_{j-1})$, $j = 2, ..., m$ is introduced. Thus, $\tau_2, ..., \tau_m$ can be expressed as

$$(\tau_2, ..., \tau_m)^T = \begin{pmatrix} \tau_1 + e^{\eta_2} \\ \tau_1 + e^{\eta_2} + e^{\eta_3} \\ \vdots \\ \tau_1 + \sum_{j=2}^{m} e^{\eta_j} \end{pmatrix} \tag{2.13}$$

which keeps the non-negativity and increasing order of the consecutive breakpoints. For the transformed variables $\eta = (\eta_2, ..., \eta_m)^T$ we assume a Multivariate Gaussian distributions as prior, i.e. $\eta \sim$ Multivariate Gaussian$(\mu, C)$ with $\mu = (\mu_2, ..., \mu_m)$ and $C = \text{diag}\{c_2^{-1}, ..., c_m^{-1}\}$.

In order to compute the joint prior distribution $p(\tau_1, \tau_2, ..., \tau_m)$, we will use the change of variable formula and the Jacobian matrix. More specifically,

- For the case with 1 break, i.e. m=2, we have that

$$\begin{aligned} p(\eta_2) &\propto e^{-\frac{c_2^2 \cdot (\eta_2 - \mu_2)^2}{2}} \\ \text{Define} \quad p(\tau_1) &\propto \tau_1^{a-1} e^{-b\tau_1} \\ \text{Thus} \quad p(\tau_1, \eta_2) &\propto e^{-\frac{c_2^2 \cdot (\eta_2 - \mu_2)^2}{2}} \cdot \tau_1^{a-1} e^{-b\tau_1} \end{aligned}$$

  If we define

$$\left.\begin{aligned} \tau_2(\tau_1, \eta_2) &= \tau_1 + e^{\eta_2} \\ T(\tau_1, \eta_2) &= \tau_1 \end{aligned}\right\} \implies \left.\begin{aligned} \tau_1(T, \tau_2) &= T \\ \eta_2(T, \tau_2) &= \log(\tau_2 - T) \end{aligned}\right\}$$

  So, the Jacobian is

$$\begin{vmatrix} \frac{\partial \tau_1(T,\tau_2)}{\partial \tau_2} & \frac{\partial \tau_1(T,\tau_2)}{\partial T} \\ \frac{\partial \eta_2(T,\tau_2)}{\partial \tau_2} & \frac{\partial \eta_2(T,\tau_2)}{\partial T} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ \frac{1}{\tau_2 - T} & \frac{-1}{\tau_2 - T} \end{vmatrix} = \frac{-1}{\tau_2 - T}$$

  Thus, we conclude that the joint prior distribution $p(\tau_1, \tau_2)$ is

$$p(\tau_1, \tau_2) \propto \|\mathrm{J}\| \cdot e^{-\frac{c_2^2 \cdot [\log(\tau_2 - \tau_1) - \mu_2]^2}{2}} \cdot \tau_1^{a-1} e^{-b\tau_1} \tag{2.14}$$
$$\propto e^{-\frac{c_2^2 \cdot [\log(\tau_2 - \tau_1) - \mu_2]^2}{2}} \cdot \tau_1^{a-1} e^{-b\tau_1} \cdot \frac{1}{\tau_2 - \tau_1}$$

- For the case with 2 breaks, i.e. m=3, we can conclude following the same logic that the joint prior distribution $p(\tau_1, \tau_2, \tau_3)$ is

$$p(\tau_1, \tau_2, \tau_3) \propto e^{-\frac{c_2^2 \cdot [\log(\tau_2 - \tau_1) - \mu_2]^2}{2}} \cdot e^{-\frac{c_3^2 \cdot [\log(\tau_3 - \tau_2) - \mu_3]^2}{2}} \cdot \tau_1^{a-1} e^{-b\tau_1} \cdot \frac{1}{\tau_2 - \tau_1} \cdot \frac{1}{\tau_3 - \tau_2} \tag{2.15}$$

If we define as $\theta = (\theta_1, ..., \theta_m)$ and $\tau = (\tau_1, ..., \tau_m)$, then the **posterior distribution** is simply the following:

$$
\begin{aligned}
&p(N, \theta, \tau, S_{obs}, Y_{obs}^{src} | n, Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs}, A_{obs}) \\
&= \frac{1}{p(n, Y_{\text{obs}}^{\text{tot}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}})} \cdot \binom{N}{n} \cdot (1 - \pi(\theta, \tau))^{N-n} \\
&\quad \cdot p(N) \cdot p(\theta) \cdot p(\tau) \cdot p(B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}} | N, \theta, \tau) \\
&\quad \cdot p(S_{\text{obs}} | \theta, \tau) \cdot p(I_{\text{obs}} \mid \gamma_{\text{obs}}, S_{\text{obs}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}) \\
&\quad \cdot p(Y_{\text{obs}}^{\text{tot}} \mid I_{\text{obs}}, S_{\text{obs}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, \gamma_{\text{obs}}) \\
&\quad \cdot p(Y_{\text{obs}}^{\text{src}} \mid Y_{\text{obs}}^{\text{tot}}, I_{\text{obs}}, S_{\text{obs}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, \gamma_{\text{obs}}) \\
&\propto \binom{N}{n} \mathbb{I}_{\{n \leq N\}} (1 - \pi(\theta, \tau))^{N-n} \\
&\cdot \binom{N + a_N - 1}{a_N - 1} \Big( \frac{1}{1 + b_N} \Big)^N \Big( \frac{b_N}{1 + b_N} \Big)^{a_N} \mathbb{I}_{\{N \in \mathbb{Z}^+\}} \\
&\cdot \prod_{j=1}^{m} \frac{b_j^{a_j}}{\Gamma(a_j)} \theta_j^{a_j - 1} e^{-b_j \theta} \mathbb{I}_{\{\theta_j > 0\}} \\
&\cdot p(\tau_1, \tau_2, ..., \tau_m) \mathbb{I}_{\{0 < \tau_1 < \tau_2 < ... < \tau_m\}} \\
&\cdot \Big[ \prod_{i=1}^{n} p(B_i, L_i, E_i) \cdot g(S_i, B_i, L_i, E_i) \\
&\cdot \sum_{j=1}^{m} \Big[ \prod_{i=1}^{j-1} \Big( \frac{\tau_{i+1}}{\tau_i} \Big)^{-\theta_i} \Big] \Big( \frac{\theta_j}{\tau_j} \Big) \Big( \frac{S}{\tau_j} \Big)^{-(\theta_j + 1)} \mathbb{I}_{\{\tau_j \leq S \leq \tau_{j+1}\}} \\
&\cdot \frac{(\lambda_i + \kappa_i)^{Y_i^{tot}}}{Y_i^{tot}!} e^{(\lambda_i + \kappa_i)} \mathbb{I}_{\{Y_i^{tot} \in \mathbb{Z}^+\}} \\
&\cdot \binom{Y_i^{tot}}{Y_i^{src}} \Big( \frac{\lambda_i}{\lambda_i + \kappa_i} \Big)^{Y_i^{src}} \Big( 1 - \frac{\lambda_i}{\lambda_i + \kappa_i} \Big)^{Y_i^{tot} - Y_i^{src}} \mathbb{I}_{\{Y_i^{src} \in \{0, 1, ..., Y_i^{tot}\}\}} \Big]
\end{aligned}
$$

This posterior distribution is not massively different from the single Pareto model posterior distribution. The conditional distributions for $N$ and $Y_{\text{obs}}^{\text{src}}$ are the same as in the single Pareto model, except for the different computation of $\pi(\theta, \tau)$ in which we replace the Pareto distribution for the flux $S$ with the broken Pareto pdf. The main computational differences in the Gibbs sampler lie in sampling from the conditional posterior distributions of $p(\tau_1 | \cdot)$, $p(\tau_2, ..., \tau_m | \cdot)$ and $p(\theta_1, ..., \theta_m | \cdot)$. More specifically, we have that:

- Sample $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, ..., \boldsymbol{\theta_m})$ as:

$$p(\theta\ |) \propto [1-\pi(\theta,\tau)]^{N-n} \cdot \prod_{j=1}^{m} \text{Gamma}\Big(\theta_j; a_j + n(j) - 1, \tag{2.16}$$
$$b_j + \mathbb{I}_{\{j\neq m\}} \log(\frac{\tau_{j+1}}{\tau_j}) \sum_{i=1}^{m} [n(i)\mathbb{I}_{\{i\geq j+1\}}] + \sum_{i\in I(j)} \log(\frac{S_i}{\tau_j})$$

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

- Sample $\boldsymbol{\tau_1}$ as:

$$p(\tau_1|\cdot) \propto [1-\pi(\theta,\tau)]^{N-n} \cdot p(\tau_1, \tau_2, ..., \tau_m)\mathbb{I}_{\{0<\tau_1<\tau_2<...<\tau_m\}} \cdot \tag{2.17}$$
$$\Big[\prod_{i=1}^{n} p(B_i, L_i, E_i) \cdot g(S_i, B_i, L_i, E_i)$$
$$\cdot \sum_{j=1}^{m} \Big[\prod_{i=1}^{j-1} \Big(\frac{\tau_{i+1}}{\tau_i}\Big)^{-\theta_i}\Big] \Big(\frac{\theta_j}{\tau_j}\Big) \Big(\frac{S}{\tau_j}\Big)^{-(\theta_j+1)} \mathbb{I}_{\{\tau_j\leq S\leq\tau_{j+1}\}}\Big]$$
$$\propto \tau^{n\theta_1+a_m-1} \cdot e^{-b_m\tau} \cdot (1-\pi(\theta,\tau))^{N-n} \cdot \prod_{j=2}^{m} e^{-\frac{c_m^2\cdot[\log(\tau_m-\tau_{(m-1)})-\mu_m]^2}{2}} \cdot$$
$$\frac{1}{\tau_m - \tau_{(m-1)}} \cdot \mathbb{I}_{\{\tau<c_m\}}$$

- Sample $(\boldsymbol{\tau_2}, \boldsymbol{...\tau_m})$ via the transformed variables $\eta_2, ..., \eta_m$ as:

$$p(\eta_2, ...\eta_m|\cdot) \propto e^{\sum_{j=2}^{m}\eta_j} \cdot [1-\pi(\theta,\tau)]^{N-n} \cdot \text{Multivariate Gaussian}(\mu, C) \tag{2.18}$$
$$\prod_{i=1}^{n} p(S_i|\theta_1, ..., \theta_m, \tau_1, ..., \tau_m) \cdot \mathbb{I}_{\{0<\tau_1<\tau_2<...<\tau_m\}}$$

## 2.2 Extending the model

### 2.2.1 Proposed extensions

In the previous section we discussed the probability modelling for both the single power law and broken power law model. We propose a series of extensions to the model proposed by Udaltsova (2014) regarding the estimation and the sampling from the joint distribution of the background noise $B$, the off axis angle $L$ and the exposure map $E$, $p(B, L, E)$, the selection of the incompleteness function $g = g(S, B, L, E)$ and the computation of the integral $\pi(\theta, \tau)$.

**Distribution of $p(B, L, E)$**: The distribution of the background noise $B$, the off axis angle $L$ and the exposure map $E$, $p(B, L, E)$, is not straightforward to define. We do have the values of those parameters for the observed sources, however we should remember that the observed sources are a biased subset of the population, so defining the distribution of $p(B, L, E)$ based on the observed values will not be accurate.

In this research, we assume a more generic approach for the probability distribution $p(B, L, E)$ that reflects the individuals characteristics of each survey. More specifically, we make the assumption that the astronomical sources are uniformly scattered in the universe. Each survey contains an exposure map and a background map. As a result, in order to draw samples from the joint distribution $p(B, L, E)$ we do the following:

- For $E$, we sample uniformly data points on the exposure map under the restriction that the value of the effective area at this datapoint is at least 10% of the maximum value in the exposure map.

- For $L$, the radius of this datapoint $r$ from the centre of the image is used in calculating the off-axis angle, i.e. $L = \frac{0.49r}{60}$.

- For $B$, the value of the background map at this datapoint is utilised as the $B$ for that source.

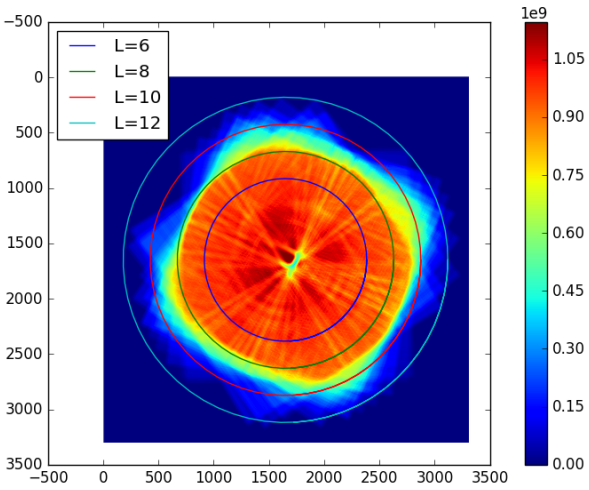



**Figure 2.1:** The exposure map from the Chandra Deep Field South survey. We see that the values of the exposure map are bigger closer to the centre of the detector as expected. The circles correspond to various off-axis angles. It is straightforward to notice that essentially when $L > 10$, the value of the exposure map is very small. In our methodology, we sample uniformly data points on the exposure map under the restriction that the value of the effective area at this datapoint is at least $10\%$ of the maximum value in the exposure map.

Figure 2.1 depicts the exposure map for the Chandra Deep Field survey. We see that the values of the exposure map are bigger closer to the centre of the detector as expected. The circles correspond to various off-axis angles. It is straightforward to notice that essentially when $L > 10$, the value of the exposure map is very small. Figure 2.2 depicts the background map from the Chandra Deep Field South survey. The background is relatively constant for off-axis angle $L < 10$.

**Choice of $g$**: The detection probability curve, or incompleteness function, is also a very important aspect of the model. Different specifications of this function will massively influence the posterior inference. In the previous subsections, we elaborated that the probability of an astronomical source being detected depends on parameters such as the background, the off-axis angle etc.. In this research, the astronomical surveys that we analyse come from NASA's flagship X-ray telescope Chandra. Thus, we use the detection probability curves proposed by Wright et al. (2015). More specifically, they propose a detection probability function of the form:

$$g(C, E, B, L) = 1 - e^{-\frac{C^{\lambda_1}}{10^{\lambda_2}}}, \tag{2.19}$$

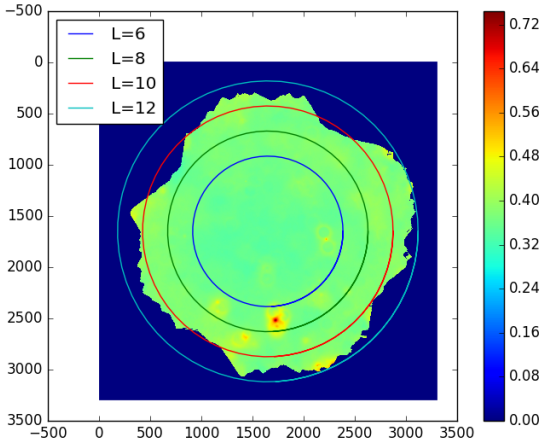



**Figure 2.2:** The background map from the Chandra Deep Field South survey. The background is relatively constant for off-axis angle $L < 10$.

in which the choice of $\lambda_1$ and $\lambda_2$ depends on the background and off-axis angle. In the $g$ function the variable $C$ corresponds to the source counts. We usually have the flux $S$ of each source and use the expected source counts as:

$$C = \frac{S \cdot E}{\gamma} \tag{2.20}$$

in the $g$ function, thus we write it as $g = g(S, B, L, E, \gamma)$. We will also refer to the function $g$ as the incompleteness function. Figure 2.3 shows source detection probability curves as a function of the number of source counts based on Wright et al. (2015). The curves are plotted for different combinations of background $B$ and off-axis angle $L$.

**Computing $\boldsymbol{\pi(\theta, \tau)}$**: One of the most subtle parts of the posterior distribution defined in the precious section is the computation of the $\pi(\theta, \tau)$, which denotes the marginal probability of observing a source as a function of the slope (or slopes) $\theta$ and $\tau$, i.e.

$$\begin{aligned} \pi(\theta, \tau) &= \int g(I|S, B, L, E) \cdot p(S, B, L, E|\theta, \tau)\, \mathrm{d}S\, \mathrm{d}B\, \mathrm{d}E\, \mathrm{d}L \\ &= \int g(I|S, B, L, E) \cdot p(S|\theta, \tau) \cdot p(B, L, E)\, \mathrm{d}S\, \mathrm{d}B\, \mathrm{d}E\, \mathrm{d}L \end{aligned} \tag{2.21}$$

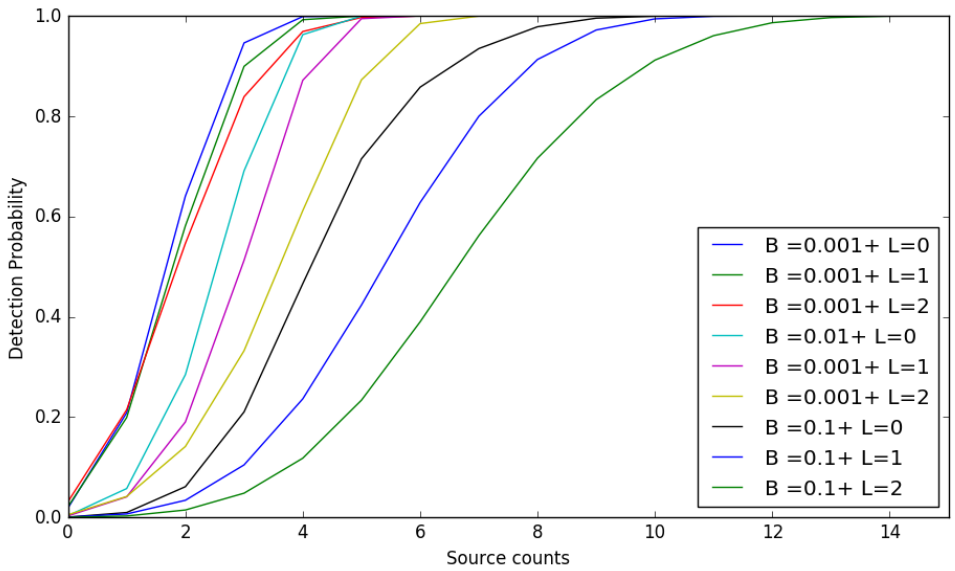



**Figure 2.3:** Source detection probability curves as a function of the number of source counts based on Wright et al. (2015). The curves are plotted for different combinations of background $B$ and off-axis angle $L$.

In order to estimate the marginal probability of observing a source, we have to evaluate the multi-dimensional integral. We can use various numerical techniques in order to evaluate this integral such as Riemann Sums or Monte Carlo integration. After numerical experiments, the Monte Carlo integration yields the best performance in terms of computational time and accuracy. More specifically, we draw samples from the empirical distribution of $B, L, E$ as we discussed above as well as samples of the flux $S$ from a Pareto distribution conditional on the values of $(\theta, \tau)$ if we assume a single power law model (or the $m$-component broken power law distribution of we assume a broken power law model). The empirical average of $g$ is the approximation of $\pi(\theta, \tau)$.

In order to speed up the MCMC, we pre-compute the $\pi(\theta, \tau)$ on a grid of values of $\theta$ and $\tau$ and then use bilinear interpolation. This strategy reduces the running time of the MCMC considerably since the pre-computed surface can be re-used in different MCMC runs. However, creating a dense enough grid is only possible for the single power law model case, in which the grid is 2-dimensional. For the broken power law model, the grid has many dimensions and creating a dense grid requires an enormous amount of time. Thus, we evaluate the integral on the fly while running the Gibbs sampler.

### 2.2.2 Alternative Sampling Methodology

It should be noted that the lower population flux limit $\tau$ is heavily correlated with $c_m$, the smallest $S_{obs}$. In order to break that correlation we develope a Partially Collapsed Gibbs sampler as well, i.e. we integrate the $S_{obs}$ from the marginal posterior distribution $p(\tau, S_{obs}|\cdot)$ in order to sample $\tau$ without conditioning on the state of $c_m$.

More specifically, for the single power law model, the marginal posterior distribution $p(\tau, S_{obs}|\cdot)$ is:

$$
\begin{aligned}
&p(\tau, S_{obs}|N, \theta, n, Y_{obs}^{tot}, Y_{obs}^{src}, B_{obs}, L_{obs}, E_{obs}, A_{obs}) \propto \\
&\propto (1-\pi(\theta,\tau))^{N-n} \cdot \frac{b_m^{a_m}}{\Gamma(a_m)} \tau^{a_m-1} e^{-b_m\tau} \mathbb{I}_{\{\tau>0\}} \\
&\cdot \prod_{i=1}^{n} p(B_i, L_i, E_i) \cdot \theta \tau^{\theta} S_i^{-(\theta+1)} \mathbb{I}_{\{\tau<S_i\}} \cdot g(S_i, B_i, L_i, E_i) \\
&\cdot \frac{(\lambda_i+\kappa_i)^{Y_i^{tot}}}{Y_i^{tot}!} e^{(\lambda_i+\kappa_i)} \mathbb{I}_{\{Y_i^{tot}\in\mathbb{Z}^+\}} \\
&\cdot \binom{Y_i^{tot}}{Y_i^{src}} \Big(\frac{\lambda_i}{\lambda_i+\kappa_i}\Big)^{Y_i^{src}} \Big(1-\frac{\lambda_i}{\lambda_i+\kappa_i}\Big)^{Y_i^{tot}-Y_i^{src}} \mathbb{I}_{\{Y_i^{src}\in\{0,1,\ldots,Y_i^{tot}\}\}} \\
&\propto (1-\pi(\theta,\tau))^{N-n} \cdot \tau^{n\theta+a_m-1} \cdot e^{-b_m\tau} \cdot \prod_{i=1}^{n} S_i^{-(\theta+1)+Y_i^{src}} \cdot e^{-\frac{S_iE_i}{\gamma}} \cdot g(S_i, B_i, L_i, E_i) \mathbb{I}_{\{\tau<S_i\}}
\end{aligned}
$$

So, the marginal posterior distribution $p(\tau|N, \theta, n, Y_{obs}^{tot}, Y_{obs}^{src}, B_{obs}, L_{obs}, E_{obs}, A_{obs})$ is computed by integrating over $S_{obs}$, i.e.

$$
\begin{aligned}
&p(\tau|N, \theta, n, Y_{obs}^{tot}, Y_{obs}^{src}, B_{obs}, L_{obs}, E_{obs}, A_{obs}) \\
&\propto (1-\pi(\theta,\tau))^{N-n} \cdot \tau^{n\theta+a_m-1} \cdot e^{-b_m\tau} \cdot \int \Big( \prod_{i=1}^{n} S_i^{-(\theta+1)+Y_i^{src}} \cdot e^{-\frac{S_iE_i}{\gamma}} \\
&\cdot g(S_i, B_i, L_i, E_i) \mathbb{I}_{\{\tau<S_i\}} \Big) \mathrm{d}S
\end{aligned}
$$

The sampling is performed using the Metropolis- Hastings Algorithm. We take a logarithmic transformation of $\tau$ in order to preserve the positivity and to avoid numerical instability since $\tau$ is a very small number ($\sim 10^{-17}$). So, the marginal posterior distribution is:

$$\eta = \log(\tau) \tag{2.22}$$

$$p(\eta|N, \theta, n, Y_{obs}^{tot}, Y_{obs}^{src}, B_{obs}, L_{obs}, E_{obs}, A_{obs}) = \tag{2.23}$$

$$\propto (1 - \pi(\theta, e^{\eta}))^{N-n} \cdot e^{(n\theta + a_m - 1)\eta} \cdot e^{-b_m e^{\eta}}$$

$$\cdot \int \Big( \prod_{i=1}^{n} S_i^{-(\theta+1)+Y_i^{src}} \cdot e^{-\frac{S_i E_i}{\gamma}}$$

$$\cdot g(S_i, B_i, L_i, E_i) \mathbb{I}_{\{\eta < \log(S_i)\}} \Big) \mathrm{d}S \tag{2.24}$$

We choose a normal distribution as a proposal distribution. The integral with respect to flux is computed using numerical integration. More specifically, we create a grid of values for the flux $S$ over which we evaluate the expression. After numerical simulations, a grid of length 50,000 suffices for the estimation of that integral.

As described in van Dyk & Jiao (2015), extra care must be taken in the order in which the parameters are sampled when we use a Partially Collapsed Gibbs with Metropolis Hastings updates so as to sample from the proper stationary distribution. In our case, we should first sample from $p(Y_i^{src}|\cdot)$, then $p(\tau|\cdot)$ and afterwards $p(S_{obs}|tau, \cdot)$. The rest of the parameters can be sampled afterwards in any order. The only concern is while sampling $p(S_{obs,i}|\tau, \cdot)$, we must ensure that $S_{obs,i} > \tau$. So, in case $S_{obs,i} < \tau$, we set $S_{obs,i} = 1.5\tau$ and then sample 10 times the $S_{obs,i}$.

In order to compare the Partially Collapsed Gibbs algorithm with the Blocked Gibbs algorithm presented in the previous section, we consider simulated data from our model with $\theta = 0.4$, $N = 40$ and $\tau = 3.5 \times 10^{-17}$. The source-specific parameters $(B, L, E)$ are sampled as described previously using the exposure map and background map of the Chandra Deep Field South survey. The energy conversion factor is held constant at $\gamma = 2.679 \times 10^{-9}$. The posterior estimates of the 3 parameters can be found at Tables 2.1 and 2.2.

**Table 2.1:** The posterior estimates of $\theta, \tau, N$ of simulated dataset using the Blocked Gibbs sampler introduced in Section 2.1.1.

| | Mean | SD | 2.5% | 97.5% |
|---|---|---|---|---|
| $\theta$ | 0.335 | 0.072 | 0.206 | 0.483 |
| $N$ | 47.7 | 11.02 | 30 | 73 |
| $\tau$ | $2.69\times10^{-17}$ | $1.22\times10^{-17}$ | $8.47\times10^{-18}$ | $5.5\times10^{-17}$ |

**Table 2.2:** The posterior estimates of $\theta, \tau, N$ of simulated dataset using the Partially Collapsed Gibbs sampler.

| | Mean | SD | 2.5% | 97.5% |
|---|---|---|---|---|
| $\theta$ | 0.334 | 0.071 | 0.204 | 0.486 |
| $N$ | 51.45 | 12.5 | 31 | 80 |
| $\tau$ | $2.06\times10^{-17}$ | $1.08\times10^{-17}$ | $5.12\times10^{-18}$ | $4.61\times10^{-17}$ |

After a series of numerical simulations, we concluded that the Partially Collapsed Gibbs sampler offers faster convergence to the target posterior distribution than the Blocked Gibbs sampler. However, this comes at a heavy computational cost. Performing numerical integration twice during each Gibbs iteration adds a lot of extra computation time, which is not compensated by the faster convergence. Thus, we suggest using the Blocked Gibbs sampler introduced in Section 2.1.1 which is 2 to 3 times faster than the Partially Collapsed Gibbs sampler.

## 2.3 Model Validation

The two samplers described in the previous sections, the Blocked Gibbs sampler and the Partially Collapsed Gibbs sampler, are implemented in Python and we have taken advantage of parallel processing in order to speed up the computations. However, since the above model contains many different levels, validating the computer software becomes a necessity. We restrain the validation analysis only for the Blocked Gibbs sampler since we will not be using the Partially Collapsed Gibbs sampler due to its high computational cost. The validation is done by generating data according to the model and check the model-fitting software for consistent posterior estimates. We also compute the marginal posterior distribution and compare it with the histograms of the marginal posterior distributions of the parameters of interest, produced from the Gibbs draws, for consistency.

### 2.3.1 Validation using posterior interval plots

A significant advantage of bayesian models is that they we can perform validation and self-consistency checks using simulated data (Cook et al. 2012, Gelman et al. 2014). This is extremely important in models such as the one presented in this chapter due its complex hierarchy, structure and the many different levels.

The validation is based on generating simulated datasets from the model generating software given some fixed values of the parameters, and then fit the simulated dataset to the model-fitting software to obtain posterior draws of the parameters. More specifically, the validation algorithm can be described as:

**Validation Algorithm:**

- Step 1: Simulate data from the model $y_{\text{obs}} \sim p(y_{\text{obs}} \mid \theta, \tau, N)$, where the true values of $(\theta, \tau, N)$ are given.

- Step 2: Run the Blocked Gibbs sampler to obtain posterior draws for the parameters of interest $(\theta, \tau, N)^{(t)} \sim p(\theta, \tau, N \mid y_{\text{obs}})$, $t = 1, ..., T$. Compute posterior quantiles using the posterior draws, i.e. for a parameter $x$, compute the $\hat{x_q} = min\{x^t \ : \ \hat{Pr}(x < x_q) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}_{\{x^t < x_q\}} \geq q\}$, for $0 < q < 1$.

- Step 3: For each parameter, evaluate posterior credible sets $C$ of level $\alpha$:

$$\int_C p(x \mid y_{\text{obs}}) \mathrm{d}x = 1 - \alpha.$$

  The estimate based on the samples from the Blocked Gibbs sampler is $(\hat{x}_L, \hat{x}_U)$ such that $q(\hat{x}_L) = \alpha/2$ and $q(\hat{x}_U) = 1 - \alpha/2$. For each parameter, record if the posterior credible set $C$ of level $\alpha$ contains the "true" value of the parameter used for generating the simulated dataset.

The steps 1,2 and 3 are repeated 20 times. We expect for each parameter that the posterior credible set $C$ of level $\alpha$ contains the "true" value of the parameter most of the times, especially if the above steps were to be repeated for a very large number

of times.

**Single Pareto Model**: For the single Pareto model, we simulated 20 datasets using the parameters $\theta = 0.8$, $N = 800$ and $\tau = 2 \times 10^{-17}$. The flux to count rate conversion factor is set constant to $\gamma = 2.679 \times 10^{-9}$. More specifically, we draw $N$ samples from a Pareto distribution for the flux of the complete source population, i.e. $S_i^{\text{tot}} \sim \text{Pareto}(\theta, \tau)$, $i = 1, ..., N$ and then draw $B_i, L_i, E_i$ from the joint distribution $p(B, L, E)$ as described in the previous section using the background map and exposure map from the Chandra Deep Field South survey. Then, we apply the incompleteness function to extract the $S^{\text{obs}}$ by computing the function $g(S_i, B_i, L_i, E_i)$ and comparing it with $u_i \sim \text{Uniform}(0,1)$; we assume that the source $i$ is observed if $u_i < g(S_i, B_i, L_i, E_i)$.

Figures 2.4, 2.5 and 2.6 show the posterior 95% interval for each of the 20 simulated datasets for the parameters $\theta$, $N$ and $\tau$ respectively. We can see that the 95% posterior intervals in the 19 out of the 20 datasets contain the "true" values of the 3 parameters.

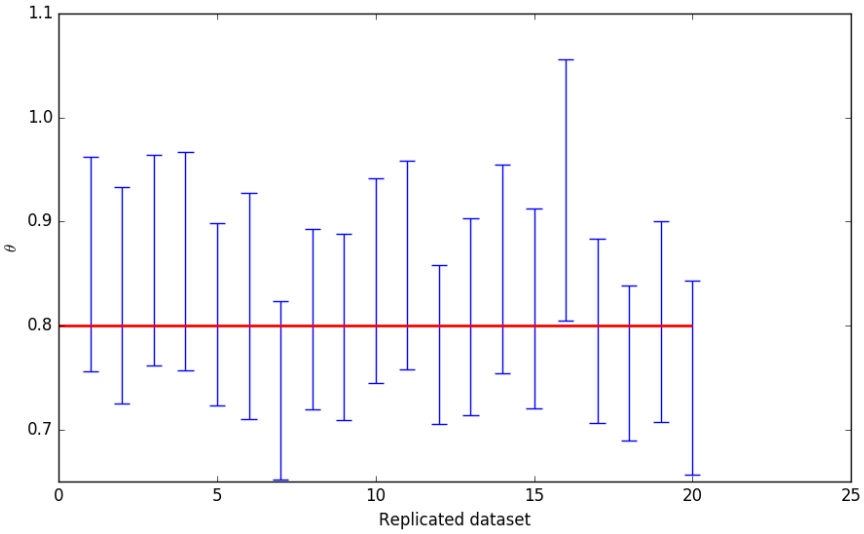



**Figure 2.4:** Posterior credible intervals of $\theta$ from 20 dataset simulations using validation process for the single Pareto model. The "true" value is $\theta = 0.8$.

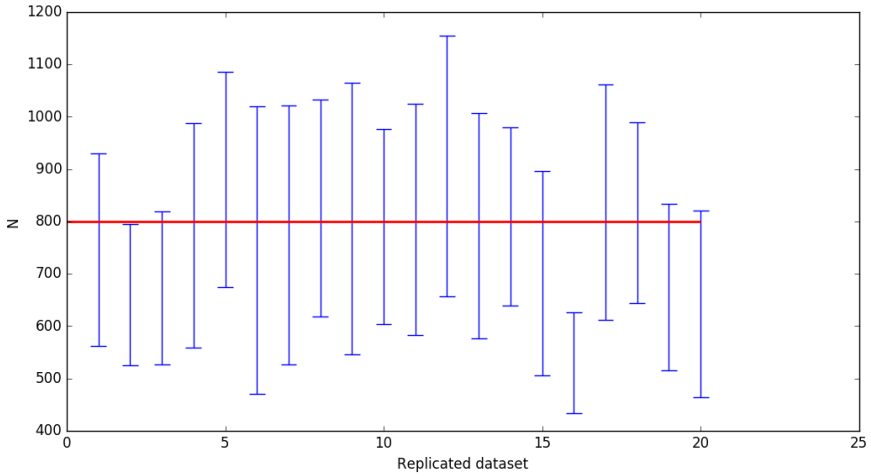



**Figure 2.5:** Posterior credible intervals of $N$ from 20 dataset simulations using validation process for the single Pareto model. The "true" value is $N = 800$.

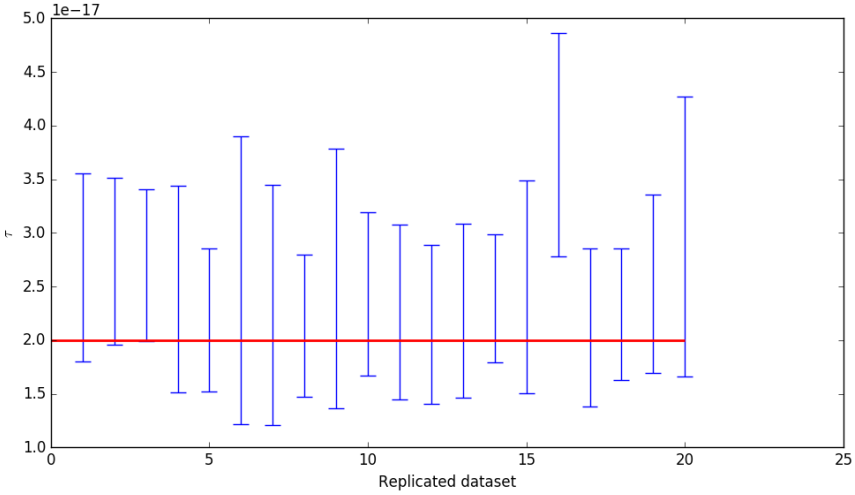



**Figure 2.6:** Posterior credible intervals of $\tau$ from 20 dataset simulations using validation process for the single Pareto model. The "true" value is $\tau = 2 \times 10^{-17}$. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

**Broken Pareto Model with 1 breakpoint**: For the broken Pareto model with 1 break, we simulated 20 datasets using parameters $\theta_1 = 0.8$, $\theta_2 = 1.2$, $N = 1000$, $\tau_1 = 1.5 \times 10^{-17}$ and $\tau = 2 \times 10^{-15}$. The flux to count rate conversion factor is set constant to $\gamma = 2.679 \times 10^{-9}$. More specifically, we draw $N$ samples from the Broken Power law distribution for the flux of the complete source population (using the inverse CDF method), and then draw $B_i, L_i, E_i$ from the joint distribution $p(B, L, E)$ as described in the previous section. Then, we applied the incompleteness function to extract the $S^{\text{obs}}$ by computing the function $g(S_i, B_i, L_i, E_i)$ and comparing it with $u_i \sim$ Uniform(0,1); we assume that the source $i$ is observed if $u_i < g(S_i, B_i, L_i, E_i)$.

Figures 2.7, 2.8, 2.9,2.10 and 2.11 show the posterior 95% interval for each of the 20 simulated datasets for the parameters $\theta_1$, $\theta_2$, $N$, $\tau_1$ and $\tau_2$ respectively.

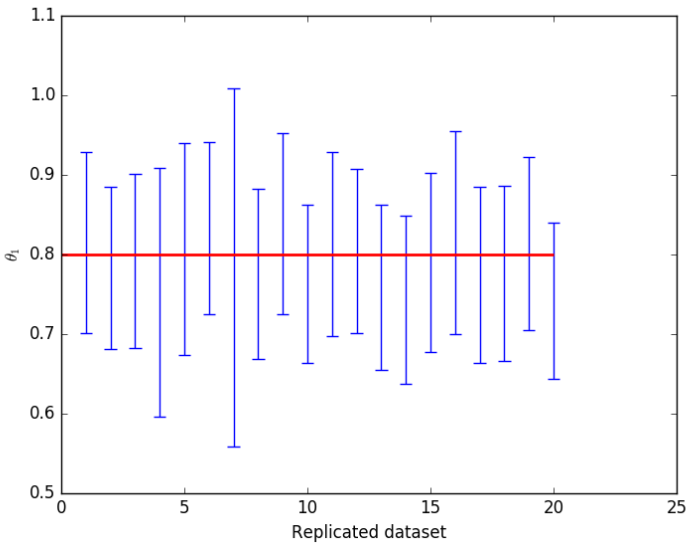



**Figure 2.7:** Posterior credible intervals of $\theta_1$ from 20 dataset simulations using validation process for the broken Pareto model with 1 break. The "true" value is $\theta_1 = 0.8$.

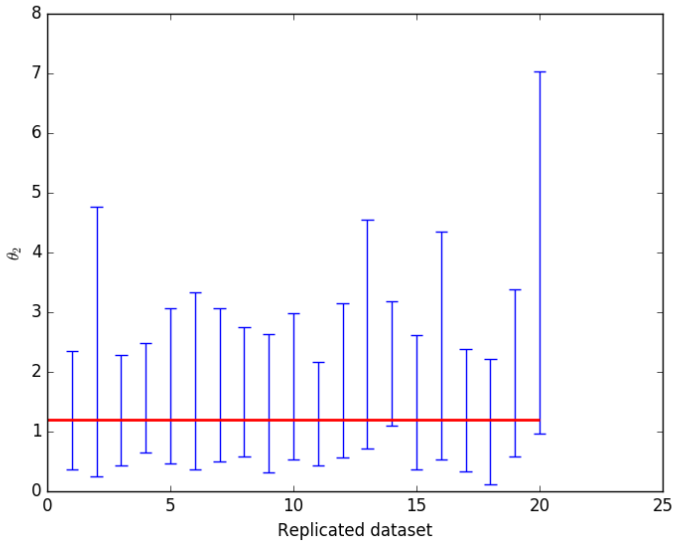



**Figure 2.8:** Posterior credible intervals of $\theta_2$ from 20 dataset simulations using validation process for the broken Pareto model with 1 break. The "true" value is $\theta_2 = 1.2$.

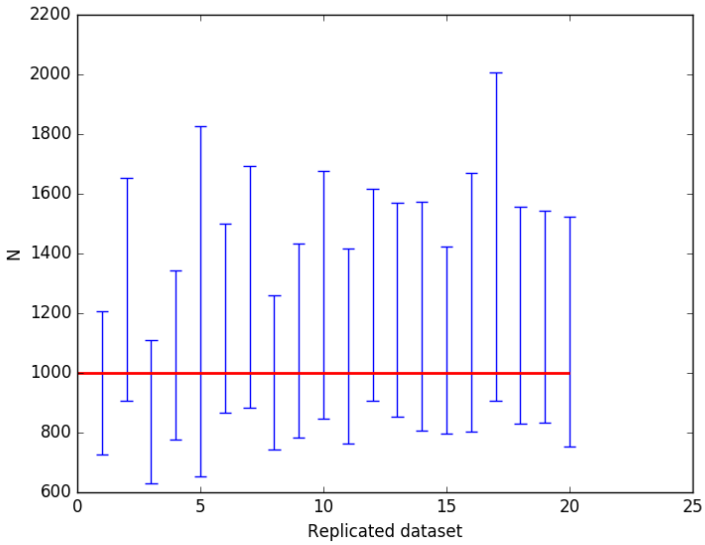



**Figure 2.9:** Posterior credible intervals of $N$ from 20 dataset simulations using validation process for the broken Pareto model with 1 break. The "true" value is $N = 1000$.

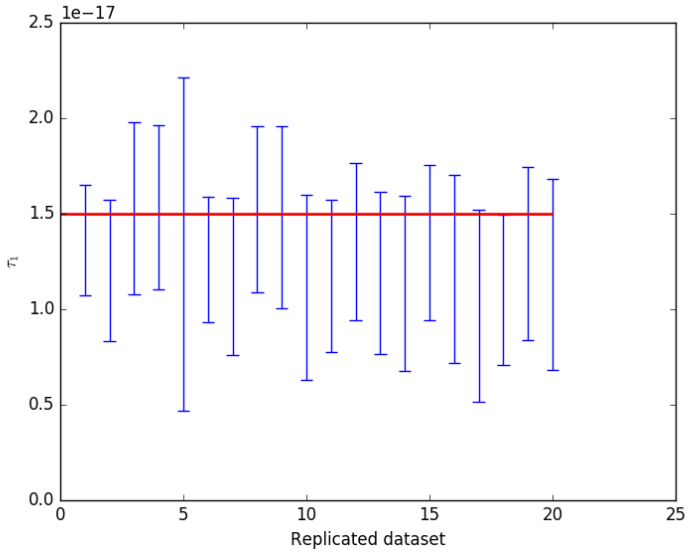



**Figure 2.10:** Posterior credible intervals of $\tau_1$ from 20 dataset simulations using validation process for the broken Pareto model with 1 break. The "true" value is $\tau_1 = 1.5 \times 10^{-17}$. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

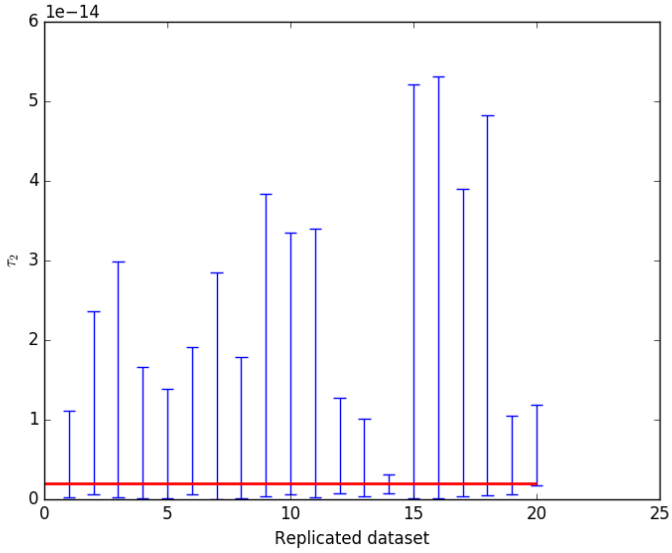



**Figure 2.11:** Posterior credible intervals of $\tau_2$ from 20 dataset simulations using validation process for the broken Pareto model with 1 break. The "true" value is $\tau_2 = 2 \times 10^{-15}$. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

Although the "true" values of the parameters are contained in the 95% posterior credible intervals for all the parameters for at least 19 out of the 20 simulated datasets, we can observe that the model exhibits biases on the estimation of $\tau_2$.

More specifically, we can observe that the posterior intervals of the breakpoint $\tau_2$ are rather wide. For high values of the flux $S$, the number of observed sources becomes small. Thus, the estimation of the breakpoint is difficult. The wide posterior interval indicates that the model doesn't not have enough data to converge to a specific breakpoint and explores many different regions for the breakpoint. Figure 2.12 depicts the 20 histograms of the marginal posterior distributions of $\tau_2$ for the simulated datasets. We can see that most of them exhibit a right fat tail and multimodality, indicating that they are exploring many different areas for $\tau_2$. If we are considering a point inference for posterior distributions that exhibit those characteristics, the posterior median or posterior mode might be a more appropriate choice than the posterior mean (Park et al. 2008 have an interesting discussion in summarising distributions with multiple modes). A more informative prior distribution would be very useful in that particular case as well as the existence of a bigger dataset.

The multimodality and the skewness are evident and in the marginal posterior distribution of $\tau_2$ that we extract from applying the same methodology to he Chandra Deep Field South survey in Section 2.5.

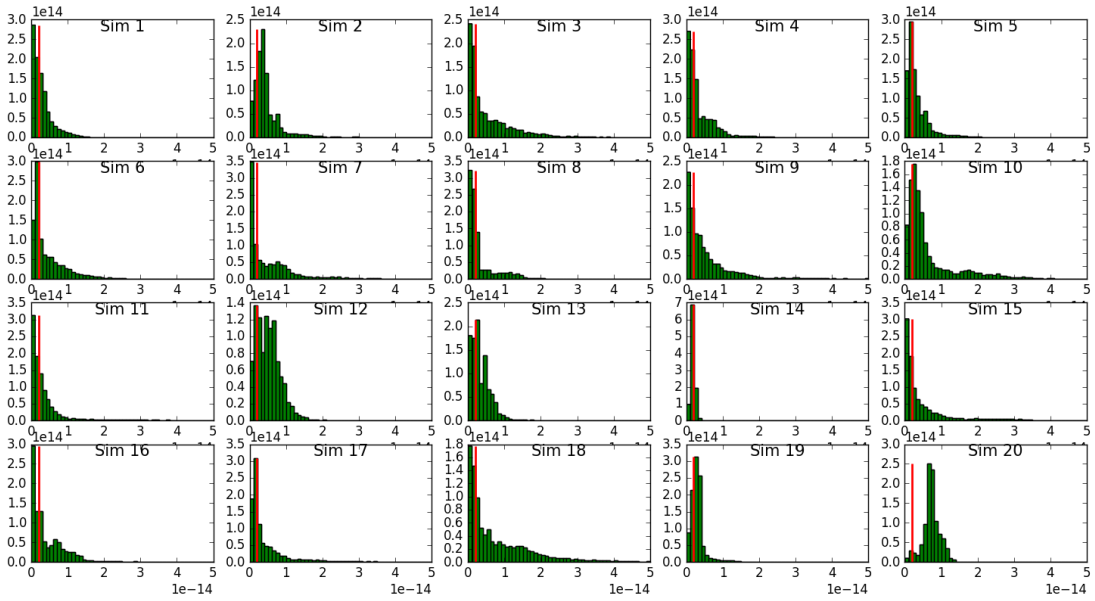



**Figure 2.12:** Posterior marginal histograms of $\tau_2$ from 20 dataset simulations using validation process for the broken Pareto model with 1 break. The "true" value is $\tau_2 = 2 \times 10^{-15}$ depicted with the vertical red line in each histogram. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

**Broken Pareto Model with 2 breakpoint**: For the broken Pareto model with 2 breaks, we simulated 20 datasets using parameters $\theta_1 = 0.5$, $\theta_2 = 0.7$, $\theta_3 = 1.3$, $N = 1000$, $\tau_1 = 1.5 \times 10^{-17}$, $\tau = 1.3 \times 10^{-15}$ and $\tau_3 = 6 \times 10^{-15}$. The flux to count rate conversion factor is set constant to $\gamma = 2.679 \times 10^{-9}$. More specifically, we draw $N$ samples from the Broken Power law distribution for the flux of the complete source population (using the inverse CDF method), and then draw $B_i, L_i, E_i$ from the joint distribution $p(B, L, E)$ as described in the previous section. Then, we applied the incompleteness function to extract the $S^{\text{obs}}$ by computing the function $g(S_i, B_i, L_i, E_i)$ and comparing it with $u_i \sim$ Uniform(0,1); we assume that the source $i$ is observed if $u_i < g(S_i, B_i, L_i, E_i)$.

Figures 2.13, 2.14, 2.15, 2.16,2.17, 2.18 and 2.19 show the posterior 95% interval for each of the 20 simulated datasets for the parameters $\theta_1$, $\theta_2$, $\theta_3$, $N$, $\tau_1$, $\tau_2$ and $\tau_3$ respectively. As in the case of the broken Pareto Model with 1 break, we can observe a bias in the estimation of $\tau_3$. More specifically, we can observe that the posterior intervals of the breakpoint $\tau_3$ are rather wide. We believe that the wide posterior interval indicates that the model doesn't not have enough data to converge to a specific breakpoint and explores many different regions for the breakpoint.

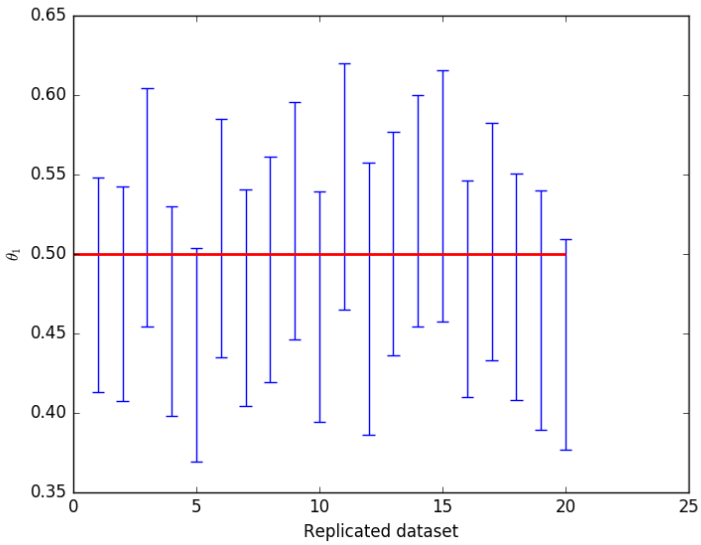



**Figure 2.13:** Posterior credible intervals of $\theta_1$ from 20 dataset simulations using validation process for the Broken Pareto model with 2 breaks. The "true" value is $\theta_1 = 0.5$.

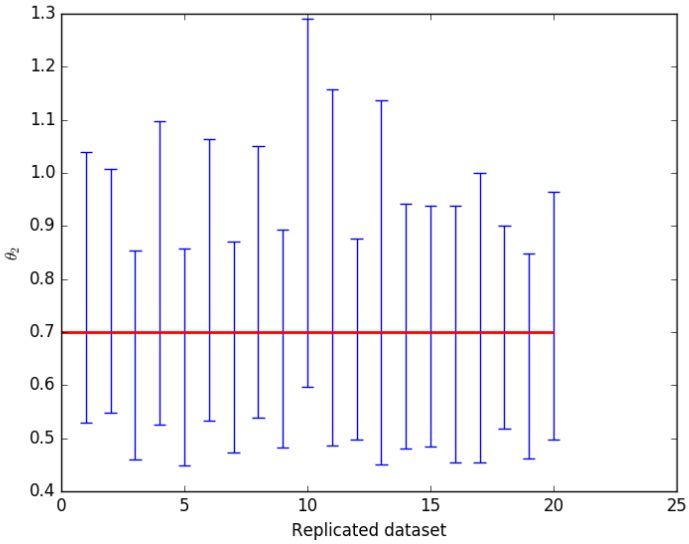



**Figure 2.14:** Posterior credible intervals of $\theta_2$ from 20 dataset simulations using validation process for the Broken Pareto model with 2 breaks. The "true" value is $\theta_2 = 0.7$.

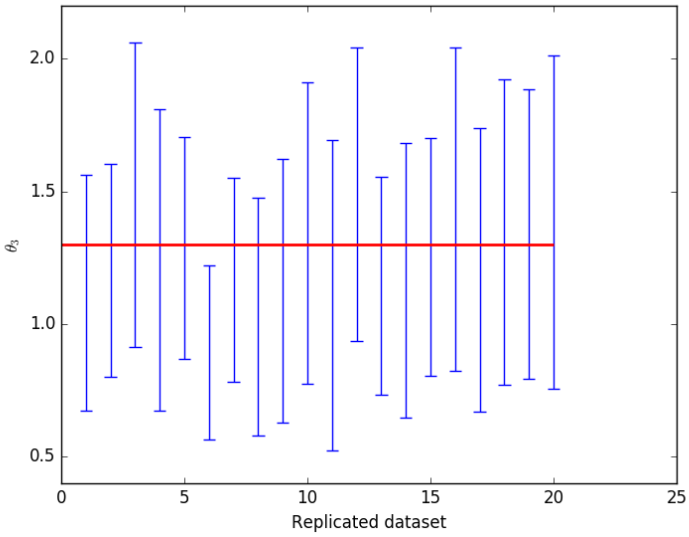



**Figure 2.15:** Posterior credible intervals of $\theta_3$ from 20 dataset simulations using validation process for the Broken Pareto model with 2 breaks. The "true" value is $\theta_3 = 1.3$.

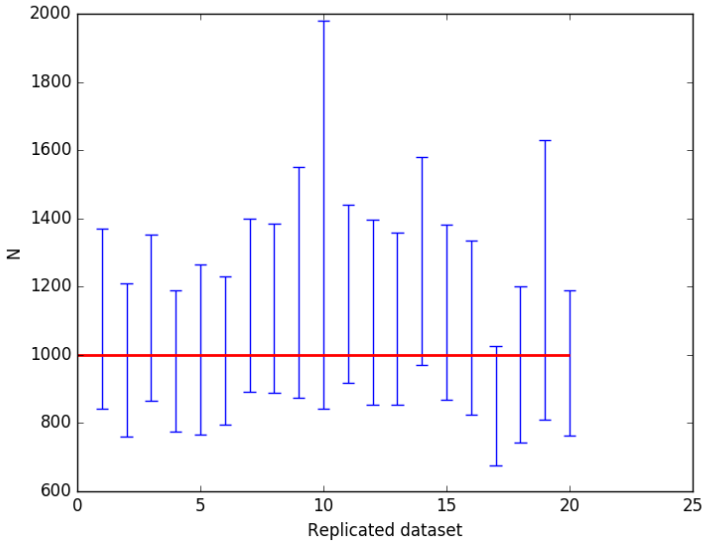



**Figure 2.16:** Posterior credible intervals of $N$ from 20 dataset simulations using validation process for the Broken Pareto model with 2 breaks. The "true" value is $N = 1000$.

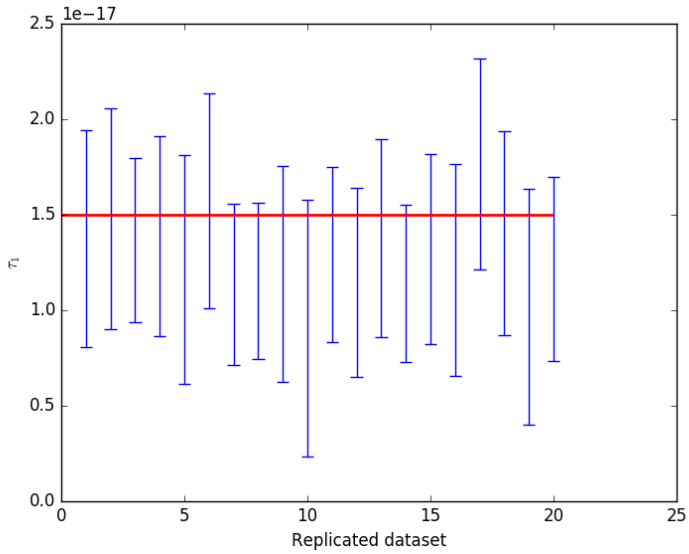



**Figure 2.17:** Posterior credible intervals of $\tau_1$ from 20 dataset simulations using validation process for the Broken Pareto model with 2 breaks. The "true" value is $\tau_1 = 1.5 \times 10^{-17}$. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

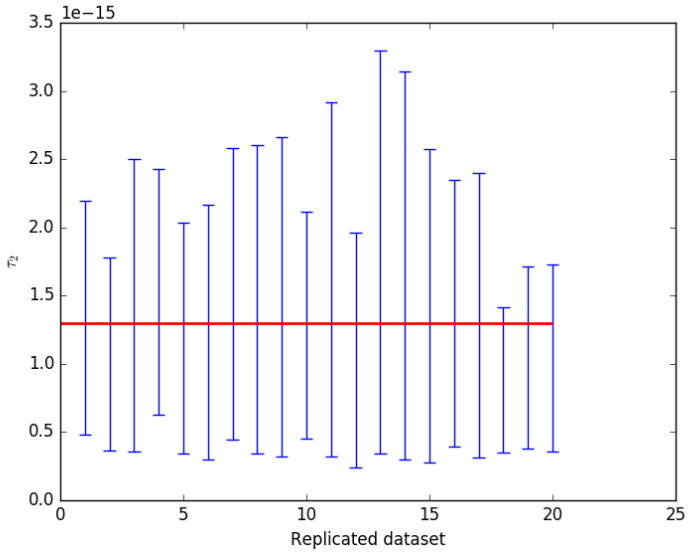



**Figure 2.18:** Posterior credible intervals of $\tau_2$ from 20 dataset simulations using validation process for the Broken Pareto model with 2 breaks. The "true" value is $\tau_2 = 1.3 \times 10^{-15}$. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

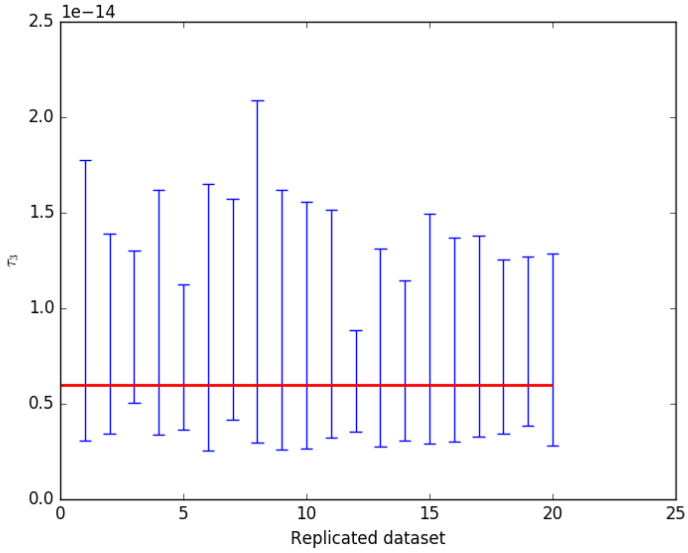



**Figure 2.19:** Posterior credible intervals of $\tau_3$ from 20 dataset simulations using validation process for the Broken Pareto model with 2 breaks. The "true" value is $\tau_3 = 6 \times 10^{-15}$. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

### 2.3.2 Evaluate the Marginal Posterior Distribution

Another way to check whether the Blocked Gibbs sampler draws samples from the correct posterior distribution is to numerically evaluate the marginal posterior distribution $p(\theta, \tau, N|\cdot)$ on a grid and then integrate in order to compute the marginal posteriors $p(\theta|\cdot)$, $p(\tau|\cdot)$, $p(N|\cdot)$ and compare them with the marginal posteriors we get from the Gibbs algorithm. More specifically, the marginal posterior distribution of $N, \theta, \tau$ (for the single Pareto model) is:

$$
\begin{aligned}
p(N, \theta, \tau | n, Y_{obs}^{tot}, \cdot) \propto & \binom{N}{n} (1 - \pi(\theta, \tau))^{N-n} \\
& \cdot \binom{N + a_N - 1}{a_N - 1} \Big( \frac{1}{1 + b_N} \Big)^N \Big( \frac{b_N}{1 + b_N} \Big)^{a_N} \\
& \cdot \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} \cdot \frac{b_m^{a_m}}{\Gamma(a_m)} \tau^{a_m - 1} e^{-b_m \tau} \\
& \cdot \prod \int \theta \tau^\theta S_i^{-(\theta+1)} \cdot g(S_i, B_i, L_i, E_i) \cdot \frac{(\lambda_i + \kappa_i)^{Y_i^{tot}}}{Y_i^{tot}!} e^{(\lambda_i + \kappa_i)} \\
& \cdot \binom{Y_i^{tot}}{Y_i^{src}} \Big( \frac{\lambda_i}{\lambda_i + \kappa_i} \Big)^{Y_i^{src}} \Big( 1 - \frac{\lambda_i}{\lambda_i + \kappa_i} \Big)^{Y_i^{tot} - Y_i^{src}} \mathrm{d}S \, \mathrm{d}Y_i^{src}
\end{aligned}
$$

Inside the integral, we have

$$
\begin{aligned}
I = & \int \theta \tau^\theta S_i^{-(\theta+1)} \cdot g(S_i, ) \cdot \frac{(\lambda_i + \kappa_i)^{Y_i^{tot}}}{Y_i^{tot}!} e^{(\lambda_i + \kappa_i)} \\
& \cdot \binom{Y_i^{tot}}{Y_i^{src}} \Big( \frac{\lambda_i}{\lambda_i + \kappa_i} \Big)^{Y_i^{src}} \Big( 1 - \frac{\lambda_i}{\lambda_i + \kappa_i} \Big)^{Y_i^{tot} - Y_i^{src}} \mathrm{d}S \mathrm{d}Y_i^{src} \\
= & \int \theta \tau^\theta S_i^{-(\theta+1)} \cdot g(S_i, ) \cdot \frac{(\lambda_i + \kappa_i)^{Y_i^{tot}}}{Y_i^{tot}!} e^{(\lambda_i + \kappa_i)} \\
& \Big[ \sum_{Y_i^{src}=0}^{Y_i^{tot}} \binom{Y_i^{tot}}{Y_i^{src}} \Big( \frac{\lambda_i}{\lambda_i + \kappa_i} \Big)^{Y_i^{src}} \Big( 1 - \frac{\lambda_i}{\lambda_i + \kappa_i} \Big)^{Y_i^{tot} - Y_i^{src}} \Big] \mathrm{d}S \\
= & \int \theta \tau^\theta S_i^{-(\theta+1)} \cdot g(S_i, ) \cdot \frac{(\lambda_i + \kappa_i)^{Y_i^{tot}}}{Y_i^{tot}!} e^{(\lambda_i + \kappa_i)} \mathrm{d}S
\end{aligned}
$$

This integral is estimated using Monte Carlo integration, i.e. drawing samples from a Pareto$(S; \theta, \tau)$ distribution and evaluating the function $g(S_i, ) \cdot \frac{(\lambda_i + \kappa_i)^{Y_i^{tot}}}{Y_i^{tot}!} e^{(\lambda_i + \kappa_i)}$

at those samples of $S$. After numerical experiments, in order to achieve sufficient accuracy we err on the side of using $5,000,000$ samples. In other words, each evaluation is very demanding on computational time, thus using this method in a big dataset with many sources, i.e. large number $N$, will require weeks of computing time, even with massive parallelisation.

As a validation example, consider data simulated from the model with $\theta = 0.9$, $\tau = 0.7 \times 10^{-17}$ and $N = 80$. After applying the incompleteness function on the data we are left with $n = 9$ observed sources. The following figure depicts the histograms we get from the Blocked Gibbs sampler after $50,000$ iterations and by removing the necessary burn-in samples for the 3 parameters of interest $\theta$, $\tau$ and $N$. The red lines correspond to the marginal posterior distributions we evaluated numerically on a grid as described above. It becomes apparent that the Blocked Gibbs algorithm samples from the correct stationary distribution.

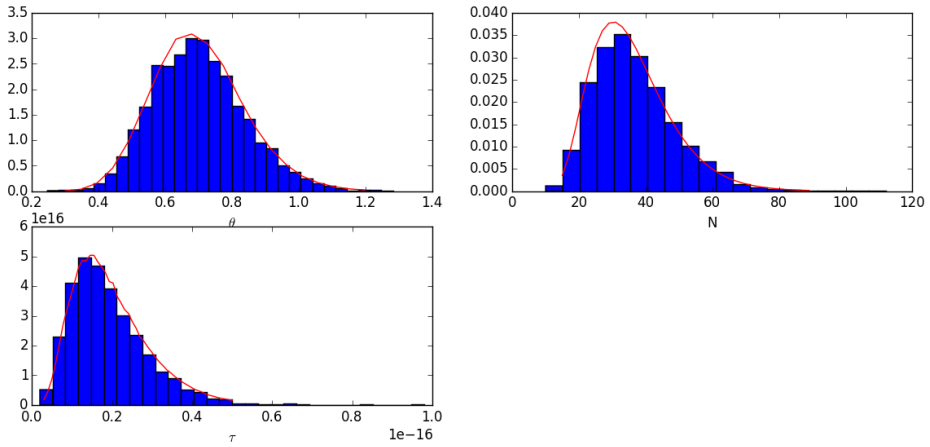



**Figure 2.20:** Posterior histograms of the three parameters $\theta$, $\tau$ and $N$ using the samples from the Gibbs sampler. The red lines correspond to the marginal posterior distributions of $\theta$, $\tau$ and $N$ evaluated numerically as described above. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

## 2.4 Posterior Inference and Model Selection

### 2.4.1 Parameter Inference

The posterior inference for the model parameters is based on the posterior MCMC draws. Given the posterior samples, we can summarise and visualise the marginal

posterior distributions of the parameters of interest. More specifically, the posterior samples are used to compute a number of different statistics for the model parameters such as the posterior mean, median and mode in order to represent the posterior point estimates. The posterior credible intervals are estimated so as to represent the uncertainty. A visual inspection of the distributions allows us to check for multimodality as well.

We estimate those posterior values after running multiple MCMC chains for a large number of iterations and using different starting values. We disregard the first half part of the chains as burn-in. The convergence of the chains is examined by looking at the trace plots. When in doubt regarding a convergence of a scalar estimate, we opt to use the $\hat{R}$ statistic, defined in Chapter 11 of Gelman et al. (2014), to monitor the convergence.

The Bayesian model which we developed allow us to look at the flux distribution in two different ways. The first characterisation of the $\log(N) - \log(S)$ relationship is through the marginal posterior distribution of the slope $\theta$ for the single power law model. Moreover, we can explore the posterior $\log(N) - \log(S)$ curve, that we plot using the different posterior samples of the observed and missing sources (given the observed sources, we can easily sample the missing sources). Examining this posterior $\log(N) - \log(S)$ curve will provide intuition about the uncertainty of the posterior estimates and the linearity of the curve. The minimum threshold $\tau$ and the total number of sources $N$ can provide us with additional information; the total number of sources $N$ is big with respect to the number of observed sources when the detection probability is low. This indicates that the posterior estimates are mostly influenced by the model assumptions and not the observed data.

Similarly, for the broken power law model, are main focus is on the broken power law slopes $\theta_1, ..., \theta_m$ and the flux breakpoints $\tau_2, ..., \tau_m$, which indicate the fluxes at which we observe a significant change in the underlying flux distribution. We also construct a posterior $\log(N) - \log(S)$ plot using the MCMC draws and sampling the fluxes for the missing sources, $S_{\text{mis}}$, for each iteration.

### 2.4.2 Sensitivity to the Choice of Priors

The choice of priors for the various parameters of interest directly affects the posterior inference. In the context of Bayesian statistics, if there is a strong prior belief about a model parameter, it should be expressed through the prior distribution. However, a strongly informative but incorrect prior can have negative impact in the quality of the posterior estimate, especially if the amount of data is not sufficient to overcome the misplaced certainty of the prior distribution.

A thorough investigation of the sensitivity of the Single Pareto model to the choice of priors is carried in Udaltsova (2014). The authors examine the impact to the posterior inference of the model parameters $\theta$, $N$ and $\tau$ of a weak prior, a moderately informative prior which is consistent with the true value and a strongly informative but incorrect prior. The conclusion of that study is that a strong but incorrect prior can adversely affect the posterior inference, thus the use of weak priors is recommended. Our numerical simulations for both the single Pareto model and the broken Power Law model verified this conclusion and thus we opt for using weakly informative priors provided to us by our collaborators from the astronomical society. However, a slight misspecification on the priors doesn't seem to affect the posterior estimates of the parameters.

### 2.4.3 Model Selection

In the previous sections we developed a series of 3 different models for estimating the flux distribution; a model with no-breaks, a model with 1-break and a model with 2-breaks. The question that arises revolves around selecting the appropriate model for a given set of data. This is not a trivial choice and the relevant literature is quite extended on the topic (see Draper 1995, O'Hagan 1995, and their discussions for an overview), although there is not a single best solution. In most cases, we expect a larger model, in our case the models with breaks, to better fit the data. However, we should examine whether the improvement in the fit is statistically significant and not just a result of over-fitting. Some of the more commonly used methods include the Bayes factors and the Deviance Information Criterion (DIC).

Bayes factors is a method for model selection in which two candidates models are

compared using the ratio of the marginal likelihood under one model to the marginal likelihood of the other model. If we define as $H_1$ and $H_2$ the two competing models, then:

$$\text{Bayes factor}(H_2; H_1) = \frac{p(y|H_2)}{p(y|H_1)}$$

Kass & Raftery (1995) provide a comprehensive overview of the method and an interpretation of the estimated Bayes factor.

The DIC (Spiegelhalter et al. 2002, see) measures the discrepancy between the data and the model. It is defined as:

$$DIC = -2 \log p(y|\theta_{\text{Bayes}}) + 2p_{DIC},$$

where $\theta_{\text{Bayes}}$ is the posterior estimate and the second term of the equation defines the bias correction term that estimates the effective number of parameters.

Udaltsova (2014) explores the performance of Bayes factor and the DIC for automatic model selection for choosing the most appropriate model, in our case the number of breaks. Their simulation results indicate that neither of those methods can lead to a reliable model selection procedure, mainly because of the complicated hierarchical structure. They also propose a new model selection method, called Bayesian Adaptive Fence Method, which besides being very computationally expensive, it stills doesn't show consistent performance in the simulation studies. We believe that further research on the topic is appropriate, which may call for the development of more problem specific methods for model selection (we discuss our proposed approach at the Discussion section of this Chapter as well as at the last Chapter of this Thesis).

## 2.5 Application: CHANDRA Deep Field South

The methodology developed in the previous sections is applied to the CHANDRA Deep Field South (CDFS) 2MS survey, one of the most sensitive X-ray surveys ever done. The CDFS is an image taken by the Chandra X-ray Observatory satellite. This survey is a deep 0.5-7.0 keV survey covering 0.11 square degrees comprised of 11 days of CHANDRA ACIS-I exposure. In our analysis we consider a sample of 358 observed sources, from which we exclude 3 sources for which we do not have spectral data (in the next section we will incorporate the uncertainty about the flux-to-count conversion factor $\gamma$ in our model which requires having the spectrum of each source; thus to facilitate the comparison of the results, we also exclude those 3 sources from the current analysis).

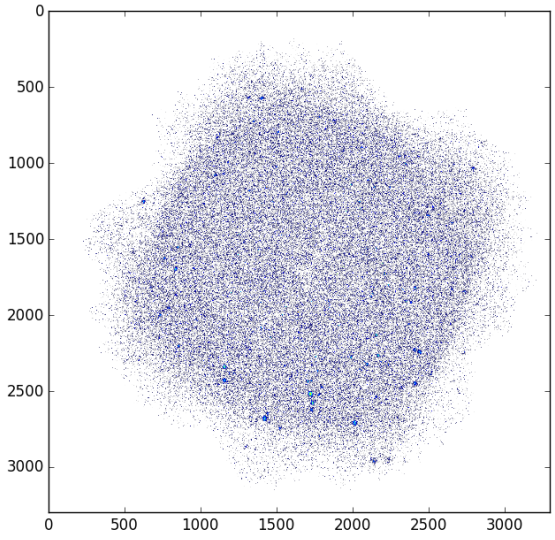



**Figure 2.21:** CHANDRA "true" colour image of the CDFS.

### 2.5.1 Single Power Law model

For the single Power Law model we assume the following priors for the parameters $(N, \theta, \tau)$:

$$\theta \sim \text{Gamma}(a = 2, b = 1)$$
$$\tau \sim \text{Gamma}(a_m = 1.38, b_m = 3.46 \times 10^{-16})$$
$$N \sim \text{Negative-Binomial}(a_N = 8.05, b_n = 0.014)$$

We ran the Blocked Gibbs sampler, as described in the previous sections, using the aforementioned priors for 60,000 iterations. As it is always a good practise with MCMC samplers, we ran 3 independent chains with different starting values, which all of them gave us very similar posterior estimates. More specifically, the first chain provided us with the posterior estimates for the parameters of interest $(N, \theta, \tau)$ that are depicted in the Table 2.3 using the last 30,000 iterations (we discarded the first 30,000 iterations as burn-in).

**Table 2.3:** Posterior estimates of major parameters for the CDFS dataset for the single Pareto model using the last 30,000 iterations for 1 of the 3 chains we ran.

| | Mean | Median | SD | 2.5% | 97.5% | Mode |
|---|---|---|---|---|---|---|
| $\theta$ | 0.963 | 0.976 | 0.03 | 0.894 | 1.00 | 0.980 |
| $N$ | 2826.3 | 2803 | 223.1 | 2472 | 3368 | 2778 |
| $\tau$ | $9.67 \times 10^{-18}$ | $1.00 \times 10^{-17}$ | $9.14 \times 10^{-19}$ | $7.72 \times 10^{-18}$ | $1.06 \times 10^{-17}$ | $1.01 \times 10^{-17}$ |

Figure 2.22 shows the trace plots for the parameters of interest $(N, \theta, \tau)$ for the first chain we ran. The convergence is quite fast as we can deduce. Figure 2.23 depicts the posterior bivariate scatter plots and 1-dimensional histograms for the parameters of interest. From both the bivariate scatter plots and the histograms of the posteriors draws we can observe that the marginal posterior distribution of the slope $\theta$ and the marginal posterior distribution of $\tau$ exhibit signs of bi-modality. This could be interpreted as an indication that the $\log(N) - \log(S)$ curve is not linear but rather piece-wise linear.

The posterior draws of the flux for the complete source population gives rise to the posterior distribution plot of the $\log(N) - \log(S)$ curve shown in Figure 2.24. Each curve in the plot corresponds to a set of fluxes for the complete source population sampled from a single iteration of Blocked Gibbs sampler scheme with observed sources shown in grey and missing sources in red. Current plot exhibits sample of 100 flux sets. The blue line is the estimated $\log(N) - \log(S)$ curve using the

posterior modes of $\theta, N, \tau$. The depicted curve does not appear to be linear. This indicates that a broken Power Law model might be a better fit, especially if we take into consideration the bi-modality in the marginal posterior distributions.

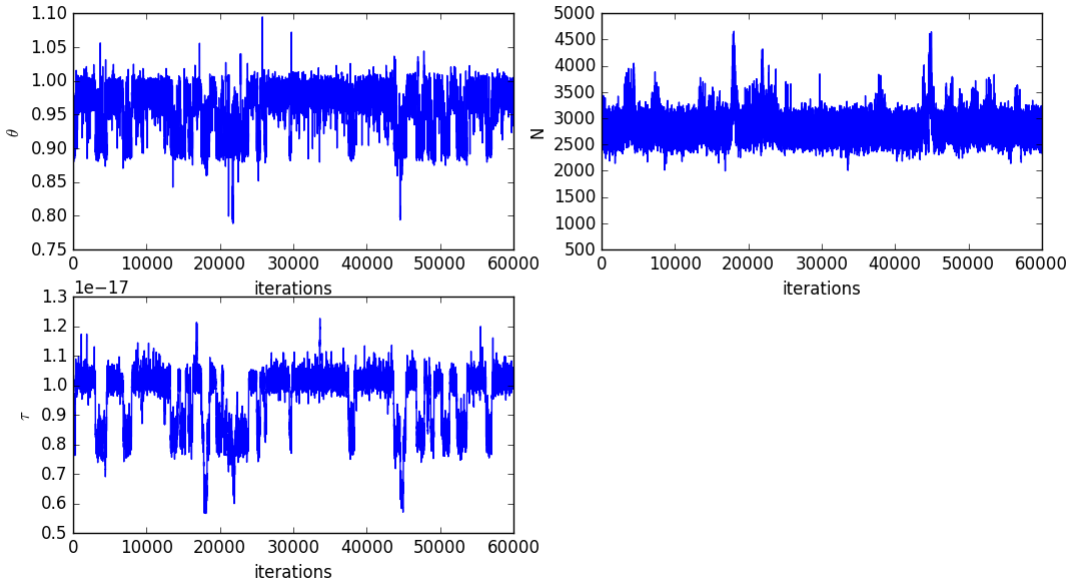



**Figure 2.22:** Trace plots of the main parameters of interest $\theta, N, \tau$ of the CDFS dataset for the Single Pareto model.

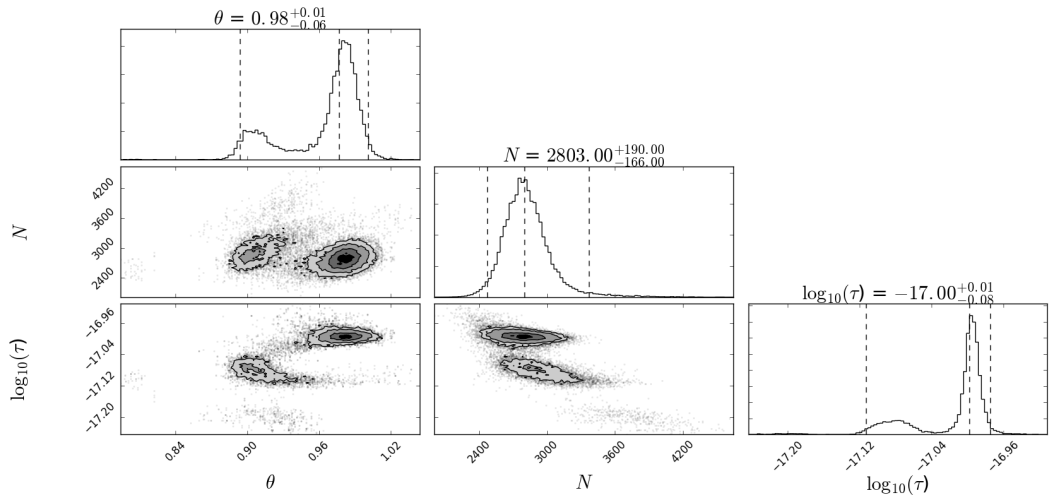



**Figure 2.23:** Bivariate scatter plots and 1-dimensional histograms of the main parameters of interest $\theta, N, \tau$ of the CDFS dataset for the Single Pareto model. The figures are plotted using the posterior draws from the Blocked Gibbs sampler after removing a burn -in sample of about 30,000 draws.

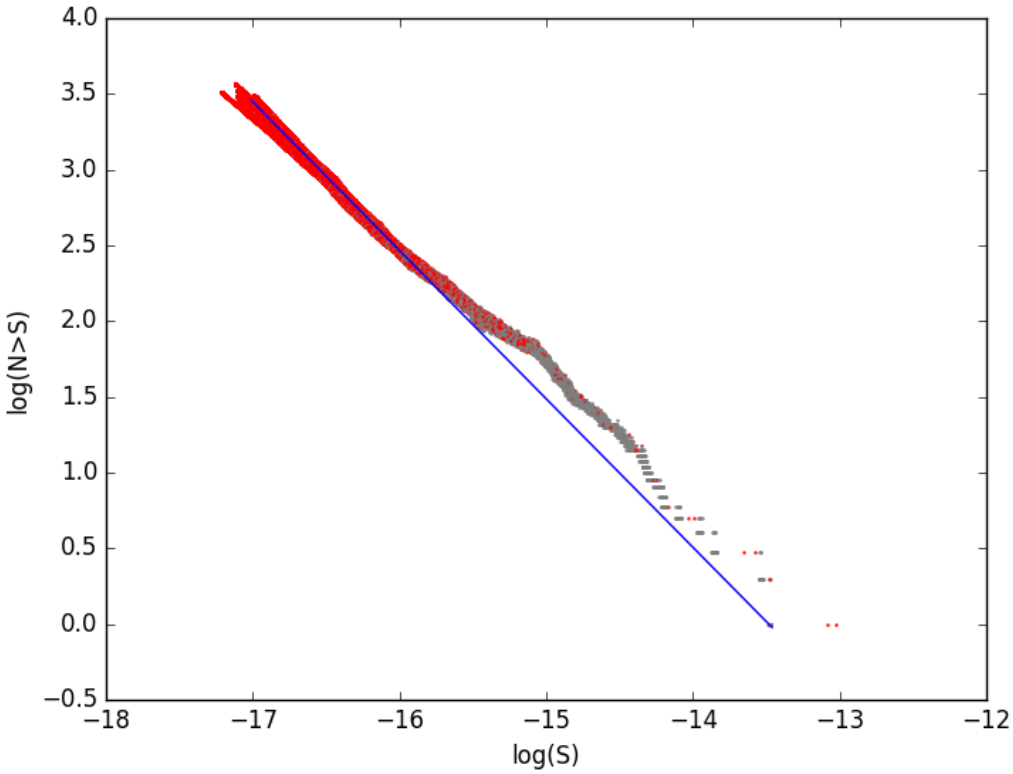



**Figure 2.24:** The posterior distribution of the $\log(N) - \log(S)$ plot of the CDFS dataset for the Single Pareto model. Each line in the plot corresponds to a set of fluxes for the complete source population sampled from a single iteration of the Blocked Gibbs sampler scheme with observed sources shown in grey and missing sources in red. Current plot exhibits sample of 100 flux sets. The depicted curve does not appear to be linear. This indicates that a Broken Power Law model might be a better fit.

### 2.5.2 Broken Power Law model with 1 break

For the Broken Power Law model with 1 break, the following priors were assumed for the parameters $(N, \theta_1, \theta_2, \tau_1, \tau_2)$:

$$\begin{aligned} \theta_1 &\sim \text{Gamma}(a = 2, b = 1) \\ \theta_2 &\sim \text{Gamma}(a = 2, b = 1) \\ \tau_1 &\sim \text{Gamma}(a_m = 1.38, b_m = 3.46 \times 10^{-16}) \\ \eta_2 = \log(\tau_2 - \tau_1) &\sim N(\mu = -35, \sigma^2 = 1) \\ N &\sim \text{Negative-Binomial}(a_N = 8.05, b_n = 0.014) \end{aligned}$$

We ran the Blocked Gibbs sampler, as described in the previous sections, using the aforementioned priors for 60,000 iterations. As it is always a good practise with MCMC samplers, we ran 3 independent chains with different starting values. More specifically, the first chain provided us with the posterior estimates for the parameters of interest $(N, \theta_1, \theta_2, \tau_1, \tau_2)$ that are depicted in the Table 2.4 using the last 30,000 iterations (we discarded the first 30,000 iterations as burn-in).

**Table 2.4:** The posterior estimates for the major parameters for the CDFS dataset using the last 30,000 iterations for the Broken Power law model with 1 break for 1 of the 3 chains we ran.

| | Mean | Median | SD | 2.5% | 97.5% | Mode |
|---|---|---|---|---|---|---|
| $\theta_1$ | 0.759 | 0.759 | 0.05 | 0.661 | 0.860 | 0.780 |
| $\theta_2$ | 1.52 | 1.27 | 0.77 | 0.677 | 3.86 | 1.13 |
| $N$ | 2139 | 2096 | 336 | 1608 | 2929 | 1879 |
| $\tau_1$ | $0.91 \times 10^{-17}$ | $0.91 \times 10^{-17}$ | $1.76 \times 10^{-18}$ | $5.73 \times 10^{-18}$ | $1.23 \times 10^{-17}$ | $0.96 \times 10^{-17}$ |
| $\tau_2$ | $5.07 \times 10^{-15}$ | $2.19 \times 10^{-15}$ | $7.01 \times 10^{-15}$ | $2.59 \times 10^{-16}$ | $2.74 \times 10^{-14}$ | $1.05 \times 10^{-15}$ |

Figure 2.25 shows the trace plots for the parameters of interest $(N, \theta_1, \theta_2, \tau_1, \tau_2)$. Figure 2.26 depicts the posterior bivariate scatter plots and 1-dimensional histograms for the parameters of interest. From both the bivariate scatter plots and the histograms of the posteriors draws we can observe that the marginal posterior distribution of the slope $\theta_1$ and the marginal posterior distribution of $\tau_1$ are unimodal distributions in contrast to the no-break case. However, the marginal posterior distributions of both $\tau_2$ and $\theta_2$ have fat tails and exhibit signs of multimodality. This behaviour is similar to the one we observed during the validation study, where we stated that this behaviour might be a result of a lack of enough sources with high values of flux. Thus, the MCMC chain explores many different areas for $\tau_2$ and subsequently for the slope $\theta_2$. If we are considering a point estimate for the parameters of interest, we suggest using the posterior mode, since it might be a more appropriate choice than the posterior mean.

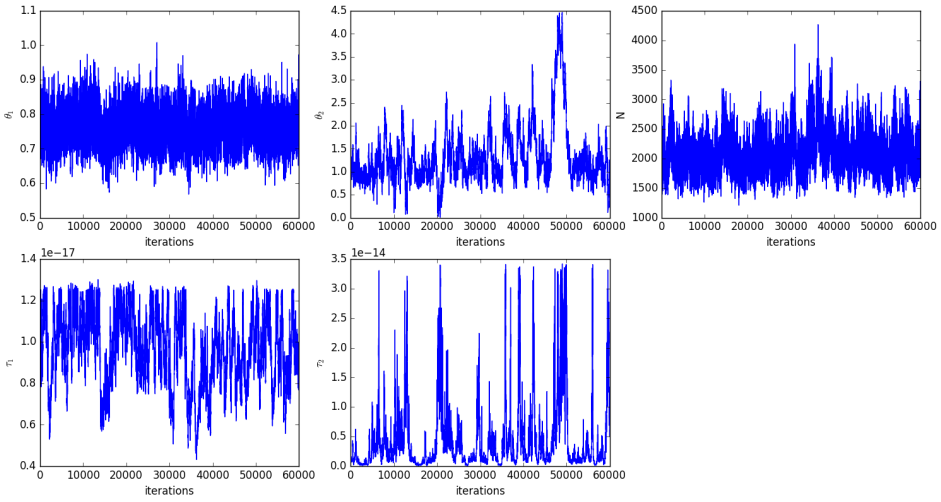



**Figure 2.25:** Trace plots of the main parameters of interest $\theta_1, \theta_2, N, \tau_1, \tau_2$ of the CDFS dataset for the broken Power law model with 1-break. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

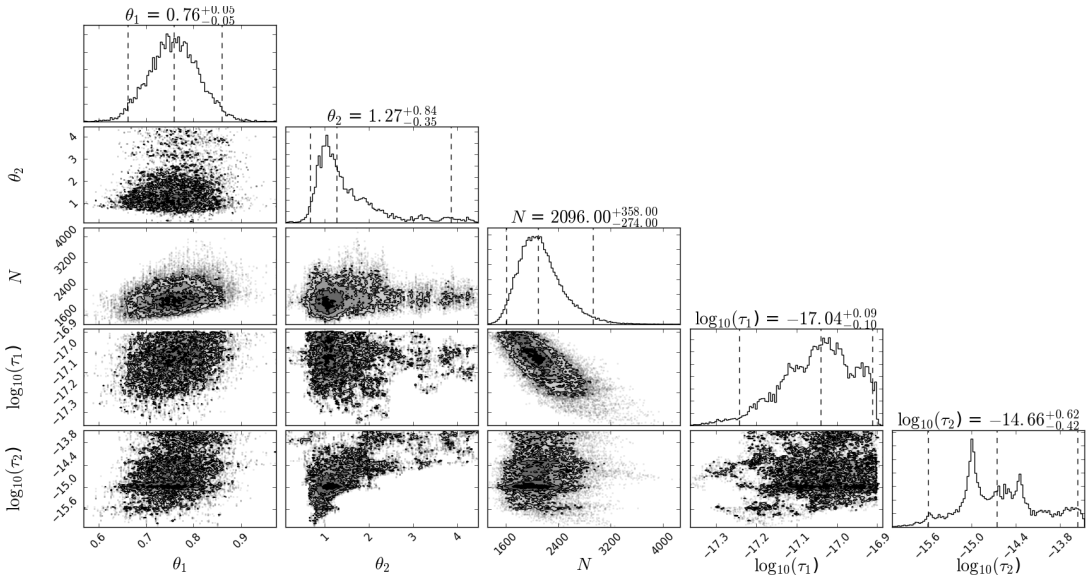



**Figure 2.26:** Bivariate scatter plots and 1-dimensional histograms of the main parameters of interest $\theta_1, \theta_2, N, \tau_1, \tau_2$ of the CDFS dataset for the broken Power law model with 1-break. The figures are plotted using the posterior draws from the Blocked Gibbs sampler after removing a burn -in sample of about 30,000 draws.

In order to examine whether this behaviour shows up as a result of lack of convergence in the MCMC chain, we compute the $\hat{R}$ statistic suggested in Gelman et al. (2014) for monitoring the convergence of scalar estimands. The value of $\hat{R}$ is 1.002 for $\tau_2$, thus we have evidence that the chains have converged. We also look at the

individual trace plots of all the 3 chains for $\tau_2$ depicted in Figure 2.27 which indicate similar behaviour. Moreover, in Figure 2.28 we plot the histograms of the marginal posterior distribution of $\tau_2$ for the 3 chains individually, and for all of the chains combined. We can observe that the histograms look very similar for all chains individually and for all of them combined.

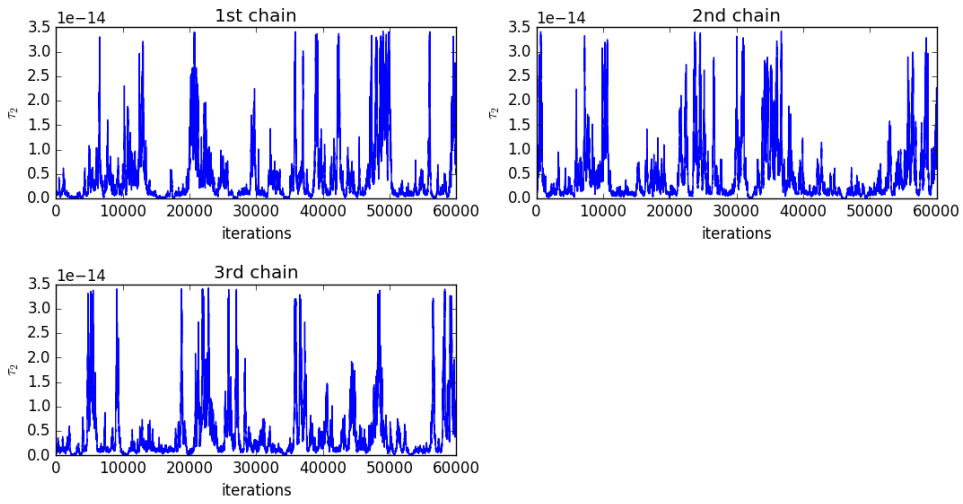



**Figure 2.27:** Trace plots of parameter $\tau_2$ of the CDFS dataset for the Broken Power law model with 1-break for all of the 3 parallel chains. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

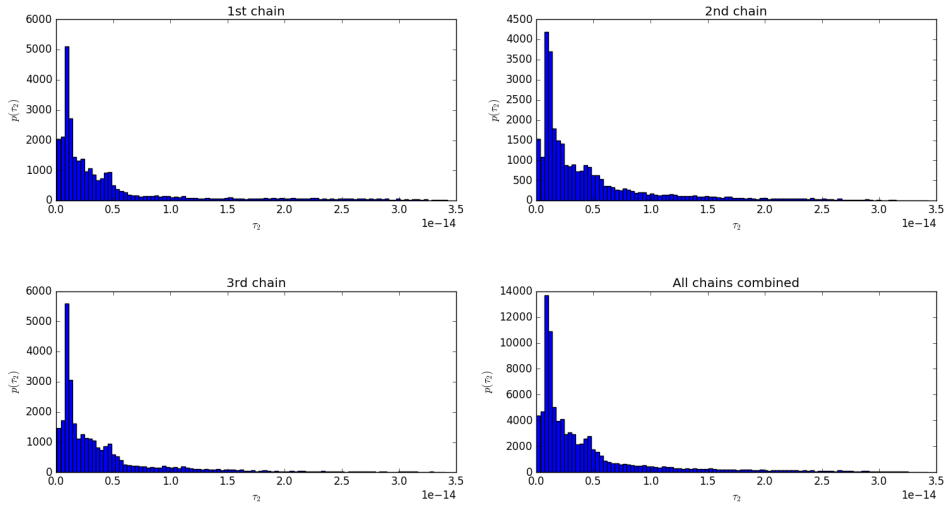



**Figure 2.28:** Histograms of the marginal posterior distribution of parameter $\tau_2$ of the CDFS dataset for the Broken Power law model with 1-break for all of the 3 parallel chains. We can observe that the histograms are very similar. The bottom right histogram is the histogram of all the posterior samples of $\tau_2$ from the 3 chains combined. It has a similar shape to the each individual histogram.

The posterior draws of the flux for the complete source population gives rise to the posterior distribution plot of the $\log(N) - \log(S)$ curve shown in Figure 2.29. Each curve in the plot corresponds to a set of fluxes for the complete source population sampled from a single iteration of Blocked Gibbs sampler scheme with observed sources shown in grey and missing sources in red. Current plot exhibits sample of 100 flux sets. The blue line is the estimated $\log(N) - \log(S)$ using the posterior modes of $\theta_2, \tau_1, \tau_2$ and the posterior medians of $\theta_1$ and $N$. The reason we choose the posterior medians of $\theta_1$ and $N$ as point estimates lies on the shape of their marginal posterior distributions. More specifically, both of those marginal posterior distributions are flat around the mode, so the actual estimation of the mode is very difficult and numerically unstable. The resulting posterior $\log(N) - \log(S)$ curve in Figure 2.29 does not appear to be linear. Thus the broken power law model with 1-break seems like a better candidate than the no break model.

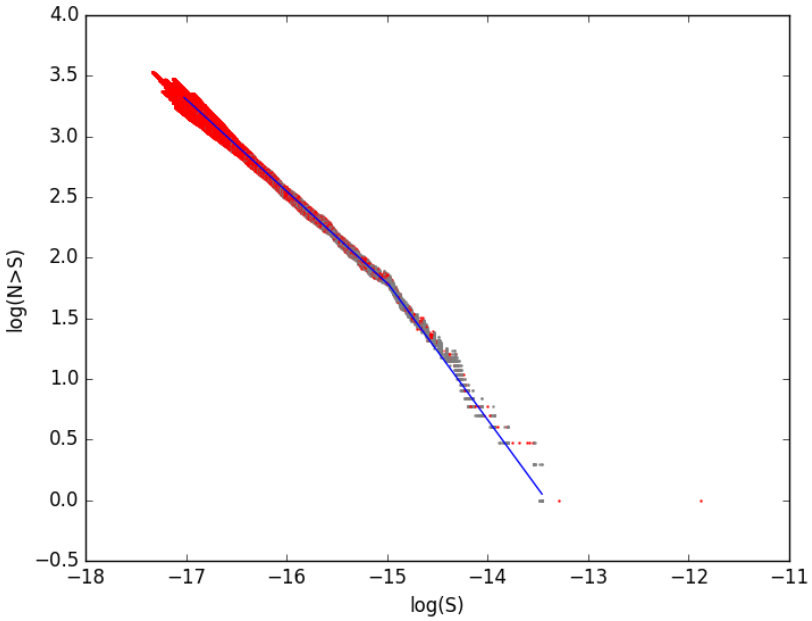



**Figure 2.29:** The posterior distribution of the $\log(N) - \log(S)$ plot of the CDFS dataset for the broken power law model with 1-break. Each line in the plot corresponds to a set of fluxes for the complete source population sampled from a single iteration of the Blocked Gibbs sampler scheme with observed sources shown in grey and missing sources in red. Current plot exhibits sample of 100 flux sets. The depicted curve does not appear to be linear, so a broken power law model seems like a better fit.

### 2.5.3 Broken Power Law model with 2 breaks

For the Broken Power Law model with 2 breaks, the following priors were assumed for the parameters $(N, \theta_1, \theta_2, \theta_3, \tau_1, \tau_2, \tau_3)$:

$$\begin{aligned}
\theta_1 &\sim \Gamma(a = 10, b = 10) \\
\theta_2 &\sim \Gamma(a = 10, b = 10) \\
\theta_2 3 &\sim \Gamma(a = 10, b = 10) \\
\tau_1 &\sim \Gamma(a_m = 1.38, b_m = 3.46 \times 10^{-16}) \\
\eta_2 = \log(\tau_2 - \tau_1) &\sim N(\mu = -35, \sigma^2 = 4) \\
\eta_3 = \log(\tau_3 - \tau_2) &\sim N(\mu = -33, \sigma^2 = 4) \\
N &\sim \text{Negative-Binomial}(a_N = 8.05, b_n = 0.014)
\end{aligned}$$

We ran the Blocked Gibbs sampler, as described in the previous sections, using the

aforementioned priors for 60,000 iterations. As for the no-break and the 1-break model, we ran 3 independent chains with different starting values. More specifically, the first chain provided us with the posterior estimates for the parameters of interest $(N, \theta_1, \theta_2, \theta_3, \tau_1, \tau_2, \tau_3)$ that are depicted in the Table 2.5 using the last 30,000 iterations (we discarded the first 30,000 iterations as burn-in).

**Table 2.5:** The posterior estimates for the major parameters for the CDFS dataset using the last 30,000 iterations for the Broken Power law model with 2-breaks for 1 of the 3 chains we ran.

| | Mean | Median | SD | 2.5% | 97.5% | Mode |
|---|---|---|---|---|---|---|
| $\theta_1$ | 0.749 | 0.749 | 0.06 | 0.644 | 0.867 | 0.750 |
| $\theta_2$ | 1.05 | 1.04 | 0.24 | 0.581 | 1.51 | 1.02 |
| $\theta_3$ | 1.16 | 1.12 | 0.40 | 0.457 | 2.14 | 1.15 |
| $N$ | 1932 | 1886 | 292 | 1480 | 2622 | 1797 |
| $\tau_1$ | $1.01 \times 10^{-17}$ | $1.04 \times 10^{-17}$ | $1.77 \times 10^{-18}$ | $6.35 \times 10^{-18}$ | $1.24 \times 10^{-17}$ | $1.21 \times 10^{-17}$ |
| $\tau_2$ | $1.06 \times 10^{-15}$ | $9.89 \times 10^{-16}$ | $5.52 \times 10^{-16}$ | $2.83 \times 10^{-16}$ | $2.56 \times 10^{-15}$ | $1.07 \times 10^{-15}$ |
| $\tau_3$ | $8.31 \times 10^{-15}$ | $7.01 \times 10^{-15}$ | $4.93 \times 10^{-15}$ | $2.53 \times 10^{-15}$ | $2.17 \times 10^{-14}$ | $5.18 \times 10^{-15}$ |

Figure 2.30 shows the trace plots for the parameters of interest $(N, \theta_1, \theta_2, \theta_3, \tau_1, \tau_2, \tau_3)$. Figure 2.31 depicts the posterior bivariate scatter plots and 1-dimensional histograms for the parameters of interest. From both the bivariate scatter plots and the histograms of the posteriors draws we can observe that the marginal posterior distribution of the slope $\theta$ and the marginal posterior distribution of $\tau$ exhibit signs of bi-modality. This could be interpreted as an indication that the $\log(N) - \log(S)$ curve is not linear but rather piece-wise linear.

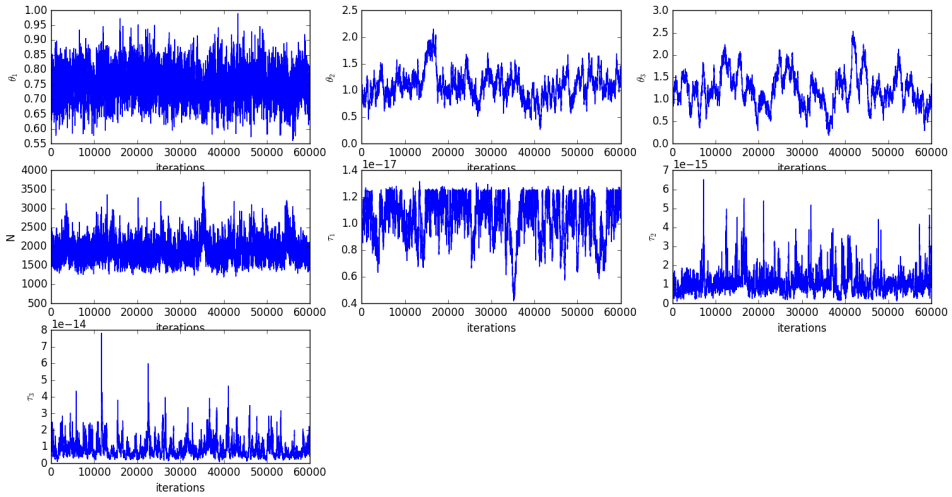



**Figure 2.30:** Trace plots of the main parameters of interest $(N, \theta_1, \theta_2, \theta_3, \tau_1, \tau_2, \tau_3)$ of the CDFS dataset for the Broken Power law model with 2-breaks. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

The posterior draws of the flux for the complete source population gives rise to the posterior distribution plot of the $\log(N) - \log(S)$ curve shown in Figure 2.32. Each curve in the plot corresponds to a set of fluxes for the complete source population sampled from a single iteration of Blocked Gibbs sampler scheme with observed sources shown in grey and missing sources in red. Current plot exhibits sample of 100 flux sets. The blue line is the estimated $\log(N) - \log(S)$ using the posterior modes of $\theta_1, \theta_2, \theta_3, \tau_1, \tau_2$ and the posterior medians of $\tau_3$ and $N$. As in the case of the model with 1-break, we err on the side of using the posterior medians as point estimates since both of the marginal posterior distributions of $\tau_3$ and $N$ are flat around the mode, so the actual estimation of the mode is very difficult and numerically unstable.The depicted curve does not appear to be linear. This indicated that a Broken Power Law model might be a better fit.

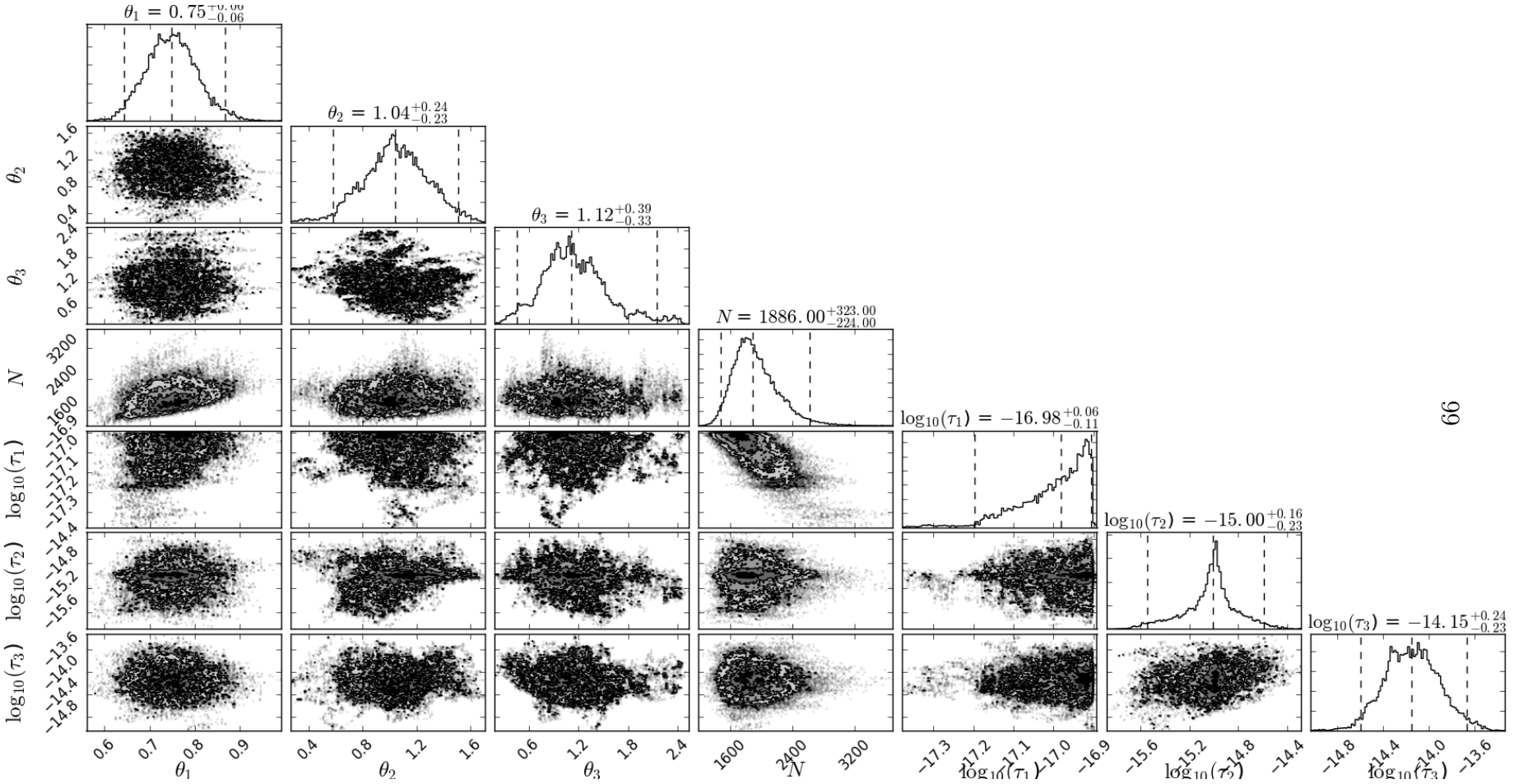



**Figure 2.31:** Bivariate scatter plots and 1-dimensional histograms of the main parameters of interest $(N, \theta_1, \theta_2, \theta_3, \tau_1, \tau_2, \tau_3)$ of the CDFS dataset for the broken Power law model with 2-breaks. The figures are plotted using the posterior draws from the Blocked Gibbs sampler after removing a burn -in sample of 30,000 draws.

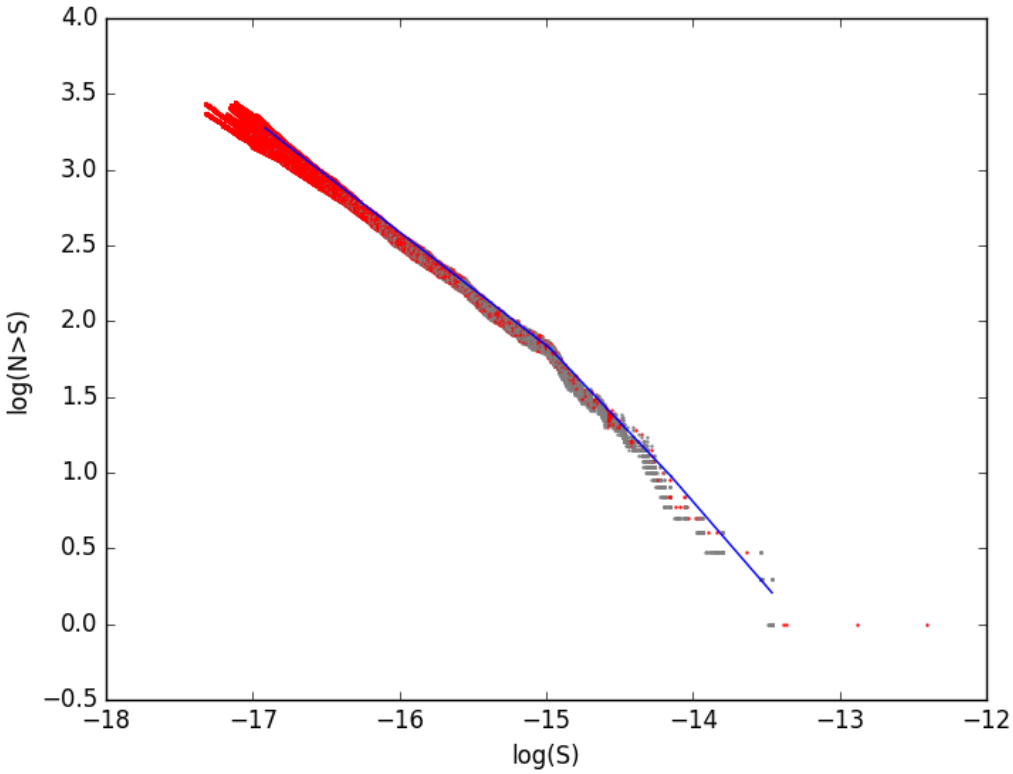



**Figure 2.32:** The posterior distribution of the $\log(N) - \log(S)$ plot of the CDFS dataset for the broken power law model with 2-breaks. Each line in the plot corresponds to a set of fluxes for the complete source population sampled from a single iteration of the Blocked Gibbs sampler scheme with observed sources shown in grey and missing sources in red. Current plot exhibits sample of 100 flux sets. The depicted curve does not appear to be linear. Thus a broken power law model seems like a better candidate than the no break model.

## 2.6 Discussion and Further Research Direction

In the previous sections, we described a hierarchical Bayesian model for estimating the $\log(N) - \log(S)$ relationship for both the linear and the piece-wise linear case based on Udaltsova (2014). This model properly accounts for the detector biases and uncertainties, and the missing data mechanism, while the relevant methods in the literature can only handle ideal, complete surveys. More specifically, it models the joint distribution of the complete data and model parameters, and then derives the posterior distribution of the model parameters marginalised across all missing

data information.

Our approach extends the work of Udaltsova (2014) in 2 ways. It employs a survey specific incompleteness function, which plays a crucial role in the posterior inference, and thus it should reflect the exact characteristics of the astronomical survey. Secondly, we utilise the survey specific background and exposure maps in order to create a joint distribution for the background contamination, the effective area and the off-axis angle. This creates a far more realistic distribution which can be used when sampling for the missing sources. Our approach provides a generalised framework for estimating the $\log(N) - \log(S)$ relationship for various different surveys. We applied our model in the CDFS dataset survey, and we present the estimated posterior $\log(N) - \log(S)$ relationship.

The flexible hierarchical Bayesian framework allows us to further extend the model at will in order to account and correct our estimates for different sources of uncertainty. During our discussion about the flux to count rate conversion, we introduced a factor $\gamma$. This parameter depends on the spectral model that is assumed and the energy band of the source. In the relevant literature about $\log(N) - \log(S)$ , this factor was considered to be constant for all the sources. In the next Chapter we extend our model to incorporate this source of uncertainty.

The main limitation of the model is the lack of a proper and effective automatic model selection process. Udaltsova (2014) conducts simulation studies to test the performance of various model selection methods, such as the Bayes factor and the DIC. However, those methods fail to show consistency in choosing the model used for simulating the data. We are suggesting two different research directions for model selection. The first approach would be focused on developing model specific heuristics and techniques for model selection. For example, we can define statistics, such as the similarity between the marginal posterior distributions of the 2 consecutive slopes, $\theta_1$ and $\theta_2$. If the differences in the two distributions are small under some metric (such as Hellinger distance), then we can assume that the there is not enough evidence about the existence of a breakpoint. However, extensive simulations would be required in order to properly test the power of such diagnostics.

A more statistically interesting approach to model selection would be the use of the

Reversible Jump MCMC (Green 1995) on the hierarchical Bayesian model. The Reversible Jump MCMC sampler is a framework for MCMC simulation that is well suited for problems in which the dimensions of the parameter space can vary between the iterations of the chain. The power of the Reversible Jump MCMC lies in the fact that, given a countable collection of candidate models, it produces a full probabilistic description of the posterior probabilities of each model. It can be employed in models like the one presented in this chapter, by applying it to the joint posterior distribution of all 3 models, namely the single Pareto, the broken power law with 1-break and the broken power law with 2-breaks.

Implementing successfully a Reversible Jump MCMC is not a trivial pursuit. The main difficulty lies in the construction of the proposal moves between different models, which is achieved via the concept of dimension matching. More specifically, in case we want to propose a move from a state $(\theta_k)$ in model $M_k$ with dimensions $n$, to a state $(\theta_k^*)$ in model $M_k^*$ with higher dimensions $m$, $m > n$, we have to generate a random vector $u$ with length $m - n$ to match the dimensions. This vector is generated from a known density $q_{d_{k->k^*}}$, and the current state along with the vector $u$ are mapped to the new state through a one-to-one mapping function $g_{k->k^*}$. Defining effective mapping functions can be challenging though, even for nested models. Thus, applying the Reversible Jump MCMC for model selection in our context is undoubtedly a promising approach, but not a straightforward one.

# 3

# Bayesian Analysis of the $\log(N) - \log(S)$ Problem with $\gamma$ Uncertainty

## 3.1 Introduction

In the previous Chapter, we developed a comprehensive method for estimating the $\log(N) - \log(S)$ relationship. This Bayesian hierarchical model properly accounts for the incompleteness of the astronomical surveys which is manifested through the missing data and the detector biases. However, due to the nature of astronomical surveys, there are other form of uncertainties associated with the measurements that should be accounted for. In this Chapter we focus our attention on the flux-to-count conversion factor $\gamma$.

As we saw in the previous chapter, we do not directly observe the flux in the astronomical surveys. Instead the data collection consists of photon counts for the observed source population. The photon counts that we measure on the detector for each source $i$ include both the counts from the actual astronomical source, $Y_i^{src}$, and the background contamination, $Y_i^{bkg}$, i.e. $Y_i^{tot} = Y_i^{src} + Y_i^{bkg}$. The photon counts from the actual astronomical source are connected to the flux $S$ through the relationship:

$$Y_i^{src}|S_i, E_i, \gamma_i \overset{\text{ind}}{\sim} Poisson\,(\,\frac{S_i \cdot E_i}{\gamma_i}\,)$$

The flux-to-count rate conversion factor $\gamma$ for a specific source, depends on the spectral model that is assumed for the source as well as the energy band of the source. More specifically, when we fit a model to the spectrum of an astronomical source, we get point estimates for the parameters of the fit, along with the uncertainty in that estimation. The estimation of the flux-to-count conversion factor $\gamma$ is based on the estimates of the parameters of the model we fit to the spectrum. Thus, since we have uncertainty in the estimation of the parameters of the model, there is a subsequent uncertainty about the value of $\gamma$.

In the vast majority of the relevant literature that tries to estimate the $\log(N) - \log(S)$ relationship, the factor $\gamma$ is assumed to be constant for all the sources. To the best of our knowledge, the only work that explores the idea of incorporating the uncertainty of the flux-to-count conversion parameter is in Zezas et al. (2007); they discuss how the uncertainty of $\gamma$ for each source can be included in a log-likelihood estimation of a $\log(N) - \log(S)$ curve by assuming that the distribution of $\gamma$ can be approximated by a multivariate normal distribution with variance taken from the covariance matrix of the spectral fit. However, they do not actually implement this methodology.

In this chapter we develop a methodology for incorporating the uncertainty of the flux-to-count rate conversion factor $\gamma$ to the Bayesian hierarchical model we developed for estimating the $\log(N) - \log(S)$ relationship. In Section 3.2 we discuss how we can construct a prior distribution for $\gamma$ for the sources of a specific astronomic survey, both for the observed and the unobserved sources. In Section 3.3 we extend the Bayesian hierarchical model by including the uncertainty of $\gamma$ and extract the joint posterior distribution of all the parameters of interest. The sampling algorithm for drawing samples from this posterior distribution is thoroughly analysed. Section 3.4 focuses on the validation of the model. In Section 3.5, we apply our methodology to the Chandra Deep Field South survey and we compare the results with those from the model that assumes constant $\gamma$ for all the sources. Finally, we summarise our findings in Section 3.6.

## 3.2 Extracting the Prior $p(\gamma)$

Extracting the uncertainty about the flux-to-count conversion factor $\gamma$ for each source we observe in an astronomical survey, expressed as a probability distribution, is the first step in modifying our Bayesian hierarchical model for estimating the $\log(N) - \log(S)$ in order to account for the uncertainty in $\gamma$. This is straightforward to accomplish using modern astronomical software.

Given the photometric data from astronomical surveys such as the Chandra Deep Field South (CDFS), we can fit a model to the spectrum of each source using the Sherpa software package (Refsdal et al. 2009, Freeman et al. 2001), and more specifically the software library `pyBLoCXS` *. The pyBLoCXS library is a python extension of Sherpa and runs a Markov chain Monte Carlo (MCMC) based algorithm designed to carry out Bayesian Low-Count X-ray Spectral (BLoCXS) analysis in the Sherpa environment (see van Dyk et al. 2001, for an analysis of the MCMC techniques applied by pyBLoCXS). The pyBLoCXS code produces conditional posterior distributions of the parameters of a predefined spectral model fit (from those available in Sherpa) to high-energy X-ray spectral data.

By choosing a spectral model and using pyBLoCXS, we can extract a posterior distribution of the $\gamma$ for each astronomical source by means of MCMC draws. This procedure though provides us with the distributions of $\gamma_i$'s for the observed sources, $\gamma_{obs}$, of an astronomical survey. Nevertheless, if we assume that the individual characteristics of the spectrum of each source that affect the distribution of $\gamma$ are independent of the missing data mechanism, then we can assume that the distributions of $\gamma_{mis}$ for the missing sources would not differ from that of the observed sources. The question that arises here is how we can fit a hierarchical prior on $\gamma = (\gamma_1, ..., \gamma_N)$ for the complete source population (hierarchical because it is specified in terms of parameters that are themselves fit to the data). Our approach on this issue is motivated by another research project on astrostatistics and the methodology we developed in that particular case.

*http://cxc.harvard.edu/sherpa4.4/threads/pyblocxs/

### 3.2.1 Influenced from Testing the Redshift Dependence in Large-Scale X-ray/Radio Emission

Astronomical jets are beams of ionised matter accelerated to high speeds close to the speed of light. While the exact mechanism that creates the jets is not fully understood, jets in active galaxies are created by supermassive black holes (SMBH). Detecting jets by X-rays is a rather challenging endeavour due to the small number of X-ray photon counts observed from jets relative to their corresponding quasar cores.

McKeough et al. (2016) work focuses on analysing 11 X-ray images from the Chandra telescope (where 25 radio jet features are present in those 11 images) and trying to detect the existence of X-ray jets. Testing for X-ray jets is based on applying a multi-scale Bayesian method known as Low Count Image Reconstruction and Analysis (LIRA, see Connors & van Dyk 2007). This method is based on testing the hypothesis that a baseline model with a flat background is insufficient to explain the observed data. The exact application of this methodology as well as the explanation of how to construct efficiently a $p$-value for the hypothesis testing is nicely presented in Stein et al. (2015).

As it was mentioned above, the generating mechanism of jets is still under debate. McKeough et al. (2016) try to test the hypothesis that there is dependence between the redshift and the X-ray to radio luminosity ratio $\rho_{x,r}$, and thus this ratio can be used as a potential diagnostic of the emission mechanism. In order to test this hypothesis, we look at the posterior distribution from LIRA of the energy flux ratio for each detected jet, and we split the distributions into two samples consisting of the jets with high redshift ($z > 3$), and with low redshift ($z < 3$). Subsequently, 2 Gaussian hierarchical priors are fitted to the the two different samples so as to examine whether the two samples differ in terms of their means.

Fitting this hierarchical prior is not trivial; a novel statistical procedure is developed that describes the process of fitting a hierarchical prior to posterior distributions that were computed using another statistical procedure (LIRA in the case of McKeough et al. 2016). This is a rather general methodology that can be easily applied to projects of similar nature. In the following subsection, we describe the method and we apply to the posterior distributions of $\gamma$ we get from pyBLoCXS for the observed

sources, in order to fit a hierarchical prior of $\gamma$ for the complete population.

### 3.2.2 Estimating the parameters of the Prior $p(\gamma)$

As we described earlier, for each observed source in a given astronomical survey, we can define a spectral model in the Sherpa software and subsequently use pyBLoCXS to draw samples from the posterior distribution $p(\gamma|\mathrm{SD})$, where SD describes the spectral data. We postulate that

$$\gamma \sim \mathrm{Gamma}(a_\gamma, b_\gamma) \tag{3.1}$$

where $\mathrm{Gamma}(a_\gamma, b_\gamma)$ defines a Gamma distribution with shape parameter $a_\gamma$ and scale parameter $b_\gamma$. The assumption for the Gamma distribution as a hierarchical prior comes from the shape of the distribution of all the samples for $\gamma$ obtained from Sherpa for all observed sources in the CDFS survey. This distribution is depicted in Figure 3.1.

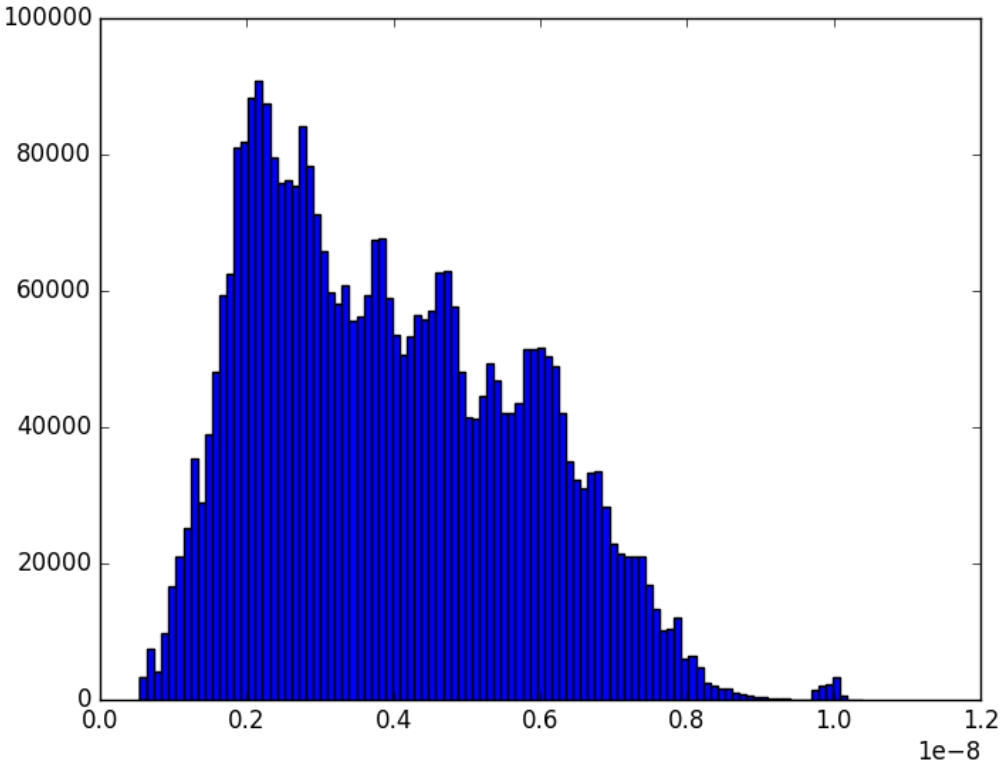



**Figure 3.1:** Histogram of all the samples of $\gamma$ obtained from Sherpa for all observed sources combined in the CDFS survey. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

We independently assume non-informative priors for the parameters in Equation 3.1, i.e., $p(a_\gamma, b_\gamma) \propto 1/b_\gamma$. The distribution in Equation 3.1 can be viewed as a hierarchical prior on $\gamma = (\gamma_1, ..., \gamma_N)$ for the complete source population; hierarchical because it is specified in terms of parameters that are themselves fit to the data. We denote this hierarchical prior distribution by $p(\gamma|a_\gamma, b_\gamma)$.

The prior distribution used in pyBLoCXS however does not coincide with that one described in Equation 3.1. More specifically, pyBLoCXS assumes flat priors on the parameters $N_H, \Gamma, A_m$ of the assumed spectral model (absorbed power law in our case). This translates to a non standard prior distribution $p_{\text{pyBLoCXS}}(\gamma)$ that can be computed numerically using a non parametric density estimator. Figure 3.2 shows the histogram of $p_{\text{pyBLoCXS}}(\gamma)$. The red curve that is over-plotted is the probability density of the distribution computed using a non parametric kernel density estimation with a Gaussian kernel.

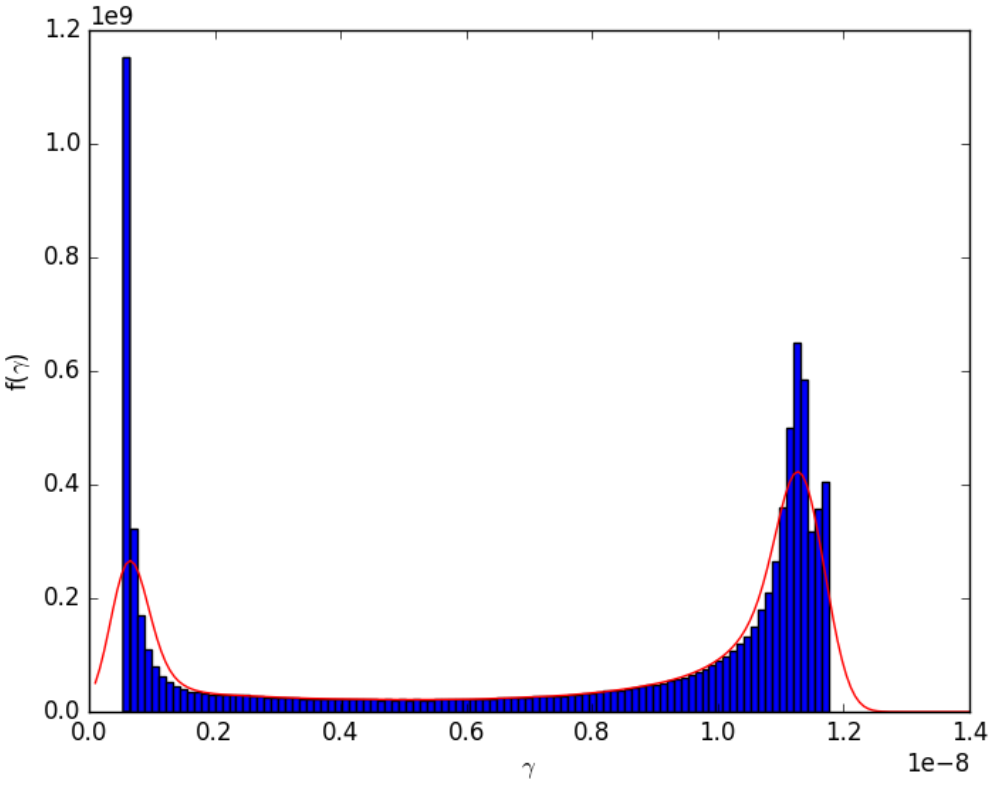



**Figure 3.2:** Histogram of the non standard prior distribution $p_{\text{pyBLoCXS}}(\gamma)$. The red line is the probability density of the distribution computed using a non parametric kernel density estimation with a Gaussian kernel. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

The difference between $p(\gamma|a_\gamma, b_\gamma)$ and $p_{\text{pyBLoCXS}}(\gamma)$ indicates that pyBLoCXS produces a Monte Carlo sample from

$$p(\gamma|\text{SD})_{\text{pyBLoCXS}} = \frac{p(\gamma|a_\gamma, b_\gamma, \text{SD}) \cdot p_{\text{pyBLoCXS}}(\gamma)}{p(\gamma|a_\gamma, b_\gamma)} \tag{3.2}$$

In order to derive Equation 3.2, we express the target posterior distribution $p(\gamma|a_\gamma, b_\gamma, \text{SD})$ (given conditional independence between $(a_\gamma, b_\gamma)$ and SD) as:

$$p(\gamma|a_\gamma, b_\gamma, \text{SD}) = \frac{p(\text{SD}|\gamma) \cdot p(\gamma|a_\gamma, b_\gamma)}{p(\text{SD})} \tag{3.3}$$

where $p(\text{SD}) = \int p(\text{SD}|\gamma) \cdot p(\gamma|a_\gamma, b_\gamma) \text{d}\gamma$. Moreover, pyBLoCXS produces a posterior

sample from

$$p(\gamma|\mathrm{SD})_{\mathrm{pyBLoCXS}} = \frac{p(\mathrm{SD}|\gamma) \cdot p_{\mathrm{pyBLoCXS}}(\gamma)}{p_{\mathrm{pyBLoCXS}}(\mathrm{SD})} \tag{3.4}$$

where $p_{\mathrm{pyBLoCXS}}(\mathrm{SD}) = \int p(\mathrm{SD}|\gamma) \cdot p_{\mathrm{pyBLoCXS}}(\gamma)\mathrm{d}\gamma$. So, by combining Equations 3.3, 3.4 we conclude the relationship 3.2.

The proposed strategy in order to draw samples from the target posterior distribution $p(\gamma|a_\gamma, b_\gamma, \mathrm{SD})$ is to use samples from the pyBLoXCS posterior as a proposal rule in a Metropolis Hastings update. The following algorithm describes the sampling scheme:

**Sampling Algorithm**

**Step 1:** Run the pyBLoXCS for all the observed sources to obtain posterior samples.

**Step 2:** Set $\gamma^{(0)}$ to a randomly selected value from the pyBLoXCS Monte Carlo sample of $\gamma$. Using standard Bayesian methods described in Son & Oh (2006), fit the $\gamma^{(0)}$ to the model in Equation (3.1) to obtain $a_\gamma^{(0)}, b_\gamma^{(0)}$.

**Step 3:** For $t = 1, \ldots, T$

**Step 3a:** Select randomly a proposal $\gamma^{\mathrm{prop}}$ from the pyBLoXCS Monte Carlo sample of $\gamma$.

**Step 3b:** Compute the $n$ Metropolis Hastings acceptance probabilities,

$$
\begin{aligned}
r_i &= \frac{p(\gamma_i^{\text{prop}} \mid a_\gamma^{(t-1)}, b_\gamma^{(t-1)}, \text{SD}) p_{\text{pyBLoXCS}}(\gamma_i^{(t-1)} \mid \text{SD})}{p(\gamma_i^{(t-1)} \mid a_\gamma^{(t-1)}, b_\gamma^{(t-1)}, \text{SD}) p_{\text{pyBLoXCS}}(\gamma_i^{\text{prop}} \mid \text{SD})} \\
&= \frac{p(\gamma_i^{\text{prop}} \mid a_\gamma^{(t-1)}, b_\gamma^{(t-1)}) \cdot p_{\text{pyBLoXCS}}(\gamma_i^{(t-1)})}{p(\gamma_i^{(t-1)} \mid a_\gamma^{(t-1)}, b_\gamma^{(t-1)}) \cdot p_{\text{pyBLoXCS}}(\gamma_i^{\text{prop}})}
\end{aligned} \tag{3.5}
$$

for $i = 1, .., n$.

**Step 3c:** For $i = 1, ..., n$ set

$$
\gamma_i^{(t)} = \begin{cases} \gamma_i^{prop} & \text{with probability } \min(1, r_i) \\ \gamma_i^{(t-1)} & \text{otherwise} \end{cases} \tag{3.6}
$$

**Step 3d:** Sample $a_\gamma^{(t)}, b_\gamma^{(t)}$ using standard Bayesian methods.

The algorithm described above is Markov chain Monte Carlo simulation, so appropriate convergence checks should be implemented as well as burn in checks.

### 3.2.3 Posterior Inference of the Parameters of the Prior $p(\gamma)$

After 12,000 iterations of the sampling algorithm, the posterior estimates for the parameters of interest, $a_\gamma$ and $b_\gamma$, are depicted in Table (3.1) (after neglecting the first 2,000 iterations as burn in).

**Table 3.1:** Posterior estimates of $a_\gamma$ and $b_\gamma$ after 12,000 iterations of the sampling algorithm (we neglect the first 2,000 iterations as burn in).

| | Mean | 2.5% | 97.5% |
|---|---|---|---|
| $a_\gamma$ | 5.58 | 4.72 | 6.48 |
| $b_\gamma$ | $7.07\times10^{-10}$ | $6.04\times10^{-10}$ | $8.38\times10^{-10}$ |

For the observed source populations, we also get from the sampling algorithm the $p(\gamma_i|a_\gamma, b_\gamma, \text{SD})$ for $i = 1, ..., n$. Figure 3.3 plots the histogram (blue colour) of all the samples of $\gamma$ obtained from Sherpa, along with the histogram of the distribution of all the samples of $\gamma$ after correcting for the difference in the prior (green colour). The

red line is the pdf of the prior Gamma distribution with parameters the posterior means of $a_\gamma, b_\gamma$ as showed in the Table 3.1.

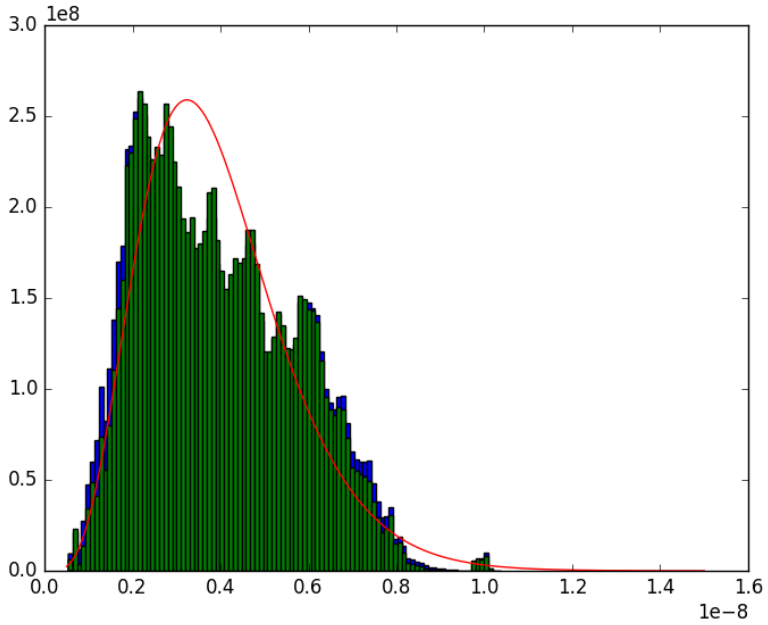



**Figure 3.3:** The blue histogram is the histogram of all the samples of $\gamma$ obtained from Sherpa. The green histogram shows the distribution of all the samples of $\gamma$ after correcting for the difference in the prior. The red line is the pdf of the prior Gamma distribution with parameters the posterior means of $a_\gamma, b_\gamma$ as showed in the above table. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

## 3.3 Probability Modelling of the $\log(N) - \log(S)$ Relationship

The main difference from the model we discussed in the previous chapter lies in the incorporation of the uncertainty about the flux-to-count conversion factor, defined as $\gamma$, into our hierarchical bayesian model. The uncertainty about the flux-to-count conversion rate $\gamma$ is expressed through the Gamma prior distribution

$$p(\gamma) \sim \text{Gamma}(\gamma; a_\gamma, b_\gamma) = \frac{b_\gamma^{a_\gamma}}{\Gamma(a_\gamma)} \theta^{a_\gamma - 1} e^{-b_\gamma \theta} \tag{3.7}$$

which was extracted following the procedure in the previous section. For the observed sources, we also have spectroscopic data, so we have the posterior distribution $p(\gamma_{obs,i}|\text{SD}_i, \alpha_\gamma, b_\gamma)$. This distribution is given to us not in a close form, but through 10,000 MCMC draws.

The distributional assumptions for other parameters of interest remain the same as in the previous chapter, and we omit there here for keeping the text more concise.

### 3.3.1 Derivation of the Joint Posterior Distribution

In the previous subsection we define the probability modelling for the flux distributions. In order to obtain the posterior distribution, we combine all the model assumptions with the prior distributions. We follow the same methodology as in the first chapter for the model without $\gamma$ uncertainty. More specifically, we define as $S_{\rm com} = (S_{\rm obs}, S_{\rm mis})$ the flux vector of observed and missing sources. Similarly $Y^{\rm tot}_{\rm obs} = (Y^{\rm tot}_{i=1}, ..., Y^{\rm tot}_{i=n})$, $Y^{\rm tot}_{\rm mis} = (Y^{\rm tot}_{i=n+1}, ..., Y^{\rm tot}_{i=N})$ where $i = 1, .., n$ corresponds to the observed sources. The complete data posterior distribution can be summarised as:

$$p(N, \theta, \tau, \gamma_{\rm com}, S_{\rm com}, I_{\rm com}, Y^{\rm src}_{\rm obs}, Y^{\rm src}_{\rm mis}, Y^{\rm tot}_{\rm mis}, B_{\rm mis}, L_{\rm mis}, E_{\rm mis}, A_{\rm mis}| \\ n, Y^{\rm tot}_{\rm obs}, B_{\rm obs}, L_{\rm obs}, E_{\rm obs}, A_{\rm obs}, {\rm SD}_{\rm obs}, a_\gamma, b_\gamma)$$

where $\rm SD_{obs}$ is the spectral data for the observed sources and $(a_\gamma, b_\gamma)$ are the parameters of the prior of $\gamma$. In order to overcome the significant computational difficulties posed by the presence of the missing data, we marginalise the full joint posterior distribution over the missing sources, i.e. we integrate out out the missing data parameters

$$(S_{\rm mis}, I_{\rm mis}, Y^{\rm src}_{\rm mis}, Y^{\rm tot}_{\rm mis}, \gamma_{\rm mis}, B_{\rm mis}, L_{\rm mis}, E_{\rm mis}, A_{\rm mis}).$$

This leaves the main parameters of interest $(N, \theta, \tau)$ and the parameters of the flux, flux to counts conversion factors and source photon counts $(S_{obs}, \gamma_{obs}, Y^{src}_{obs})$ of the observed sources in the marginalised joint posterior. By using this sampling scheme, the dimension of the sampled quantities is kept constant.

The marginalised joint-posterior distribution of the parameters of interest is (see the

Appendix A for detailed derivation):

$$p(\gamma_{\text{obs}}, N, \theta, \tau, S_{\text{obs}}, Y_{\text{obs}}^{\text{src}} \mid n, Y_{\text{obs}}^{\text{tot}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}}, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma) \tag{3.8}$$

$$= \frac{1}{p(n, Y_{\text{obs}}^{\text{tot}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}}, \text{SD}_{\text{obs}}, \mu, \sigma)} \cdot \binom{N}{n} \cdot (1 - \pi(\theta, \tau))^{N-n} \tag{3.9}$$

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

$$\cdot\, p(\gamma_{\text{obs}} \mid \text{SD}_{\text{obs}}, a_\gamma, b_\gamma) \cdot p(S_{\text{obs}} | \theta, \tau)$$

$$\cdot\, p(I_{\text{obs}} \mid \gamma_{\text{obs}}, S_{\text{obs}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}})$$

$$\cdot\, p(Y_{\text{obs}}^{\text{tot}} \mid I_{\text{obs}}, S_{\text{obs}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, \gamma_{\text{obs}})$$

$$\cdot\, p(Y_{\text{obs}}^{\text{src}} \mid Y_{\text{obs}}^{\text{tot}}, I_{\text{obs}}, S_{\text{obs}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, \gamma_{\text{obs}})$$

where

$$\pi(\theta, \tau) = \int g(S, B, L, E, \gamma) \cdot p(\gamma) \cdot p(S|\theta, \tau) \cdot p(B, L, E)\, \mathrm{d}S\, \mathrm{d}B\, \mathrm{d}E\, \mathrm{d}L\, \mathrm{d}\gamma$$

- **Single power law model**: the Equation (3.9) becomes:

$$
\begin{aligned}
&p(\gamma_{\text{obs}}, N, \theta, \tau, S_{\text{obs}}, Y_{\text{obs}}^{\text{src}} \mid n, Y_{\text{obs}}^{\text{tot}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}}, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma) \\
&\propto \frac{1}{p(n, Y_{\text{obs}}^{\text{tot}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}}, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma)} \cdot \binom{N}{n} \cdot (1 - \pi(\theta, \tau))^{N-n} \\
&\quad \cdot \binom{N + a_N - 1}{a_N - 1} \Big(\frac{1}{1 + b_N}\Big)^N \Big(\frac{b_N}{1 + b_N}\Big)^{a_N} \mathbb{I}_{\{N \in \mathbb{Z}^+\}} \\
&\quad \cdot \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} \mathbb{I}_{\{\theta > 0\}} \\
&\quad \cdot \frac{b_m^{a_m}}{\Gamma(a_m)} \tau^{a_m - 1} e^{-b_m \tau} \mathbb{I}_{\{\tau > 0\}} \\
&\quad \cdot \prod_{i=1}^{n} p(\gamma_i | \text{SD}_i, a_\gamma, b_\gamma) \cdot \theta \tau^\theta S_i^{-(\theta+1)} \mathbb{I}_{\{\tau < S_i\}} \cdot g(S_i, B_i, L_i, E_i, \gamma_i) \\
&\quad \cdot \frac{(\frac{S_i E_i}{\gamma_i} + B_i A_i)^{Y_i^{\text{tot}}}}{Y_i^{\text{tot}}!} e^{(\frac{S_i E_i}{\gamma_i} + B_i A_i)} \mathbb{I}_{\{Y_i^{tot} \in \mathbb{Z}^+\}} \\
&\quad \cdot \binom{Y_i^{\text{tot}}}{Y_i^{\text{src}}} \Big(\frac{\frac{S_i E_i}{\gamma_i}}{\frac{S_i E_i}{\gamma_i} + B_i A_i}\Big)^{Y_i^{\text{src}}} \Big(1 - \frac{\frac{S_i E_i}{\gamma_i}}{\frac{S_i E_i}{\gamma_i} + B_i A_i}\Big)^{Y_i^{\text{tot}} - Y_i^{\text{src}}} \mathbb{I}_{\{Y_i^{\text{src}} \in \{0, 1, \ldots, Y_i^{\text{tot}}\}\}}
\end{aligned}
$$

- **Broken power law model**: the Equation (3.9) becomes:

$$
\begin{aligned}
&p(\gamma_{\text{obs}}, N, \theta, \tau, S_{\text{obs}}, Y_{\text{obs}}^{\text{src}} \mid n, Y_{\text{obs}}^{\text{tot}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}}, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma) \\
&\propto \frac{1}{p(n, Y_{\text{obs}}^{\text{tot}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}}, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma)} \cdot \binom{N}{n} \cdot (1 - \pi(\theta, \tau))^{N-n} \\
&\quad \cdot \binom{N + a_N - 1}{a_N - 1} \Big(\frac{1}{1 + b_N}\Big)^N \Big(\frac{b_N}{1 + b_N}\Big)^{a_N} \mathbb{I}_{\{N \in \mathbb{Z}^+\}} \\
&\quad \cdot \prod_{j=1}^{m} \Big[\frac{b_j^{a_j}}{\Gamma(a_j)} \theta_j^{a_j - 1} e^{-b_j \theta} \mathbb{I}_{\{\theta_j > 0\}}\Big] \cdot p(\tau_1, \tau_2, ..., \tau_m) \mathbb{I}_{\{0 < \tau_1 < \tau_2 < \ldots < \tau_m\}} \\
&\quad \cdot \prod_{i=1}^{n} \Big[p(\gamma_i | \text{SD}_i, a_\gamma, b_\gamma) \cdot g(S_i, B_i, L_i, E_i, \gamma_i) \\
&\quad \cdot \sum_{j=1}^{m} \Big[\prod_{i=1}^{j-1} \Big(\frac{\tau_{i+1}}{\tau_i}\Big)^{-\theta_i}\Big] \Big(\frac{\theta_j}{\tau_j}\Big) \Big(\frac{S}{\tau_j}\Big)^{-(\theta_j + 1)} \mathbb{I}_{\{\tau_j \le S \le \tau_{j+1}\}} \\
&\quad \cdot \frac{(\frac{S_i E_i}{\gamma_i} + B_i A_i)^{Y_i^{\text{tot}}}}{Y_i^{\text{tot}}!} e^{(\frac{S_i E_i}{\gamma_i} + B_i A_i)} \mathbb{I}_{\{Y_i^{tot} \in \mathbb{Z}^+\}} \\
&\quad \cdot \binom{Y_i^{\text{tot}}}{Y_i^{\text{src}}} \Big(\frac{\frac{S_i E_i}{\gamma_i}}{\frac{S_i E_i}{\gamma_i} + B_i A_i}\Big)^{Y_i^{\text{src}}} \Big(1 - \frac{\frac{S_i E_i}{\gamma_i}}{\frac{S_i E_i}{\gamma_i} + B_i A_i}\Big)^{Y_i^{\text{tot}} - Y_i^{\text{src}}} \mathbb{I}_{\{Y_i^{\text{src}} \in \{0, 1, \ldots, Y_i^{\text{tot}}\}\}}\Big]
\end{aligned}
$$

### 3.3.2 Derivations of the Conditional Posterior Distributions

The sampling of such a complicated posterior distribution is not trivial. Following the same logic as in the previous chapter, we utilise a Blocked Gibbs sampler in which each conditional posterior distribution requires a different sampling strategy. The derivation of the conditional posterior distributions can be found in detail at the Appendix. Here we present each conditional posterior distribution and the strategy used for sampling from that distribution.

#### Conditional Posterior Distributions of the Single Power Law model

For the Single Power Law model, we have:

**Conditional distribution of $\boldsymbol{Y_{obs}^{src}}$**: The full conditional distribution for $Y_{obs}^{src}$ is:

$$\begin{aligned} p(Y_{obs}^{src}|\cdot) &\propto p(Y_{obs}^{src}|Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs}, I_{obs}, S_{obs}, \gamma_{\text{obs}}) \\ &= \prod_{i=1}^{n} Binomial\Big(Y_i^{src}; Y_i^{tot}, \frac{\lambda_i}{\lambda_1 + \kappa_i}\Big), \end{aligned}$$

where $\lambda_i = S_i E_i / \gamma_i$ and $k_i = B_i A_i$. Since the observed sources are independent we can sample the vector $Y_{obs}^{src}$ component-wise for $i = 1, ..., n$ as

$$p(Y_i^{src}|\cdot) \sim Binomial\Big(Y_i^{src}; Y_i^{tot}, \frac{\lambda_i}{\lambda_1 + \kappa_i}\Big).$$

**Conditional distribution of $\boldsymbol{S_{obs}}$**: The full conditional distribution for $S_{obs}$ is:

$$\begin{aligned} p(S_{obs}|\cdot) \propto\ & p(S_{obs}|N, \theta, \tau) \cdot p(I_{obs}|S_{obs}, B_{obs}, L_{obs}, E_{obs}, \gamma_{\text{obs}}) \\ & \cdot p(Y_{obs}^{tot}|B_{obs}, L_{obs}, E_{obs}, I_{obs}, S_{obs}, \gamma_{\text{obs}}) \cdot p(Y_{obs}^{src}|Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs}, I_{obs}, S_{obs}, \gamma_{\text{obs}}) \end{aligned}$$

where $\lambda(S_i, E_i, B_i, L_i) = S_i E_i/\gamma_i$ and $k(B_i, A_i) = B_i A_i$. By independence of the $S_{obs}$ we can sample component-wise for $i = 1, ..., n$ as

$$p(S_i|\cdot) \sim Pareto(S_i|N, \theta, \tau)\mathbb{I}_{\{\tau<S_i\}} \cdot g(S_i, B_i, L_i, E_i, \gamma_i) \cdot Poisson(Y_i^{tot}; \lambda_i + \kappa_i)$$
$$\cdot Binomial\Big(Y_i^{src}; Y_i^{tot}, \frac{\lambda_i}{\lambda_1 + \kappa_i}\Big).$$

In the datasets we are interested in, the values of the flux $S$ are really small $\sim 10^{-18} - 10^{-12}$. In order to avoid numerical overflows we take a logarithmic transformation

$$z_i = log(S_i)$$
$$p(z_i|\cdot) = e^{z_i} \cdot p(S_i = e^{z_i})\cdot$$

We use the Metropolis- Hastings Algorithm in order to sample each $z_i$. A truncated normal distribution is used as a proposal distribution since we have a lower bound $log(\tau)$, and we tune the variance of the proposal distribution adaptively during the first 300 iterations of the MCMC so as to achieve an acceptance ratio between 20% to 60%.

**Conditional distribution of $\boldsymbol{\theta}$**: The full conditional distribution for $\theta$ is:

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

We use the Metropolis- Hastings Algorithm in order to sample $\theta$. A symmetric normal distribution is used as a proposal distribution, i.e. $N(\theta^{prop}; \theta^{curr}, \sigma_\theta^2)$. The variance of the proposal distribution $\sigma_\theta^2$ is chosen so as to achieve an acceptance ratio between 20% to 60%.

**Conditional distribution of $\boldsymbol{N}$**: The full conditional distribution for $N$, the total unknown number of sources, is:

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

We draw samples from this conditional distribution by the inverse CDF method. The CDF is computed analytically.

**Conditional distribution of $\boldsymbol{\tau}$**: The full conditional distribution for $\tau$ is:

$$\begin{aligned} p(\tau|\cdot) &\propto p(\tau) \cdot (1 - \pi(\theta, \tau))^{N-n} \cdot p(S_{obs}|N, \theta, \tau) \\ &\propto \tau^{n\theta + a_m - 1} \cdot e^{-b_m \tau} \cdot (1 - \pi(\theta, \tau))^{N-n} \cdot \mathbb{I}_{\{\tau < c_m\}}, \quad \text{where } c_m = min(S_1, ..., S_n) \end{aligned}$$

The $\tau$ is the most challenging parameter to sample since it is in the deepest level in the model hierarchy. Since $\tau$ is smaller than the minimum of $S$, its value is very small. Thus, we take a logarithmic transformation in order to avoid underflowing and to increase the numerical stability, i.e.:

$$\begin{aligned} \eta &= log(\tau) \\ p(\eta|\cdot) &= e^{\eta} \cdot p(\tau = e^{\eta}) \end{aligned}$$

We use the Metropolis- Hastings Algorithm in order to sample $\tau$. A truncated normal distribution is used as a proposal distribution since we have an upper bound $log(c_m)$, the minimum value of the observed fluxes. We tune the variance of the proposal distribution so as to achieve an acceptance ratio between 20% to 60%.

**Conditional posterior distribution of $\boldsymbol{\gamma_{obs,i}}$**: For the observed sources $i = 1, ..., n$ we have the spectral information $\text{SD}_i$. So, the full conditional posterior

distribution is

$$
\begin{aligned}
p(\gamma_{obs}|\cdot) \propto & \prod_{i=1}^{n} p(\gamma_i|SD_i, a_\gamma, b_\gamma) \cdot g(S_i, B_i, L_i, E_i, \gamma_i) \\
& \cdot \frac{(\frac{S_i E_i}{\gamma_i} + B_i A_i)^{Y_i^{tot}}}{Y_i^{tot}!} e^{(\frac{S_i E_i}{\gamma_i} + B_i A_i)} \mathbb{I}_{\{Y_i^{tot} \in \mathbb{Z}^+\}} \\
& \cdot \binom{Y_i^{tot}}{Y_i^{src}} \Big( \frac{\frac{S_i E_i}{\gamma_i}}{\frac{S_i E_i}{\gamma_i} + B_i A_i} \Big)^{Y_i^{src}} \Big( 1 - \frac{\frac{S_i E_i}{\gamma_i}}{\frac{S_i E_i}{\gamma_i} + B_i A_i} \Big)^{Y_i^{tot} - Y_i^{src}} \mathbb{I}_{\{Y_i^{src} \in \{0,1,\ldots,Y_i^{tot}\}\}}
\end{aligned}
$$

By independence of the observed sources, we can sample component-wise for $i = 1, ..., n$ as:

$$
\begin{aligned}
p(\gamma_i|\cdot) \propto & \, p(\gamma_i|SD_i, a_\gamma, b_\gamma) \cdot g(S_i, B_i, L_i, E_i, \gamma_i) \cdot Poisson(Y_i^{tot}; \frac{S_i E_i}{\gamma_i} + B_i A_i) \\
& \cdot Binomial\Big(Y_i^{src}; Y_i^{tot}, \frac{\frac{S_i E_i}{\gamma_i}}{\frac{S_i E_i}{\gamma_i} + B_i A_i}\Big).
\end{aligned}
$$

We use the Metropolis- Hastings Algorithm in order to sample each $\gamma_i$, using $p(\gamma_i|SD_i, a_\gamma, b_\gamma)$ as proposal distribution. The algorithm would be as follows:

**Step 1:** Sample a proposal state randomly from $p(\gamma_i|SD_i, a_\gamma, b_\gamma)$, i.e. $\gamma_i^{prop} \sim p(\gamma_i|SD_i, a_\gamma, b_\gamma)$.

**Step 2:** Compute the Metropolis Hastings ratio :

$$\begin{aligned}\alpha_i &= \frac{p(\gamma_i^{prop}|\cdot)\cdot p(\gamma_i^{curr}|SD_i, a_\gamma, b_\gamma)}{p(\gamma_i^{curr}|\cdot)\cdot p(\gamma_i^{prop}|SD_i, a_\gamma, b_\gamma)} \\ &= \frac{g(S_i, B_i, L_i, E_i, \gamma_i^{prop})\cdot \text{Poisson}(Y_i^{tot}; \frac{S_i E_i}{\gamma_i^{prop}} + B_i A_i)}{g(S_i, B_i, L_i, E_i, \gamma_i^{curr})\cdot Poisson(Y_i^{tot}; \frac{S_i E_i}{\gamma_i^{curr}} + B_i A_i)} \\ &\quad \cdot \frac{Binomial\Big(Y_i^{src}; Y_i^{tot}, \frac{\frac{S_i E_i}{\gamma_i^{prop}}}{\frac{S_i E_i}{\gamma_i^{prop}} + B_i A_i}\Big)}{Binomial\Big(Y_i^{src}; Y_i^{tot}, \frac{\frac{S_i E_i}{\gamma_i^{curr}}}{\frac{S_i E_i}{\gamma_i^{curr}} + B_i A_i}\Big)}\end{aligned}$$

**Step 3:** Set

$$\gamma_i^{(new)} = \begin{cases} \gamma_i^{prop} & \text{with probability } \min(1, \alpha_i) \\ \gamma_i^{\text{curr}} & \text{otherwise} \end{cases}$$

The limitation of sampling $p(\gamma_i|\cdot)$ in that manner is the lack of flexibility in tuning the acceptance ration. However, in our numerical experiments for both simulated data and the CDFS data, we haven't come across a case for which the acceptance ratio was below 10%.

**Computing $\boldsymbol{\pi(\theta, \tau)}$**: The marginal probability of observing a source as a function of the slope (or slopes) $\theta$ and $\tau$, i.e.

$$\pi(\theta, \tau) = \int g(S, B, L, E, \gamma)\cdot p(\gamma)\cdot p(S|\theta, \tau)\cdot p(B, L, E)\,\mathrm{d}S\,\mathrm{d}B\,\mathrm{d}E\,\mathrm{d}L\,\mathrm{d}\gamma$$

is computed using Monte Carlo integration. More specifically, we draw samples from the empirical distribution of $B, L, E$ as we discussed in the first chapter. We draw samples of the flux $S$ from a Pareto distribution conditional on the values of $(\theta, \tau)$ if we assume a single power law model (or the $m$-component broken power law distribution if we assume a broken power law model). We also sample $\gamma$ from the prior $p(\gamma|a_\gamma, b_\gamma)$. The empirical average of $g$ is the approximation of $\pi(\theta, \tau)$.

### Conditional Posterior Distributions of the Broken Power Law model

For the Broken Power Law model, the conditional distributions for $N$ and $Y_{\text{obs}}^{\text{src}}$ are the same as in the Single Power Law model, except for the different computation of $\pi(\theta, \tau)$ in which we replace the Pareto distribution for the flux $S$ with the broken Pareto pdf. The main computational differences in the Gibbs sampler lie in sampling from the conditional posterior distributions of $p(\tau_1|\cdot)$, $p(\tau_2, ..., \tau_m|\cdot)$ and $p(\theta_1, ..., \theta_m|\cdot)$. More specifically, we have that

**Conditional posterior distribution of $\boldsymbol{\theta = (\theta_1, ..., \theta_m)^T}$**: The full conditional posterior distribution for $\theta = (\theta_1, ..., \theta_m)$ is:

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

where $I(j) = \{i \ : \ \tau_j \leq S_i \leq \tau_{j+1}\}$ denotes the existence of sources with flux contained in the interval corresponding to the $j$ -th mixture component, and $n(j)$ is the cardinality of $I(j)$ (the number of sources in that interval). The sampling of the vector $\theta$ is done using the Metropolis-Hastings algorithm. A multivariate normal distribution is used as a proposal distribution with the covariance matrix chosen so as to achieve an acceptance ratio between 20% to 60%.

**Conditional posterior distribution of $\boldsymbol{\tau_1}$**: The full conditional posterior distribution for $\tau_1$ is:

$$p(\tau_1|\cdot) \propto \tau^{n\theta_1+a_m-1} \cdot e^{-b_m\tau} \cdot (1-\pi(\theta,\tau))^{N-n} \cdot \prod_{j=2}^{m} e^{-\frac{c_m^2 \cdot [\log(\tau_m - \tau_{(m-1)}) - \mu_m]^2}{2}}$$
$$\cdot \frac{1}{\tau_m - \tau_{(m-1)}} \cdot \mathbb{I}_{\{\tau < c_m\}}$$

The sampling of $\tau_1$ is performed similarly to the case of $\tau$ in the Single Power Law model. In other words, we take a logarithmic transformation $\eta_1 = log(\tau_1)$ and we use the Metropolis- Hastings Algorithm in order to sample $\tau_1$. A truncated normal distribution is used as a proposal distribution since we have an upper bound $log(c_m)$, the minimum value of the observed fluxes. We tune the variance of the proposal distribution so as to achieve an acceptance ratio between 20% to 60%.

**Conditional posterior distribution of $(\boldsymbol{\tau_2, ...\tau_m})$**: We sample $(\tau_2, ...\tau_m)$ via the full joint conditional posterior distribution of the transformed variables $\eta_2, ..., \eta_m$. We remind that $\eta_j = h_j(\tau_j \mid \tau_{j-1}) = \log(\tau_j - \tau_{j-1})$, $j = 2, ..., m$. Thus, after applying a change of variables we have:

$$p(\eta_2, ...\eta_m|\cdot) \propto e^{\sum_{j=2}^{m} \eta_j} \cdot [1-\pi(\theta,\tau)]^{N-n} \cdot \text{Multivariate Gaussian}(\mu, C)$$
$$\cdot \prod_{j=1}^{m} \Big\{ \prod_{l=1}^{j-1} \Big( \frac{\tau_{l+1}}{\tau_l} \Big)^{-\theta_l} \Big\}^{n(j)} \cdot \prod_{i \in I(j)} \Big( \frac{\theta_j}{\tau_j} \Big) \Big( \frac{S}{\tau_j} \Big)^{-(\theta_j+1)}$$

The sampling of the vector $(\eta_2, ...\eta_m)$ is done using the Metropolis-Hastings algorithm. A multivariate normal distribution is used as a proposal distribution with the covariance matrix chosen so as to achieve an acceptance ratio between 20% to 60%.

### 3.3.3 Parameter Inference

The posterior inference for the model parameters is based on the posterior MCMC draws. For the single power law model, our main parameter of interest is the slope $\theta$, while the minimum threshold $\tau$ and the total number of sources $N$ are of secondary interest. Similarly, for the broken power law model, are main focus is on the broken power law slopes $\theta_1, ..., \theta_m$ and the flux breakpoints $\tau_2, ..., \tau_m$.

We use the posterior mean, median and mode to represent the posterior point estimates and the 95% posterior credible intervals to represent the uncertainty. We estimate those values after running the MCMC chains for a large number of iterations. We disregard the first half part of the chain as burn-in. The convergence of the chains is examined by looking at the trace plots.

We also construct a posterior $\log(N) - \log(S)$ plot using the MCMC draws and sampling the $S_{\text{mis}}$ for each iteration. Examining this posterior $\log(N) - \log(S)$ curve is helpful in assessing the parameter estimates from a visual perspective.

## 3.4 Model Validation

As in the previous chapter, it is crucial to perform a validation of our model and our software, since the model is complex and contains many different levels. We should also exercise caution in the development of the code, since we come across relatively small numbers for some of the parameters, like the flux, and thus we should consider the problems with arithmetic underflowing. The validation is done by generating data according to the model and check the model-fitting software for consistent posterior estimates.

### 3.4.1 Validation using posterior interval plots

We follow the same methodology for performing the validation as in the previous chapter. As a reminder, we generate simulated datasets from the model generating software given some fixed values of the parameters, and then fit the simulated dataset to the model-fitting software to obtain posterior draws of the parameters. For each

parameter, we evaluate posterior credible sets $C$ of level $\alpha$:

$$\int_C p(x \mid y_{\text{obs}})\mathrm{d}x = 1 - \alpha.$$

The estimate based on the samples from the Blocked Gibbs sampler is $(\hat{x}_L, \hat{x}_U)$ such that $q(\hat{x}_L) = \alpha/2$ and $q(\hat{x}_U) = 1 - \alpha/2$. For each parameter, we record if the posterior credible set $C$ of level $\alpha$ contains the "true" value of the parameter used for generating the simulated dataset.

The steps 1,2 and 3 are repeated 20 times. We expect for each parameter that the posterior credible set $C$ of level $\alpha$ contains the "true" value of the parameter most of the times, especially if the above steps were to be repeated for a very large number of times.

**Single Pareto Model**: For the single Pareto model, we simulated 20 datasets using parameters $\theta = 0.8$, $N = 1000$ and $\tau = 2 \times 10^{-17}$. For the flux to count rate conversion factor distributions of the observed sources, $p(\gamma|\text{SD})$, we sample with replacement from the set of the 355 distributions we have estimated from the CDFS dataset as described in previous subsection. More specifically, we draw $N$ samples from a Pareto distribution for the flux of the complete source population, i.e. $S_i^{\text{tot}} \sim \text{Pareto}(\theta, \tau)$, $i = 1, ..., N$ and then draw $B_i, L_i, E_i$ from the joint distribution $p(B, L, E)$ as described in the first Chapter. using the background map and exposure map from the Chandra Deep Field South survey. Then, we apply the incompleteness function to extract the $S^{\text{obs}}$ by computing the function $g(C_i = \frac{S_i * E_i}{\gamma_i}, B_i, L_i, E_i)$ and comparing it with $u_i \sim \text{Uniform}(0,1)$ and assuming the source $i$ is observed if $u_i < g(S_i, B_i, L_i, E_i)$. For the $\gamma_i$ in the incompleteness function, we use the mean of the $p(\gamma_i|\text{SD})$ as a point estimate.

Figures 3.4, 3.5 and 3.6 show the posterior 95% interval for each of the 20 simulated datasets for the parameters $\theta$, $N$ and $\tau$ respectively. At least 19 out of the 20 posterior intervals contain the values of the parameters that were used to generate the simulated datasets.

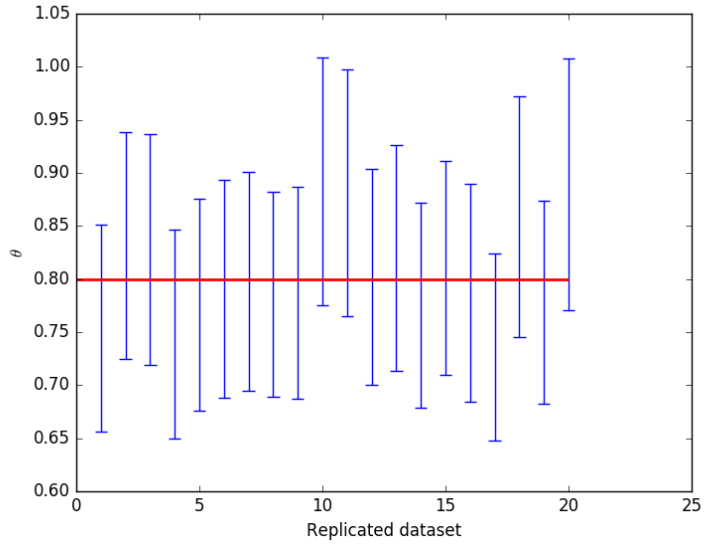



**Figure 3.4:** Posterior credible intervals of $\theta$ from 20 dataset simulations using validation process for the Single Pareto model. The "true" value is $\theta = 0.8$.

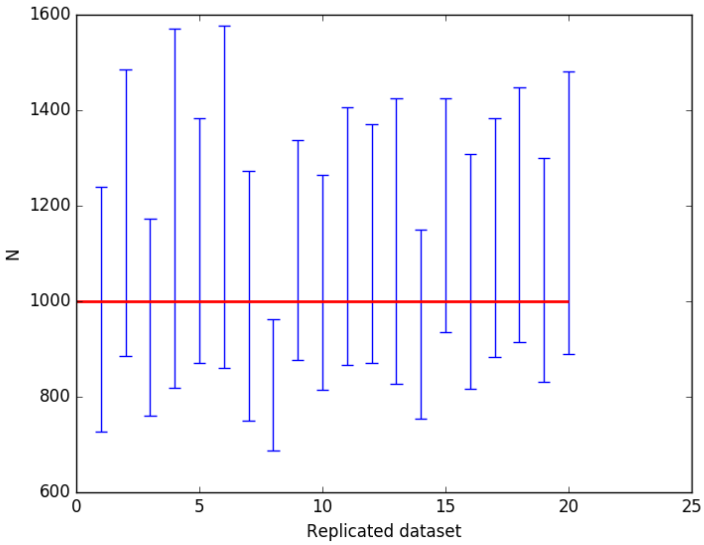



**Figure 3.5:** Posterior credible intervals of $N$ from 20 dataset simulations using validation process for the Single Pareto model. The "true" value is $N = 1000$.

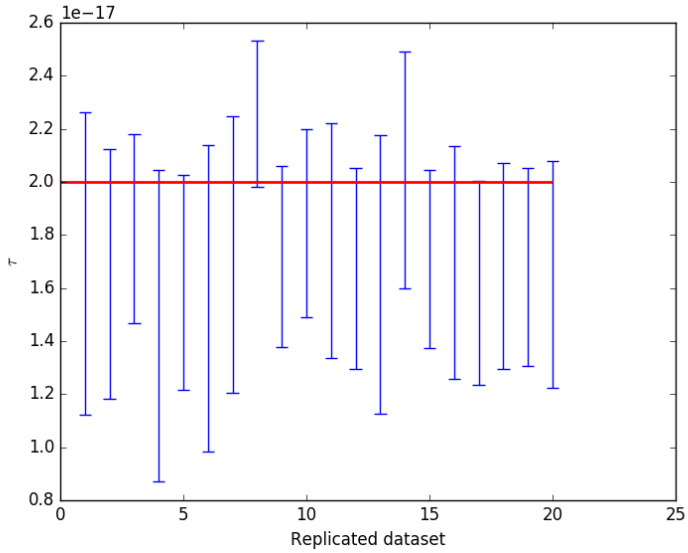



**Figure 3.6:** Posterior credible intervals of $\tau$ from 20 dataset simulations using validation process for the Single Pareto model. The "true" value is $\tau = 2 \times 10^{-17}$. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

**Broken Pareto Model with 1 breakpoint**: For the Broken Pareto model with 1 break, we simulated 20 datasets using parameters $\theta_1 = 0.8$, $\theta_2 = 1.2$, $N = 2000$, $\tau_1 = 1.5 \times 10^{-17}$ and $\tau = 2 \times 10^{-15}$. The flux to count rate conversion factor distributions of the observed sources, $p(\gamma|\text{SD})$, were sampled with replacement from the set of the 355 distributions of the CDFS dataset. More specifically, we draw $N$ samples from the Broken Power law distribution for the flux of the complete source population (using the inverse CDF method), and then draw $B_i, L_i, E_i$ from the joint distribution $p(B, L, E)$ as described in the first Chapter. Then, we applied the incompleteness function to extract the $S^{\text{obs}}$ by computing the function $g(C_i = \frac{S_i * E_i}{\gamma_i}, B_i, L_i, E_i)$ and comparing it with $u_i \sim$ Uniform(0,1) and assuming the source $i$ is observed if $u_i < g(S_i, B_i, L_i, E_i)$. For the $\gamma_i$ in the incompleteness function, we use the mean of the $p(\gamma_i|\text{SD})$ as a point estimate.

Figures 3.7, 3.8, 3.9,3.10 and 3.11 show the posterior 95% interval for each of the 20 simulated datasets for the parameters $\theta_1$, $\theta_2$, $N$, $\tau_1$ and $\tau_2$ respectively. All the 20 marginal posterior intervals of all 5 parameters contain the values of the parameters that were used to generate the simulated datasets. As in Chapter 2, we can observe that the posterior intervals of the breakpoint $\tau_2$ are rather wide. We observe the same multimodality and skewness in the marginal posterior distribution of $\tau_2$ for the CDFS dataset. We believe that this multimodality is an evidence of a lack of

enough data, and thus the MCMC explores many different regions. We refer the reader to the corresponding subsection in the Chapter 2 for more details.

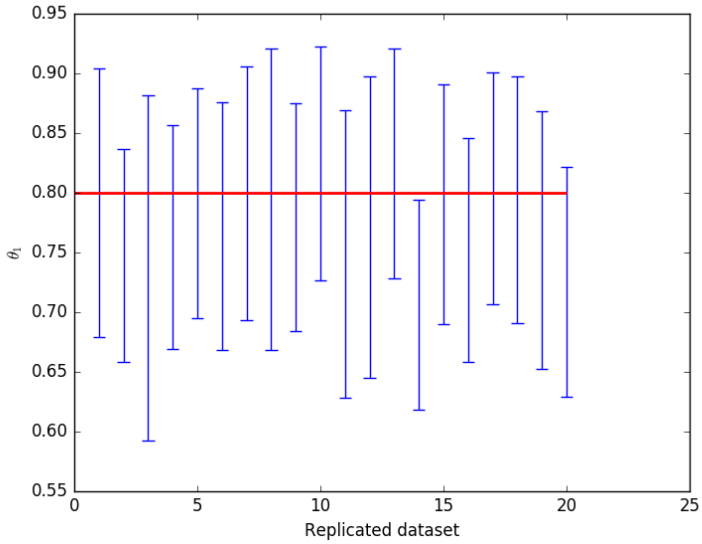



**Figure 3.7:** Posterior credible intervals of $\theta_1$ from 20 dataset simulations using validation process for the Broken Pareto model with 1 break. The "true" value is $\theta_1 = 0.8$.

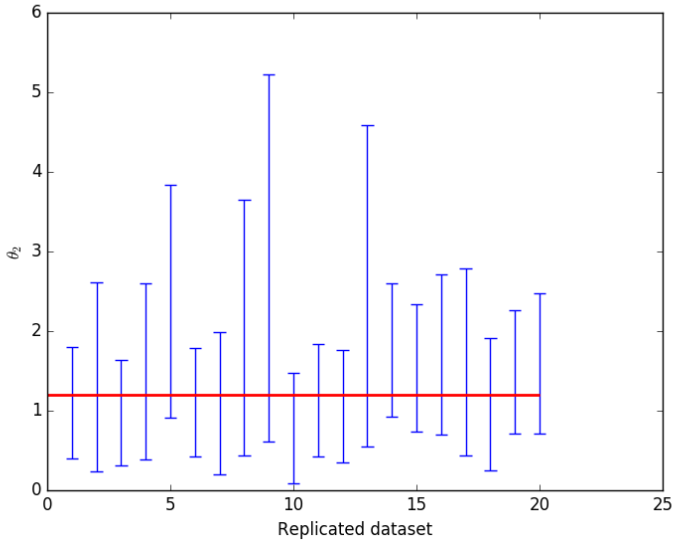



**Figure 3.8:** Posterior credible intervals of $\theta_2$ from 20 dataset simulations using validation process for the Broken Pareto model with 1 break. The "true" value is $\theta_2 = 1.2$.

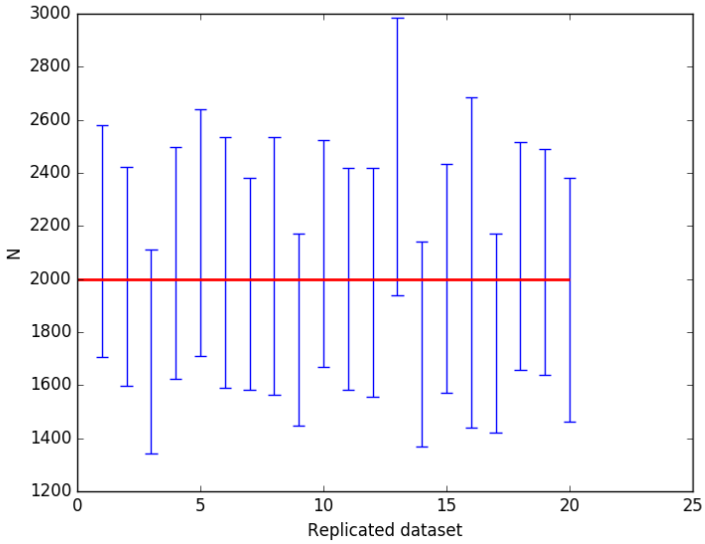



**Figure 3.9:** Posterior credible intervals of $N$ from 20 dataset simulations using validation process for the Broken Pareto model with 1 break. The "true" value is $N = 2000$.

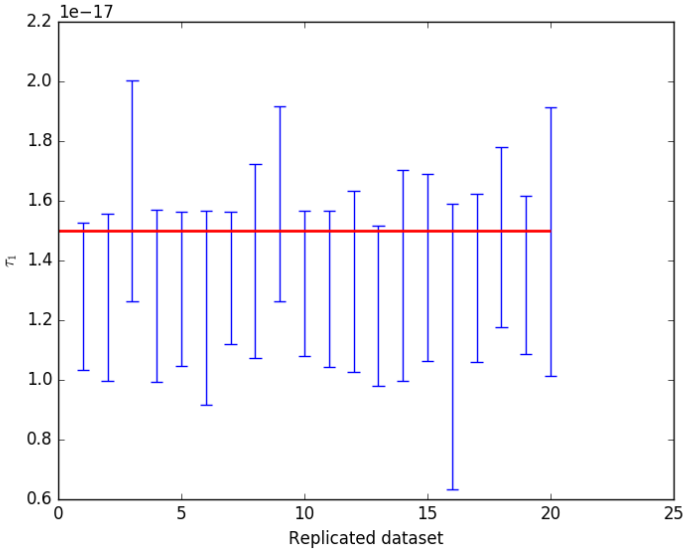



**Figure 3.10:** Posterior credible intervals of $\tau_1$ from 20 dataset simulations using validation process for the Broken Pareto model with 1 break. The "true" value is $\tau_1 = 1.5 \times 10^{-17}$. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

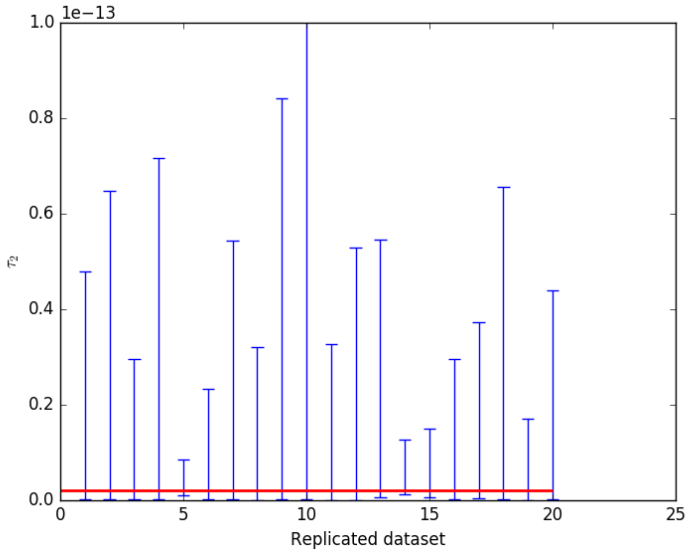



**Figure 3.11:** Posterior credible intervals of $\tau_2$ from 20 dataset simulations using validation process for the Broken Pareto model with 1 break. The "true" value is $\tau_2 = 2 \times 10^{-15}$. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

**Broken Pareto Model with 2 breakpoint**: For the Broken Pareto model with 2 breaks, we simulated 20 datasets using parameters $\theta_1 = 0.5$, $\theta_2 = 0.7$, $\theta_3 = 1.3$, $N = 1000$, $\tau_1 = 1.5 \times 10^{-17}$, $\tau = 1.3 \times 10^{-15}$ and $\tau_3 = 6 \times 10^{-15}$. Similarly as in the 1-breakpoint model, the flux to count rate conversion factor distributions of the observed sources, $p(\gamma|\text{SD})$, were sampled with replacement from the set of the 355 distributions of the CDFS dataset. More specifically, we draw $N$ samples from the Broken Power law distribution for the flux of the complete source population (using the inverse CDF method), and then draw $B_i, L_i, E_i$ from the joint distribution $p(B, L, E)$. Then, we applied the incompleteness function to extract the $S^{\text{obs}}$ by computing the function $g(C_i = \frac{S_i * E_i}{\gamma_i}, B_i, L_i, E_i)$ and comparing it with $u_i \sim$ Uniform(0,1); we assume that the source $i$ is observed if $u_i < g(S_i, B_i, L_i, E_i)$. For the $\gamma_i$ in the incompleteness function, we use the mean of the $p(\gamma_i|\text{SD})$ as a point estimate.

Figures 3.12, 3.13, 3.14, 3.15, 3.16, 3.17 and 3.18 show the posterior 95% interval for each of the 20 simulated datasets for the parameters $\theta_1$, $\theta_2$, $\theta_3$, $N$, $\tau_1$, $\tau_2$ and $\tau_3$ respectively. At least 19 out of the 20 marginal posterior intervals of all 7 parameters of interest contain the values of the parameters that were used to generate the simulated datasets. As in the case of the Broken Pareto Model with 1 break, we can observe a bias in the estimation of $\tau_3$. More specifically, we can observe that the

posterior intervals of the breakpoint $\tau_3$ are rather wide (as well as for $\theta_3$, since those 2 parameters are heavily correlated). We believe that the wide posterior interval indicates that the model doesn't not have enough data to converge to a specific breakpoint and explores many different regions for the breakpoint.

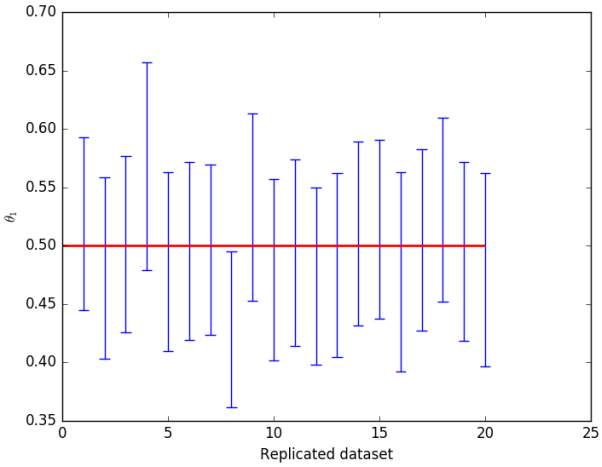



**Figure 3.12:** Posterior credible intervals of $\theta_1$ from 20 dataset simulations using validation process for the Broken Pareto model with 2 breaks. The "true" value is $\theta_1 = 0.5$.

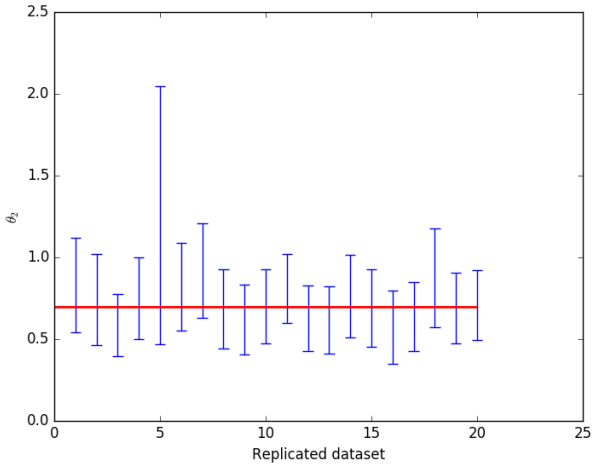



**Figure 3.13:** Posterior credible intervals of $\theta_2$ from 20 dataset simulations using validation process for the Broken Pareto model with 2 breaks. The "true" value is $\theta_2 = 0.7$.

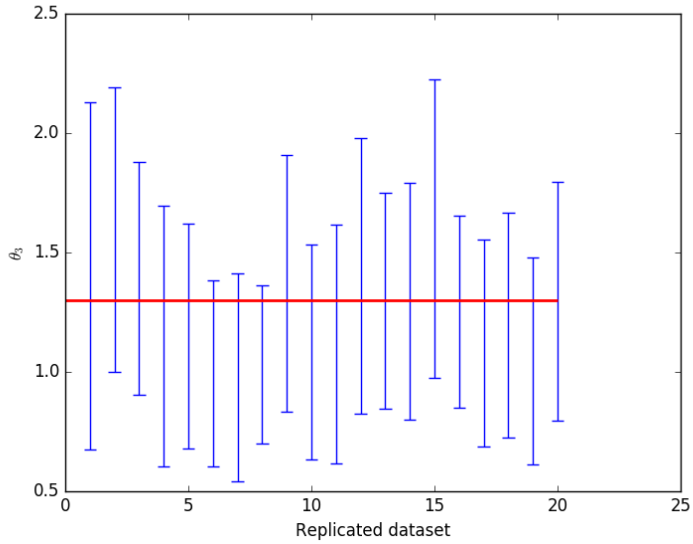



**Figure 3.14:** Posterior credible intervals of $\theta_3$ from 20 dataset simulations using validation process for the Broken Pareto model with 2 breaks. The "true" value is $\theta_3 = 1.3$.

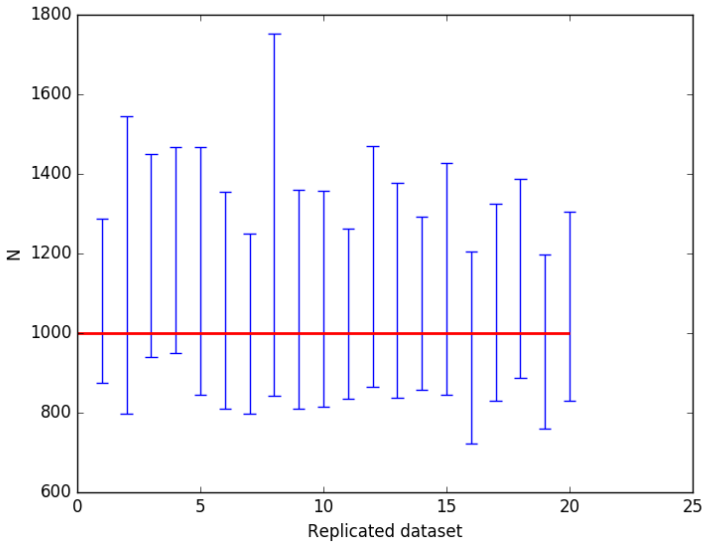



**Figure 3.15:** Posterior credible intervals of $N$ from 20 dataset simulations using validation process for the Broken Pareto model with 2 breaks. The "true" value is $N = 1000$.

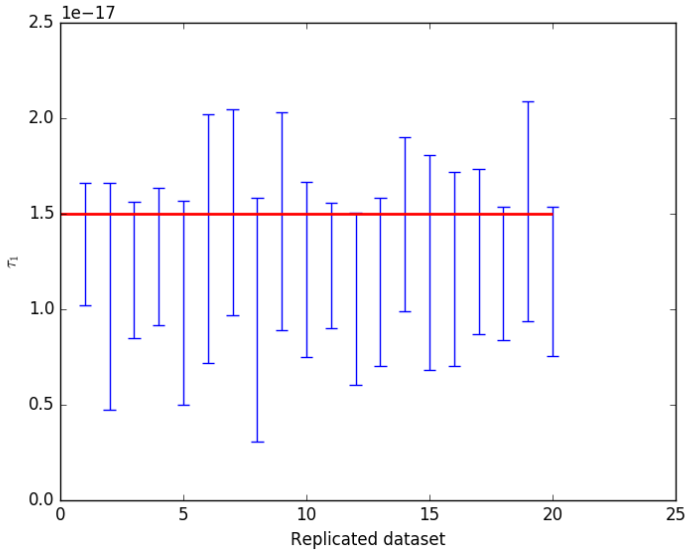



**Figure 3.16:** Posterior credible intervals of $\tau_1$ from 20 dataset simulations using validation process for the Broken Pareto model with 2 breaks. The "true" value is $\tau_1 = 1.5 \times 10^{-17}$. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

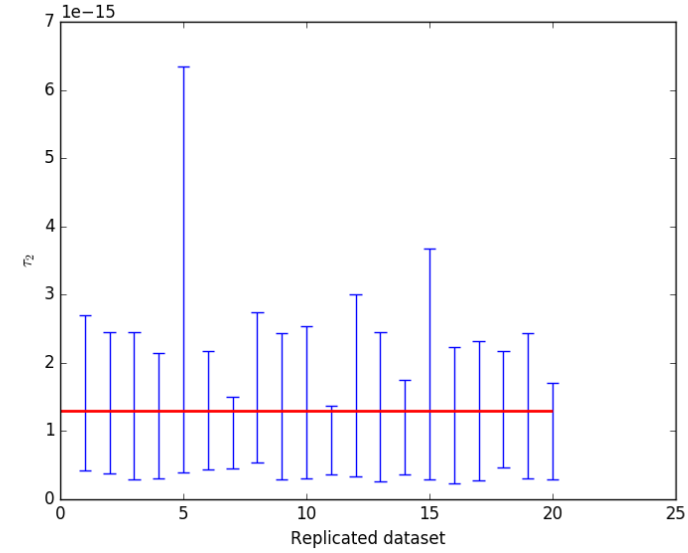



**Figure 3.17:** Posterior credible intervals of $\tau_2$ from 20 dataset simulations using validation process for the Broken Pareto model with 2 breaks. The "true" value is $\tau_2 = 1.3 \times 10^{-15}$. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

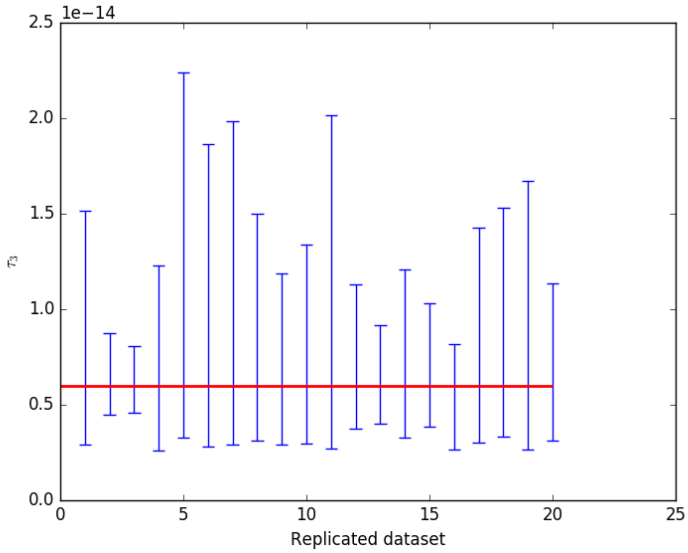



**Figure 3.18:** Posterior credible intervals of $\tau_3$ from 20 dataset simulations using validation process for the Broken Pareto model with 2 breaks. The "true" value is $\tau_3 = 6 \times 10^{-15}$. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

## 3.5 Application: CHANDRA Deep Field South

We apply the hierarchical Bayesian model that incorporates the uncertainty about $\gamma$ to the CHANDRA Deep Field South (CDFS) survey which was introduced in the first chapter. As a reminder, in our analysis we consider a sample of 358 observed sources, from which we exclude 3 sources for which we do not have spectral data

### 3.5.1 Single Power Law model

For the Single Power Law model with $\gamma$ uncertainty, we assumed the following priors for the parameters $(N, \theta, \tau)$:

$$\theta \sim \Gamma(a = 2, b = 1)$$
$$\tau \sim \Gamma(a_m = 1.38, b_m = 3.46 \times 10^{-16})$$
$$N \sim \text{Negative-Binomial}(a_N = 8.05, b_n = 0.014)$$

We ran the Blocked Gibbs sampler, as described in the previous sections, using the aforementioned priors for 60,000 iterations. We simulated 3 independent chains with different starting values, which all of them gave us very similar posterior estimates. More specifically, the first chain provided us with the posterior estimates for the parameters of interest $(N, \theta, \tau)$ that are depicted in the Table 3.2 using the last 30,000 iterations (we discarded the first 30,000 iterations as burn-in).

**Table 3.2:** The posterior estimates for the major parameters for the CDFS dataset using the last 30,000 iterations for the single Pareto model with $\gamma$ uncertainty for 1 of the 3 chains we ran.

| | Mean | Median | SD | 2.5% | 97.5% | Mode |
|---|---|---|---|---|---|---|
| $\theta$ | 0.713 | 0.711 | 0.05 | 0.626 | 0.805 | 0.682 |
| $N$ | 2240 | 2176 | 348 | 1759 | 3139 | 2019 |
| $\tau$ | $9.32\times10^{-18}$ | $9.77\times10^{-18}$ | $1.69\times10^{-18}$ | $4.90\times10^{-18}$ | $1.13\times10^{-17}$ | $1.12\times10^{-17}$ |

Figure 3.19 shows the trace plots for the parameters of interest $(N, \theta, \tau)$. The convergence is quite fast as we can deduce. Figure 3.20 depicts the posterior bivariate scatter plots and 1-dimensional histograms for the parameters of interest. If we compare the results with those from the single Pareto model with constant $\gamma$, we notice that the model with $\gamma$ uncertainty converges to a smaller value for the slope and as a result assumes a smaller number of sources for the complete population. Moreover, we do not observe the bi-modality that was evident in the marginal posterior distributions of the slope $\theta$ and the marginal posterior distribution of $\tau$ for the single Pareto model with constant $\gamma$.

The posterior draws of the flux for the complete source population gives rise to the posterior distribution plot of the $\log(N) - \log(S)$ curve shown in Figure 3.21. Each curve in the plot corresponds to a set of fluxes for the complete source population sampled from a single iteration of Blocked Gibbs sampler scheme with observed sources shown in grey and missing sources in red. Current plot exhibits sample of 100 flux sets. The blue line is the estimated $\log(N) - \log(S)$ curve using the poste-

rior medians of $\theta, N, \tau$. The depicted curve does not appear to be linear.

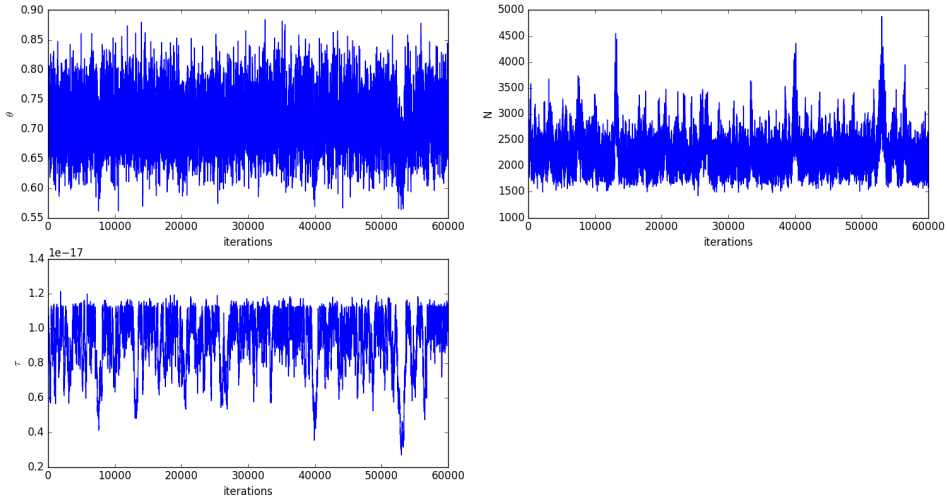



**Figure 3.19:** Trace plots of the main parameters of interest $\theta, N, \tau$ of the CDFS dataset for the Single Pareto model with $\gamma$ uncertainty. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

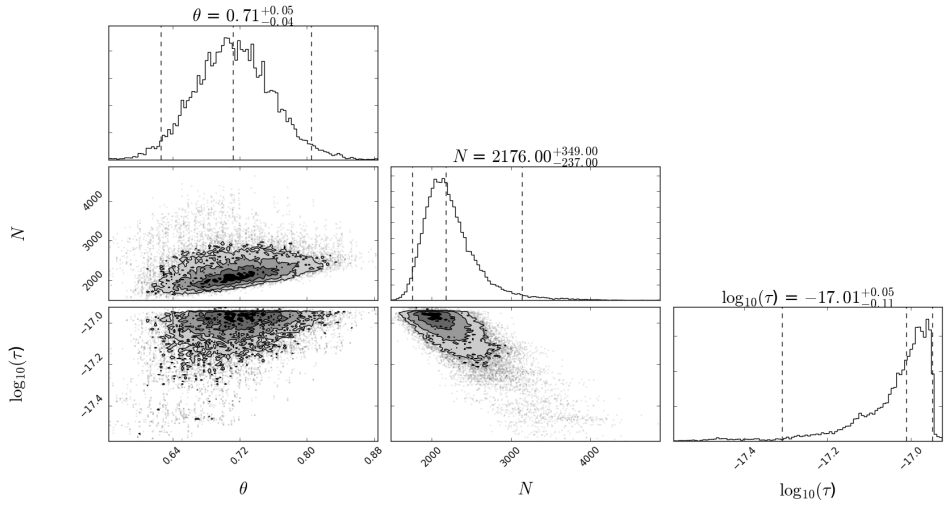



**Figure 3.20:** Bivariate scatter plots and 1-dimensional histograms of the main parameters of interest $\theta, N, \tau$ of the CDFS dataset for the Single Pareto model with $\gamma$ uncertainty. The figures are plotted using the posterior draws from the Blocked Gibbs sampler after removing a burn -in sample of about 30,000 draws.

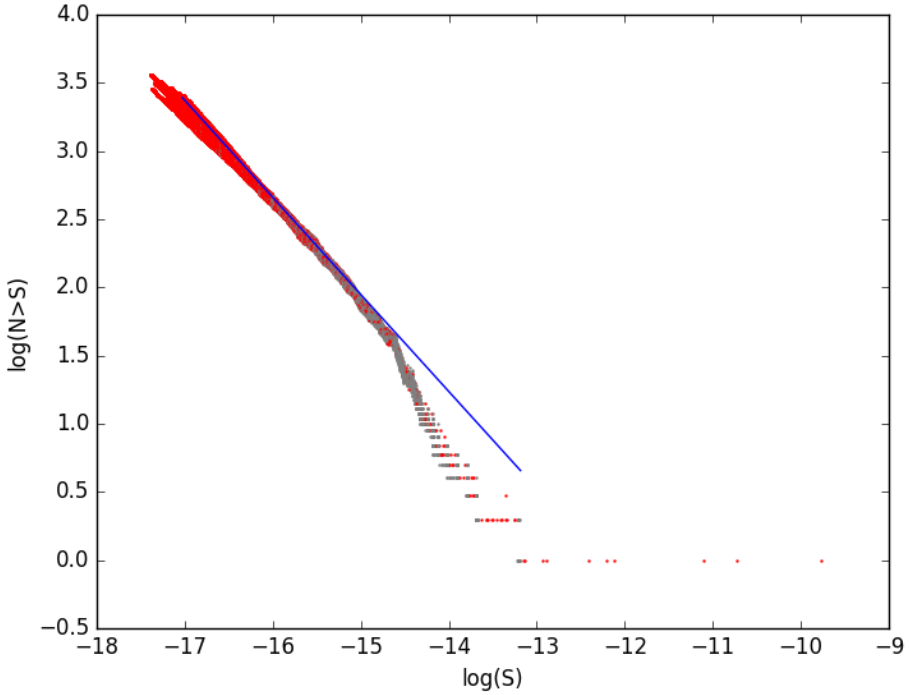



**Figure 3.21:** The posterior distribution of the $\log(N) - \log(S)$ plot of the CDFS dataset for the Single Pareto model with $\gamma$ uncertainty. Each line in the plot corresponds to a set of fluxes for the complete source population sampled from a single iteration of the Blocked Gibbs sampler scheme with observed sources shown in grey and missing sources in red. Current plot exhibits sample of 100 flux sets. The depicted curve does not appear to be linear.

### 3.5.2 Broken Power Law model with 1 break

For the Broken Power Law model with 1 break and with $\gamma$ uncertainty, the following priors were assumed for the parameters $(N, \theta_1, \theta_2, \tau_1, \tau_2)$:

$$\begin{aligned} \theta_1 &\sim \Gamma(a = 2, b = 1) \\ \theta_2 &\sim \Gamma(a = 2, b = 1) \\ \tau_1 &\sim \Gamma(a_m = 1.38, b_m = 3.46 \times 10^{-16}) \\ \eta_2 = \log(\tau_2 - \tau_1) &\sim N(\mu = -35, \sigma^2 = 1) \\ N &\sim \text{Negative-Binomial}(a_N = 8.05, b_n = 0.014) \end{aligned}$$

We ran the Blocked Gibbs sampler, as described in the previous sections, using the aforementioned priors for 60,000 iterations. We simulated 3 independent chains with

different starting values. More specifically, the first chain provided us with the posterior estimates for the parameters of interest $(N, \theta_1, \theta_2, \tau_1, \tau_2)$ that are depicted in the Table 3.3 using the last 30,000 iterations (we discarded the first 30,000 iterations as burn-in).

By comparing the posterior statistics with those from the broken power law model with 1 break and with constant $\gamma$, we mainly see some differences in the marginal posterior distribution of $\theta_1$ which has a smaller posterior mean and the slightly higher posterior mode of $\tau_2$.

**Table 3.3:** The posterior estimates for the major parameters for the CDFS dataset using the last 30,000 iterations for the broken power law model with 1 break and with $\gamma$ uncertainty for 1 of the 3 chains we ran.

| | Mean | Median | SD | 2.5% | 97.5% | Mode |
|---|---|---|---|---|---|---|
| $\theta_1$ | 0.669 | 0.669 | 0.05 | 0.574 | 0.762 | 0.714 |
| $\theta_2$ | 1.24 | 1.17 | 0.36 | 0.742 | 2.11 | 1.04 |
| $N$ | 2140 | 2075 | 352 | 1649 | 3068 | 2016 |
| $\tau_1$ | $8.92\times10^{-18}$ | $9.29\times10^{-18}$ | $1.85\times10^{-18}$ | $4.13\times10^{-18}$ | $1.22\times10^{-17}$ | $9.13\times10^{-18}$ |
| $\tau_2$ | $1.73\times10^{-15}$ | $1.53\times10^{-15}$ | $1.57\times10^{-15}$ | $2.69\times10^{-16}$ | $4.55\times10^{-15}$ | $2.33\times10^{-15}$ |

Figure 3.22 shows the trace plots for the parameters of interest $(N, \theta_1, \theta_2, \tau_1, \tau_2)$. Figure 3.23 depicts the posterior bivariate scatter plots and 1-dimensional histograms for the parameters of interest. We observe that the marginal posterior distribution of $\tau_2$ is relatively flat except for a spike. Thus, if we are considering a point estimate for the parameters of interest, we suggest using the posterior mode for $\tau_2$ and the $\tau_1$, since it might be a more appropriate choice than the posterior median.

The posterior draws of the flux for the complete source population gives rise to the posterior distribution plot of the $\log(N) - \log(S)$ curve shown in Figure 3.24. Each curve in the plot corresponds to a set of fluxes for the complete source population sampled from a single iteration of Blocked Gibbs sampler scheme with observed sources shown in grey and missing sources in red. Current plot exhibits sample of 100 flux sets. The blue line is the estimated $\log(N) - \log(S)$ using the posterior modes of $\tau_1$ and $\tau_2$ and the posterior medians of $\theta_1, \theta_2$ and $N$. The resulting posterior $\log(N) - \log(S)$ curve in Figure 3.24 does not appear to be linear. Thus the broken power law model with 1-break seems like a better candidate than the no break model.

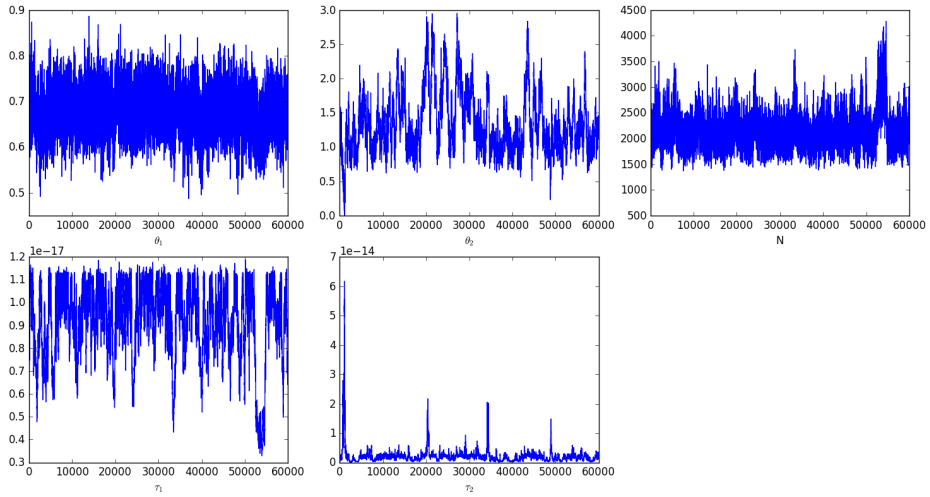



**Figure 3.22:** Trace plots of the main parameters of interest $(N, \theta_1, \theta_2, \tau_1, \tau_2)$ of the CDFS dataset for the broken power law model with 1 break and with $\gamma$ uncertainty. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

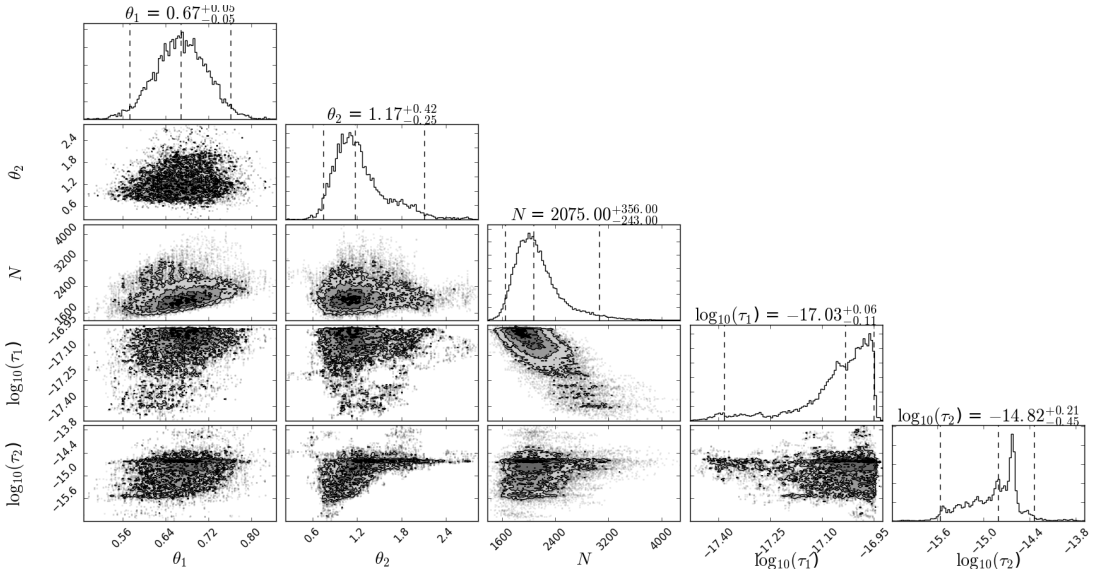



**Figure 3.23:** Bivariate scatter plots and 1-dimensional histograms of the main parameters of interest $(N, \theta_1, \theta_2, \tau_1, \tau_2)$ of the CDFS dataset for the broken power law model with 1 break and with $\gamma$ uncertainty. The figures are plotted using the posterior draws from the Blocked Gibbs sampler after removing a burn -in sample of about 30,000 draws.

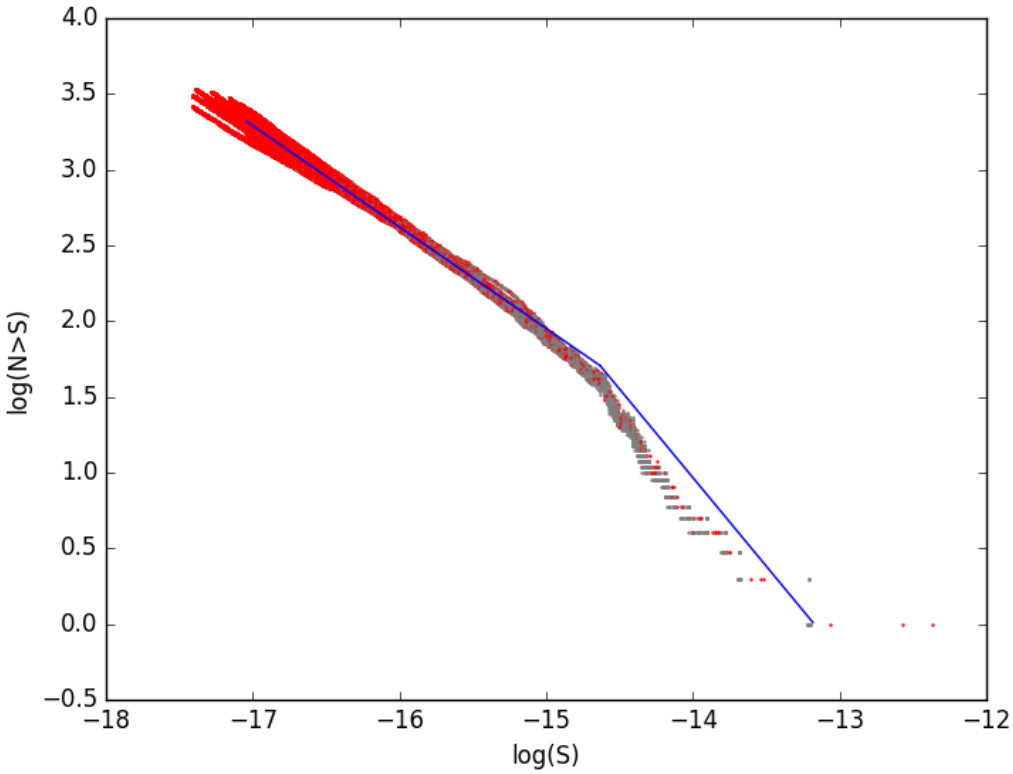



**Figure 3.24:** The posterior distribution of the $\log(N) - \log(S)$ plot of the CDFS dataset for the broken power law model with 1 break and with $\gamma$ uncertainty. Each line in the plot corresponds to a set of fluxes for the complete source population sampled from a single iteration of the Blocked Gibbs sampler scheme with observed sources shown in grey and missing sources in red. Current plot exhibits sample of 100 flux sets. The depicted curve does not appear to be linear. This indicated that a broken power Law model might be a better fit.

### 3.5.3 Broken Power Law model with 2 breaks

For the Broken Power Law model with 2 breaks, the following priors were assumed for the parameters $(N, \theta_1, \theta_2, \theta_3, \tau_1, \tau_2, \tau_3)$:

$$
\begin{aligned}
\theta_1 &\sim \Gamma(a = 10, b = 10) \\
\theta_2 &\sim \Gamma(a = 10, b = 10) \\
\theta_2 3 &\sim \Gamma(a = 10, b = 10) \\
\tau_1 &\sim \Gamma(a_m = 1.38, b_m = 3.46 \times 10^{-16}) \\
\eta_2 = \log(\tau_2 - \tau_1) &\sim N(\mu = -35, \sigma^2 = 4) \\
\eta_3 = \log(\tau_3 - \tau_2) &\sim N(\mu = -33, \sigma^2 = 4) \\
N &\sim \text{Negative-Binomial}(a_N = 8.05, b_n = 0.014)
\end{aligned}
$$

We ran the Blocked Gibbs sampler, as described in the previous sections, using the aforementioned priors for 60,000 iterations. As for the no-break and the 1-break model, we ran 3 independent chains with different starting values. More specifically, the first chain provided us with the posterior estimates for the parameters of interest $(N, \theta_1, \theta_2, \theta_3, \tau_1, \tau_2, \tau_3)$ that are depicted in the Table 3.4 using the last 30,000 iterations (we discarded the first 30,000 iterations as burn-in).

The posterior statistics are similar with those from the broken power law model with 2-breaks and with constant $\gamma$, except for $\theta_1$, $\tau_2$ and $\tau_3$. For $\theta_1$, we mainly see some differences in the marginal posterior distribution of $\theta_1$ which has a smaller posterior mean. However, the posterior mode of $\tau_3$ is identical with the posterior mode of $\tau_2$ for the broken power law model with 2-breaks and with constant $\gamma$ (the posterior mode of $\tau_2$ is smaller). This might be an indication that the model does not considers the existence of another break. This distance between posterior modes could be a useful heuristic for model selection.

**Table 3.4:** The posterior estimates for the major parameters for the CDFS dataset using the last 30,000 iterations for the broken power law model with 2-breaks and with $\gamma$ uncertainty for 1 of the 3 chains we ran.

| | Mean | Median | SD | 2.5% | 97.5% | Mode |
|---|---|---|---|---|---|---|
| $\theta_1$ | 0.661 | 0.661 | 0.05 | 0.561 | 0.757 | 0.650 |
| $\theta_2$ | 1.04 | 1.01 | 0.23 | 0.662 | 1.57 | 1.00 |
| $\theta_3$ | 1.22 | 1.21 | 0.34 | 0.588 | 1.91 | 1.18 |
| $N$ | 2003 | 1969 | 268 | 1569 | 2618 | 1914 |
| $\tau_1$ | $9.39\times10^{-18}$ | $9.75\times10^{-18}$ | $1.49\times10^{-18}$ | $5.96\times10^{-18}$ | $1.13\times10^{-17}$ | $1.08\times10^{-17}$ |
| $\tau_2$ | $1.01\times10^{-15}$ | $8.2\times10^{-16}$ | $5.87\times10^{-16}$ | $2.87\times10^{-16}$ | $2.41\times10^{-15}$ | $5.79\times10^{-16}$ |
| $\tau_3$ | $6.62\times10^{-15}$ | $5.47\times10^{-15}$ | $4.45\times10^{-15}$ | $2.06\times10^{-15}$ | $1.89\times10^{-14}$ | $2.45\times10^{-15}$ |

Figure 3.25 shows the trace plots for the parameters of interest $(N, \theta_1, \theta_2, \theta_3, \tau_1, \tau_2, \tau_3)$. The convergence is quite fast. Figure 3.26 depicts the posterior bivariate scatter plots and 1-dimensional histograms for the parameters of interest. From both the bivariate scatter plots and the histograms of the posteriors draws we can observe that the marginal posterior distributions of the $\tau_2$ and $\tau_3$ are multi-modals. Thus, we suggest using the posterior median as point estimate, since the estimation of the mode would be numerically unstable.

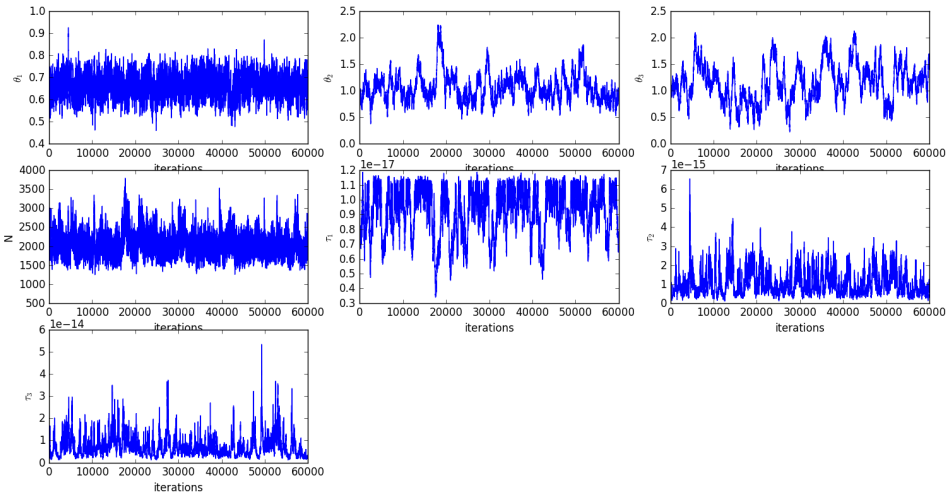



**Figure 3.25:** Trace plots of the main parameters of interest $(N, \theta_1, \theta_2, \theta_3, \tau_1, \tau_2, \tau_3)$ of the CDFS dataset for the broken power law model with 2-breaks and with $\gamma$ uncertainty. Note that the symbol "e" on the axis values denotes the scientific notation for 10 to the respective power.

The posterior draws of the flux for the complete source population gives rise to the posterior distribution plot of the $\log(N) - \log(S)$ curve shown in Figure 3.27. Each

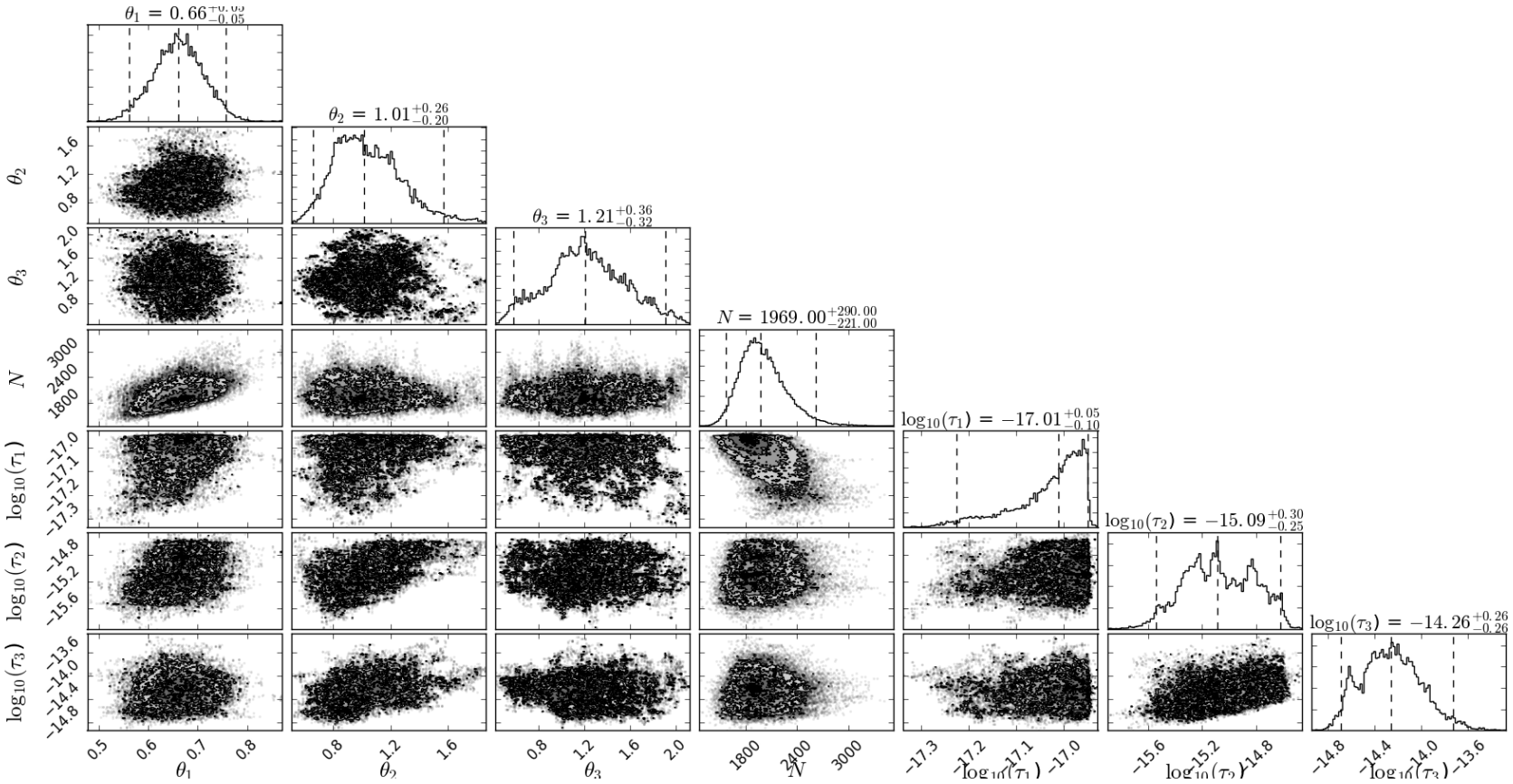



**Figure 3.26:** Bivariate scatter plots and 1-dimensional histograms of the main parameters of interest $(N, \theta_1, \theta_2, \theta_3, \tau_1, \tau_2, \tau_3)$ of the CDFS dataset for the broken power law model with 2-breaks and with $\gamma$ uncertainty. The figures are plotted using the posterior draws from the Blocked Gibbs sampler after removing a burn -in sample of about 30,000 draws.

curve in the plot corresponds to a set of fluxes for the complete source population sampled from a single iteration of Blocked Gibbs sampler scheme with observed sources shown in grey and missing sources in red. Current plot exhibits sample of 100 flux sets. The blue line is the estimated $\log(N) - \log(S)$ curve using the posterior medians of $N, \theta_1, \theta_2, \theta_3, \tau_1, \tau_2$, and $\tau_3$. The depicted curve does not appear to be linear. This indicated that a broken power law model might be a better fit.

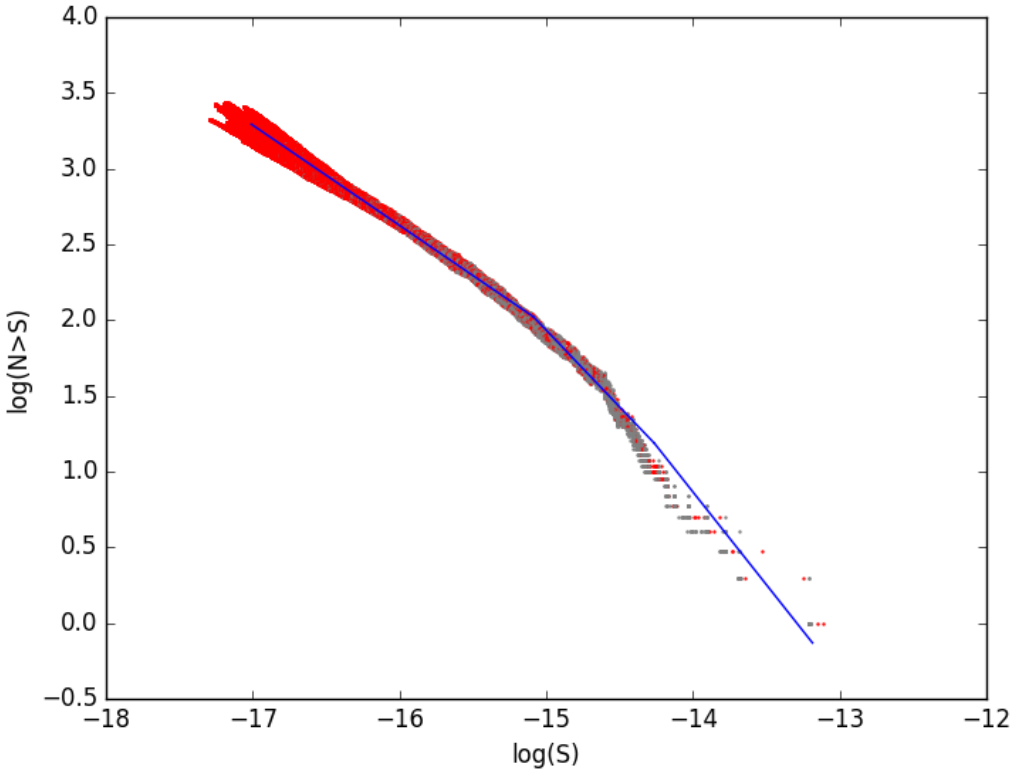



**Figure 3.27:** The posterior distribution of the $\log(N) - \log(S)$ plot of the CDFS dataset for the broken power law model with 2-breaks and with $\gamma$ uncertainty. Each line in the plot corresponds to a set of fluxes for the complete source population sampled from a single iteration of the Blocked Gibbs sampler scheme with observed sources shown in grey and missing sources in red. Current plot exhibits sample of 100 flux sets. The depicted curve does not appear to be linear. This indicated that a broken power law model might be a better fit.

### 3.5.4 Comparison with the Model without $\gamma$ uncertainty

In the previous subsections we obtained posterior estimates for the $\log(N) - \log(S)$ curve for different number of breaks, while accounting for the uncertainty in the flux-to-count conversion factor $\gamma$. Naturally, the first question that arises revolves

around whether incorporating this level of uncertainty affects significantly the posterior estimates we obtained in the previous chapter by using the simpler model that assumed a constant value for $\gamma$.

In the case of the model with no breaks, we can see that the simpler model suggests a steeper slope $\theta$ than the model with $\gamma$ uncertainty, although the 95% posterior intervals do overlap. The posterior estimates for the lower threshold $\tau$ are very close for both models, although the bi-modality that we observed in the simpler model for both the slope and the lower threshold are not present in the model with $\gamma$ uncertainty, possibly because the uncertainty in $\gamma$ creates a more diffused posterior distribution. For the $\log(N) - \log(S)$ with 1-break, the posterior estimates for both model are relatively close. What pops out is that the model with $\gamma$ uncertainty seems to exhibit less variance in the marginal posterior distributions of $\theta_2$ (the slope of the second line), and $\tau_2$ (the location of the break-point). We can hypothesise that this behaviour might be a result of the uncertainty in $\gamma$ that creates a more diffused posterior distribution. In the case of the $\log(N) - \log(S)$ with 2-breaks, the posterior estimates of the parameters of interest are very close.

Thus, although we believe that further research is required in terms of applying the two models to more surveys, including the uncertainty about the flux-to-count conversion factor doesn't seem to dramatically affect the posterior estimates, except for the case of the model with no breaks.

## 3.6 Discussion

In this Chapter we extended the hierarchical Bayesian model for estimating the $\log(N) - \log(S)$ relationship, by properly incorporating the uncertainty about the flux-to-count conversion factor $\gamma$. This constitutes a very innovative approach on the $\log(N) - \log(S)$ estimation, since the methods in the relevant $\log(N) - \log(S)$ estimation assume that $\gamma$ is constant for all the sources. However, the value of $\gamma$ depends on the spectral model that is assumed for the source as well as the energy band of the source. Thus, it should be properly accounted for when estimating the

parameters of the $\log(N) - \log(S)$ curve.

In order to account for this uncertainty, we have to extract the uncertainty about $\gamma$, expressed as a different probability distribution for each observed source, using modern astronomical software. Given those distributions and assuming that the individual characteristics of the spectrum of each source that affect the distribution of $\gamma$ are independent of the missing data mechanism, we fit a hierarchical prior for the complete source population (hierarchical because it is specified in terms of parameters that are themselves fit to the data).

In order to fit this prior, we use an innovative statistical methodology that was initially developed to tackle another astrophysical problem (McKeough et al. 2016). Fitting a hierarchical prior based on posterior distributions is undoubtedly a very useful concept with many applications. Hence, the flexibility of the method we developed -using an MCMC sampler to account for the difference between the prior we are trying to fit and the prior used in order to get the existing posteriors- is well suited to the task.

The resulting methodology about estimating $\log(N) - \log(S)$ offers to the astronomical community a very powerful and at the same time versatile tool. It can be applied to different astronomical surveys with little effort; it only requires the relevant incompleteness function, the background and exposure maps, and the uncertainty in the estimation of $\gamma$ extracted using relevant astronomical software. Furthermore, the hierarchical structure allows for easy extensions of the model for any other source of uncertainty.

# 4

# Classifying Galaxies Using a Data Driven Approach - Connection to the $\log(N) - \log(S)$

In the previous Chapters, we presented a method for estimating the density of the flux for a given astronomical population. We also emphasised that the density function will not be the same for different astronomical populations. This assumes that we have a method to classify a priori astronomical populations to different classes. However, accurately classifying astronomical sources is a particularly difficult problem with many complexities.

One long and heavily researched classification problem in Astronomy is the assignment of galaxies to different classes based on their activity levels. Despite the extensive literature on this problem, the existing classification schemes are mostly purely theoretical in nature and do not offer a robust classification methodology. This Chapter discusses a novel data-driven approach that aims to classify galaxies to different activity classes. This classification can be used in order to produce different $\log(N) - \log(S)$ curves for the different galaxy populations.

This chapter is organised as follows. In section 1 we describe the scientific problem and the mathematical background. Section 2 discusses the implementation of the

method on galaxy spectra and Section 3 compares our multidimensional data driven classification scheme with an existing scheme. Section 4 introduces multidimensional linear decision boundaries that we compare in terms of their prediction accuracy with both our new method and the existing scheme. In Section 5 we review our results and discuss further research directions.

## 4.1 Introduction

### 4.1.1 The Scientific Problem

Spectroscopy is the study of the measurement of radiation intensity as a function of wavelength. When electrons orbiting atoms fall from energised orbits to rest orbits, they emit light photons with wavelengths resulting from the excess energy given up by the electrons. The atoms and molecules have unique spectra. Consequently, these spectra can be used to detect, identify and quantify information about the atoms and molecules.

Spectroscopy has been used extensively in Astronomy, since Astronomers can use the spectrum of a luminous body to help determine its composition (elements) and temperature (degree of excitement). Spectroscopy has also been utilised in identifying the main power source in active galaxies. The energy output of galaxies is dominated by two main processes: star-formation and/or accretion onto a supermassive central black-hole, the latter witnessed as an Active Galactic Nucleus (AGN). The result of those two processes is to heat and excite their surrounding gas, making it to glow in specific wavelengths corresponding to emission from specific atoms or ions. The intensity of this emission is a tell-tale signature of the conditions in the gas, but most importantly of the energy source that is heating the gas.

The interplay between those two processes -star formation and/or accretion onto a black hole- is key for understanding the demographics of galactic activity and the co-evolution of nuclear black-holes and their host galaxies (e.g. Kormendy & Ho 2013). The main tool we have for characterising the type of activity in galaxies is its imprint on the emerging spectrum of the photo-ionised interstellar medium (ISM). AGN generally produce harder ionising continua which result in stronger high-excitation lines that we can obtain from reprocessing of the spectrum of young stellar populations (e.g. Ferland 2003).

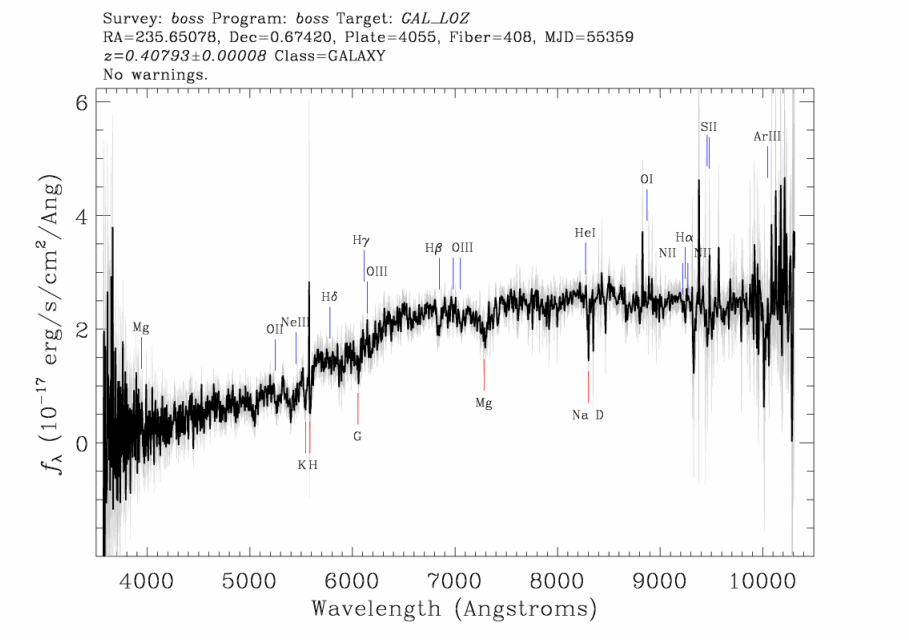



**Figure 4.1:** A selected spectrum from the DR10 BOSS data, showing absorption (red) and emission (blue) lines (http://www.sdss.org).

The importance of characterising the ionising source of emission-line regions was recognised early on and led to the first systematic presentation of optical emission-line diagnostic tools by Baldwin et al. (1981). This work introduced three diagrams based on four emission-line intensity ratios: $\log([N_{II}]/H\alpha)$, $\log([S_{II}]/H\alpha)$, $\log(O_I/H\alpha)$ and $\log(O_{III}/H\beta)$. These diagrams, known as Baldwin-Phillips-Terlevich (BPT) diagrams, were able to discriminate between star-forming galaxies (SFGs) and galaxies dominated by AGN activity. At the same time, a third class of galaxies was recognized by Heckman (1980) on the basis of their relatively stronger lower-ionisation lines (Low-Ionisation Nuclear Emission line Regions; LINERs). The format of the BPT diagrams that are typically used today was refined by Veilleux & Osterbrock (1987), and they include all three classes of objects (SFGs, LINERS, AGN).

However, the exact demarcation between SFGs and AGNs is generally defined empirically and hence it is subject to considerable uncertainty. Based on stellar population synthesis and photoinization models Kewley et al. (2001) introduced a maximum 'starburst' line on the BPT diagrams which defines the upper bound for the SFGs. Driven by the fact that AGN and SFGs observed in the Sloan Digital Sky Survey

(SDSS; York et al. 2000) show two distinct loci extending below the demarcation line of Kewley et al. (2001), a new empirical upper bound for the SFGs was put forward by Kauffmann et al. (2003) in order to distinguish the pure SFGs. The objects between this new empirical SFG line and the demarcation line of Kewley et al. (2001) belong to the class of Composite galaxies (also referred to as Transition objects in previous studies; e.g. Ho et al. 1997). The spectra of these Composite galaxies have been traditionally interpreted as the result of significant contributions from both AGN and star-forming activity, although, more recently it has been proposed that their strong high-excitation lines could be the result of shocks (e.g. Rich et al. 2014).

Subsequently, Kewley et al. (2006) introduced another empirical line for distinguishing Seyferts and LINERs. More recently, Shi et al. (2015) explored other emission-line intensity ratios that could improve the classification. They used support vector machines to test the classification accuracy using a dataset of galaxies classified as either SFG, AGN, or Composite based on Kauffmann et al. (2003).

Figure 4.2 depicts a sample from the SDSS DR8 survey on the BPT diagnostic diagrams; the red line is the maximum 'starburst' line. Galaxies below the red line are classified as belonging to the SFG class and galaxies above the line are classified as AGNs. By examining the first diagnostic diagram, we can identify two different loci of sources: a stream moving from the bottom middle to the top left and another fuzzier stream moving from the bottom middle to the top right. The first stream corresponds to the SFGs and the latter to the AGN. The other two diagnostics verify that the AGN are made up of two groups, the Seyferts and the LINERs. Nevertheless, it is apparent that the SFGs in the first 2 diagrams are considerably below the red line and the empirical blue line between the LINERs and the Seyferts seems slightly inconsistent.

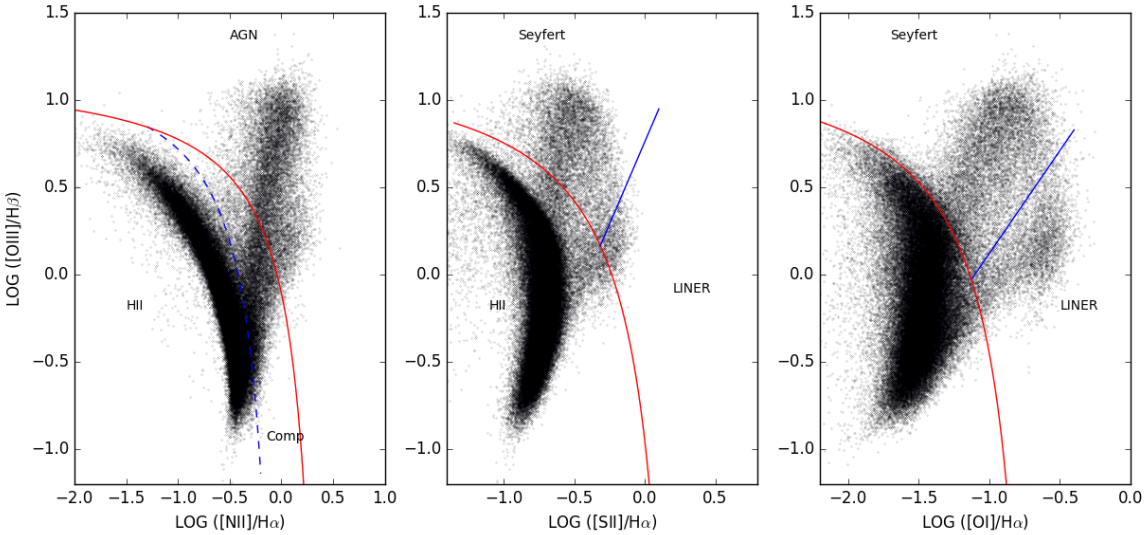



**Figure 4.2:** Example diagnostic diagrams (BPT) based on a sample from the SDSS DR8.

The currently used classification scheme suffers from a significant drawback. The use of multiple diagnostic diagrams independently of one another often gives contradicting classifications for the same galaxies (e.g. Ho et al. 1997). According to Kewley et al. (2006), 8% of the galaxies in their sample are characterised as ambiguous in that they were classified as belonging to different classes based on at least two diagnostic diagrams (for clarity we use the term contradicting to emphasise that the different 2-dimensional diagnostics can give different classifications). Such contractions arise because BPT diagrams are projections of a complex multi-dimensional space onto 2-dimensional planes. This limits the power of this diagnostic tool and may lead to inconsistencies between the different diagnostic diagrams.

Moreover, the number of extragalactic emission-line objects for which accurate spectra are available has grown rapidly in recent years, especially with the advent of the SDSS. This massive dataset reveals inconsistencies between the theoretical and empirical upper bounds and the actual distribution of the observed line ratios for the different classes (e.g. Kauffmann et al. 2003).

The inefficiency of the existing approach gives rise to the question whether we could use a more data-driven method for effectively classifying the galaxies. In this chapter we propose a classification scheme, the soft allocation data driven (SoDDA) method, which is based on the clustering of galaxy emission-line ratios in the 4-dimensional

space defined by the $[N_{II}]/H\alpha$, $[S_{II}]/H\alpha$, $O_I/H\alpha$ and $O_{III}/H\beta$ ratios. This is motivated by the clustering of the SFG, AGN and LINER loci on the 2-dimensional projections of the emission line diagnostic diagrams. Our classification scheme arises from a model that specifies the joint distribution of the emission line ratios of each galaxy class to be a finite mixture of multivariate Gaussian (MG) distributions. Given the emission line ratios of each galaxy, we compute the posterior probability of each galaxy belonging to each galaxy class. This allows us to achieve a soft clustering. A similar approach was successfully implemented by Mukherjee et al. (1998) in another clustering problem in which they used a mixture of MG distributions to discriminate between distinct classes of gamma-ray bursts.

### 4.1.2 Statistical Background

#### Expectation Maximisation Algorithm

The Expectation Maximisation algorithm (commonly referred to as EM algorithm) is an iterative optimisation algorithm used for estimating the maximum likelihood estimates (or posterior mode finding in a Bayesian context) in statistical models that involve missing or unobserved latent data. It was formalised in the very influential paper of Dempster et al. (1977), although the authors discuss that it was used previously in various applications. Since then, the EM algorithm has been used with success in a great number of problems across many different scientific fields. The EM is unique in contrast to other optimisation algorithms, such as Newton -type algorithms or Gauss Seidel methods, since it is formulated in statistical terms.

More specifically, suppose that we have a statistical model with observed data $x$, a vector of unknown parameters $\theta$ and unobserved data $z$. Suppose that we are interested in the marginal likelihood $p(x|\theta)$, but this likelihood is hard to maximise directly using another optimisation algorithm. However, if we can work more easily with the conditional densities $p(z, x|\theta)$ and $p(\theta|z, x)$, then we can use the EM algorithm. More specifically, the EM alternates between performing an expectation (E) step and a maximisation (M) step as follows

*E-step*: Compute $Q(\theta|\theta^{(t)}) = E_z[log\, p(x, z|\theta)|x, \theta^{(t)}] = \int \log[p(z, x|\theta)]p(z|\theta^{(t)}, x)dz$,

*M-step*: Set $\theta^{(t+1)} = \text{argmax}_\theta \, Q(\theta|\theta^{(t)})$,

where the subscript $t$ indexes the iteration. In the E step we essentially replace the missing values by their expected value given the current state of the unknown parameters $\theta^{(t)}$, and in the M step we estimate the parameters assuming the missing data are equal to their estimated values.

The EM algorithm enjoys stable convergence properties; by iterating between those 2 steps, the likelihood $p(x|\theta)$ increases in each iteration and furthermore the algorithm converges to a stationary point of $p(x|\theta)$. There is no guarantee though that the EM converges to the MLE, thus it is advised to routinely re-run the algorithm with different starting values, especially for multimodal cases.

The power of the EM algorithm lies on the fact that many models, such as mixture models or hierarchical models, can be expressed as probability models on an augmented parameter space. In order to describe the concept of data augmentation, we define as augmented data, $x_{aug}$, the combination of the observed data $x$ and any latent variables or missing data, $x_{mis}$. A data augmentation scheme is a model that satisfies the constraint:

$$p(x|\theta) = \int_{x_{mis}} p(x_{aug}|\theta) \, dx_{mis} \tag{4.1}$$

The necessity of this requirement is obvious because $p(x_{aug}|\theta)$ is introduced for computational purposes and thus the marginal distribution of $x$ implied by $p(x_{aug}|\theta)$ must be the original model $p(x|\theta)$. The utility of data augmentation lies on the fact that a good choice of $p(\theta|x_{aug})$ and $p(x_{aug}|x, \theta)$ can divide the initial problem into two simpler conditional models. The concept of data augmentation is highly desirable on many occasions in which direct maximisation of the likelihood is challenging. The added parameters can be thought as missing data, and thus, we can apply the EM algorithm for maximum likelihood estimation. A classic example of data augmentation is the finite mixture models that we explore in this chapter.

### Finite Mixture Models

Cluster analysis is a statistical method that aims to partition a dataset into subgroups so that the members within each subgroup are more homogeneous (according to some criterion) than the population as a whole. In this Chapter, we use a class of probabilistic (model-based) algorithms that assumes that the data are an identically and independently distributed (i.i.d.) sample from a population described by a density function, which is taken to be a mixture of component density functions. Formally a mixture model corresponds to the mixture distribution that represents the probability distribution of observations in the overall population. Finite mixture models have been studied extensively as a clustering technique (Wolfe 1970). It is common to assume that the mixture components are all from the same parametric family, such as the Gaussian. The use of mixture models arises naturally in our problem, since the population of galaxies is made up of several homogeneous subgroups: SFGs, Seyferts, LINERs and Composites.

Fraley & Raftery (2002) proposed a general framework to model a population as a mixture of $K$ subpopulations. Specifically, let $x_i$ be a vector of length $p$ containing measurements of object $i$ $(i = 1, ..., n)$ from a population. In our application the $x_i$ tabulates the $p = 4$ emission line ratios for galaxy $i$. A finite mixture model expresses the likelihood of $x_i$ as:

$$p(x_i|\theta, \pi) = \sum_{k=1}^{K} \pi_k f_k(x_i|\theta_k), \tag{4.2}$$

where $f_k$ and $\theta_k$ are the probability density and parameters for the distribution of subpopulation $k$, and $\pi_k$ is the relative size of subpopulation $k$, with $\pi_k \geq 0$ and $\sum_{i=1}^{K} \pi_k = 1$. Given a sample of n independent galaxies $x = (x_1, x_2, ..., x_n)$, the joint density can be expressed as:

$$p(x|\theta, \pi) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(x_i|\theta_k), \tag{4.3}$$

where $\theta = (\theta_1, ..., \theta_K)$ and $\pi = (\pi_1, ..., \pi_K)$.

### Estimating the parameters of a finite mixture model

Dempster, Laird & Rubin (1977) propose a framework that can be used to compute the maximum likelihood estimators (MLE) in finite mixture models using the Expectation-Maximization (EM) algorithm. Define the unknown parameters as $\phi = (\theta, \pi)$. The MLE is $\phi^\star = \text{argmax}_\phi p(x \mid \phi)$, where $\text{argmax}_\phi$ is an operator that extracts the value of $\phi$ that maximises the likelihood function, $p(x \mid \phi)$.

In the context of finite mixture models, Dempster et al. (1977) introduced an unobserved vector $z$ $(n \times K)$, where $z_{i\bullet}$ is the indicator vector of length $K$ with $z_{ik} = 1$ if object $i$ belongs to subpopulation $k$ and 0 otherwise. Because the $z_{i\bullet}$ are not observable, they are called latent variables. In this case they specify to which subpopulation each galaxy belongs. Given a statistical model consisting of observed data $x$, a set of unobserved latent data $z$, and a vector of unknown parameters $\phi = (\theta, \pi)$, the EM algorithm iteratively performs alternating expectation (E) and maximisation (M) steps:

E-step: Compute $Q(\phi|\phi^{(t)}) = \text{E}[\log\ p(x, z|\phi)|x, \phi^{(t)}]$,

M-step: Set $\phi^{(t+1)} = \text{argmax}_\phi\, Q(\phi|\phi^{(t)})$,

where the superscript $t$ indexes the iteration, and E[.] is the weighted mean evaluated by marginalising over all possible values of $z$.

The joint distribution $p(x, z|\theta, \pi)$ can be factorised as $p(x, z|\theta, \pi) = p(z|\theta, \pi) \cdot p(x|z, \theta, \pi)$, where $p(z|\theta, \pi)$ is a product of $n$ multinomial distribution $p(z|\theta, \pi) = \prod_{i=1}^{n} \prod_{k=1}^{K} \pi_k^{z_{ik}}$. Conditional on $z_{ik} = 1$, $p(x_i) = f_k(x_i|\theta_k)$. The logarithm of the conditional distribution of $x$ and $z$ given $(\theta, \pi)$, i.e. the log-likelihood, is:

$$\ell(\theta, \pi|x, z) = \log p(x, z \mid \theta, \pi) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log[\pi_k f_k(x_i|\theta_k)]. \tag{4.4}$$

The E-step requires us to compute the conditional expectation of Equation 4.4 given $(\theta^{(t)}, \pi^{(t)})$. Because Equation 4.4 is linear in the components of each $z_{i\bullet}$, it suffices

to compute the conditional expectation of the components of each $z_{i\bullet}$ given $x$ and $(\theta^{(t)}, \pi^{(t)})$. This is the conditional probabilities of $i$ belonging to subpopulation $k$ given $(\theta^{(t)}, \pi^{(t)})$. More specifically:

$$\mathrm{E}[z_{ik}|\theta^{(t)}, \pi^{(t)}, x] = \frac{\pi_k^{(t)} f_k(x_i|\theta_k^{(t)})}{\sum_{k=1}^{K} \pi_k^{(t)} f_k(x_i|\theta_k^{(t)})} = \gamma(z_{ik}) \tag{4.5}$$

The M-step requires us to maximise the conditional expectation of Equation 4.4 with respect to $\pi$ and $\theta$, i.e. to maximise $\sum_{i=1}^{n}\sum_{k=1}^{K}\gamma(z_{ik})\log[\pi_k f_k(x_i|\theta_k)]$. The particular form of the M-step depends on the choice of density distributions, $f_k$, for the subpopulations. Here we assume Multivariate Gaussian distributions for each subpopulation.

Multivariate Gaussian (MG) mixture models can be used for data with varying structures due to the flexibility in the definition of variance matrices. The density of the MG distribution for subpopulation $k$ is:

$$f_x(x_i) = \frac{1}{\sqrt{(2\pi)^p|\Sigma_k|}} \exp\Big( -\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k)\Big). \tag{4.6}$$

The EM formulation for an MG mixture is presented in detail in Dempster et al. (1977). The E-step has the same formulation as in Equation 4.5, with $f_k$ given in Equation 4.6 with $\theta_k = (\mu_k, \Sigma_k)$, where $\mu_k$ represent the means and $\Sigma_k$ the covariance matrices of the $x_i$ line ratios for galaxies in subpopulation $k$. For the M-step, the updates of the parameters have closed form solutions (Bilmes et al. 1998),

$$\pi_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\gamma(z_{ik}) \tag{4.7}$$

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} x_i\gamma(z_{ik})}{\sum_{i=1}^{n}\gamma(z_{ik})} \tag{4.8}$$

$$\Sigma_k^{(t+1)} = \frac{\sum_{i=1}^{n}\gamma(z_{ik})(x_i - \mu_k^{(t+1)})(x_i - \mu_k^{(t+1)})^T}{\sum_{i=1}^{n}\gamma(z_{ik})}. \tag{4.9}$$

### Bayesian Information Criterion

The Bayesian Information Criterion (BIC) (Schwarz et al. 1978) is a model selection criterion and is used in choosing among competing models. When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. The BIC resolves this problem by introducing a penalty term for the number of parameters in the model. The BIC is based on the maximum log-likelihood obtained with each candidate model, and penalised by the increased complexity associated with more parameters. More specifically, the BIC is computed as

$$\text{BIC}(K) = -2 \cdot \log L^{\star}(d) + d \log(n),$$

where $\log L^{\star}(d)$ is the maximised value of the likelihood when the number of free parameters to be estimated is $d$. We prefer the model with the smallest BIC.

The BIC is an asymptotic result derived under the assumptions that the data distribution is in the exponential family. A concise derivation of the BIC can be found in Bhat & Kumar (2010). A very interesting result of BIC is that it is asymptotically consistent (Friedman et al. 2001). In other words, if we are examining a family of models that include the true model, then BIC will select the correct model as the sample size $n \to \infty$.

### Gap Statistic

The gap statistic (Tibshirani et al. 2001) is a model selection criterion used for estimating the number of clusters in a set of data (find the value of $K$ in a finite mixture model approach). The gap statistic compares the normalised intra-cluster distances between points in a given cluster, $W_K$, for different total number of clusters $K$, with a null reference distribution obtained assuming data with no obvious clustering. The null reference distribution is generated by sampling uniformly from the original datasets bounding box multiple times. The estimate for the optimal number of clusters $K$ is the value for which the $W_K$ falls the farthest below the

reference curve.

### Support Vector Machines

A Support Vector Machines (SVM) (Cortes & Vapnik 1995) is a discriminative classifier formally defined by a separating hyperplane. An SVM model is a representation of the data as points in space, mapped so that the data of the separate categories are divided by a clear gap that is as wide as possible. New data are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. In other words, given classified galaxies, the algorithm outputs an optimal hyperplane which can be used to categorise new unlabelled galaxies. The SVM algorithm has been widely applied in many different scientific fields. In our research, it is very useful in order to introduce hard decision boundaries after classifying the galaxies in their respective classes.

Following the notation from Friedman et al. (2001), suppose we have $N$ pairs $(x_1, y_1)$, $(x_2, y_2), ..., (x_N, y_N)$, where $x_i \in \mathbb{R}$ and $y_i \in \{-1, 1\}$ represents the class (2 classes for simplicity, can be easily extended to the multi class case). The goal of SVMs is to find an optimal separating hyperplane between the two classes. Let us define a hyperplane by:

$$x: \quad f(x) = x^T\beta + \beta_0 = 0 \tag{4.10}$$

where $\beta$ is a unit vector. If the classes were fully separable, we could define the function $y_i f(x_i) > 0$ for every $i$ and we would try to find the hyperplane that creates the biggest margin $M$ between the classes. This optimisation problem can be written as:

$$\begin{aligned} & max_{\beta,\beta_0,||\beta||=1} \quad M \\ \text{s.t.} \quad & y_i(x_i^T\beta + \beta_0) \geq M, \quad i = 1, ..., N \end{aligned} \tag{4.11}$$

It can be shown that this optimization problem can be formulated without the norm

constraint

$$\begin{aligned} & max_{\beta,\beta_0} \quad ||\beta|| \\ \text{s.t.} \quad & y_i(x_i^T\beta + \beta_0) \geq 1, \quad i = 1, ..., N \end{aligned} \tag{4.12}$$

where $M = 1/||\beta||$. In more realistic applications, the classes are not fully separable. In order to maintain the concept of maximising the margin, we allow some points to be on the wrong side of the separating hyperplane by defining the slack variables $(\xi_1, \xi_2, ..., \xi_N)$ so that the constraint becomes $y_i(x_i^T\beta + \beta_0) \geq M(1 - \xi_i)$ for $\xi_i \geq 0$ and $\sum_{i=1}^{N} \xi_i \leq$ constant. By defining as above $M = 1/||\beta||$ we can rewrite the optimisation problem as:

$$\begin{aligned} & max_{\beta,\beta_0} \quad ||\beta|| \qquad (4.13) \\ \text{s.t.} \quad & y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \quad \forall i \\ & \xi_i \geq 0, \quad \sum_{i=1}^{N} \xi_i \leq \text{ constant} \qquad (4.14) \end{aligned}$$

This formulation shows a very important aspect of SVM; the points that are well inside their class boundary do not play a major role in defining the boundary. The most important points are the ones close to the boundary and whose on the wrong side of the separating hyperplane.This optimisation problem is quadratic with linear inequality constraints, so it is a convex optimisation problem. Further details can be found at Friedman et al. (2001).

A very interesting feature of SVM is the so called "kernel trick". Whereas the original problem may be stated in a finite dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason, it was proposed that the original finite-dimensional space be mapped into a much higher-dimensional space, presumably making the separation easier in that space. To keep the computational load reasonable, the mappings used by SVM schemes are designed to ensure that dot products may be computed easily in terms of the variables in the original space, by defining them in terms of a kernel function $k(x, y)$ selected to suit the problem. Therefore, instead of linear boundaries, SVMs can produce more com-

plex decision boundaries in the original dimensional space by enlarging the feature space using basis expansions such as polynomials or splines. Some commonly used kernel functions are the polynomial kernel (of $d$th degree) and the gaussian kernel.

## 4.2 The classification Scheme

As it was discussed in the previous section, a mixture model is a probabilistic model for representing the presence of subpopulations within an overall population, without requiring that an observed data set should identify the sub-population to which an individual observation belongs. The use of mixture models arises naturally in our problem, since the population of galaxies is made up of several homogeneous subgroups: SFGs, Seyferts, LINERs and Composites.

Fraley & Raftery (2002) point out that mixtures of Multivariate Gaussian (MG) distributions are appropriate if the subpopulations are centred at the means, $\mu_k$, with increased density for data closer to the means. As a result, the practical use of MG mixture models could be limited if the data exhibit non-Gaussian features, including asymmetry, multi-modality and/or heavy tails. In the SDSS DR8 dataset that we examine, it is apparent that the subpopulations exhibit non Gaussians characteristics such as convexity, skewness and multimodality. In order to account for these non-Gaussian features, we use a mixture of MG distributions with $K$ considerably larger than the actual number of galaxy classes. In this way, we represent each galaxy class by a mixture of several MG distributions. This allows a great deal of flexibility in the class-specific distributions of emission line ratios. With the fitted (large $K$) MG mixture in hand we can then perform hyper-clustering of the $K$ MG distributions so as to concatenate them into subpopulations representing the four desired galaxy classes. The number ($K >> 4$) of MG distributions that we fit to our data is chosen using the Bayesian Information Criterion (BIC) of Schwarz et al. (1978) and the gap statistic (Tibshirani et al. 2001).

Our Soft Data Driven Allocation (SoDDA) scheme accomplishes the hyper-clustering of the $K$ subpopulations into the four galaxy classes using the classification scheme of Kewley et al. (2006). More specifically, we treat the fitted subpopulations means $(\mu_1^\star, ..., \mu_K^\star)$ as a dataset and classify them into the four galaxy classes. For example, suppose we fit 10 MG distributions and the means of the distributions $1, 3$ and $5$ are

classified by Kewley et al. (2006) as SFGs, then the distribution of the SFGs under SoDDA would be

$$f_{\text{SFG}}(x_i) = \frac{\pi_1^\star f_1(x|\theta_1^\star, \pi_1^\star) + \pi_3^\star f_3(x|\theta_3^\star, \pi_3^\star) + \pi_5^\star f_5(x|\theta_5^\star, \pi_5^\star)}{\pi_1^\star + \pi_3^\star + \pi_5^\star}. \tag{4.15}$$

Via the allocations of the means of the $K$ subpopulations into the four galaxy classes, we have defined the distribution of the emission line ratios for each galaxy class as a finite mixture of MG distributions. Specifically, let $f_{\text{SFG}}(x)$, $f_{\text{LINER}}(x)$,

$f_{\text{Seyfert}}(x)$, and $f_{\text{Comp}}(x)$ be the distributions under SoDDA of the emission line ratios of SFGs, LINERs, Seyferts and Composites galaxies respectively. Then, given the four emission line ratios $x_i$ of a galaxy i, the posterior probability of galaxy $i$ being of type $c$ is:

$$\rho_{ic} = P(\text{galaxy } i \text{ is of type } c) \tag{4.16}$$

$$= \frac{f_c(x)}{\sum_c f_c(x)}, \quad \text{for } c \text{ in } \{\text{SFG}, \text{LINER}, \text{Seyfert}, \text{Comp}\}. \tag{4.17}$$

### 4.2.1 Implementation

The SDSS provides an excellent resource of nuclear spectra of galaxies covering all different activity types (e.g. Kauffmann et al. 2003). For the definition of our multi-dimensional activity diagnostics we use the "galspec" database of spectral-line measurements from the Max-Plank Institute for Astronomy and Johns Hopkins University group. We used the version of the catalog made publicly available through the SDSS Data Release 8 (Aihara et al. 2011, Eisenstein et al. 2011), which contains 1,843,200 objects. The spectral-line measurements are based on single Gaussian fits to star-light subtracted spectra, and they are corrected for foreground Galactic absorption (Tremonti et al. 2004, Kauffmann et al. 2003, Brinchmann et al. 2004). Since the same catalog has been used for the definition of the two-dimensional and multi-dimensional diagnostics of Kauffmann et al. (2003) and Vogt et al. (2014) respectively, it is the best benchmark for testing the SoDDA. From this catalog we selected all objects which have line flux measurements for the 8 diagnostic lines we consider here, with a signal-to-noise ratio (SNR) greater than 6, which ensures the

use of reliable line flux measurements for our analysis. The final sample consists of 90,388 galaxies.

We implement the fitting of the mixture of $K$ MG distributions using the `scikit-learn` Python library * under the constraint that the covariance matrices are full rank, and the diagonal elements cannot be smaller than $10^{-3}$ to avoid overestimation, i.e. converging to a small number of data points. Because this algorithm can be sensitive to the choice of starting values, we routinely rerun it with 5 different randomly selected sets of starting values. We choose the value among the 5 converged points with the largest likelihood to be the MLE, denoted $(\pi^\star, \mu^\star, \Sigma^\star)$.

We apply the BIC and gap statistic for values of $K$ ranging from 5 to 50 in increments of 5. Figures 4.3 and 4.4 plot the BIC and gap statistics. BIC suggests an optimum value of around $K = 20$, while the gap statistic suggests a value of $K = 10$. Since we are ultimately concatenating the clusters, we err on the side of large $K$, with $K = 20$, so as to capture as much detail in the data as possible without overfitting.

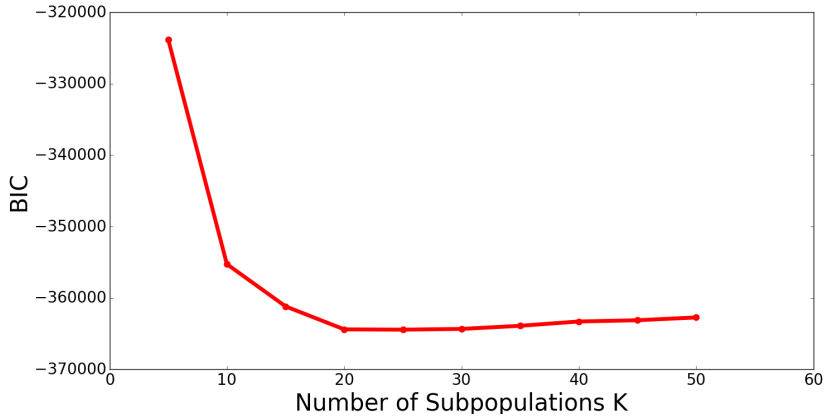



**Figure 4.3:** The Bayesian Information Criterion (BIC) computed over a grid of values of $K$ (in increments of 5) using the data of the SDSS DR8. The BIC is a model selection criterion based on the log-likelihood; the model with the lowest BIC value is preferred, indicating that in this case the optimal number of subpopulations is $K = 20$.

*http://scikit-learn.org/stable/

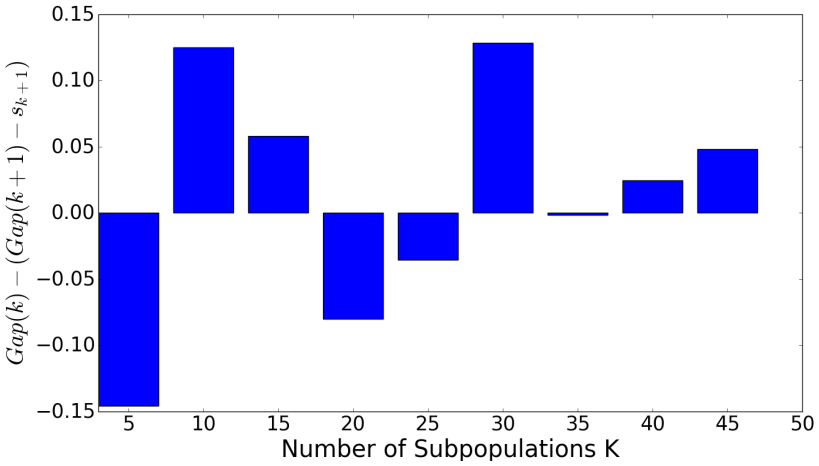



**Figure 4.4:** The Gap statistic computed over a grid of values of $K$ (in increments of 5) using the data of the SDSS DR8. The Gap statistic compares the intra-subpopulation distances between points in a given subpopulation with a null reference distribution of the data, i.e., a distribution with no obvious clustering. This figure shows that the smallest value of $K$ for which the data measure exceeds the randomly generated measure is $K = 10$.

Figure 4.5 displays the BPT diagnostic diagrams for SDSS DR8 with each point colour coded according to its most probable subpopulation among the $K = 20$ fit. The means of the subpopulations are plotted for $k = 1, .., 20$. To visualise the spacial extent of each of the 20 subpopulation, Figure 4.6 plots the $[N_{II}]/H\alpha$ vs $O_{III}/H\beta$ diagnostic diagram for each subpopulation. We emphasise that the full 4-dimensional geometry of the subpopulations cannot be seen in the 2-dimensional projections.

Subpopulation 4 is located in a different region in each of the three diagnostic diagrams in Figure 4.5. Furthermore, the 3-dimensional distribution of Subpopulation 4 is fuzzy, distorted, and totally disjoint from the distribution of the other subpopulations (Figure 4.7). Inspection of the optical spectra of several of the sources allocated to Subpopulation 4, shows broad emission lines with complex structure. Because these lines cannot be well modelled with the single Gaussians used, the resulting line measurements are unreliable. Therefore, we discard Subpopulation 4 from our analysis. In order to normalize the probabilities, we divide all the subpopulations weights by $(1 - \pi_4^\star)$.

SoDDA associates each of the 19 clusters with one activity class based on the projection of their mean on the 2-dimensional BPT diagnostic diagrams, and their location with respect to the activity class separating lines reported in Kewley et al.

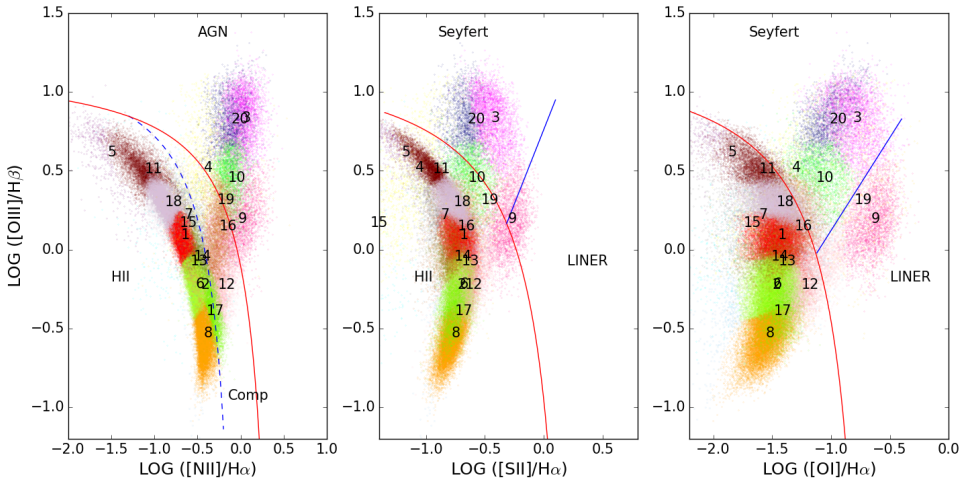



**Figure 4.5:** The BPT diagnostic diagrams for the SDSS DR8 sample; each datapoint is coloured according to its most probable allocation to one of the 20 multivariate Gaussian Distributions. The maximum 'starburst' line of Kewley et al. (2001) is shown as a solid red line and the empirical upper bound on SFG of Kauffmann et al. (2003) is plotted as dashed blue line. The empirical line for distinguishing Seyferts and LINERs of Kewley et al. (2006) is depicted by the solid blue line.

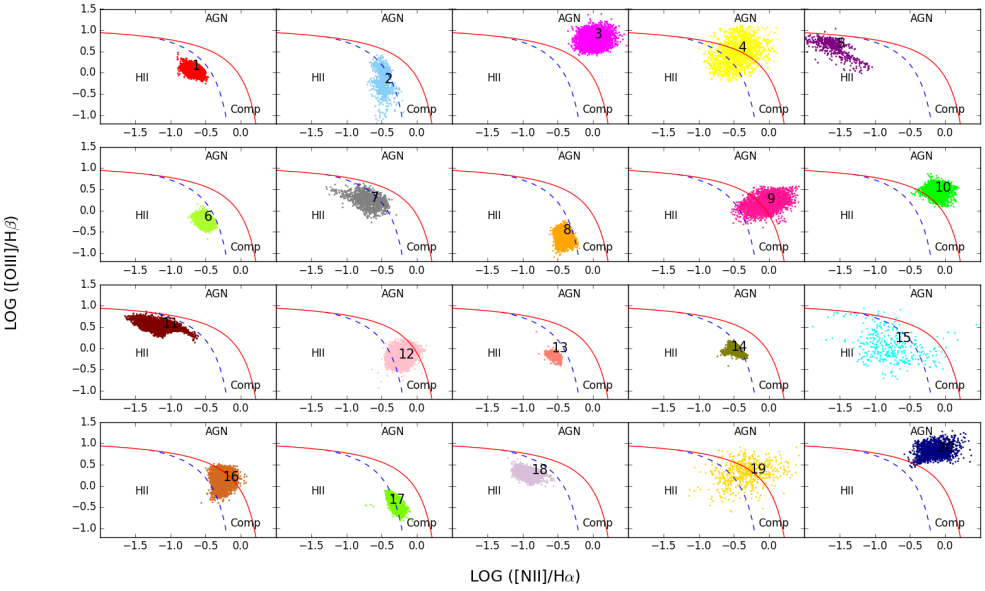



**Figure 4.6:** The 20 subpopulations plotted on the $[N_{II}]/H\alpha$ vs $O_{III}/H\beta$ projection of the 4-dimensional diagnostic diagram. The subpopulations are numbered following the scheme in Figure 4.5. This figure shows the spatial extent of each subpopulation and their location with respect to the standard diagnostic lines in the $O_{III}/H\beta$ diagram. Since these are 2-dimensional projection of the 4-dimensional distribution in each subpopulation, they only give an indication of the extent and location of each subpopulation.

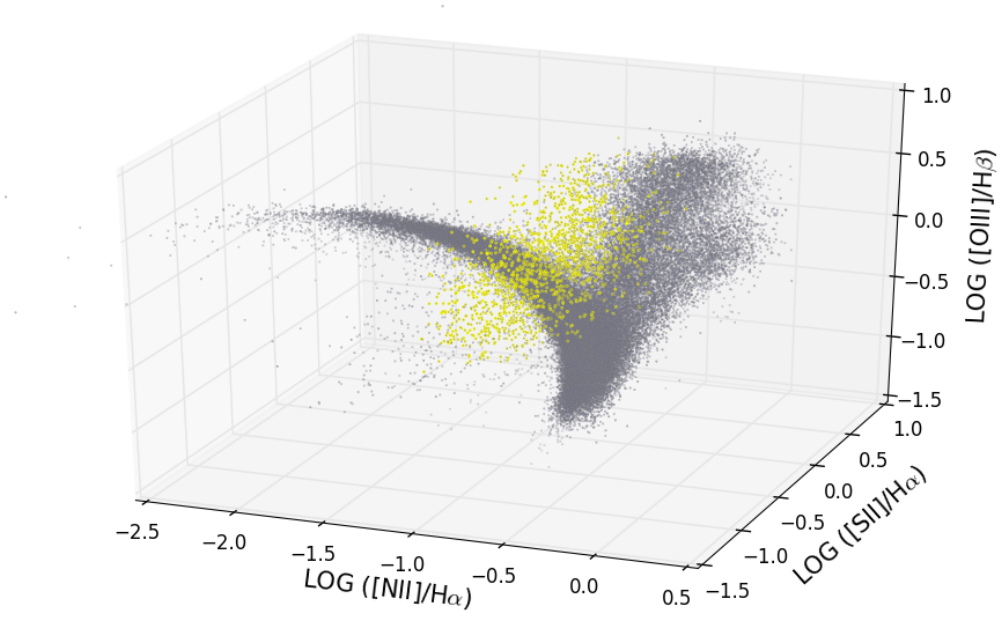



**Figure 4.7:** A 3-dimensional projection of the SDSS DR8 sample on the ([N$_{II}$]/H$\alpha$, [S$_{II}$]/H$\alpha$, [O$_{III}$]/H$\beta$) volume, showing the locus of the data points allocated to subpopulation 4 (yellow) in comparison to all other subpopulations (grey). The distribution of Subpopulation 4 is clearly distinct from the that of the other data. The 3 dimensional structure of the joint distribution of the 19 other subpopulations (in grey) shows complexities that are lost in its 2-dimensional projections.

(2006). The allocations are given in Table 4.1 for the 19 subpopulations means. Subpopulation 10 transcends the Composite and Seyfert classes. The main discriminator between Composite galaxies and Seyferts is the $[N_{II}]/H\alpha$ diagnostic and the mean of subpopulation 10 is clearly above the maximum 'starburst' line on the BPT diagrams introduced by Kewley et al. (2001) as an upper bound of SFGs. Thus, we include subpopulation 10 in the Seyfert class. After combining the 19 subpopulations to form the 4 galaxy classes as described in Table 4.1, we compute the posterior probability of each galaxy being a SFG, Seyfert, LINER, or Composite using Equation 4.17. The second row in Figure 4.8 shows the BPT diagnostic diagrams for SDSS DR8 with each galaxy colour coded according to its most probable galaxy class (red for SFGs, yellow for Seyferts, blue for LINERs, and green for the Composites) under SoDDA. To highlight the spatial extent of each cluster, we plot the BPT diagrams for each activity class (SFGs, Seyferts, LINERs and Composites) individually in Figure 4.9. Figure 4.10 depicts a 3-dimensional projection of the SDSS DR8 sample on the ($[N_{II}]/H\alpha$, $[S_{II}]/H\alpha$, $[O_{III}]/H\beta$) volume.

SoDDA provides a robust classification for the vast majority of the galaxies in the SDSS DR8 sample. For 90.6% of the galaxies, $\max_c \rho_{ic}$ is greater than 75%. That is, the most probably class for each of 90.6% of the galaxies has a posterior probability greater than 75%, indicating strong confidence in the adopted classification. Furthermore, the difference between the largest and the second largest $\rho_{ic}$ (among the classes) is smaller than 1% for only 0.17% of the galaxies, which indicates that the classification is uncertain for very few galaxies. This is illustrated in Figure 4.11 which plots $\max_c \rho_{ic}$, against the difference between $\max_c \rho_{ic}$ and the second largest $\rho_{ic}$ among the classes. The red line denotes a difference between the two highest values of $\rho_{ic}$ (among the classes) of 1%. There are only a few galaxies with a most probably class that is less that 1% (or even 10%) more probably than the second most probably class.

In order to assess the stability of the classification we randomly select a bootstrap sample consisting of 90% of the SDSS DR8 data (sampled without replacement and excluding Subpopulation 4). Using the bootstrap sample, we retune the classifier by estimating the means, weights, and covariance matrices for the 19 subpopulations, assigning each to one of the 4 activity classes, and recalculating the probability that each galaxy (in the SDSS DR8 sample we used for our original analysis excluding Subpopulation 4) belongs to each of the 4 classes. We denote these probabilities,

**Table 4.1:** The suggested classification of the 19 subpopulations means.

| Class | Subpopulation ID |
|---|---|
| SFG | 1,2,5,6,7,8,11,13,14,15,17,18 |
| Seyferts | 3,10,20 |
| LINER | 9 |
| Composites | 12,16,19 |

$\rho_{ic}^{\rm boot}$, to distinguish them from those computed with the full SDSS DR8 sample, namely $\rho_{ic}$. There is excellent agreement between the original classification and that obtained using the bootstrap sample. Specifically, 99.2% of the galaxies are classified into the same activity type with both classifiers. Similarly, 95.7% of the galaxies classified as Composites using the original classifier are classified in the same way using the set of parameters obtained from using the bootstrap sample. The figures are 96.3% for Seyferts, 97.9% for LINERs, and 99.9% for SFGs.

Overall there is little difference between the class probabilities of the individual galaxies computed with the full data and with the bootstrap sample. To illustrate this, we plot $\max_c \rho_{ic} - \max_c \rho_{ic}^{\rm boot}$ against $\max_c \rho_{ic}$ in Figure 4.12. Galaxies that are classified differently by the two classifiers are plotted in red. Again, there is excellent agreement: Not only is the classification of the vast majority of galaxies the same for both classifiers, but the probabilities of belonging to the chosen class are both similar and high. Of the few galaxies (0.8%) that are classified differently, 86% have $\max_c \rho_{ic} < 60\%$, meaning their classification was not clear to begin with. Overall, our classifier appears robust to the choice of sample used for tuning.

## 4.3 Comparing with Existing Classification Scheme

In order to show the advantages of our approach, we compare our method with the scheme proposed by Kewley et al. (2006). In contrast to the standard approach of using hard thresholds to define the different classes, SoDDA uses soft clustering rather than hard thresholds. We thus calculate the posterior probability of each galaxy belonging to each activity class. Moreover, SoDDA is not based on any particular set of 2 dimensional projections of the distributions of emission line ratios, but rather takes into account the joint distribution of all 4 emission-line ratios. Thus, the main difference between the two schemes is that SoDDA does not produce contradictory

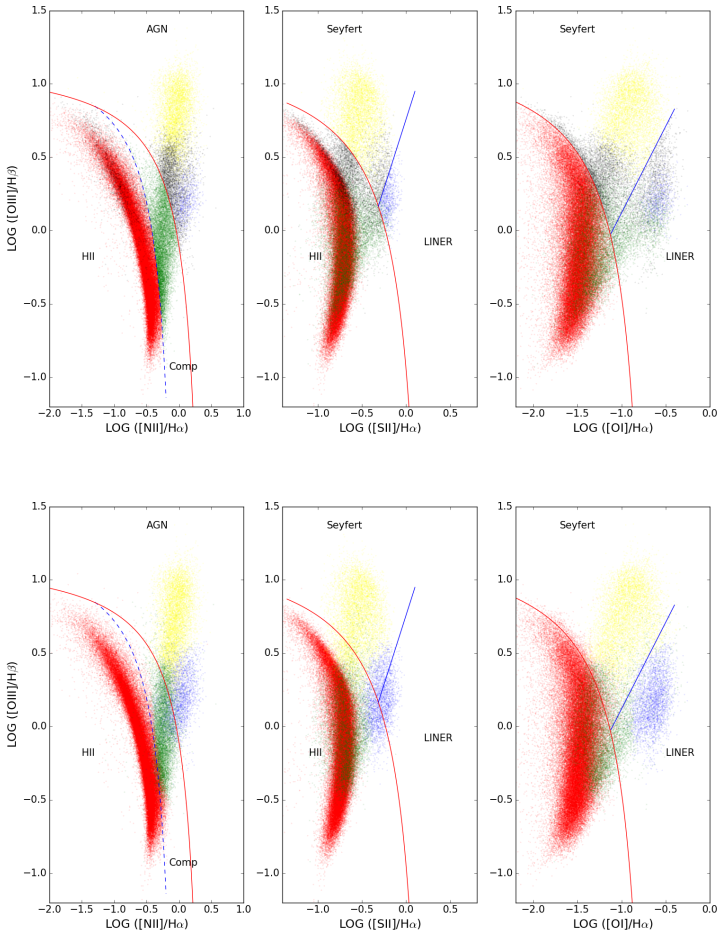



**Figure 4.8:** The BPT diagrams for the galaxies in the SDSS DR8 sample, based on the Kewley et al. (2006) scheme (top) and SoDDA (bottom). Each galaxy is colour coded according to its classification: red for SFGs, yellow for Seyferts, blue for LINERs, green for the Composite galaxies, and black for the Contradicting classifications. Note the lack of any contradicting classifications (black points) in the SoDDA results (bottom). For reference we also plot the the maximum 'starburst' line of Kewley et al. (2001) (solid red), the empirical upper bound on SFG of Kauffmann et al. (2003) (dashed blue), and the empirical line distinguishing Seyferts and LINERs (Kewley et al. 2006; solid blue).

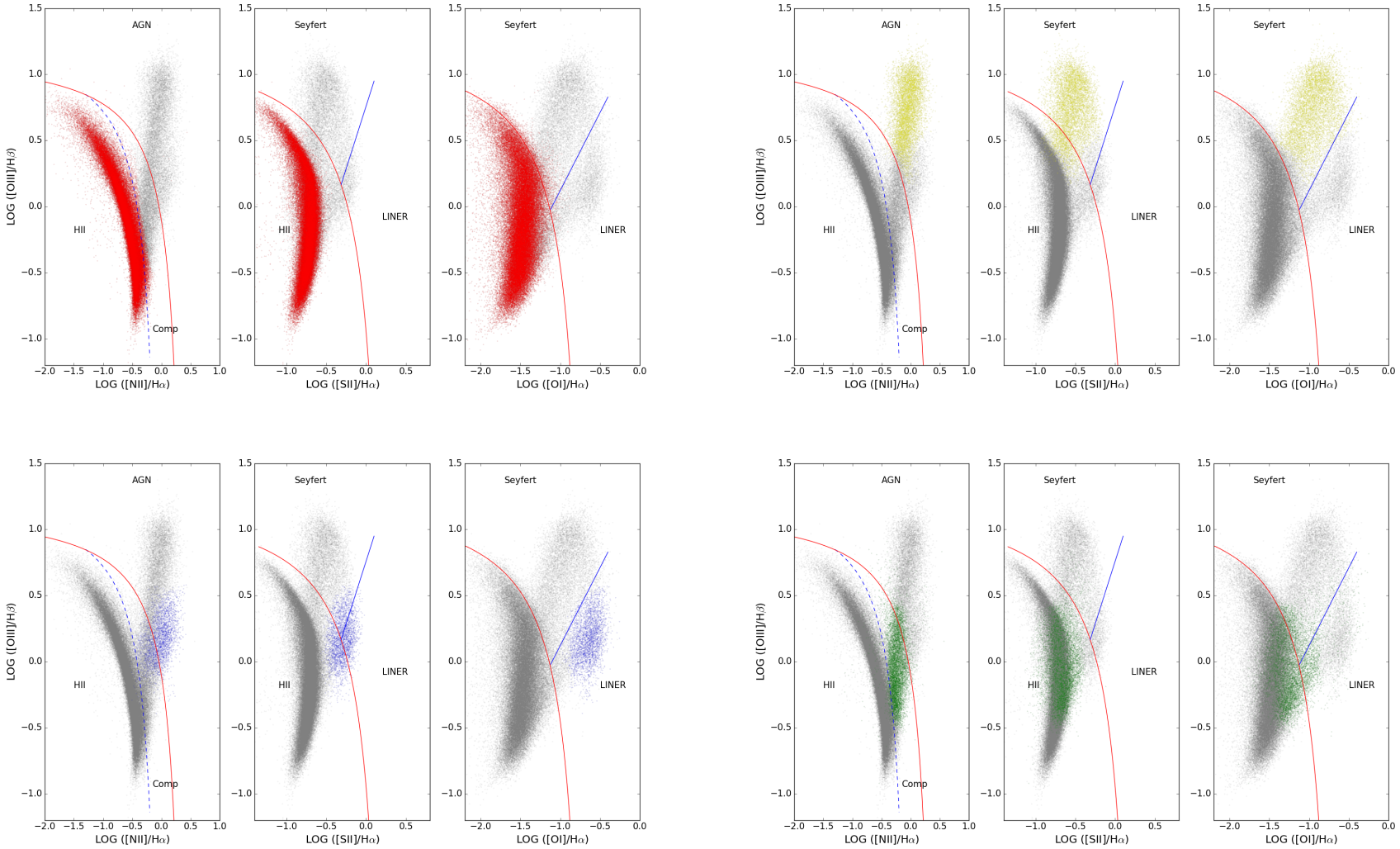



**Figure 4.9:** The locus of galaxies classified into the different activity types using SoDDA plotted on the three BPT diagrams. Each set of panels shows a different class (clockwise from top left): (a) SFGs (red), (b) Seyfert (yellow); (c) LINERs (blue), (d) Composite (green). For reference the full sample is also plotted in grey. The maximum 'starburst' line of Kewley et al. (2001) is plotted as a solid red line, the empirical upper bound on SFG of Kauffmann et al. (2003) is plotted as a dashed blue line, and the empirical line distinguishing Seyferts and LINERs (Kewley et al. 2006) is plotted as a solid blue line.

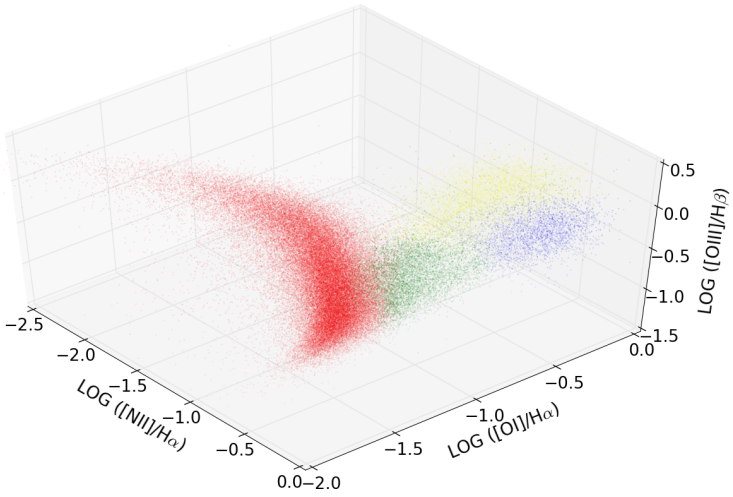



**Figure 4.10:** A 3-dimensional projection of the SDSS DR8 sample on the ([$N_{II}$]/H$\alpha$, [$S_{II}$]/H$\alpha$, [$O_{III}$]/H$\beta$) volume, in which each datapoint is plotted with different colour according to the allocation from SoDDA classification scheme (red for SFGs, yellow for Seyferts, blue for LINERs, green for the Composites and black for the Ambiguous galaxies) This 3-dimensional projections allows us to observe the complex structure of the 4 galaxy activity classes.

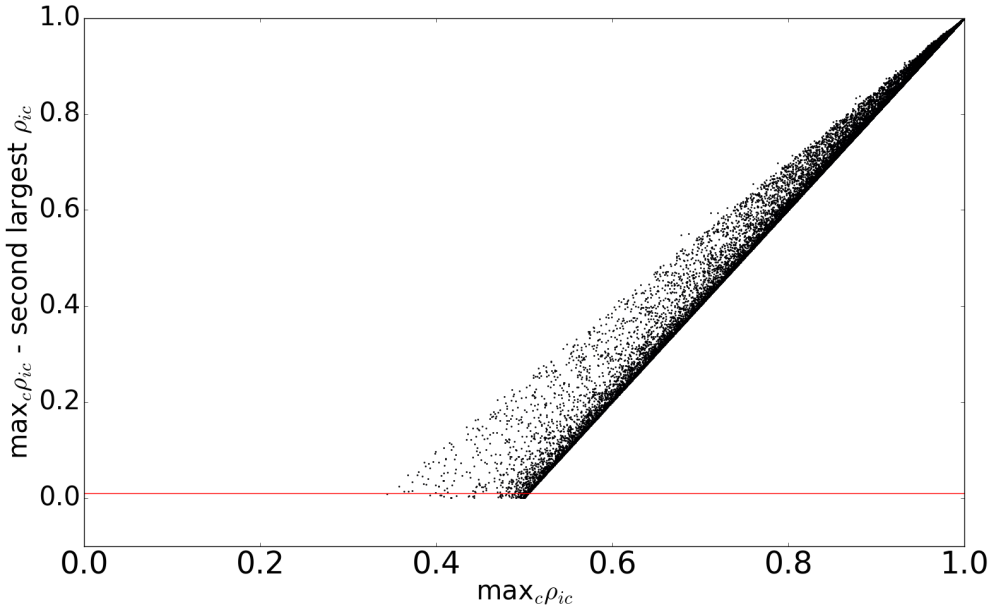



**Figure 4.11:** The difference between the SoDDA probabilities of the most likely and second most likely class for each galaxy in the SDSS D8 sample. The difference is plotted against the probability of the most likely class. The red line corresponds to a difference of $1\%$. Only 0.17% of the galaxies exhibit a difference between the probabilities of the most and second most likely classes of less than 1%. $90.6\%$ of the galaxies have $\max_c \rho_{ic} > 75\%$, indicating a highly confident classification.

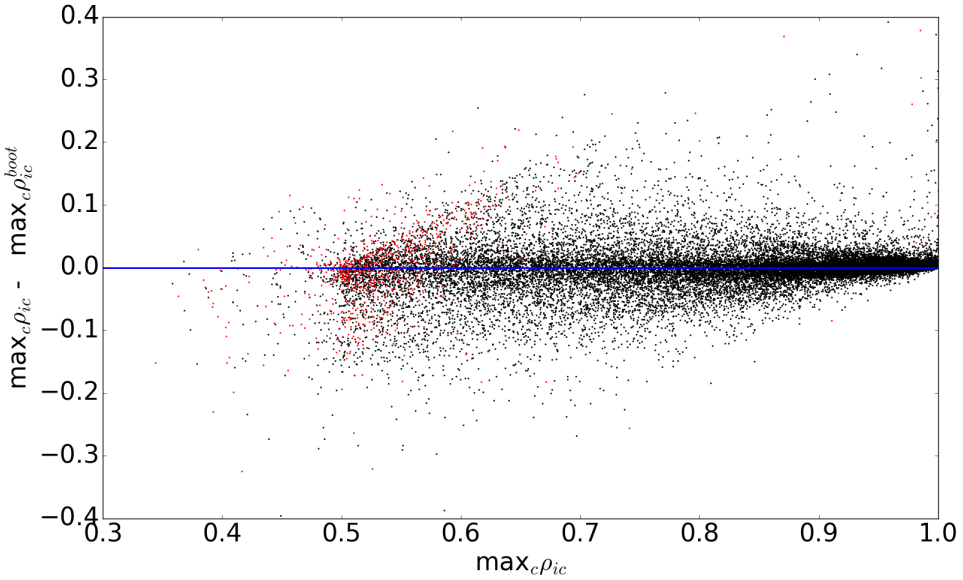



**Figure 4.12:** A plot of the difference between the class probabilities of the individual galaxies computed with the full data and with the bootstrap sample, namely a plot of $\max_c \rho_{ic} - \max_c \rho_{ic}^{\rm boot}$ against $\max_c \rho_{ic}$. Galaxies that are classified differently by the two schemes are plotted in red. The vast majority of galaxies have the same classification under both schemes; those that do not (only 0.8% of the full sample) have $\max_c \rho_{ic} < 60\%$ (86% of them), meaning they lie close to the iso-probability surface between two or more classes.

classifications for the same galaxy; there is no contradictory classification from the different diagnostics, because SoDDA provides a single coherent summary based on *all* diagnostic line ratios: a posterior membership probability for each galaxy.

A 3-way classification table that compares SoDDA with the commonly used scheme proposed by Kewley et al. (2006) appears in Table 4.2. Each cell has 3 values: the number of galaxies with (i) $\rho_{ic} \geq 75\%$, (ii) $50\% \leq \rho_{ic} < 75\%$, and (iii) $\rho_{ic} < 50\%$, where $\rho_{ic}$ is the posterior probability that galaxy $i$ belongs to galaxy class $c$. For example, the cell in the first row and first column shows that of the galaxies that both SoDDA and the Kewley et al. (2006) method classify as SFG, 65,080 are SFGs under SoDDA with probability greater than 75%, 946 with probability between 50% and 75%, and only 3 with probability less than 50%. On the other hand, 1,744 of the galaxies that are characterised as ambiguous by Kewley et al. (2006) are estimated with SoDDA to be SFGs with probability over 75%, a robust classification.

The first row in Figure 4.8 shows the classification suggested by Kewley et al. (2006), using the same colour coding as in the second row of Figure 4.8 (which shows the

**Table 4.2:** A 3-way classification table that compares the SoDDA classification with the standard, 2-dimensional classification scheme (Kewley et al. 2006). Each cell has 3 values: the number of galaxies with (i) $\rho_{ic} \geq 75\%$, (ii) $50\% \leq \rho_{ic} < 75\%$, and (iii) $\rho_{ic} < 50\%$, where $\rho_{ic}$ is the posterior probability that galaxy $i$ belongs to galaxy class $c$ under SoDDA. Contradictory classifications are called ambiguous classifications by Kewley et al. (2006).

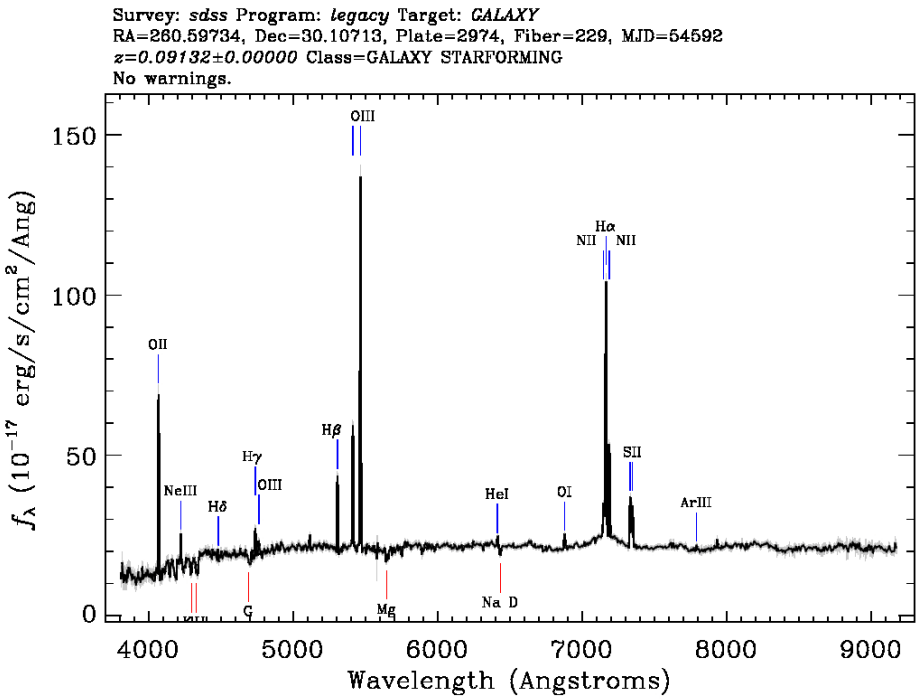
| | | Kewley et al. (2006) | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | SFGs | | | Seyferts | | | LINERs | | | Comp | | | Contradictory | | | Total | | |
| SoDDA | SFGs | 65080 | 946 | 3 | 6 | 0 | 0 | 0 | 0 | 0 | 927 | 1343 | 69 | 1744 | 62 | 8 | 67757 | 2351 | 80 |
| | Seyferts | 0 | 0 | 0 | 5471 | 262 | 8 | 0 | 0 | 0 | 2 | 28 | 10 | 349 | 1131 | 45 | 5822 | 1421 | 63 |
| | LINERs | 0 | 0 | 0 | 9 | 33 | 5 | 778 | 4 | 0 | 700 | 181 | 15 | 891 | 234 | 44 | 2378 | 452 | 64 |
| | Comp | 57 | 251 | 4 | 32 | 40 | 4 | 0 | 0 | 0 | 4211 | 2668 | 103 | 258 | 801 | 38 | 4558 | 3760 | 149 |

classification suggested by SoDDA) but with galaxies that are characterised as having contradictory classification plotted in black. The overlap between the composite galaxies (green) and the SFGs (red) is clear in the SoDDA classification (middle and right panels of Figure 4.8), indicating that the 2-dimensional projection of this 4-dimensional parameter space is insufficient for capturing its complex structure and accurately classifying the galactic activity. The use of hard boundaries defined independently in the 2-dimensional projections is responsible for those galaxies with contradictory classification. On the other hand the probabilistic approach of SoDDA simultaneously accounts for the 4-dimensional structure of the data space and inherently alleviates these inconsistent classifications, while at the same time giving a confident classification of the galaxies to activity classes.

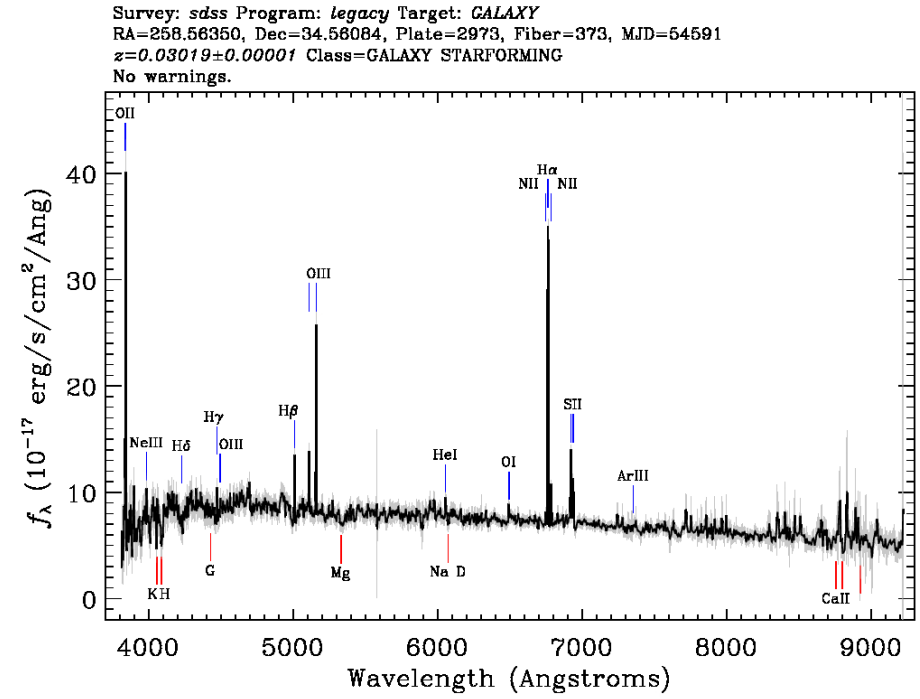
In order to offer some connection between the classification result and the underlying spectra, we plot in Figure 4.13 the spectra from 2 galaxies that have contradicting classification by Kewley et al. (2006), but the galaxy in the upper panel is classified as Seyfert with probability 87.4% by SoDDA, while the galaxy in the lower panel is classified as SFG with probability 99.9% by SoDDA. Similarly, in Figure 4.14, the upper panel shows a spectra from a galaxy that is classified as LINER by both Kewley et al. (2006) and SoDDA (with probability 99.4%), and the lower panel a spectra from a galaxy that is classified as having contradictory classification by Kewley et al. (2006), but SoDDA classifies it as Composite with probability 91.6%.

## 4.4 Multidimensional Decision Boundaries

In order to provide a more immediately usable diagnostic in the spirit of the classification lines of Kauffmann et al. (2003) and Kewley et al. (2006), which however, *simultaneously* employ the information in all diagnostic lines, we use a support vector machine (SVM) (Cortes & Vapnik 1995) to obtain multidimensional decision boundaries based on the SoDDA results. A SVM is a discriminative classifier formally defined by a separating hyperplane. In other words, given classified galaxies, the algorithm outputs an optimal hyperplane which can be used to categorise new unlabelled galaxies.

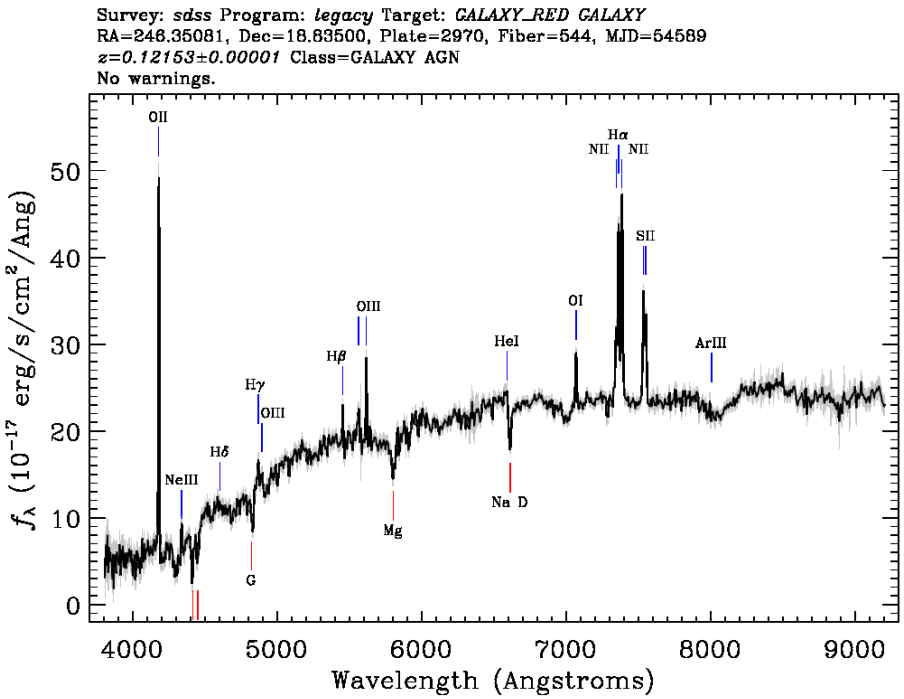



**(a)** The galaxy with SpecObjId 3348489347031656448 from SDDS DR8 (http://www.sdss.org).

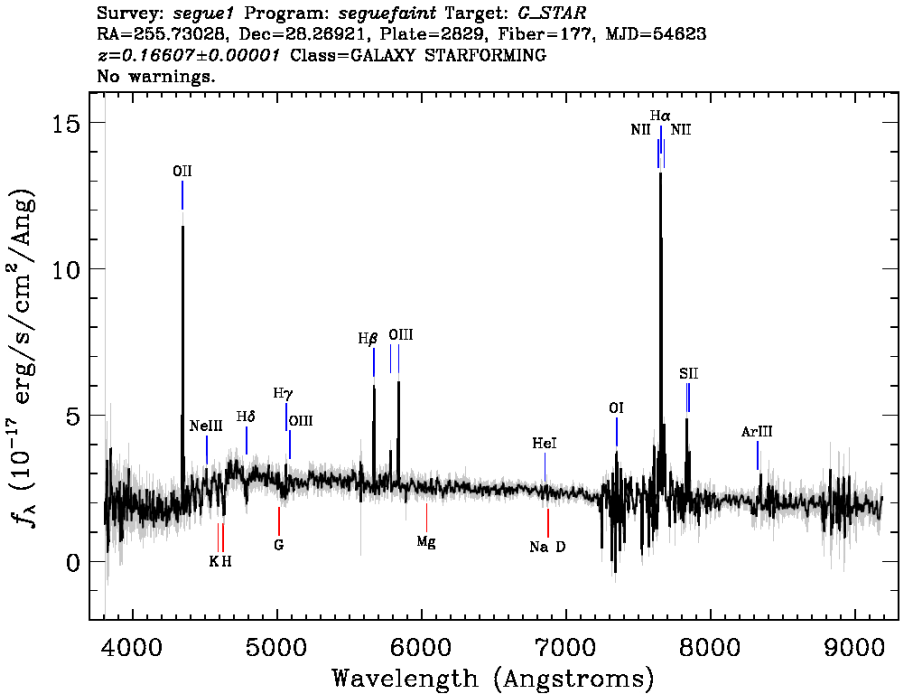



**(b)** The galaxy with SpecObjId 3347403029526636544 from SDDS DR8 (http://www.sdss.org).

**Figure 4.13:** The upper subplot shows the spectra of the galaxy with SpecObjId 3348489347031656448 from SDDS DR8. This galaxy is characterised as having contradicting classification by Kewley et al. (2006), but SoDDA classify it as Seyfert with probability 87.4%. The lower subplot shows the spectra of the galaxy with SpecObjId 3347403029526636544 from SDDS DR8. This galaxy is characterised as having contradicting classification by Kewley et al. (2006), but SoDDA classify it as SFG with probability 99.9%.




**(a)** The galaxy with SpecObjId 3344072333894641664 from SDDS DR8 (http://www.sdss.org).




**(b)** The galaxy with SpecObjId 3347403029526636544 from SDDS DR8 (http://www.sdss.org).

**Figure 4.14:** The upper subplot shows the spectra of the galaxy with SpecObjId 3344072333894641664 from SDDS DR8. This galaxy is characterised as LINER by both Kewley et al. (2006) and SoDDA (with probability 99.4%). The lower subplot shows the spectra of the galaxy with SpecObjId 3185219567408408576 from SDDS DR8. This galaxy is characterised as having contradicting classification by Kewley et al. (2006), but SoDDA classify it as Composite with probability 91.6%.

### 4.4.1 4-dimensional Decision Boundaries

The input data for the derivation of the boundaries implied by the SVM are the 4 emission line ratios for the galaxies in SDSS DR8 (i.e. $x$), and the classification for each galaxy $y_i$ as obtained with SoDDA. We use the `scikit-learn` Python library to fit the SVM model. We employ a linear kernel function; a more complex function did not provide an improvement significant enough to justify its use, especially given the simplicity of a linear kernel. The SVM algorithm requires tuning the cost factor parameter $C$, that sets the width of the margin between hyperplanes separating different classes of objects. After a grid search in a range of values for $C$, we suggest a value of $C = 1$ based on 10-fold cross-validation. $\mathcal{K}$-fold cross-validation is a model validation method for estimating the performance of the model. The data is split in $\mathcal{K}$ roughy equal parts. For each $\kappa \in (1, ..., \mathcal{K})$ we fit the model in the other $\mathcal{K}-1$ parts of the data and calculate the prediction error of the fitted model when predicting the $\kappa$th part of the data. By repeating this procedure in a range of values for the model parameters, we choose the values of the parameters that give us the model with the minimum expected prediction error.

Using the SoDDA classification, we employ a SVM approach to define multidimensional surfaces separating the galaxy activity classes. More specifically, we find an optimal separation hyperplane using the 4 emission line ratios for the galaxies from SDSS DR8 and their most probable classification obtained by SoDDA as inputs. The 4-dimensional linear decision boundaries for the four galaxy classes are defined as:

**SFG:**

$$\begin{aligned} -5.964\log([\mathrm{N_{II}}]/\mathrm{H}\alpha) - 1.487\log([\mathrm{S_{II}}]/\mathrm{H}\alpha) - 0.048\log([\mathrm{O_I}]/\mathrm{H}\alpha) \\ - 5.447\log([\mathrm{O_{III}}]/\mathrm{H}\beta) > 1.562 \end{aligned} \tag{4.18}$$

$$\begin{aligned} -3.202\log([\mathrm{N_{II}}]/\mathrm{H}\alpha) - 3.363\log([\mathrm{S_{II}}]/\mathrm{H}\alpha) - 5.613\log([\mathrm{O_I}]/\mathrm{H}\alpha) \\ + 0.275\log([\mathrm{O_{III}}]/\mathrm{H}\beta) > 8.072 \end{aligned} \tag{4.19}$$

$$\begin{aligned} -19.83\log([\mathrm{N_{II}}]/\mathrm{H}\alpha) - 1.679\log([\mathrm{S_{II}}]/\mathrm{H}\alpha) - 5.916\log([\mathrm{O_I}]/\mathrm{H}\alpha) \\ - 6.140\log([\mathrm{O_{III}}]/\mathrm{H}\beta) > 16.98 \end{aligned} \tag{4.20}$$

**Seyferts:**

$$-5.964\log([\mathrm{N_{II}}]/\mathrm{H}\alpha) - 1.487\log([\mathrm{S_{II}}]/\mathrm{H}\alpha) - 0.048\log([\mathrm{O_I}]/\mathrm{H}\alpha) - 5.447\log([\mathrm{O_{III}}]/\mathrm{H}\beta) < 1.562 \quad (4.21)$$

$$0.42\log([\mathrm{N_{II}}]/\mathrm{H}\alpha) - 5.391\log([\mathrm{S_{II}}]/\mathrm{H}\alpha) - 6.899\log([\mathrm{O_I}]/\mathrm{H}\alpha) + 11.90\log([\mathrm{O_{III}}]/\mathrm{H}\beta) > 11.92 \quad (4.22)$$

$$6.724\log([\mathrm{N_{II}}]/\mathrm{H}\alpha) + 4.065\log([\mathrm{S_{II}}]/\mathrm{H}\alpha) - 2.521\log([\mathrm{O_I}]/\mathrm{H}\alpha) + 10.19\log([\mathrm{O_{III}}]/\mathrm{H}\beta) > 2.832 \quad (4.23)$$

**LINERs:**

$$-3.202\log([\mathrm{N_{II}}]/\mathrm{H}\alpha) - 3.363\log([\mathrm{S_{II}}]/\mathrm{H}\alpha) - 5.613\log([\mathrm{O_I}]/\mathrm{H}\alpha) + 0.275\log([\mathrm{O_{III}}]/\mathrm{H}\beta) < 8.072 \quad (4.24)$$

$$0.420\log([\mathrm{N_{II}}]/\mathrm{H}\alpha) - 5.391\log([\mathrm{S_{II}}]/\mathrm{H}\alpha) - 6.899\log([\mathrm{O_I}]/\mathrm{H}\alpha)) + 11.90\log([\mathrm{O_{III}}]/\mathrm{H}\beta) < 11.92 \quad (4.25)$$

$$2.753\log([\mathrm{N_{II}}]/\mathrm{H}\alpha) + 11.77\log([\mathrm{S_{II}}]/\mathrm{H}\alpha) + 5.280\log([\mathrm{O_I}]/\mathrm{H}\alpha) - 1.647\log([\mathrm{O_{III}}]/\mathrm{H}\beta) > -10.11 \quad (4.26)$$

**Composites:**

$$-19.83\log([\mathrm{N_{II}}]/\mathrm{H}\alpha) - 1.679\log([\mathrm{S_{II}}]/\mathrm{H}\alpha) - 5.916\log([\mathrm{O_I}]/\mathrm{H}\alpha) - 6.140\log([\mathrm{O_{III}}]/\mathrm{H}\beta) < 16.98 \quad (4.27)$$

$$6.724\log([\mathrm{N_{II}}]/\mathrm{H}\alpha) + 4.065\log([\mathrm{S_{II}}]/\mathrm{H}\alpha) - 2.521\log([\mathrm{O_I}]/\mathrm{H}\alpha) + 10.19\log([\mathrm{O_{III}}]/\mathrm{H}\beta) < 2.832 \quad (4.28)$$

$$2.753\log([\mathrm{N_{II}}]/\mathrm{H}\alpha) + 11.77\log([\mathrm{S_{II}}]/\mathrm{H}\alpha) + 5.280\log([\mathrm{O_I}]/\mathrm{H}\alpha) - 1.647\log([\mathrm{O_{III}}]/\mathrm{H}\beta) < -10.11 \quad (4.29)$$

These multidimensional decision boundaries achieve a mean classification accuracy of about 98% based on a 10-fold cross validation with respect to the SoDDA classification. Table 4.3 compares the SoDDA classification with the proposed classification from the SVM, while Table 4.4 compares the scheme from Kewley et al. (2006) with the SVM. We see excellent agreement between the SoDDA and the SVM based classification. More specifically, 99% of the galaxies classified as SFGs by SoDDA are classified in the same way the SVM-based classification. The figures are 96% for Seyferts, 93% for LINERs, and 88% for Composites. On the other hand, the comparison with the traditional 2-dimensional diagnostics reveals a larger discrepancy owing mainly to the inconsistent classifications between each of the three different diagnostic diagrams.

**Table 4.3:** Comparison of the SoDDA classification with that of the 4-dimensional SVM ($[N_{II}]/H\alpha$, $[S_{II}]/H\alpha$, $O_I/H\alpha$ and $O_{III}/H\beta$ space).

| | | SoDDA | | | | |
|---|---|---|---|---|---|---|
| | | SFGs | Seyferts | LINERs | Composites | Total |
| SVM | SFGs | 69794 | 0 | 0 | 573 | 70367 |
| | Seyferts | 3 | 7033 | 59 | 273 | 7368 |
| | LINERs | 0 | 20 | 2703 | 143 | 2866 |
| | Composites | 391 | 253 | 132 | 7478 | 8254 |
| | Total | 70188 | 7306 | 2894 | 8467 | |

**Table 4.4:** Comparison of the classifications of a 4-dimensional SVM with that of the method by Kewley et al. (2006) ($[N_{II}]/H\alpha$, $[S_{II}]/H\alpha$, $O_I/H\alpha$ and $O_{III}/H\beta$ space). Contradictory classifications are called ambiguous classifications by Kewley et al. (2006).

| | | Kewley et al. (2006) | | | | | |
|---|---|---|---|---|---|---|---|
| | | SFGs | Seyferts | LINERs | Composites | Contradictory | Total |
| SVM | SFGs | 66199 | 6 | 0 | 2348 | 1814 | 70367 |
| | Seyferts | 0 | 5831 | 0 | 4 | 1533 | 7368 |
| | LINERs | 0 | 22 | 782 | 873 | 1189 | 2866 |
| | Composites | 142 | 11 | 0 | 7032 | 1069 | 8254 |
| | Total | 66341 | 5870 | 782 | 10257 | 5605 | |

### 4.4.2 3-dimensional Decision Boundaries

Because the $[O_I]$ line is generally hard to observe (very weak and hence difficult to measure), it is common to use measurements of $\log([N_{II}]/H\alpha)$, $\log([S_{II}]/H\alpha)$ and $\log([O_{III}]/H\beta)$, but not for $\log([O_I]/H\alpha)$. Thus, we derive decision boundaries by fitting the SVM algorithm to the SDSS DR8 dataset using the classifications from SoDDA and only the 3 emission line ratios ($\log([N_{II}]/H\alpha)$, $\log([S_{II}]/H\alpha)$ and $\log([O_{III}]/H\beta)$) as inputs. The resulting 3-dimensional decision surfaces for the four galaxy classes are defined as:

**SFG:**

$$-5.989\log([\mathrm{N_{II}}]/\mathrm{H}\alpha) - 1.534\log([\mathrm{S_{II}}]/\mathrm{H}\alpha) - 5.465\log([\mathrm{O_{III}}]/\mathrm{H}\beta) > 1.543 \quad (4.30)$$

$$-6.307\log([\mathrm{N_{II}}]/\mathrm{H}\alpha) - 8.721\log([\mathrm{S_{II}}]/\mathrm{H}\alpha) - 1.184\log([\mathrm{O_{III}}]/\mathrm{H}\beta) > 6.781 \quad (4.31)$$

$$-19.42\log([\mathrm{N_{II}}]/\mathrm{H}\alpha) - 6.912\log([\mathrm{S_{II}}]/\mathrm{H}\alpha) - 6.415\log([\mathrm{O_{III}}]/\mathrm{H}\beta) > 12.62 \quad (4.32)$$

**Seyferts:**

$$-5.989\log([\mathrm{N_{II}}]/\mathrm{H}\alpha) - 1.534\log([\mathrm{S_{II}}]/\mathrm{H}\alpha) - 5.465\log([\mathrm{O_{III}}]/\mathrm{H}\beta) < 1.543 \quad (4.33)$$

$$0.112\log([\mathrm{N_{II}}]/\mathrm{H}\alpha) - 10.74\log([\mathrm{S_{II}}]/\mathrm{H}\alpha) + 10.13\log([\mathrm{O_{III}}]/\mathrm{H}\beta) > 8.089 \quad (4.34)$$

$$5.918\log([\mathrm{N_{II}}]/\mathrm{H}\alpha) + 1.422\log([\mathrm{S_{II}}]/\mathrm{H}\alpha) + 9.623\log([\mathrm{O_{III}}]/\mathrm{H}\beta) > 1.611 \quad (4.35)$$

**LINERs:**

$$-6.307\log([\mathrm{N_{II}}]/\mathrm{H}\alpha) - 8.721\log([\mathrm{S_{II}}]/\mathrm{H}\alpha) - 1.184\log([\mathrm{O_{III}}]/\mathrm{H}\beta) < 6.781 \quad (4.36)$$

$$0.112\log([\mathrm{N_{II}}]/\mathrm{H}\alpha) - 10.74\log([\mathrm{S_{II}}]/\mathrm{H}\alpha) + 10.13\log([\mathrm{O_{III}}]/\mathrm{H}\beta) < 8.089 \quad (4.37)$$

$$2.383\log([\mathrm{N_{II}}]/\mathrm{H}\alpha) + 14.56\log([\mathrm{S_{II}}]/\mathrm{H}\alpha) + 0.378\log([\mathrm{O_{III}}]/\mathrm{H}\beta) > -6.724 \quad (4.38)$$

**Table 4.5:** Comparison of the SoDDA classification with that of the 3-dimensional SVM ([$N_{II}$]/H$\alpha$, [$S_{II}$]/H$\alpha$ and $O_{III}$/H$\beta$ space).

| | | SoDDA | | | | |
|---|---|---|---|---|---|---|
| | | SFGs | Seyferts | LINERs | Composites | Total |
| SVM | SFGs | 69746 | 0 | 7 | 808 | 70561 |
| | Seyferts | 5 | 7010 | 99 | 278 | 7392 |
| | LINERs | 0 | 66 | 2574 | 154 | 2794 |
| | Composites | 437 | 230 | 214 | 7227 | 8108 |
| | Total | 70188 | 7306 | 2894 | 8467 | |

**Composites:**

$$-19.42\log([N_{II}]/H\alpha) - 6.912\log([S_{II}]/H\alpha) - 6.415\log([O_{III}]/H\beta) < 12.62 \quad (4.39)$$

$$5.918\log([N_{II}]/H\alpha) + 1.422\log([S_{II}]/H\alpha) + 9.623\log([O_{III}]/H\beta) < 1.611 \quad (4.40)$$

$$2.383\log([N_{II}]/H\alpha) + 14.56\log([S_{II}]/H\alpha) + 0.378\log([O_{III}]/H\beta) < -6.724 \quad (4.41)$$

The multidimensional decision boundaries achieve a mean classification accuracy of about 97% based on a 10-fold cross validation with respect to the SoDDA classification. Table 4.3 compares the SoDDA classification with the proposed classification from the SVM, while Table 4.4 compares the scheme from Kewley et al. (2006) with the SVM. Similarly as in the case of the 4-dimensional based SVM classification, we have excellent agreement with the SoDDA classification and slightly worse agreement with the traditional 2-dimensional diagnostics. Surprisingly we find excellent agreement between the 3-dimensional and the 4- dimensional SVM diagnostics indicating that the lack of the 4th dimension ([$O_I$]/$H_\alpha$) does not significantly affect the quality of the classification. More specifically, the 3-dimensional SVM has 99% accuracy for SFGs with the SoDDA, 96% for the Seyferts, 89% for the LINERs and 85% for the Composites. In other words, removing the [$O_I$]/H$\alpha$) dimension has no impact on the classification error for SFGs and the Seyferts, and affects the error by 4% for the LINERs and by 3% for the Composites.

**Table 4.6:** Comparison of the classifications of a 3-dimensional SVM with that of the method by Kewley et al. (2006) ($[N_{II}]/H\alpha$, $[S_{II}]/H\alpha$ and $O_{III}/H\beta$ space). Contradictory classifications are called ambiguous classifications by Kewley et al. (2006).

| | | Kewley et al. (2006) | | | | | |
|---|---|---|---|---|---|---|---|
| | | SFGs | Seyferts | LINERs | Composites | Ambiguous | Total |
| SVM | SFGs | 66274 | 5 | 0 | 2383 | 1899 | 70561 |
| | Seyferts | 0 | 5782 | 0 | 5 | 1605 | 7392 |
| | LINERs | 0 | 81 | 782 | 838 | 1093 | 2794 |
| | Composites | 67 | 2 | 0 | 7031 | 1008 | 8108 |
| | Total | 66341 | 5870 | 782 | 10257 | 5605 | |

## 4.5 Discussion and Connection with the $log(N) - log(S)$

We propose a new soft clustering scheme, the soft allocation data driven method (SoDDA), for classifying galaxies using emission-line ratios. Our method uses an optimal number of MG distributions in order to capture the multi-dimensional structure of the dataset and afterwards concatenate those MG distributions into groups by assigning them to different activity types, based on the location of their means with respect to the loci of the activity types as defined by Kewley et al. (2006).

### 4.5.1 Comparison with standard diagnostic

The main advantages of this method are the use of all four optical-line ratios simultaneously, thus maximising the available information and avoiding contradicting classifications, and treating each class as a distribution resulting into soft classification boundaries.

One of the issues with data-driven classification is whether the data have enough discriminating power for distinguishing the different activity classes. A strong indication in this direction comes from the fact that the original BPT diagnostic Baldwin et al. (1981) and its more recent redefinition by Kauffmann et al. (2003) and Kewley et al. (2006) was driven by the clustering of the activity classes in different loci on the 2-dimensional line-ratio diagrams. Furthermore, this distinction was supported by photoionisation models (Kewley et al. 2001) which indicate that while there is a continuous evolution of the location of sources on the 2-dimensional diagnostic diagrams as a function of their metallicity and hardness of the ionising continuum, star-forming galaxies occupy a distinct region of this diagram. In our analysis we fol-

low a hybrid approach in which we identify clusters based on the multi-dimensional distribution of the object line-ratios, and we associate the clusters with activity types based on their location of the standard 2-dimensional diagnostic diagrams. This way we give a physical interpretation to each cluster, while tracing the multi-dimensional distribution of the line ratios.

The fact that our analysis identifies multiple clusters within each activity class could indicate that there are subclasses that merit special attention. An indication for this is cluster 4. The morphology of this cluster is distinct from the remaining population (Figure 4.7), which was attributed to the complex spectra of the objects in this cluster.

The approach followed in this paper treats the multi-dimensional emission-line diagnostic diagram as a mixture of different classes. This is a more realistic approach as it does not assume fixed boundaries between the activity classes. Instead, it takes into account the fact that the emission-line ratios of the different activity classes may overlap, which is reflected on the probabilities for an object to belong to a given class. This in fact is reflected in the often inconsistent classification between different 2-dimensional diagnostics (Ho et al. 1997, Yuan et al. 2010). Therefore, the optimal way to characterize a galaxy is by calculating the probability that it belongs to each of the activity classes. This also gives us the possibility to define samples of galaxies in the different classes at various confidence levels.

Another advantage of this approach is that we take into account all available information for the activity classification of galactic nuclei. This is important given the complex shape of the multi-dimensional distributions of the emission line ratios (e.g. online 3-dimensional rotating diagnostics, see also Vogt et al. (2014)). This way we increase the power of the 2-dimensional diagnostic tools, and eliminate the contradicting classifications they often give. This is demonstrated by the excellent agreement between the classification of the 4-dimensional diagnostic ($O_{III}/H\beta$, $O_I/H\alpha$, $N_{II}/H\alpha$, $S_{II}/H\alpha$) with the 3-dimensional diagnostic excluding the often weak and hard to detect $O_I$ line ($O_{III}/H\beta$, , $N_{II}/H\alpha$, $S_{II}/H\alpha$); see 4.4.2. This agreement indicates that the loss of the diagnostic power of the $O_I/H\alpha$ line (which the main discriminator between LINERs and other activity types (e.g. Kewley et al. (2006)) in the 4-dimensional diagnostic, can be compensated by the structure of the locus of the different activity types which allows their distinction even in the 3-dimensional

diagnostic.

Although we believe that the probabilistic clustering contains more information about the classification of each active galaxy, the use of hard decision boundaries for classification is effective and closer to the standard approach used in the literature. Therefore, we also present hard classification criteria by employing SVM on the distribution of line-ratios of objects assigned to each activity class. The classification accuracy with these hard criteria is $\sim 98\%$ when compared to the soft classification (SoDDA). This indicates that the extended tails of the line-ratio distributions of the different activity classes result in only a small degree of overlap and hence misclassification.

### 4.5.2 Connection to $\log(N) - \log(S)$

Having a classification scheme in order to distinguish between AGNs and SFGs will open new horizons in the $\log(N) - \log(S)$ research. Most of the sources we observe in the deep field images (extended observations of the sky that try to measure distant astronomical sources) are galaxies. So, using SoDDA, the soft-clustering scheme we propose, we will be able to create individual $\log(N) - \log(S)$ curves for the four classes of galaxies, i.e. SFGs, LINERs, Seyferts and Composites and examine the different parameters of the distribution of the flux.

As it was stated in the previous chapters, the parameters of the $\log(N) - \log(S)$ curve are informative about the type of population. In the literature there have been attempts to create different $\log(N) - \log(S)$ curves for the different type of galaxies. Lehmer et al. (2012) focuses on the CDFS survey and uses Kewley et al. (2006) classification scheme in order to cluster galaxies and then build $\log(N) - \log(S)$ curves for each type. This approach is limited by the fact that the existing classification scheme provides hard-clustering and fails to classify many galaxies due to ambiguity.

The classification scheme we propose gives posterior probabilities for each galaxy of belonging to one of the four categories. Thus if we define the matrix that contains the posterior probabilities as $m$ and define an indicator matrix $J$, where $J_{i,j} = 1$ if the galaxy $i$ is of type $j$ and zero otherwise, then our Gibbs Sampler could be written as:

Step 1: Sample $J^{t+1} \sim Multinomial(m)$.

Step 2: Cluster using $J^{t+1}$ the galaxies and for each of the four categories run one iteration of the Gibbs sampler using the galaxies of each type as the observed sources.

This method will yield four different $log(N) - log(S)$ curves for the different types of galaxies, allowing Astronomers to study the differences of those populations using the parameters of the curves.

# 5

# Discussion

In this work, we present a comprehensive and innovative approach to estimating the $\log(N) - \log(S)$ relationship, both for the linear and the piece-wise linear case with a known number of pieces. More specifically, we develop a hierarchical Bayesian model that properly accounts for the missing data mechanism and other sources of uncertainty, such as the uncertainty about the flux-to-count conversion factor. By using a Bayesian approach, our method produces a posterior distribution for the $\log(N) - \log(S)$ curve instead of a best-fit estimate, and the uncertainty of the posterior estimates of the parameters of interest can be closely examined through their respective marginal posterior distributions.

Our work extends a recently proposed Bayesian approach for estimating the $\log(N) - \log(S)$ relationship (Udaltsova 2014). The author proposes a hierarchical Bayesian model in order to take into consideration the measurement and detector biases, as well as the missing data mechanism. Our approach extends the work of Udaltsova (2014) by employing a survey specific incompleteness function, and by utilising the survey specific background and exposure maps in order to create a joint distribution for the background contamination, the effective area and the off-axis angle.

We further extend the hierarchical Bayesian model for estimating the $\log(N) - \log(S)$ relationship, by properly incorporating the uncertainty about the flux-to-count conversion factor $\gamma$. This constitutes a very innovative approach on the $\log(N) - \log(S)$

estimation, since the methods in the relevant $\log(N) - \log(S)$ estimation assume that $\gamma$ is constant for all the sources. We extract the uncertainty about $\gamma$, expressed as a different probability distribution for each observed source, using modern astronomical software, and then we fit a hierarchical prior for the complete source population. In order to fit this prior, we use an innovative statistical methodology that can be applied to multiple statistical problems of similar nature.

The resulting methodology about estimating $\log(N) - \log(S)$ offers to the astronomical community a very powerful and at the same time versatile tool, while the hierarchical structure allows for easy extensions of the model for any other source of uncertainty.

## 5.1 Creating a Soft Clustering Scheme for Classifying Galaxies

Studying the $\log(N) - \log(S)$ relationship for different source populations can give us further insight into the differences between the various types of astronomical populations. Exploiting this idea, we delved into a long and heavily researched classification problem in Astronomy, which is the classification of galaxies to different activity classes (Star Forming Galaxies, LINERs, Seyferts and Composites).

Therefore, we propose a new soft clustering scheme, the soft allocation data driven method (SoDDA), for classifying galaxies using emission-line ratios. Our method utilises a big number of Multivariate Gaussian (MG) distributions in order to capture the multi-dimensional structure of the dataset and afterwards concatenate those MG distributions into groups by assigning them to different activity types. The main advantage of this method is the use of all four optical-line ratios simultaneously, thus maximising the available information and avoiding contradicting classifications. We also present hard classification criteria by employing SVM on the distribution of line-ratios of objects assigned to each activity class.

### 5.1.1 Connection to $\log(N) - \log(S)$

Having a classification scheme that enables us to distinguish between AGNs and SFGs will undoubtedly open new horizons in the $\log(N) - \log(S)$ research. The

use of SoDDA, the soft-clustering scheme we propose, will allow for the creation of individual $\log(N) - \log(S)$ curves for the four classes of galaxies, i.e. SFGs, LINERs, Seyferts and Composites, offering the opportunity to examine the different parameters of the 4 distributions of the flux. In the last Section of Chapter 4, we discuss a framework for estimating the 4 different $\log(N) - \log(S)$ curves using the hierarchical Bayesian model.

## 5.2 Limitations

The main limitation of the hierarchical Bayesian model is the lack of a proper and effective automatic model selection process. Despite a series of simulation studies by Udaltsova (2014), neither of the standard model selection methods, such as the Bayes factor and the DIC, manage to exhibit consistency in choosing the model used for simulating the data. As a result, further research should be conducted on the model selection procedure.

Towards this end, we suggest two different research directions; the development of model specific heuristics and techniques, and the implementation of a Reversible Jump MCMC sampler. For the first case, we can define statistics, such as the similarity between the marginal posterior distributions of the 2 consecutive slopes, $\theta_1$ and $\theta_2$, measured by some metric. If the distance between the two distributions is small under that predefined measure, then we can assume that there is not enough evidence to support the hypothesis of the existence of a breakpoint. Extensive simulations would be required in order to properly test the power of such diagnostics.

A more statistically interesting approach to model selection would be the use of the Reversible Jump MCMC (Green 1995) on the hierarchical Bayesian model. The Reversible Jump MCMC sampler is a framework for MCMC simulation that is well suited for problems in which the dimensions of the parameter space can vary between the iterations of the chain. However, implementing successfully a Reversible Jump MCMC is not a trivial task. The main difficulty lies in the construction of the proposal moves between different models .Thus, applying the Reversible Jump MCMC for model selection in our context is undoubtedly a promising approach, but not a straightforward one.

# References


Aihara, H., Prieto, C. A., An, D., Anderson, S. F., Aubourg, É., Balbinot, E., Beers, T. C., Berlind, A. A., Bickerton, S. J., Bizyaev, D. et al. (2011), 'The eighth data release of the sloan digital sky survey: first data from sdss-iii', *The Astrophysical Journal Supplement Series* **193**(2), 29.

Baldwin, J., Phillips, M. & TERLEVICH, R. (1981), 'Classification parameters for the emission-line spectra of extragalactic objects', *Publications of the Astronomical Society of the Pacific* pp. 5–19.

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

Bhat, H. S. & Kumar, N. (2010), 'On the derivation of the bayesian information criterion', *School of Natural Sciences, University of California* .

Bilmes, J. A. et al. (1998), 'A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models', *International Computer Science Institute* **4**(510), 126.

Brinchmann, J., Charlot, S., White, S., Tremonti, C., Kauffmann, G., Heckman, T. & Brinkmann, J. (2004), 'The physical properties of star-forming galaxies in the low-redshift universe', *Monthly Notices of the Royal Astronomical Society* **351**(4), 1151–1179.

Brooks, S., Gelman, A., Jones, G. & Meng, X.-L. (2011), *Handbook of markov chain monte carlo*, CRC press.

Buchner, J., Georgakakis, A., Nandra, K., Brightman, M., Menzel, M.-L., Liu, Z., Hsu, L.-T., Salvato, M., Rangel, C., Aird, J., Merloni, A. & Ross, N. (2015), 'Obscuration-dependent Evolution of Active Galactic Nuclei', *apj* **802**, 89.

Connors, A. & van Dyk, D. A. (2007), How To Win With Non-Gaussian Data: Poisson Goodness-of-Fit, *in* G. J. Babu & E. D. Feigelson, eds, 'Statistical Challenges in Modern Astronomy IV', Vol. 371 of *Astronomical Society of the Pacific Conference Series*, p. 101.

Cook, S. R., Gelman, A. & Rubin, D. B. (2012), 'Validation of software for bayesian models using posterior quantiles', *Journal of Computational and Graphical Statistics* .

Cortes, C. & Vapnik, V. (1995), 'Support-vector networks', *Machine learning* **20**(3), 273–297.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the royal statistical society. Series B (methodological)* pp. 1–38.

Draper, D. (1995), 'Assessment and propagation of model uncertainty', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 45–97.

Efron, B. & Morris, C. (1973), 'Stein's estimation rule and its competitors—an empirical bayes approach', *Journal of the American Statistical Association* **68**(341), 117–130.

Efron, B. & Morris, C. (1975), 'Data analysis using stein's estimator and its generalizations', *Journal of the American Statistical Association* **70**(350), 311–319.

Eisenstein, D. J., Weinberg, D. H., Agol, E., Aihara, H., Prieto, C. A., Anderson, S. F., Arns, J. A., Aubourg, É., Bailey, S., Balbinot, E. et al. (2011), 'Sdss-iii: Massive spectroscopic surveys of the distant universe, the milky way, and extra-solar planetary systems', *The Astronomical Journal* **142**(3), 72.

Ferland, G. J. (2003), 'Quantitative spectroscopy of photoionized clouds', *Annual Review of Astronomy and Astrophysics* **41**(1), 517–554.

Fraley, C. & Raftery, A. E. (2002), 'Model-based clustering, discriminant analysis, and density estimation', *Journal of the American statistical Association* **97**(458), 611–631.

Freeman, P., Doe, S. & Siemiginowska, A. (2001), Sherpa: a mission-independent data analysis application, *in* 'International Symposium on Optical Science and Technology', International Society for Optics and Photonics, pp. 76–87.

Friedman, J., Hastie, T. & Tibshirani, R. (2001), *The elements of statistical learning*, Vol. 1, Springer series in statistics Springer, Berlin.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2014), *Bayesian data analysis*, Vol. 2, Chapman & Hall/CRC Boca Raton, FL, USA.

Gelman, A. & Rubin, D. B. (1992), 'Inference from iterative simulation using multiple sequences', *Statistical science* pp. 457–472.

Geman, S. & Geman, D. (1984), 'Stochastic relaxation, gibbs distributions, and the bayesian restoration of images', *IEEE Transactions on pattern analysis and machine intelligence* (6), 721–741.

Georgakakis, A., Nandra, K., Laird, E., Aird, J. & Trichas, M. (2008), 'A new method for determining the sensitivity of x-ray imaging observations and the x-ray number counts', *Monthly Notices of the Royal Astronomical Society* **388**(3), 1205–1213.

Gilks, W. R., Best, N. & Tan, K. (1995), 'Adaptive rejection metropolis sampling within gibbs sampling', *Applied Statistics* pp. 455–472.

Green, P. J. (1995), 'Reversible jump markov chain monte carlo computation and bayesian model determination', *Biometrika* **82**(4), 711–732.

Hastings, W. K. (1970), 'Monte carlo sampling methods using markov chains and their applications', *Biometrika* **57**(1), 97–109.

Heckman, T. M. (1980), 'An optical and radio survey of the nuclei of bright galaxies-activity in normal galactic nuclei', *Astronomy and Astrophysics* **87**, 152–164.

Ho, L. C., Filippenko, A. V., Sargent, W. L. & Peng, C. Y. (1997), 'A search for "dwarf" seyfert nuclei. iv. nuclei with broad h$\alpha$ emission', *The Astrophysical Journal Supplement Series* **112**(2), 391.

Jóhannesson, G., Björnsson, G. & Gudmundsson, E. H. (2006), 'Afterglow light curves and broken power laws: a statistical study', *The Astrophysical Journal Letters* **640**(1), L5.

Kass, R. E. & Raftery, A. E. (1995), 'Bayes factors', *Journal of the american statistical association* **90**(430), 773–795.

Kass, R. E. & Wasserman, L. (1996), 'The selection of prior distributions by formal rules', *Journal of the American Statistical Association* **91**(435), 1343–1370.

Kauffmann, G., Heckman, T. M., Tremonti, C., Brinchmann, J., Charlot, S., White, S. D., Ridgway, S. E., Brinkmann, J., Fukugita, M., Hall, P. B. et al. (2003), 'The host galaxies of active galactic nuclei', *Monthly Notices of the Royal Astronomical Society* **346**(4), 1055–1077.

Kelly, B. C. (2007), 'Some Aspects of Measurement Error in Linear Regression of Astronomical Data', *apj* **665**, 1489–1506.

Kewley, L. J., Dopita, M., Sutherland, R., Heisler, C. & Trevena, J. (2001), 'Theoretical modeling of starburst galaxies', *The Astrophysical Journal* **556**(1), 121.

Kewley, L. J., Groves, B., Kauffmann, G. & Heckman, T. (2006), 'The host galaxies and classification of active galactic nuclei', *Monthly Notices of the Royal Astronomical Society* **372**(3), 961–976.

Kormendy, J. & Ho, L. C. (2013), 'Coevolution (or not) of supermassive black holes and host galaxies', *Annual Review of Astronomy and Astrophysics* **51**, 511–653.

Lehmer, B. D., Xue, Y., Brandt, W., Alexander, D., Bauer, F., Brusa, M., Comastri, A., Gilli, R., Hornschemeier, A., Luo, B. et al. (2012), 'The 4? ms chandra deep field-south number counts apportioned by source class: Pervasive active galactic nuclei and the ascent of normal galaxies', *The Astrophysical Journal* **752**(1), 46.

Liu, J. S., Wong, W. H. & Kong, A. (1994), 'Covariance structure of the gibbs sampler with applications to the comparisons of estimators and augmentation schemes', *Biometrika* **81**(1), 27–40.

Maccacaro, T., Gioia, I., Zamorani, G., Feigelson, E., Fener, M., Giacconi, R., Griffiths, R., Murray, S., Stocke, J. & 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**, 504–511.

Maccacaro, T., Romaine, S. & Schmitt, J. H. (1987), Log n-log s slope determination in imaging x-ray astronomy, *in* 'Symposium-International Astronomical Union', Vol. 124, Cambridge University Press, pp. 597–600.

Malmquist, G. K. (1920), 'A study of the stars of spectral type a', *Meddelanden fran Lunds Astronomiska Observatorium Serie II* **22**, 3–69.

McKeough, K., Siemiginowska, A., Cheung, C. C., Stawarz, Ł., Kashyap, V. L., Stein, N., Stampoulis, V., van Dyk, D. A., Wardle, J. F. C., Lee, N. P., Harris, D. E., Schwartz, D. A., Donato, D., Maraschi, L. & Tavecchio, F. (2016), 'Detecting Relativistic X-Ray Jets in High-redshift Quasars', *apj* **833**, 123.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953), 'Equation of state calculations by fast computing machines', *The journal of chemical physics* **21**(6), 1087–1092.

Metropolis, N. & Ulam, S. (1949), 'The monte carlo method', *Journal of the American statistical association* **44**(247), 335–341.

Mukherjee, S., Feigelson, E. D., Babu, G. J., Murtagh, F., Fraley, C. & Raftery, A. (1998), 'Three types of gamma-ray bursts', *The Astrophysical Journal* **508**(1), 314.

Murdoch, H. S., Crawford, D. F. & 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**, 1–14.

O'Hagan, A. (1995), 'Fractional bayes factors for model comparison', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 99–138.

Park, T. & van Dyk, D. A. (2009), 'Partially collapsed gibbs samplers: Illustrations and applications', *Journal of Computational and Graphical Statistics* **18**(2), 283–305.

Park, T., van Dyk, D. A. & Siemiginowska, A. (2008), 'Searching for narrow emission lines in x-ray spectra: Computation and methods', *The Astrophysical Journal* **688**(2), 807.

Refsdal, B. L., Doe, S. M., Nguyen, D. T., Siemiginowska, A. L., Bonaventura, N. R., Burke, D., Evans, I. N., Evans, J. D., Fruscione, A., Galle, E. C. et al. (2009), Sherpa: 1d/2d modeling and fitting in python, *in* 'Proceedings of the 8th Python in Science Conference, Pasadena, CA, 2009, edited by G. Varoquaux, S. van der Walt and J. Millman', Vol. 1, p. 51.

Rich, J. A., Kewley, L. J. & Dopita, M. A. (2014), 'Composite spectra in merging u/lirgs caused by shocks', *The Astrophysical Journal Letters* **781**(1), L12.

Roberts, G. O. & Sahu, S. K. (1997), 'Updating schemes, correlation structure, blocking and parameterization for the gibbs sampler', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**(2), 291–317.

Rosenthal, J. S. (1995), 'Minorization conditions and convergence rates for markov chain monte carlo', *Journal of the American Statistical Association* **90**(430), 558–566.

Rubin, D. B. (1996), 'Multiple imputation after 18+ years', *Journal of the American statistical Association* **91**(434), 473–489.

Schmitt, J. H. & Maccacaro, T. (1986), 'Number-counts slope estimation in the presence of poisson noise', *The Astrophysical Journal* **310**, 334–342.

Schwarz, G. et al. (1978), 'Estimating the dimension of a model', *The annals of statistics* **6**(2), 461–464.

Shi, F., Liu, Y.-Y., Sun, G.-L., Li, P.-Y., Lei, Y.-M. & Wang, J. (2015), 'A support vector machine for spectral classification of emission-line galaxies from the sloan digital sky survey', *Monthly Notices of the Royal Astronomical Society* **453**(1), 122–127.

Son, Y. S. & Oh, M. (2006), 'Bayesian estimation of the two-parameter gamma distribution', *Communications in Statistics—Simulation and Computation®* **35**(2), 285–293.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & 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.

Stein, N. M., van Dyk, D. A., Kashyap, V. L. & Siemiginowska, A. (2015), 'Detecting Unspecified Structure in Low-count Images', *apj* **813**, 66.

Teerikorpi, P. (2004), 'Influence of a generalized eddington bias on galaxy counts', *Astronomy & Astrophysics* **424**(1), 73–78.

Tibshirani, R., Walther, G. & Hastie, T. (2001), 'Estimating the number of clusters in a data set via the gap statistic', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**(2), 411–423.

Tremonti, C. A., Heckman, T. M., Kauffmann, G., Brinchmann, J., Charlot, S., White, S. D., Seibert, M., Peng, E. W., Schlegel, D. J., Uomoto, A. et al. (2004), 'The origin of the mass-metallicity relation: insights from 53,000 star-forming galaxies in the sloan digital sky survey', *The Astrophysical Journal* **613**(2), 898.

Udaltsova, I. S. (2014), The Universe at Your Fingertips: Bayesian Modeling and Computation in Problems of Observational Cosmology, PhD thesis, University of California Davis.

van Dyk, D. A., Connors, A., Kashyap, V. L. & Siemiginowska, A. (2001), 'Analysis of Energy Spectra with Low Photon Counts via Bayesian Posterior Simulation', *apj* **548**, 224–243.

van Dyk, D. A. & Jiao, X. (2015), 'Metropolis-hastings within partially collapsed gibbs samplers', *Journal of Computational and Graphical Statistics* **24**(2), 301–327.

van Dyk, D. A. & Meng, X.-L. (2001), 'The art of data augmentation', *Journal of Computational and Graphical Statistics* **10**(1), 1–50.

van Dyk, D. A. & Park, T. (2008), 'Partially collapsed gibbs samplers: Theory and methods', *Journal of the American Statistical Association* **103**(482), 790–796.

Veilleux, S. & Osterbrock, D. E. (1987), 'Spectral classification of emission-line galaxies', *The Astrophysical Journal Supplement Series* **63**, 295–310.

Vogt, F. P., Dopita, M. A., Kewley, L. J., Sutherland, R. S., Scharwächter, J., Basurah, H. M., Ali, A. & Amer, M. A. (2014), 'Galaxy emission line classification using three-dimensional line ratio diagrams', *The Astrophysical Journal* **793**(2), 127.

Wolfe, J. H. (1970), 'Pattern clustering by multivariate mixture analysis', *Multivariate Behavioral Research* **5**(3), 329–350.

Wong, R. K., Baines, P., Aue, A., Lee, T. C., Kashyap, V. L. et al. (2014), 'Automatic estimation of flux distributions of astrophysical source populations', *The Annals of Applied Statistics* **8**(3), 1690–1712.

Wright, N. J., Drake, J. J., Guarcello, M. G., Kashyap, V. L. & Zezas, A. (2015), 'Simulating the sensitivity to stellar point sources of chandra x-ray observations', *arXiv preprint arXiv:1511.03943* .

York, D. G., Adelman, J., Anderson Jr, J. E., Anderson, S. F., Annis, J., Bahcall, N. A., Bakken, J., Barkhouser, R., Bastian, S., Berman, E. et al. (2000), 'The sloan digital sky survey: Technical summary', *The Astronomical Journal* **120**(3), 1579.

Yuan, T.-T., Kewley, L. & Sanders, D. (2010), 'The role of starburst-active galactic nucleus composites in luminous infrared galaxy mergers: Insights from the new optical classification scheme', *The Astrophysical Journal* **709**(2), 884.

Zezas, A. & Fabbiano, G. (2002), 'Chandra observations of "the antennae" galaxies (ngc 4038/4039). iv. the x-ray source luminosity function and the nature of ultraluminous x-ray sources', *The Astrophysical Journal* **577**(2), 726.

Zezas, A., Fabbiano, G., Baldi, A., Schweizer, F., King, A. R., Rots, A. H. & Ponman, T. J. (2007), 'Chandra Monitoring Observations of the Antennae Galaxies. II. X-Ray Luminosity Functions', *apj* **661**, 135–148.

# A

## A.1 Bayesian Modelling for $\log(N)-\log(S)$ with $\gamma$ (flux-to-count conversion rate) uncertainty

### A.1.1 Model Assumptions

The distributional assumptions for the parameters of interest are the following:

- **The flux S:**
  - For the **Single power law model** we have for the pdf of the flux $S$ and the associated priors:

$$p(S) \sim \text{Pareto}(S;\theta,\tau) = \theta\tau^{\theta}S^{-(\theta+1)} \tag{A.1}$$

$$p(\theta) \sim \text{Gamma}(\theta;a,b) = \frac{b^{a}}{\Gamma(a)}\theta^{a-1}e^{-b\theta} \tag{A.2}$$

$$p(\tau) \sim \text{Gamma}(\tau;a_m,b_m) = \frac{b_m^{a_m}}{\Gamma(a_m)}\tau^{a_m-1}e^{-b_m\tau} \tag{A.3}$$

  - For the **Broken power-law model** with $(m-1)$ breaks we have for the flux $S$ :

$$f(S) = p(S|\theta,\tau) = \sum_{j=1}^{m} \Big[ \prod_{i=1}^{j-1} \Big( \frac{\tau_{i+1}}{\tau_i} \Big)^{-\theta_i} \Big] \Big( \frac{\theta_j}{\tau_j} \Big) \Big( \frac{S}{\tau_j} \Big)^{-(\theta_j+1)} \mathbb{I}_{\{\tau_j \le S \le \tau_{j+1}\}}, \tag{A.4}$$

where $\theta = (\theta_1, ..., \theta_m)$ are the $m$ power-law slopes, $\tau_1$ is the flux population minimum threshold, $(\tau_2, ..., \tau_m)$ are the consequent breakpoints and $\prod_{i=1}^{0} \Big( \frac{\tau_{i+1}}{\tau_i} \Big)^{-\theta_i} = 1$.

$$\theta_j \sim \text{Gamma}(a_j, b_j), \quad j = 1, ..., m \tag{A.5}$$

$$\tau_1 \sim \text{Gamma}(a_\tau, b_\tau) \tag{A.6}$$

$$\eta_j = h_j(\tau_j \mid \tau_{j-1}) = \log(\tau_j - \tau_{j-1}), \quad j = 2, ..., m \tag{A.7}$$

$$\eta \sim \text{Multivariate Gaussian}(\mu, C) \tag{A.8}$$

- **Total number of sources N (observed and unobserved)**:

$$p(N) \sim \text{Negative-Binomial}(N; a_N, b_N) = \binom{N + a_N - 1}{a_N - 1} \Big( \frac{1}{1+b_N} \Big)^N \Big( \frac{b_N}{1+b_N} \Big)^{a_N} \tag{A.9}$$

- **The flux-to-count conversion rate $\gamma$**:

$$p(\gamma) \sim \text{Gamma}(\gamma; a_\gamma, b_\gamma) = \frac{b_\gamma^{a_\gamma}}{\Gamma(a_\gamma)} \theta^{a_\gamma - 1} e^{-b_\gamma \theta} \tag{A.10}$$

- **The photon counts for each source**, $i = 1, ..., N$:

$$Y_i^{\text{tot}} = Y_i^{\text{src}} + Y_i^{\text{bgr}} \tag{A.11}$$

$$Y_i^{src} | S_i, E_i, \gamma_i \overset{\text{ind}}{\sim} Poisson\ (\ \lambda(S_i, E_i, \gamma_i) = \frac{S_i E_i}{\gamma_i} = \lambda_i\ ) = \frac{e^{-\lambda_i} \lambda_i^{Y_i^{\text{src}}}}{Y_i^{\text{src}}!} \tag{A.12}$$

$$Y_i^{bkg} | B_i, A_i \overset{\text{ind}}{\sim} Poisson\ (\ k(B_i, A_i) = B_i A_i = k_i\ ) = \frac{e^{-k_i} k_i^{Y_i^{\text{bgr}}}}{Y_i^{\text{bgr}}!} \tag{A.13}$$

- **The incompleteness function**:

$$g(S, B, L, E, \gamma) = P(I = 1 \mid S, B, L, E, \gamma) \tag{A.14}$$

- **The marginal probability of detection**

$$\begin{aligned}\pi(\theta, \tau) &= \int P(I = 1 \mid S, B, L, E, \gamma) \cdot p(S, B, L, E, \gamma | \theta, \tau) \, \mathrm{d}S \, \mathrm{d}B \, \mathrm{d}E \, \mathrm{d}L \, \mathrm{d}\gamma \\ &= \int g(S, B, L, E, \gamma) \cdot p(S|\theta, \tau) \cdot p(B, L, E) \cdot p(\gamma) \, \mathrm{d}S \, \mathrm{d}B \, \mathrm{d}E \, \mathrm{d}L \, \mathrm{d}\gamma\end{aligned} \tag{A.15}$$

$$\begin{aligned}P(S_{\text{obs}} \mid n, \theta, \tau, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, \gamma_{\text{obs}}) &= \prod_{i=1}^{n} P(S_{\text{obs},i} \mid n, \theta, \tau, B_{\text{obs},i}, L_{\text{obs},i}, E_{\text{obs},i}, \\ &\qquad \gamma_{\text{obs},i}) \\ &= \prod_{i=1}^{n} p(S_i, I_i = 1 \mid \theta, \tau, B_i, L_i, E_i, \gamma_i) \\ &= \prod_{i=1}^{n} p(S_i \mid \theta, \tau) p(I_i = 1 \mid B_i, L_i, E_i, \gamma_i) \\ &= \prod_{i=1}^{n} \text{Pareto}(S_i; \theta, \tau) g(S_i, B_i, L_i, E_i, \gamma_i)\end{aligned} \tag{A.16}$$

$$\begin{aligned}P(S_{\text{mis}} \mid n, N, \theta, \tau, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, \gamma_{\text{obs}}) &= \prod_{i=n+1}^{N} P(S_{\text{mis},i} \mid n, \theta, \tau, B_{\text{obs},i}, L_{\text{obs},i}, E_{\text{obs},i}, \\ &\qquad \gamma_{\text{obs},i}) \\ &= \prod_{i=n+1}^{N} p(S_i, I_i = 0 \mid \theta, \tau, B_i, L_i, E_i, \gamma_i) \\ &= \prod_{i=n+1}^{N} p(S_i \mid \theta, \tau) p(I_i = 0 \mid B_i, L_i, E_i, \gamma_i) \\ &= \prod_{i=n+1}^{N} \text{Pareto}(S_i; \theta, \tau)(1 - g(S_i, B_i, L_i, E_i, \gamma_i))\end{aligned} \tag{A.17}$$

Let $\lambda_i = S_i E_i/\gamma_i$ and $k_i = B_i A_i$. Then

$$\begin{aligned} p(Y^{\text{tot}}_{\text{obs}} \mid n, S_{\text{obs}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, \gamma_{\text{obs}}) &= \prod_{i=1}^{n} p(Y^{\text{tot}}_{\text{obs},i} \mid n, S_{\text{obs},i}, B_{\text{obs},i}, L_{\text{obs},i}, E_{\text{obs},i}, \gamma_{\text{obs},i}) \\ &= \prod_{i=1}^{n} \frac{(\lambda_i + k_i)^{Y^{\text{tot}}_{\text{obs},i}} e^{-(\lambda_i + k_i)}}{Y^{\text{tot}}_{\text{obs},i}} \\ &= \prod_{i=1}^{n} \text{Poisson}(Y^{\text{tot}}_{\text{obs},i}; \lambda_i + k_i) \end{aligned} \tag{A.18}$$

We know that if $X \sim \text{Poisson}(\lambda_i)$, $Y \sim \text{Poisson}(k_i)$ and $W = X + Y$, then $X \mid W = w \sim \text{Binomial}(X; w, \frac{\lambda_i}{\lambda_i + k_i})$. Thus,

$$p(Y^{\text{src}}_{\text{obs}} \mid Y^{\text{tot}}_{\text{obs}}, n, N, S_{\text{obs}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, \gamma_{\text{obs}}) = \prod_{i=1}^{n} p(Y^{\text{src}}_{\text{obs,i}} \mid Y^{\text{tot}}_{\text{obs,i}}, n, N, S_{\text{obs,i}}, B_{\text{obs,i}}, \tag{A.19}$$

$$L_{\text{obs,i}}, E_{\text{obs,i}}, \gamma_{\text{obs,i}})$$

$$= \prod_{i=1}^{n} \text{Binomial}(Y^{\text{src}}_{\text{obs,i}}; Y^{\text{tot}}_{\text{obs,i}}, \frac{\lambda_i}{\lambda_i + k_i}) \tag{A.20}$$

### A.1.2 Derivation of Posterior Distribution

The complete data posterior distribution is

$$\begin{aligned} p(N, \theta, \tau, \gamma_{\text{com}}, S_{\text{com}}, I_{\text{com}}, Y^{\text{src}}_{\text{obs}}, Y^{\text{src}}_{\text{mis}}, Y^{\text{tot}}_{\text{mis}}, B_{\text{mis}}, L_{\text{mis}}, E_{\text{mis}}, A_{\text{mis}} \mid & \\ n, Y^{\text{tot}}_{\text{obs}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}}, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma) & \end{aligned} \tag{A.21}$$

where $\text{SD}_{\text{obs}}$ is the spectral data for the observed sources and $(a_\gamma, b_\gamma)$ are the parameters of the prior of $\gamma$. We integrate out the missing data parameters

$$(S_{\text{mis}}, I_{\text{mis}}, Y^{\text{src}}_{\text{mis}}, Y^{\text{tot}}_{\text{mis}}, \gamma_{\text{mis}}, B_{\text{mis}}, L_{\text{mis}}, E_{\text{mis}}, A_{\text{mis}}).$$

This leaves the main parameters of interest $(N, \theta, \tau)$ and the parameters of the flux, flux to counts conversion factors and source photon counts $(S_{obs}, \gamma_{obs}, Y^{src}_{obs})$ of the

observed sources in the marginalised joint posterior. By using this sampling scheme, the dimension of the sampled quantities is kept constant.

The marginalised joint-posterior distribution of the parameters of interest is:

$$
\begin{aligned}
&p(\gamma_{\text{obs}}, N, \theta, \tau, S_{\text{obs}}, Y^{\text{src}}_{\text{obs}} \mid n, Y^{\text{tot}}_{\text{obs}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}}, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma) \\
&= \frac{1}{p(n, Y^{\text{tot}}_{\text{obs}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}}, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma)} \cdot \\
&\int p(\gamma_{\text{com}}, I_{\text{com}}, S_{\text{com}}, Y^{\text{src}}_{\text{com}}, Y^{\text{tot}}_{\text{com}}, B_{\text{com}}, L_{\text{com}}, E_{\text{com}}, A_{\text{com}}, n | N, \theta, \tau, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma) \\
&\cdot p(N, \theta, \tau) \text{d}\gamma_{\text{mis}} \text{d}I_{\text{mis}} \, \text{d}S_{\text{mis}} \, \text{d}B_{\text{mis}} \, \text{d}L_{\text{mis}} \, \text{d}E_{\text{mis}} \, \text{d}A_{\text{mis}} \, \text{d}Y^{\text{src}}_{\text{mis}} \, \text{d}Y^{\text{tot}}_{\text{mis}} \\
&= \frac{1}{p(n, Y^{\text{tot}}_{\text{obs}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}}, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma)} \cdot p(N, \theta, \tau) \cdot \\
&\int p(\gamma_{\text{com}}, I_{\text{com}}, S_{\text{com}}, Y^{\text{src}}_{\text{com}}, Y^{\text{tot}}_{\text{com}}, B_{\text{com}}, L_{\text{com}}, E_{\text{com}}, A_{\text{com}} | N, \theta, \tau, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma) \mathbb{I}_{\{\sum_{i=1}^N I_i = n\}} \\
&\text{d}\gamma_{\text{mis}} \text{d}I_{\text{mis}} \, \text{d}S_{\text{mis}} \, \text{d}B_{\text{mis}} \, \text{d}L_{\text{mis}} \, \text{d}E_{\text{mis}} \, \text{d}A_{\text{mis}} \, \text{d}Y^{\text{src}}_{\text{mis}} \, \text{d}Y^{\text{tot}}_{\text{mis}} \\
&= \frac{1}{p(n, Y^{\text{tot}}_{\text{obs}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}}, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma)} \cdot p(N, \theta, \tau) \cdot \\
&\int_A p(\gamma_{\text{com}}, I_{\text{com}}, S_{\text{com}}, Y^{\text{src}}_{\text{com}}, Y^{\text{tot}}_{\text{com}}, B_{\text{com}}, L_{\text{com}}, E_{\text{com}}, A_{\text{com}} | N, \theta, \tau, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma) \\
&\text{d}\gamma_{\text{mis}} \text{d}I_{\text{mis}} \, \text{d}S_{\text{mis}} \, \text{d}B_{\text{mis}} \, \text{d}L_{\text{mis}} \, \text{d}E_{\text{mis}} \, \text{d}A_{\text{mis}} \, \text{d}Y^{\text{src}}_{\text{mis}} \, \text{d}Y^{\text{tot}}_{\text{mis}}
\end{aligned}
$$

where $A = \{\text{all permutations of vector } I \text{ of length } N \text{ with entries } I_j \in \{0, 1\} \text{ such that } \sum_{j=1}^N I_j = n\}$. Since we do not know which components $S_i$ of the vector $S_{\text{com}}$ are missing, in order to integrate with respect to $I_{\text{mis}}$, we integrate over all the $N-n$ components for which $I_i = 0$. This is done by integrating over all permutations of the vector $I_{\text{com}}$ in which $\sum_{i=1}^N I_i = n$.

Consider the simplest case where $N - n = 1$, i.e. there is only 1 missing value and for the rest components we have that $I_i = 1$. In that case there are only $N$ combinations in which there is a source with $I_i = 0$. Then:

$$\int_A p(\gamma_{\text{com}}, I_{\text{com}}, S_{\text{com}}, Y^{\text{src}}_{\text{com}}, Y^{\text{tot}}_{\text{com}}, B_{\text{com}}, L_{\text{com}}, E_{\text{com}}, A_{\text{com}} | N, \theta, \tau, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma)$$
$$\mathrm{d}\gamma_{\text{mis},1} \mathrm{d}I_{\text{mis},1}\, \mathrm{d}S_{\text{mis},1}\, \mathrm{d}B_{\text{mis},1}\, \mathrm{d}L_{\text{mis},1}\, \mathrm{d}E_{\text{mis},1}\, \mathrm{d}A_{\text{mis},1}\, \mathrm{d}Y^{\text{src}}_{\text{mis},1}\, \mathrm{d}Y^{\text{tot}}_{\text{mis},1}$$

$$= \int_A \prod_{i=1}^{N} p(\gamma_i, I_i, S_i, Y^{\text{src}}_i, Y^{\text{tot}}_i, B_i, L_i, E_i, A_i | N, \theta, \tau, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma)$$
$$\mathrm{d}\gamma_{\text{mis},1}\, \mathrm{d}I_{\text{mis},1}\, \mathrm{d}S_{\text{mis},1}\, \mathrm{d}B_{\text{mis},1}\, \mathrm{d}L_{\text{mis},1}\, \mathrm{d}E_{\text{mis},1}\, \mathrm{d}A_{\text{mis},1}\, \mathrm{d}Y^{\text{src}}_{\text{mis},1}\, \mathrm{d}Y^{\text{tot}}_{\text{mis},1}$$

$$= N \cdot \int \prod_{i=2}^{N} p(I_i = 1, \gamma_i, S_i, Y^{\text{src}}_i, Y^{\text{tot}}_i, B_i, L_i, E_i, A_i | N, \theta, \tau, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma)$$
$$p(I_1 = 0, \gamma_1, S_1, Y^{\text{src}}_1, Y^{\text{tot}}_1, B_1, L_1, E_1, A_1 | N, \theta, \tau, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma)$$
$$\mathrm{d}\gamma_1\, \mathrm{d}S_1\, \mathrm{d}B_1\, \mathrm{d}L_1\, \mathrm{d}E_1\, \mathrm{d}A_1\, \mathrm{d}Y^{\text{src}}_1\, \mathrm{d}Y^{\text{tot}}_1$$

$$= N \cdot p(\gamma_{\text{obs}}, I_{\text{obs}}, S_{\text{obs}}, Y^{\text{src}}_{\text{obs}}, Y^{\text{obs}}_{\text{obs}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}} | N, \theta, \tau, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma)$$
$$\int p(I_1 = 0, \gamma_1, S_1, Y^{\text{src}}_1, Y^{\text{tot}}_1, B_1, L_1, E_1, A_1 | N, \theta, \tau, a_\gamma, b_\gamma)$$
$$\mathrm{d}\gamma_1\, \mathrm{d}S_1\, \mathrm{d}B_1\, \mathrm{d}L_1\, \mathrm{d}E_1\, \mathrm{d}A_1\, \mathrm{d}Y^{\text{src}}_1\, \mathrm{d}Y^{\text{tot}}_1$$

$$= N \cdot p(\gamma_{\text{obs}}, I_{\text{obs}}, S_{\text{obs}}, Y^{\text{src}}_{\text{obs}}, Y^{\text{obs}}_{\text{obs}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}} | N, \theta, \tau, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma)$$
$$\int p(B_1, L_1, E_1, A_1) \cdot p(\gamma_1 \mid a_\gamma, b_\gamma) \Big\{ \int p(I_1 = 0 \mid S_1, \gamma_1, B_1, L_1, E_1) \cdot p(S_1 \mid \theta, \tau) \mathrm{d}S_1$$
$$\cdot \int p(Y^{tot}_1 | N, \theta, S_1, B_1, L_1, E_1)\, \mathrm{d}Y^{tot}_1 \cdot \int p(Y^{src}_1 | Y^{tot}_1, N, \theta, S_1, B_1, L_1, E_1)\, \mathrm{d}Y^{src}_1 \Big\}$$
$$\mathrm{d}\gamma_1\, \mathrm{d}B_1\, \mathrm{d}L_1\, \mathrm{d}E_1\, \mathrm{d}A_1$$

$$= N \cdot p(\gamma_{\text{obs}}, I_{\text{obs}}, S_{\text{obs}}, Y^{\text{src}}_{\text{obs}}, Y^{\text{obs}}_{\text{obs}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}} | N, \theta, \tau, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma)$$
$$\int (1 - g(S_1, \gamma_1, B_1, L_1, E_1)) \cdot p(S_1 \mid \theta, \tau) \cdot p(\gamma_1 \mid a_\gamma, b_\gamma) \cdot p(B_1, L_1, E_1)\, \mathrm{d}S_1\, \mathrm{d}\gamma_1$$
$$\mathrm{d}B_1\, \mathrm{d}L_1\, \mathrm{d}E_1$$

$$= N \cdot p(\gamma_{\text{obs}}, I_{\text{obs}}, S_{\text{obs}}, Y^{\text{src}}_{\text{obs}}, Y^{\text{obs}}_{\text{obs}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}} | N, \theta, \tau, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma)$$
$$\cdot (1 - \pi(\theta, \tau))$$

For the case of $N - n = 2$ we have $\binom{N}{2}$ combinations and by independence we can conclude that the integral over the missing values will simplify to $(1 - \pi(\theta, \tau))^2$.

Following the same logic, for the general $\binom{N}{n}$ we have that the posterior distribution over the parameters of interest would be:

$$
\begin{aligned}
&p(\gamma_{\text{obs}}, N, \theta, \tau, S_{\text{obs}}, Y_{\text{obs}}^{\text{src}} \mid n, Y_{\text{obs}}^{\text{tot}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}}, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma) \qquad \text{(A.22)}\\
&= \frac{1}{p(n, Y_{\text{obs}}^{\text{tot}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}}, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma)} \cdot p(N, \theta, \tau) \cdot \\
&\quad \int_A p(\gamma_{\text{com}}, I_{\text{com}}, S_{\text{com}}, Y_{\text{com}}^{\text{src}}, Y_{\text{com}}^{\text{tot}}, B_{\text{com}}, L_{\text{com}}, E_{\text{com}}, A_{\text{com}} | N, \theta, \tau, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma) \\
&\quad \text{d}\gamma_{\text{mis}} \text{d}I_{\text{mis}}\, \text{d}S_{\text{mis}}\, \text{d}B_{\text{mis}}\, \text{d}L_{\text{mis}}\, \text{d}E_{\text{mis}}\, \text{d}A_{\text{mis}}\, \text{d}Y_{\text{mis}}^{\text{src}}\, \text{d}Y_{\text{mis}}^{\text{tot}} \\
&= \frac{1}{p(n, Y_{\text{obs}}^{\text{tot}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}}, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma)} \cdot p(N, \theta, \tau) \cdot \\
&\quad \binom{N}{n} \cdot p(\gamma_{\text{obs}}, I_{\text{obs}}, S_{\text{obs}}, Y_{\text{obs}}^{\text{src}}, Y_{\text{obs}}^{\text{obs}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}} | N, \theta, \tau, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma) \\
&\quad \cdot (1 - \pi(\theta, \tau))^{N-n} \\
&= \frac{1}{p(n, Y_{\text{obs}}^{\text{tot}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}}, \text{SD}_{\text{obs}}, \mu, \sigma)} \cdot \binom{N}{n} \cdot (1 - \pi(\theta, \tau))^{N-n} \\
&\quad \cdot p(N) \cdot p(\theta) \cdot p(\tau) \cdot p(B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}} | N, \theta, \tau) \\
&\quad \cdot p(\gamma_{\text{obs}} \mid \text{SD}_{\text{obs}}, a_\gamma, b_\gamma) \cdot p(S_{\text{obs}} | \theta, \tau) \\
&\quad \cdot p(I_{\text{obs}} \mid \gamma_{\text{obs}}, S_{\text{obs}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}) \\
&\quad \cdot p(Y_{\text{obs}}^{\text{tot}} \mid I_{\text{obs}}, S_{\text{obs}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, \gamma_{\text{obs}}) \\
&\quad \cdot p(Y_{\text{obs}}^{\text{src}} \mid Y_{\text{obs}}^{\text{tot}}, I_{\text{obs}}, S_{\text{obs}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, \gamma_{\text{obs}})
\end{aligned}
$$

where

$$\pi(\theta, \tau) = \int g(S, B, L, E, \gamma) \cdot p(\gamma) \cdot p(S | \theta, \tau) \cdot p(B, L, E)\, \text{d}S\, \text{d}B\, \text{d}E\, \text{d}L\, \text{d}\gamma$$

- **Single power law model**: the Equation A.23 becomes:

$$
\begin{aligned}
&p(\gamma_{\text{obs}}, N, \theta, \tau, S_{\text{obs}}, Y_{\text{obs}}^{\text{src}} \mid n, Y_{\text{obs}}^{\text{tot}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}}, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma) \\
&\propto \frac{1}{p(n, Y_{\text{obs}}^{\text{tot}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}}, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma)} \cdot \binom{N}{n} \cdot (1 - \pi(\theta, \tau))^{N-n} \\
&\quad \cdot \binom{N + a_N - 1}{a_N - 1} \Big(\frac{1}{1 + b_N}\Big)^N \Big(\frac{b_N}{1 + b_N}\Big)^{a_N} \mathbb{I}_{\{N \in \mathbb{Z}^+\}} \\
&\quad \cdot \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} \mathbb{I}_{\{\theta > 0\}} \\
&\quad \cdot \frac{b_m^{a_m}}{\Gamma(a_m)} \tau^{a_m - 1} e^{-b_m \tau} \mathbb{I}_{\{\tau > 0\}} \\
&\quad \cdot \prod_{i=1}^{n} p(\gamma_i | \text{SD}_i, a_\gamma, b_\gamma) \cdot \theta \tau^\theta S_i^{-(\theta+1)} \mathbb{I}_{\{\tau < S_i\}} \cdot g(S_i, B_i, L_i, E_i, \gamma_i) \\
&\quad \cdot \frac{(\frac{S_i E_i}{\gamma_i} + B_i A_i)^{Y_i^{\text{tot}}}}{Y_i^{\text{tot}}!} e^{(\frac{S_i E_i}{\gamma_i} + B_i A_i)} \mathbb{I}_{\{Y_i^{tot} \in \mathbb{Z}^+\}} \\
&\quad \cdot \binom{Y_i^{\text{tot}}}{Y_i^{\text{src}}} \Big(\frac{\frac{S_i E_i}{\gamma_i}}{\frac{S_i E_i}{\gamma_i} + B_i A_i}\Big)^{Y_i^{\text{src}}} \Big(1 - \frac{\frac{S_i E_i}{\gamma_i}}{\frac{S_i E_i}{\gamma_i} + B_i A_i}\Big)^{Y_i^{\text{tot}} - Y_i^{\text{src}}} \mathbb{I}_{\{Y_i^{\text{src}} \in \{0, 1, \ldots, Y_i^{\text{tot}}\}\}}
\end{aligned}
\tag{A.23}
$$

- **Broken power law model**: the Equation A.23 becomes:

$$
\begin{aligned}
&p(\gamma_{\text{obs}}, N, \theta, \tau, S_{\text{obs}}, Y_{\text{obs}}^{\text{src}} \mid n, Y_{\text{obs}}^{\text{tot}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}}, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma) \\
&\propto \frac{1}{p(n, Y_{\text{obs}}^{\text{tot}}, B_{\text{obs}}, L_{\text{obs}}, E_{\text{obs}}, A_{\text{obs}}, \text{SD}_{\text{obs}}, a_\gamma, b_\gamma)} \cdot \binom{N}{n} \cdot (1 - \pi(\theta, \tau))^{N-n} \\
&\cdot \binom{N + a_N - 1}{a_N - 1} \Big(\frac{1}{1 + b_N}\Big)^N \Big(\frac{b_N}{1 + b_N}\Big)^{a_N} \mathbb{I}_{\{N \in \mathbb{Z}^+\}} \\
&\cdot \prod_{j=1}^{m} \Big[\frac{b_j^{a_j}}{\Gamma(a_j)} \theta_j^{a_j - 1} e^{-b_j \theta} \mathbb{I}_{\{\theta_j > 0\}}\Big] \cdot p(\tau_1, \tau_2, ..., \tau_m) \mathbb{I}_{\{0 < \tau_1 < \tau_2 < ... < \tau_m\}} \\
&\cdot \prod_{i=1}^{n} \Big[p(\gamma_i | \text{SD}_i, a_\gamma, b_\gamma) \cdot g(S_i, B_i, L_i, E_i, \gamma_i) \\
&\cdot \sum_{j=1}^{m} \Big[\prod_{i=1}^{j-1} \Big(\frac{\tau_{i+1}}{\tau_i}\Big)^{-\theta_i}\Big] \Big(\frac{\theta_j}{\tau_j}\Big) \Big(\frac{S}{\tau_j}\Big)^{-(\theta_j + 1)} \mathbb{I}_{\{\tau_j \leq S \leq \tau_{j+1}\}} \\
&\cdot \frac{(\frac{S_i E_i}{\gamma_i} + B_i A_i)^{Y_i^{\text{tot}}}}{Y_i^{\text{tot}}!} e^{(\frac{S_i E_i}{\gamma_i} + B_i A_i)} \mathbb{I}_{\{Y_i^{tot} \in \mathbb{Z}^+\}} \\
&\cdot \binom{Y_i^{\text{tot}}}{Y_i^{\text{src}}} \Big(\frac{\frac{S_i E_i}{\gamma_i}}{\frac{S_i E_i}{\gamma_i} + B_i A_i}\Big)^{Y_i^{\text{src}}} \Big(1 - \frac{\frac{S_i E_i}{\gamma_i}}{\frac{S_i E_i}{\gamma_i} + B_i A_i}\Big)^{Y_i^{\text{tot}} - Y_i^{\text{src}}} \mathbb{I}_{\{Y_i^{\text{src}} \in \{0,1,...,Y_i^{\text{tot}}\}\}}\Big]
\end{aligned}
\tag{A.24}
$$

### A.1.3 Derivations of the Conditional Posterior Distributions for Single Power Law Model

**Conditional distribution of $Y_{obs}^{src}$**: The full conditional distribution for $Y_{obs}^{src}$ is:

$$
\begin{aligned}
p(Y_{obs}^{src} | \cdot) &\propto p(Y_{obs}^{src} | Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs}, I_{obs}, S_{obs}, \gamma_{\text{obs}}) \\
&\propto \prod_{i=1}^{n} \binom{Y_i^{tot}}{Y_i^{src}} \Big(\frac{\lambda_i}{\lambda_i + \kappa_i}\Big)^{Y_i^{src}} \Big(1 - \frac{\lambda_i}{\lambda_i + \kappa_i}\Big)^{Y_i^{tot} - Y_i^{src}} \mathbb{I}_{\{Y_i^{src} \in \{0,1,...,Y_i^{tot}\}\}} \\
&= \prod_{i=1}^{n} Binomial\Big(Y_i^{src}; Y_i^{tot}, \frac{\lambda_i}{\lambda_1 + \kappa_i}\Big),
\end{aligned}
$$

where $\lambda_i = S_i E_i / \gamma_i$ and $k_i = B_i A_i$. Since the observed sources are independent we can sample the vector $Y_{obs}^{src}$ component-wise for $i = 1, ..., n$ as

$$p(Y_i^{src}|\cdot) \sim Binomial\Big(Y_i^{src}; Y_i^{tot}, \frac{\lambda_i}{\lambda_1 + \kappa_i}\Big).$$

**Conditional distribution of $S_{obs}$**: The full conditional distribution for $S_{obs}$ is:

$$\begin{aligned}
p(S_{obs}|\cdot) &\propto p(S_{obs}|N, \theta, \tau) \cdot p(I_{obs}|S_{obs}, B_{obs}, L_{obs}, E_{obs}, \gamma_{\text{obs}}) \\
&\cdot p(Y_{obs}^{tot}|B_{obs}, L_{obs}, E_{obs}, I_{obs}, S_{obs}, \gamma_{\text{obs}}) \cdot p(Y_{obs}^{src}|Y_{obs}^{tot}, B_{obs}, L_{obs}, E_{obs}, I_{obs}, S_{obs}, \\
&\quad \gamma_{\text{obs}}) \\
&\propto \prod_{i=1}^{n} \theta \tau^{\theta} S_i^{-(\theta+1)} \mathbb{I}_{\{\tau < S_i\}} \cdot g(S_i, B_i, L_i, E_i, \gamma_i) \cdot \frac{(\lambda_i + \kappa_i)^{Y_i^{tot}}}{Y_i^{tot}!} e^{(\lambda_i + \kappa_i)} \mathbb{I}_{\{Y_i^{tot} \in \mathbb{Z}^+\}} \\
&\cdot \binom{Y_i^{tot}}{Y_i^{src}} \Big(\frac{\lambda_i}{\lambda_i + \kappa_i}\Big)^{Y_i^{src}} \Big(1 - \frac{\lambda_i}{\lambda_i + \kappa_i}\Big)^{Y_i^{tot} - Y_i^{src}} \mathbb{I}_{\{Y_i^{src} \in \{0,1,\ldots,Y_i^{tot}\}\}}
\end{aligned}$$

where $\lambda(S_i, E_i, B_i, L_i) = S_i E_i / \gamma_i$ and $k(B_i, A_i) = B_i A_i$. By independence of the $S_{obs}$ we can sample component-wise for $i = 1, ..., n$ as

$$\begin{aligned}
p(S_i|\cdot) &\sim Pareto(S_i|N, \theta, \tau) \mathbb{I}_{\{\tau < S_i\}} \cdot g(S_i, B_i, L_i, E_i, \gamma_i) \cdot Poisson(Y_i^{tot}; \lambda_i + \kappa_i) \\
&\cdot Binomial\Big(Y_i^{src}; Y_i^{tot}, \frac{\lambda_i}{\lambda_1 + \kappa_i}\Big).
\end{aligned}$$

**Conditional distribution of $\boldsymbol{\theta}$**: The full conditional distribution for $\theta$ is:

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

**Conditional distribution of $\boldsymbol{N}$**: The full conditional distribution for $N$, the total unknown number of sources, is:

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

**Conditional distribution of $\boldsymbol{\tau}$**: The full conditional distribution for $\tau$ is:

$$\begin{aligned}
p(\tau|\cdot) &\propto p(\tau) \cdot (1-\pi(\theta,\tau))^{N-n} \cdot p(S_{obs}|N,\theta,\tau) \\
&\propto \frac{b_m^{a_m}}{\Gamma(a_m)}\tau^{a_m-1}e^{-b_m\tau}\mathbb{I}_{\{\tau>0\}} \cdot (1-\pi(\theta,\tau))^{N-n} \cdot \Big[\prod_{i=1}^{n}\theta\tau^{\theta}S_i^{-(\theta+1)}\mathbb{I}_{\{\tau<S_i\}}\Big] \\
&\propto \tau^{n\theta+a_m-1} \cdot e^{-b_m\tau} \cdot (1-\pi(\theta,\tau))^{N-n} \cdot \mathbb{I}_{\{\tau<c_m\}}, \quad \text{where } c_m = min(S_1,...,S_n)
\end{aligned}$$

**Conditional posterior distribution of $\boldsymbol{\gamma_{obs,i}}$**: For the observed sources $i = 1, ..., n$ we have the priors $p(\gamma_{obs,i}|\cdot)$. So, the full conditional posterior distribution is

$$
\begin{aligned}
p(\gamma_{obs}|\cdot) \propto & \prod_{i=1}^{n} p(\gamma_i|SD_i, a_\gamma, b_\gamma) \cdot g(S_i, B_i, L_i, E_i, \gamma_i) \\
& \cdot \frac{(\frac{S_i E_i}{\gamma_i} + B_i A_i)^{Y_i^{tot}}}{Y_i^{tot}!} e^{(\frac{S_i E_i}{\gamma_i} + B_i A_i)} \mathbb{I}_{\{Y_i^{tot} \in \mathbb{Z}^+\}} \\
& \cdot \binom{Y_i^{tot}}{Y_i^{src}} \Big( \frac{\frac{S_i E_i}{\gamma_i}}{\frac{S_i E_i}{\gamma_i} + B_i A_i} \Big)^{Y_i^{src}} \Big( 1 - \frac{\frac{S_i E_i}{\gamma_i}}{\frac{S_i E_i}{\gamma_i} + B_i A_i} \Big)^{Y_i^{tot} - Y_i^{src}} \mathbb{I}_{\{Y_i^{src} \in \{0,1,...,Y_i^{tot}\}\}}
\end{aligned}
$$

By independence of the observed sources, we can sample component-wise for $i = 1, ..., n$ as:

$$
\begin{aligned}
p(\gamma_i|\cdot) \propto & \, p(\gamma_i|SD_i, a_\gamma, b_\gamma) \cdot g(S_i, B_i, L_i, E_i, \gamma_i) \cdot Poisson(Y_i^{tot}; \frac{S_i E_i}{\gamma_i} + B_i A_i) \\
& \cdot Binomial\Big( Y_i^{src}; Y_i^{tot}, \frac{\frac{S_i E_i}{\gamma_i}}{\frac{S_i E_i}{\gamma_i} + B_i A_i} \Big).
\end{aligned}
$$

### A.1.4 Derivations of the Conditional Posterior Distributions for Broken Power Law Model

**Conditional posterior distribution of $\boldsymbol{\theta = (\theta_1, ..., \theta_m)^T}$**: The full conditional posterior distribution for $\theta = (\theta_1, ..., \theta_m)$ is:

$$p(\theta \mid) \propto [1 - \pi(\theta, \tau)]^{N-n} \cdot \prod_{j=1}^{m} \theta_j^{a_j - 1} e^{-b_j \theta_j} \mathbb{I}_{\{\theta_j > 0\}} \cdot \prod_{i=1}^{n} \Big[ \sum_{j=1}^{m} \Big[ \prod_{l=1}^{j-1} \Big( \frac{\tau_{l+1}}{\tau_l} \Big)^{-\theta_l} \Big] \Big( \frac{\theta_j}{\tau_j} \Big) \Big( \frac{S}{\tau_j} \Big)^{-(\theta_j + 1)} \mathbb{I}_{\{\tau_j \leq S \leq \tau_{j+1}\}} \Big]$$

The indicators $\mathbb{I}_{\{\tau_j \leq S \leq \tau_{j+1}\}}$ eliminate all the terms in the sum but one. Thus, we have that

$$p(\theta \mid) \propto [1 - \pi(\theta, \tau)]^{N-n} \cdot \prod_{j=1}^{m} \theta_j^{a_j - 1} e^{-b_j \theta_j} \mathbb{I}_{\{\theta_j > 0\}} \cdot \prod_{j=1}^{m} \Big\{ \prod_{l=1}^{j-1} \Big( \frac{\tau_{l+1}}{\tau_l} \Big)^{-\theta_l} \Big\}^{n(j)} \cdot \prod_{i \in I(j)} \Big( \frac{\theta_j}{\tau_j} \Big) \Big( \frac{S}{\tau_j} \Big)^{-(\theta_j + 1)}$$

where $I(j) = \{i \ : \ \tau_j \leq S_i \leq \tau_{j+1}\}$ denotes the existence the sources with flux contained in the interval corresponding to the $j$ -th mixture component, and $n(j)$ is the cardinality of $I(j)$ (the number of sources in that interval). By re-writing the above equation, we get

$$
\begin{aligned}
p(\theta\mid) &\propto [1-\pi(\theta,\tau)]^{N-n}\cdot\prod_{j=1}^{m}\theta_j^{a_j+n(j)-1}\mathbb{I}_{\{\theta_j>0\}}\cdot\prod_{j=1}^{m}e^{-b_j\theta_j}\\
&\quad\cdot\prod_{j=1}^{m}\Big\{\prod_{l=1}^{j-1}\Big(\frac{\tau_{l+1}}{\tau_l}\Big)^{-\theta_l}\Big\}^{n(j)}\cdot\prod_{j=1}^{m}e^{-\theta_j\sum_{i\in I(j)}\log(\frac{S_i}{\tau_j})}\\
&\propto [1-\pi(\theta,\tau)]^{N-n}\cdot\prod_{j=1}^{m}\theta_j^{a_j+n(j)-1}\mathbb{I}_{\{\theta_j>0\}}\\
&\quad\cdot\exp\Big(-\Big[\theta_j b_j+n(j)\mathbb{I}_{\{j>0\}}\sum_{l=1}^{j-1}\theta_l\log(\frac{\tau_{l+1}}{\tau_l})+\theta_j\sum_{i\in I(j)}\log(\frac{S_i}{\tau_j})\Big]\Big)\\
&\propto [1-\pi(\theta,\tau)]^{N-n}\cdot\prod_{j=1}^{m}\mathrm{Gamma}\Big(\theta_j;a_j+n(j),\\
&\quad b_j+\mathbb{I}_{\{j\neq m\}}\log(\frac{\tau_{j+1}}{\tau_j})\sum_{i=1}^{m}[n(i)\mathbb{I}_{\{i\geq j+1\}}]+\sum_{i\in I(j)}\log(\frac{S_i}{\tau_j})\Big)
\end{aligned}
$$

**Conditional posterior distribution of $\boldsymbol{\tau_1}$**: The full conditional posterior distribution for $\tau_1$ is:

$$
\begin{aligned}
p(\tau_1|\cdot) &\propto [1-\pi(\theta,\tau)]^{N-n}\cdot p(\tau_1,\tau_2,...,\tau_m)\mathbb{I}_{\{0<\tau_1<\tau_2<...<\tau_m\}}\cdot\Big[\prod_{i=1}^{n}p(B_i,L_i,E_i)\\
&\quad\cdot g(S_i,B_i,L_i,E_i)\cdot\sum_{j=1}^{m}\Big[\prod_{i=1}^{j-1}\Big(\frac{\tau_{i+1}}{\tau_i}\Big)^{-\theta_i}\Big]\Big(\frac{\theta_j}{\tau_j}\Big)\Big(\frac{S}{\tau_j}\Big)^{-(\theta_j+1)}\mathbb{I}_{\{\tau_j\leq S\leq\tau_{j+1}\}}\Big]\\
&\propto [1-\pi(\theta,\tau)]^{N-n}\cdot\tau_1^{a_m-1}e^{-b_m\tau_1}\cdot\prod_{j=2}^{m}e^{-\frac{c_m^2\cdot[\log(\tau_m-\tau_{(m-1)})-\mu_m]^2}{2}}\cdot\frac{1}{\tau_m-\tau_{(m-1)}}\\
&\quad\cdot\mathbb{I}_{\{\tau<c_m\}}\cdot\prod_{j=1}^{m}\Big\{\prod_{l=1}^{j-1}\Big(\frac{\tau_{l+1}}{\tau_l}\Big)^{-\theta_l}\Big\}^{n(j)}\cdot\prod_{i\in I(j)}\Big(\frac{\theta_j}{\tau_j}\Big)\Big(\frac{S}{\tau_j}\Big)^{-(\theta_j+1)}\\
&\propto \tau^{n\theta_1+a_m-1}\cdot e^{-b_m\tau}\cdot(1-\pi(\theta,\tau))^{N-n}\cdot\prod_{j=2}^{m}e^{-\frac{c_m^2\cdot[\log(\tau_m-\tau_{(m-1)})-\mu_m]^2}{2}}\\
&\quad\cdot\frac{1}{\tau_m-\tau_{(m-1)}}\cdot\mathbb{I}_{\{\tau<c_m\}}
\end{aligned}
$$

**Conditional posterior distribution of $(\tau_2, ...\tau_m)$**: We sample $(\tau_2, ...\tau_m)$ via the full joint conditional posterior distribution of the transformed variables $\eta_2, ..., \eta_m$. We remind that $\eta_j = h_j(\tau_j \mid \tau_{j-1}) = \log(\tau_j - \tau_{j-1})$, $j = 2, ..., m$. Thus, after applying a change of variables we have:

$$
\begin{aligned}
p(\eta_2, ...\eta_m|\cdot) &\propto p(\tau_2, ...\tau_m|\cdot) \cdot e^{\sum_{j=2}^{m} \eta_j} \\
&\propto e^{\sum_{j=2}^{m} \eta_j} \cdot [1 - \pi(\theta, \tau)]^{N-n} \cdot p(\eta_2, ...\eta_m) \cdot \prod_{j=1}^{m} \Big\{ \prod_{l=1}^{j-1} \Big( \frac{\tau_{l+1}}{\tau_l} \Big)^{-\theta_l} \Big\}^{n(j)} \\
&\quad \cdot \prod_{i \in I(j)} \Big( \frac{\theta_j}{\tau_j} \Big) \Big( \frac{S}{\tau_j} \Big)^{-(\theta_j + 1)} \\
&\propto e^{\sum_{j=2}^{m} \eta_j} \cdot [1 - \pi(\theta, \tau)]^{N-n} \cdot \text{Multivariate Gaussian}(\mu, C) \\
&\quad \cdot \prod_{j=1}^{m} \Big\{ \prod_{l=1}^{j-1} \Big( \frac{\tau_{l+1}}{\tau_l} \Big)^{-\theta_l} \Big\}^{n(j)} \cdot \prod_{i \in I(j)} \Big( \frac{\theta_j}{\tau_j} \Big) \Big( \frac{S}{\tau_j} \Big)^{-(\theta_j + 1)}
\end{aligned}
$$