

THE UNIVERSITY OF CHICAGO

POINT PROCESS MODELS FOR ASTRONOMY:
QUASARS, CORONAL MASS EJECTIONS, AND SOLAR FLARES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
ANGELA BETH HUGEBACK

CHICAGO, ILLINOIS

AUGUST 2007

This thesis is dedicated to my mom, Judy Snyder.

ABSTRACT

This thesis presents a statistical analysis of two interesting astronomical applications that involve point process data.

The first chapter presents an application in solar physics that looks at two types of solar events: solar flares and coronal mass ejections (CMEs). The data are viewed as a marked point process in time, and the analysis seeks to determine whether there is a local temporal correlation between the two event types. A key statistical complication is that the two processes follow an inhomogeneous long term rate function that changes with the solar cycle. Our goal is to evaluate the local temporal correlation, adjusting for the fluctuation in intensity due to changes in the long term solar cycle. We present a novel approach that begins by aggregating the relative onset times for CME events relative to solar flare events in a small local window. Any local temporal correlation between the two processes would be represented by a dip or a bump in the histogram of those relative time points. We present two functional forms to model the distribution of the relative time points, and model parameters are estimated by the method of maximum likelihood. Statistical significance of the local temporal correlation was assessed via a bootstrap approach that attempts to mimic the long term trends in the data while removing any local dependence that exists. We find that there is indeed a statistically significant increase in the probability that a CME will occur in the minutes preceding the onset of a solar flare. A second bootstrap approach was implemented in order to construct confidence intervals for the fitted parameter estimates in order to provide a descriptive model for the nature of the local temporal correlation.

The second chapter presents a statistical analysis of the Quasar Luminosity Function. This application falls under the subfield of cosmology, and the thrust of the analysis is density estimation, model selection, and the detection of interaction effects in a two-dimensional point process. Following the work of Richards et al. [24], this paper presents a flexible parametric model of quasars as a non-homogeneous Poisson process over the space of luminosity and redshift. Our contribution to this work is two-fold: first, to allow

the introduction of higher-order terms as warranted by the data, and second, to extend the model in order to detect and describe any interaction effect between luminosity and redshift present in the data. Major complicating factors in this analysis are data truncation that occurs when quasars are too faint or too far away to be detected from Earth, and sampling bias due to the fact that quasars in the data set were not selected uniformly at random from the population of all quasars in the universe.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my thesis advisor, Dr. Marc A. Coram, for doing his best to teach me to think like a statistician, and in particular for making the entire experience so much fun. It is through Marc's consistent encouragement and endless enthusiasm that I have come to discover my own research interests and a true love for the field of statistics. I am forever indebted and will be forever thankful.

I would like to thank my doctoral committee members, Dr. Michael L. Stein and Dr. Steven P. Lalley for their generous support, excellent statistical advice, and careful critique of my work throughout the dissertation process.

I would like to extend a special thank you to Dr. C. Alex Young for giving me the special opportunity to spend the summer of 2006 conducting statistical research with him at the NASA Goddard Space Flight Center. Alex introduced me to the solar flare and coronal mass ejection data analyzed in this thesis, and his support and expert knowledge of solar and space physics have made this research possible.

I would also like to thank Dr. Sebastian Jester – formerly of the Experimental Astrophysics Group at FermiLab and currently at the Max Planck Institute for Astronomy in Heidelberg, Germany – for inspiring my interest in the estimation of the Quasar Luminosity Function, and for providing me with the data and background knowledge in astronomy that have made that work possible.

I would like to thank Dr. G Jogesh Babu and Dr. Eric D. Feigelson of Penn State for coordinating the Statistical Challenges in Modern Astronomy conference series and tutorials which introduced me to numerous friendly folks in both astronomy and statistics and gave me a sense of community.

And finally, I would like to thank my family and good friends who have kept me happy and healthy throughout the years: Justin, Jeff, Mom and Dad, Mark, Oli and Tinna, Sarah, Elliot, David, Kris and Jon, Christina, Glenn, and Claire. Thanks guys!

TABLE OF CONTENTS

ABSTRACT	iii
ACKNOWLEDGEMENTS	v
LIST OF FIGURES	viii
LIST OF TABLES	x
Chapter	
1 DETECTING A TEMPORAL CORRELATION BETWEEN SOLAR FLARES AND CORONAL MASS EJECTIONS	1
1.1 Introduction	1
1.2 Literature Review	2
1.3 The Data	2
1.3.1 The GOES Solar Flare Catalog	2
1.3.2 The Catholic University CDAW Coronal Mass Ejection Catalog	4
1.4 Preliminary Investigation	4
1.5 Methods	6
1.5.1 Detecting a Temporal Correlation Between Solar Flares and CMEs	6
1.5.2 Two Models for the Distribution of Relative Onset Times	8
1.5.3 A Numerical Approach to Parameter Estimation	11
1.5.4 Assessing the Significance of a Local Temporal Correlation	12
1.5.5 Estimating the Uncertainty in the Parameter Estimates	16
1.6 Summary	19

2	ESTIMATING THE QUASAR LUMINOSITY FUNCTION	20
2.1	Introduction	20
2.1.1	Background	20
2.1.2	Scientific Goals	22
2.1.3	Overview of Methods	23
2.2	The Data	23
2.2.1	The Sloan Digital Sky Survey	23
2.2.2	Additional Astronomical Considerations	24
2.2.3	Observational Probabilities	28
2.2.4	Final Data for Analysis	29
2.3	Methods	29
2.3.1	A Poisson Process Model for the Quasar Luminosity Function	29
2.3.2	Estimation of Poisson Intensities on a Grid	35
2.3.3	A Flexible Parameterization for the Quasar Luminosity	38
2.3.4	Maximizing the Likelihood	40
2.3.5	Fitted Models for the Quasar Luminosity Function	46
2.3.6	Fitted Models for Partitioned Quasar Data	51
2.4	Simulation Studies	59
2.4.1	Technique for Simulating Quasar Data	59
2.4.2	Performance of Model Fitting Methodology on Quasar Data Simulated under Randomly Generated Models	59
2.4.3	Performance of Model Fitting Methodology on Quasar Data Simulated under SDSS Fitted Model	60
2.5	Summary	60
Appendix		
A	MATHEMATICAL DETAILS	68
A.1	Poisson Process on a Transformed Space	68
B	ADDITIONAL TABLES	72
C	INDEX OF SYMBOLIC NOTATION	80
	REFERENCES	82

LIST OF FIGURES

1.1	GOES solar flare data with associated intensity threshold	3
1.2	Sliding window estimates of CME and flare rates over time	5
1.3	A typical 100 day interval of CDAW coronal mass ejections and GOES solar flare events	7
1.4	Histogram and fitted densities for the relative time points $\tau_D \in \mathbb{T}$ of CDAW CME onset times relative to GOES flare onset times	9
1.5	Diagram of the bootstrap procedure for evaluating the significance of the local temporal correlation	15
1.6	Diagram of the bootstrap procedure for constructing confidence intervals for the parameter estimates	18
2.1	Hubble telescope images of a quasar	21
2.2	The K correction as a function of redshift	25
2.3	The distance modulus DM as a function of redshift	26
2.4	The volume differential γ as a function of redshift	27
2.5	Heat map of the estimated observational probabilities on the apparent magnitude scale	30
2.6	Heat map of the estimated observational probabilities on the absolute magnitude scale	31
2.7	Scatterplot of SDSS quasars on the apparent magnitude scale	32
2.8	Scatterplot of SDSS quasars on the absolute magnitude scale	33
2.9	Preliminary Analysis: Maximum likelihood estimates of quasar rate assuming independent grid squares	37
2.10	Heat map of fitted density under the no-interaction model	48
2.11	Deviance residuals for fitted density under the no-interaction model	50
2.12	Heat map of fitted density under the model that allows for an interaction between absolute magnitude M and redshift z	52
2.13	Heat map of second-best fitted density under the model that allows for an interaction between absolute magnitude M and redshift z	53
2.14	Deviance residuals for fitted density under the model that allows for an interaction between absolute magnitude M and redshift z	54
2.15	Side-by-side heat map of fitted densities for partitioned quasar data under the model that allows for an interaction between absolute magnitude M and redshift z	56
2.16	Scatterplot of simulated quasar dataset #1 under a randomly generated model	61

2.17	Scatterplot of simulated quasar dataset #2 under a randomly generated model	62
2.18	Scatterplot of simulated quasar dataset #1 under the fitted model for the SDSS quasar data	65
2.19	Scatterplot of simulated quasar dataset #2 under the fitted model for the SDSS quasar data	66

LIST OF TABLES

1.1	Maximum likelihood estimates of parameters under two different functional forms for the distribution of \mathbb{T}	12
1.2	Maximum likelihood estimates and bootstrapped 95% confidence intervals for model parameters	18
2.1	BIC table for complete data under the no-interaction model	47
2.2	BIC table for complete data under the model which allows for an interaction between absolute magnitude M and redshift z	49
2.3	BIC table for low-redshift group in the partitioned analysis	57
2.4	BIC table for high-redshift group in the partitioned analysis	58
2.5	Fitted model complexities for simulated quasar data	63
2.6	Comparison of true model complexities to fitted model complexities for simulated quasar data	64
2.7	Fitted model complexities for simulated quasar data under SDSS fitted model	64
B.1	Example GOES soft X-ray flare data	73
B.2	Example Catholic University CDAW coronal mass ejection data	74
B.3	Example quasar data from the SDSS Data Release III	75
B.4	Example entries in the distance modulus DM table	76
B.5	Example entries in the K correction table	77
B.6	Example entries in the volume differential $\gamma(z)$ table	78
B.7	Example entries in the observation probabilities $\psi(m, z)$ table	79
C.1	Symbolic Notation for Chapter 1	81
C.2	Symbolic Notation for Chapter 2	81

CHAPTER 1

DETECTING A TEMPORAL CORRELATION BETWEEN SOLAR FLARES AND CORONAL MASS EJECTIONS

1.1 Introduction

Solar flares [10] [14] are enormous explosions that occur on the surface of the Sun and usually last for a few minutes. Coronal mass ejections (CMEs) [13] are the expulsion of tremendous bubbles of plasma from the solar corona that may last for several hours. Both types of events are of interest to solar and space physicists [11] [15], as they have the potential to result in geomagnetic storms that cause electrical disturbances on or near Earth.

There is currently widespread debate in the solar physics community about whether solar flares and CMEs are temporally associated. Better descriptions of the time trends in and relationships between these two processes would increase scientists' understanding of the way in which magnetic energy is released in the solar corona and allow for more accurate predictions of CMEs and flares.

Our analysis demonstrates that there is a statistically significant local temporal correlation between the onset times of coronal mass ejections and solar flares, and provides some insight into the nature of this correlation. A novel statistical approach is presented for assessing local dependence between two types of events in a marked point process [22] [4] with an inhomogeneous large-scale intensity function. Block bootstrap techniques [7] are implemented in order to assess the significance of the local temporal correlation and to form confidence intervals for parameter estimates in a descriptive model of that dependence.

1.2 Literature Review

There is currently no published work that attempts to analyze the marked point process [22] [4] that is obtained by superimposing the onset times of both solar flares and CMEs. Wheatland, Sturrock, and McTiernan [28] used the Bayesian blocks technique proposed by Scargle [26] to fit a piecewise constant intensity function to solar flare data, and discovered an overabundance of short waiting times than would be expected under the independent Poisson hypothesis.

In later works, Wheatland examined the local behavior of solar flares [30] and CMEs [31] and hypothesized that there may be an additional obscuration effect [29] in the solar flare data, whereby the instrument experiences a delay in the ability to detect small flares after a large flare is observed. In his 2004 paper, Wheatland [32] concludes that a piecewise-constant Poisson process provides a good model for solar flaring activity associated with a particular active region.

1.3 The Data

1.3.1 The GOES Solar Flare Catalog

The solar flare data analyzed in this paper consists of 6,506 soft X-ray flares from the Geostationary Operational Environmental Satellites (GOES) catalog that were observed between April 1, 1999 and December 31, 2003. The solar flare is presented in Figure 1.1 according to onset time and soft X-ray intensity level. The catalog also include measures of peak time, end time, and physical location on the Sun.

In order to avoid any bias that could result from fluctuations in background levels of solar X-Ray emission that interfere with the detection of small solar flares, only those flares with a peak flux greater than a threshold value of -5.6 on the \log_{10} scale ($10^{-5.6} W/m^2$, corresponding to a GOES C2.51 class flare) were included in the study. This threshold is indicated by a red horizontal line in Figure 1.1.

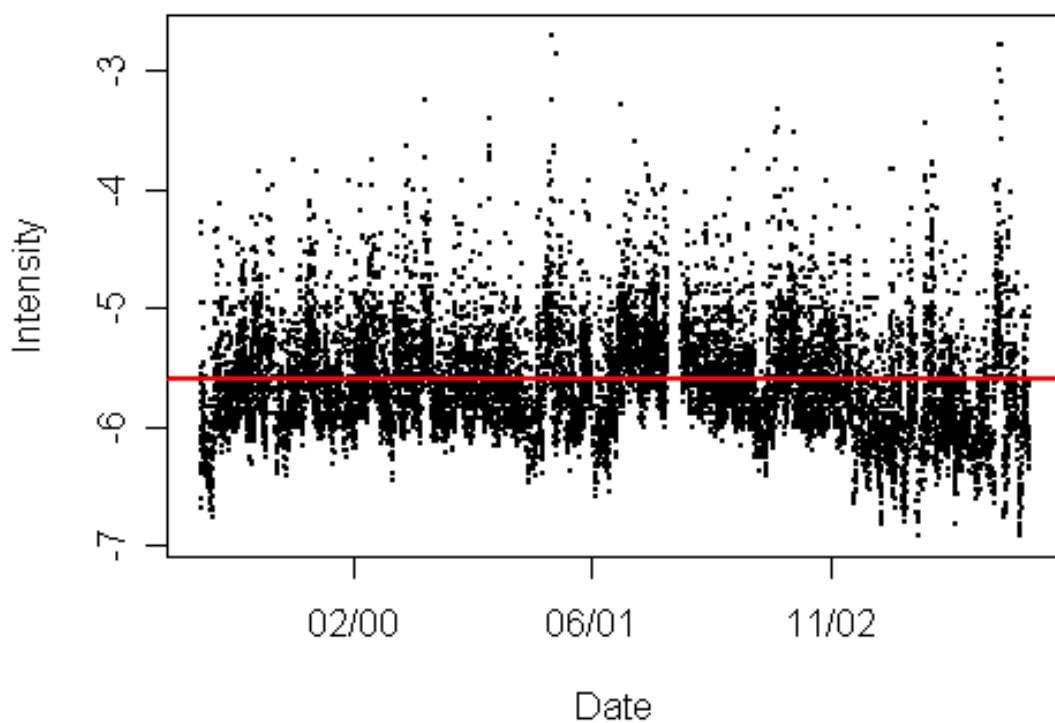


Figure 1.1: GOES solar flare events with onset time on the horizontal axis and \log_{10} soft X-ray intensity on the vertical axis. The red horizontal line indicates the intensity threshold of -5.6 on the \log_{10} scale ($10^{-5.6} \text{ W/m}^2$, corresponding to a GOES C2.51 class flare). Only those flares with a \log_{10} intensity above -5.6 were used in this analysis.

1.3.2 The Catholic University CDAW Coronal Mass Ejection Catalog

The coronal mass ejection data used in this analysis comes from the CDAW CME catalog which is generated and maintained by NASA and The Catholic University of America. This catalog contains all CMEs manually identified since 1996 from the Large Angle and Spectrometric Coronagraph (LASCO) on board the Solar and Heliospheric Observatory (SOHO) mission.

The portion of the CDAW CME catalog used in this analysis contains 6,604 CME events that were observed between April 1, 1999 and December 31, 2003. The catalog contains time stamps for the first time of appearance of the CME in the C2 field of view of the SOHO-LASCO instrument, as well as an estimate of the actual onset time for each CME. The estimated onset times were obtained using a linear extrapolation based on the observed velocity of each CME at the time of first observation. All time stamps have a one-minute temporal resolution, but it should be noted that the LASCO instrument only takes an image of the Sun about once every 20 minutes, and therefore the times of first appearance for CMEs in the C2 field of view are recorded in 20 minute intervals. The catalog also contains measurements of CME velocity (measured in km/s) and principle angle (the positional angle on the Sun measured in degrees counterclockwise from the north pole).

1.4 Preliminary Investigation

In order to gain some intuition about the behavior of the two solar processes over time, their Poisson intensity functions were estimated using a 100-day sliding window to estimate the average number of events per day for each type of solar event. The results are shown in Figure 1.2. It is known that both CMEs and solar flares follow trends in the solar cycle, and therefore it is expected that we would observe a tendency for the rates of the two processes to follow each other. An interesting observation is that the flaring rate increased sharply during the solar maximum which occurred at the end of 2001, while the CME rate remained relatively stable during that time.

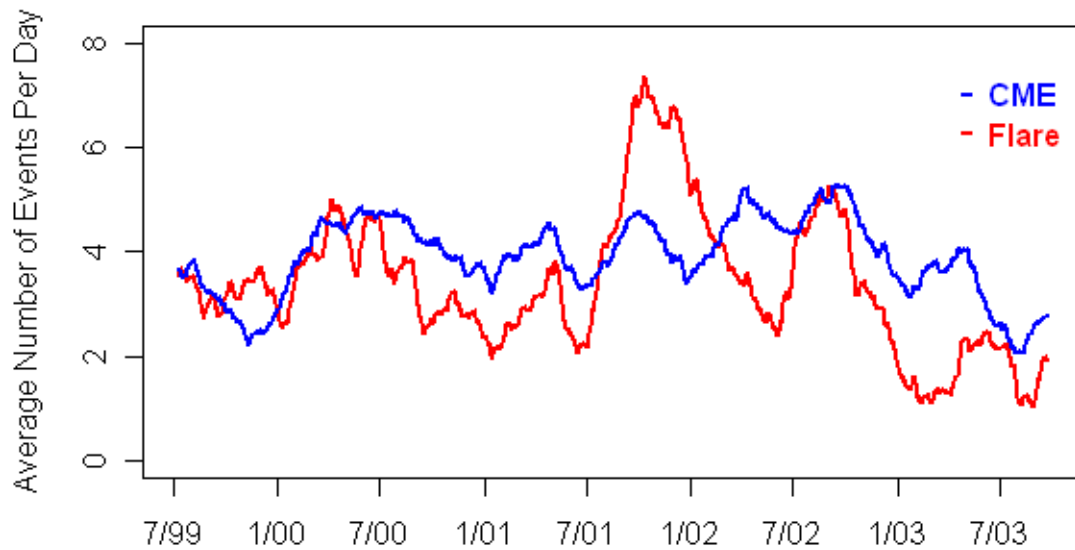


Figure 1.2: Sliding window estimates of CME and flare rates over time. This plot contains information about the long-term trends in the rates of the two processes. Note that the flare rate increased during the solar maximum which occurred at the end of 2001, while the CME rate remained relatively stable during that time. This plot was created by computing the average number of events per day that occurred over a 100 day sliding window. That is, for each day between April 1, 1999 and December 31, 2003, we counted the number of CMEs and the number of flares that occurred within ± 50 days of that date. The counts were then divided by 100 so that the estimates represent the average number of events per day for the time period from -50 to $+50$ days about the given date.

To get a glimpse into the nature of the data on a smaller time scale, a typical 100 day interval of GOES solar flare events and CDAW CME events is presented in Figure 1.3. It is clear that a simple homogeneous Poisson process [16] does not provide an adequate description of the univariate point processes underlying either data set. In addition, the data shown in Figure 1.3 may hint at a tendency for CME and flare events to clump together on a local time scale. It is this phenomena that is the central focus for our research.

1.5 Methods

1.5.1 Detecting a Temporal Correlation Between Solar Flares and CMEs

It is known that occurrence rates of solar flares and coronal mass ejections both follow the long-term trends in the solar cycle. The main goal of this study is to determine whether these two types of events are in fact correlated on a much smaller time scale. That is, we are interested in answering the question of whether the occurrence of a single solar flare event makes the occurrence of a CME more or less likely in the surrounding minutes or hours, and vice versa.

In order to access the significance of a local temporal correlation between the two processes, we create a data set of secondary measurements which represent the relative onset times between the two types of events. For each flare onset time τ_F and each CME onset time τ_C , we compute the difference

$$\tau_D = \tau_C - \tau_F \tag{1.1}$$

Next, we let \mathbb{T} be the set of all differences τ_D such that $\tau_D \in [-W, W]$, where W is the half-width of the temporal window. To clarify, we restrict our collection \mathbb{T} to be only those differences between CMEs and flares that occurred within W minutes of one another. For this analysis we chose $W=500$, so that only those CMEs and flares that occurred within $W=500$ minutes of each other were included in \mathbb{T} .

Once these differences are collected, a histogram of the relative time points can provide insight into the nature of the relationship between the two processes. The intuition here is that if the two processes are indeed independent in a local sense, then we would expect

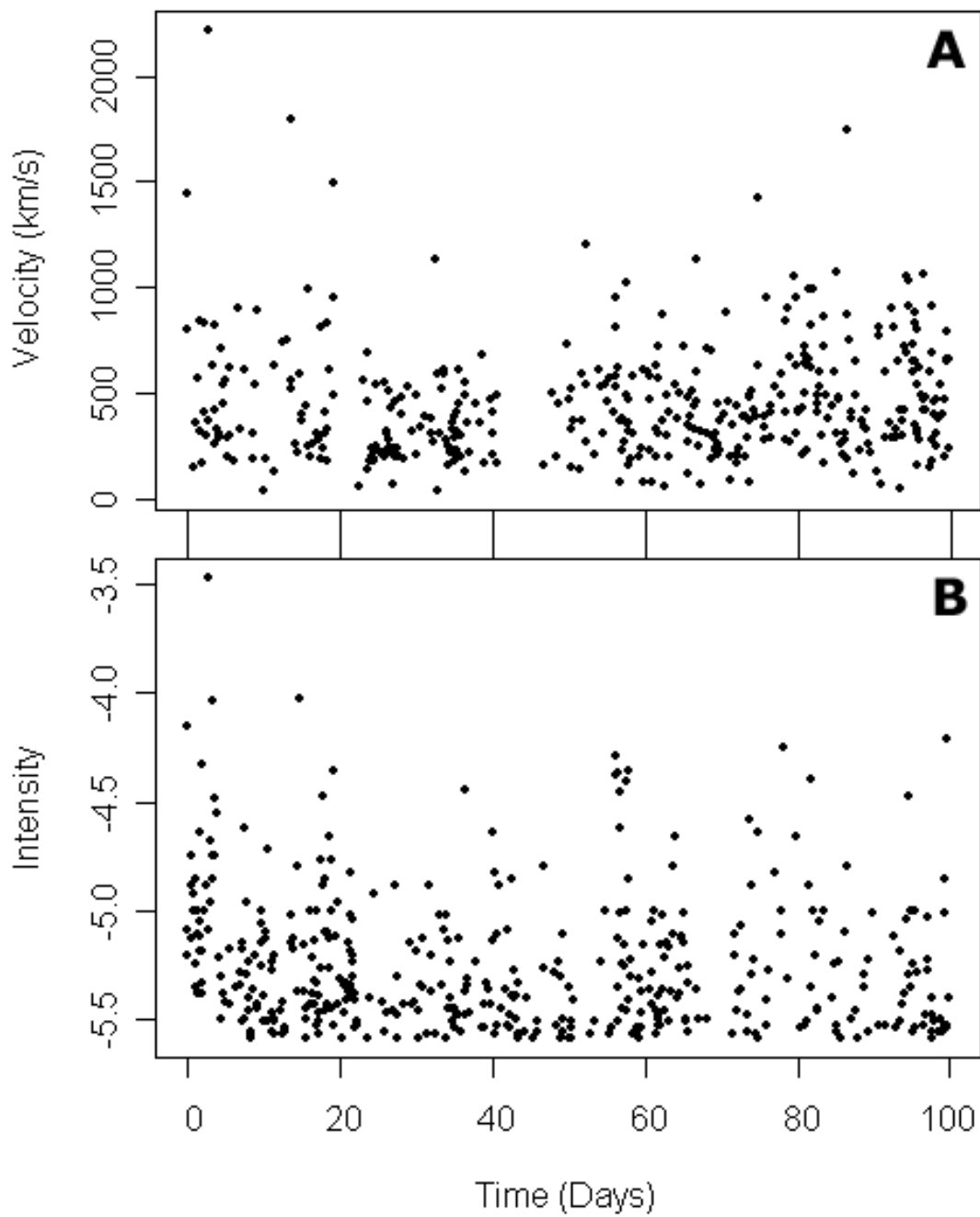


Figure 1.3: A typical 100 day interval of (A) CDAW coronal mass ejections and (B) GOES solar flare events. The onset time is presented on the horizontal axis and a relevant measure of event size for each type of event is presented on the vertical axes. For CMEs the event size was measured by velocity (in km/s), and for flares the event size was measured in X-ray intensity on the \log_{10} scale.

a histogram of the values $\tau_D \in \mathbb{T}$ to appear flat, or uniform. If instead this histogram was found to include a bump of increased density, then this bump would represent an increase or decrease in the probability that a CME event will occur relative to the onset time of a flare event. Note that this analysis does not take into account the physical proximity of a particular CME and flare event.

A histogram of the $|\mathbb{T}| = 17,356$ relative time points τ_D for the CDAW coronal mass ejection onset times relative to the GOES solar flare onset times is given in Figure 1.4. One interesting observation is that there appears to be a bump to the left of time $t = 0$, which suggests that there may be increased probability that a coronal mass ejection will occur immediately preceding a solar flare event.

However, there are two key considerations in our use of the measurements τ_D that we must be careful to state clearly. First, we can not consider the measurements $\tau_D \in \mathbb{T}$ to be independent samples, because it is possible for (and this indeed occurs in our data) a single CME event to be counted twice within \mathbb{T} whenever it occurs within $W=500$ minutes of two different flare events. Second, we know that the CME and flare processes both follow the solar cycle, and so there will be long-term correlations in the overall intensity function for which we must account.

1.5.2 Two Models for the Distribution of Relative Onset Times

In order to assess the correlation between the two processes, two functional forms for the density of the relative onset times \mathbb{T} were considered. Each of the functional forms was intended to allow for a bump of *increased* probability of the occurrence of a CME relative to the occurrence of a flare, as this was the direction of the relationship observed in the data. One model is a simple piecewise-uniform density with a uniform bump, and the other is a smoother mixture model comprised of an overall uniform density with a Gaussian bump.

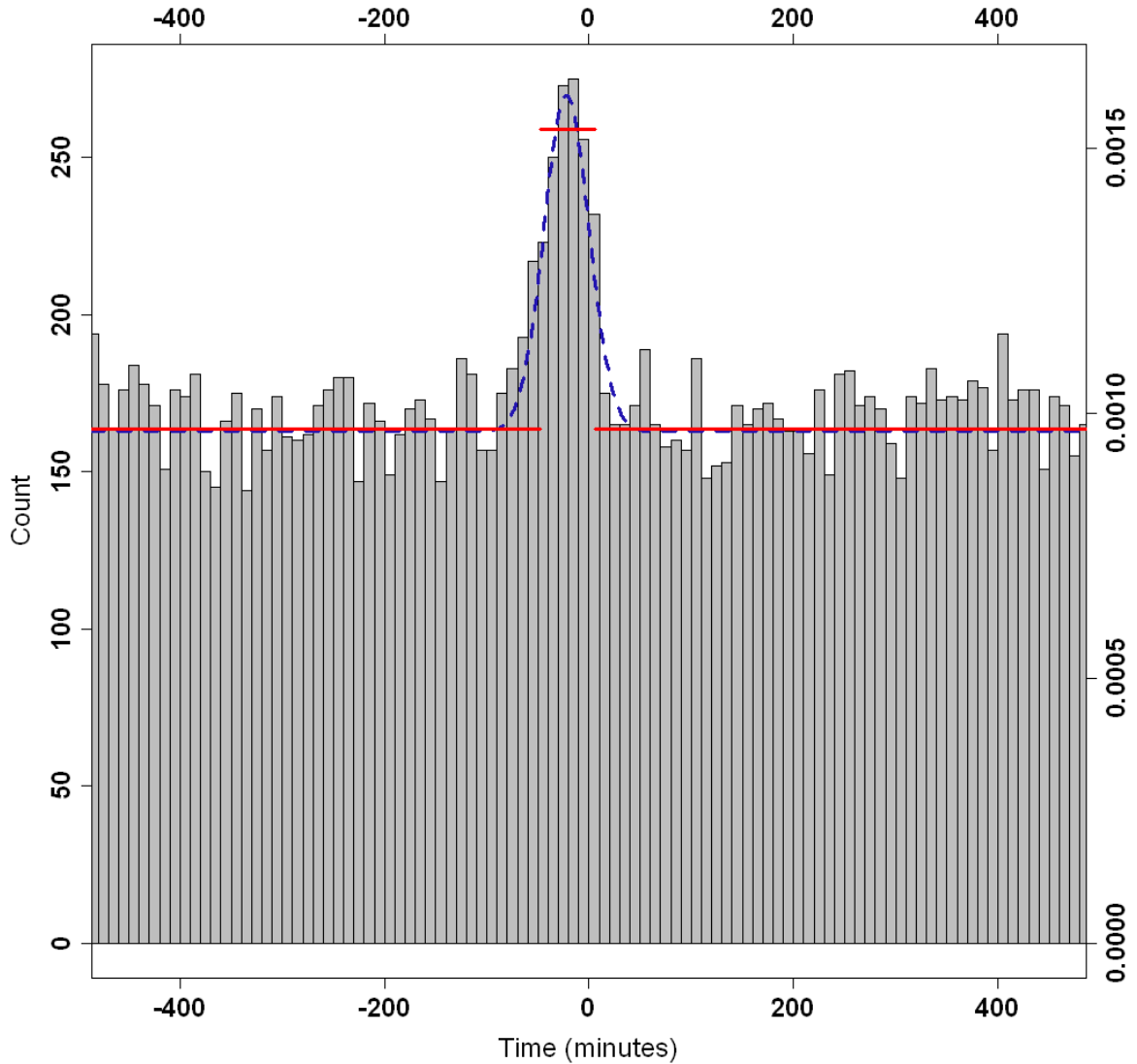


Figure 1.4: Histogram and fitted densities for the relative time points $\tau_D \in \mathbb{T}$ of CDAW CME onset times relative to GOES flare onset times. There are $|\mathbb{T}| = 17,356$ relative time points in this data set. The histogram counts are shown in gray, with a bin width 10 minutes. The fitted density under the piecewise-uniform model is shown in red, and the fitted density under the Gaussian-uniform mixture model is shown in blue. The vertical axes ticks represent counts on the left and fitted density on the right. Note that although the histogram displays binned counts for the data, the maximum likelihood fitting of the two densities was performed in a continuous (non-binned) setting under the resolution of the catalog time stamps, which are in minutes.

A Piecewise-Uniform Model

The first model for the relative times points \mathbb{T} is a simple piecewise-uniform density

$$f(t; \gamma, \mu, \sigma) = \begin{cases} \frac{\gamma}{2W} + \frac{1-\gamma}{2\sigma} & , \text{ if } t \in [\mu - \sigma, \mu + \sigma] \\ \frac{\gamma}{2W} & , \text{ otherwise} \end{cases} \quad (1.2)$$

where γ is the mixture parameter, μ represents the center of the bump, σ represents the half-width of the bump, and W is the half-width of the temporal window. The uniform bump is restricted to lie fully within the range $[-W, W]$, and the density must integrate to one. If the observations were strictly independent, the likelihood would be

$$\begin{aligned} \mathcal{L}(\gamma, \mu, \sigma | \mathbb{T}) &= \prod_{i=1}^n P(T = t_i | \gamma, \mu, \sigma) \\ &= \left(\frac{\gamma}{2W}\right)^{|\mathbb{T}_0|} \cdot \left[\frac{\gamma}{2W} + \frac{1-\gamma}{2\sigma}\right]^{|\mathbb{T}_1|} \end{aligned} \quad (1.3)$$

where $|\mathbb{T}_0|$ is the number of relative time points τ_D that lie outside the bump, $|\mathbb{T}_1|$ is the number of relative time points τ_D that lie inside the bump, and the support set for the parameters is as described above. Note that the resolution of onset time for both CMEs and flares is in minutes, so we will restrict σ and μ to be in increments of 0.5 minutes. The log-likelihood is

$$\begin{aligned} \ell(\gamma, \mu, \sigma | \mathbb{T}) &= \ln \left[\left(\frac{\gamma}{2W}\right)^{|\mathbb{T}_0|} \cdot \left[\frac{\gamma}{2W} + \frac{1-\gamma}{2\sigma}\right]^{|\mathbb{T}_1|} \right] \\ &= |\mathbb{T}_0| \cdot \ln \left(\frac{\gamma}{2W}\right) + |\mathbb{T}_1| \cdot \ln \left[\frac{\gamma}{2W} + \frac{1-\gamma}{2\sigma}\right] \end{aligned} \quad (1.4)$$

A Gaussian-Uniform Mixture Model

An examination of the data suggests that a smoother model of density for the relative onset times \mathbb{T} might better fit the data. To this end, a uniform density with a Gaussian ‘‘bump’’ was considered. The general functional form for this density is

$$\mathbf{f}(t; \gamma, \mu, \sigma) = \frac{\gamma}{2W} + \frac{1-\gamma}{I} \cdot \mathbf{g}(t; \mu, \sigma) \quad (1.5)$$

where μ and σ are the mean and standard deviation of the Gaussian bump, $\mathbf{g}(t; \mu, \sigma)$ is the Gaussian density with mean μ and standard deviation σ , and I is the integral of the Gaussian density restricted to the interval $[-W, W]$. For identifiability, model parameters were restricted to $\gamma \in [0, 1]$, $\mu \in [-W, W]$, and $\sigma \in [0, W]$ with μ and σ integer valued. The overall density \mathbf{f} is by default restricted to integrate to one. If the observations were strictly independent, the likelihood would be

$$\begin{aligned} \mathcal{L}(\gamma, \mu, \sigma | \mathbb{T}) &= \prod_{i=1}^n P(T = t_i | \gamma, \mu, \sigma) \\ &= \prod_{i=1}^n \left[\frac{\gamma}{2W} + \frac{1-\gamma}{I} \cdot \mathbf{g}(t_i; \mu, \sigma) \right] \end{aligned} \quad (1.6)$$

1.5.3 A Numerical Approach to Parameter Estimation

For each of the two functional forms considered above, maximizing the likelihood is mathematically intractable. In addition, the likelihood surfaces is non-convex, so it is not possible to obtain maximum likelihood estimates for the parameters using a simple gradient search algorithm. As an alternative, a semi-greedy [8] grid search was utilized to improve the ability to locate the true maximum likelihood parameter values under each data set.

The procedure begins by evaluating P points in the likelihood surface determined by a coarse grid over the parameter space. From within this grid, the M parameter values that led to the maximal function evaluations are selected. A finer grid of P points is then placed over each of the M locations in the parameter space, and the function is evaluated

Model for Distribution of \mathbb{T}	$\hat{\gamma}$	$\hat{\mu}$	$\hat{\sigma}$	log-likelihood
Piecewise-Uniform Model	0.970	-21	26.5	-119792
Gaussian-Uniform Mixture Model	0.965	-22	22.0	-119780

Table 1.1: Maximum likelihood estimates of parameters under two different functional forms for the distribution of \mathbb{T} . For both distributional models, γ is the mixture parameter between the overall uniform density and the “bump” of increased probability, and μ and σ represent the center and spread of the bump.

at all $M \times P$ grid points. Once again, M optimal locations are then selected from those $M \times P$ locations, and the process is repeated until two conditions are met: (1) the desired resolution is achieved for each parameter, and (2) at the given stage in the process, the M selected maximal locations are neighboring. There are three unknown parameters under each of the two densities considered, so for convenience the results presented in this paper were obtained using the above procedure with $P=10^3=1000$ and $M=10$.

The fitted densities under the piecewise-uniform model and the Gaussian-uniform mixture model are plotted along with the histogram of relative time points in Figure 1.4, and parameter estimates are presented in Table 1.1. Note that under the piecewise-uniform model, the period of increased probability for a CME to occur is during the time interval of $\tau_D \in [-47.5, 5.5]$ minutes relative to the onset time of a solar flare. Also note that the log-likelihood is larger under the Gaussian-uniform mixture model, which suggests that the smoother model provides a better fit for the data.

1.5.4 Assessing the Significance of a Local Temporal Correlation

A block-of-blocks bootstrap [7] [17] [20] technique was implemented in order to assess whether an observed increase (or “bump”) in the fitted density was in fact statistically significant. Our bootstrap method creates resampled data sets from the original data in such a way that the overall temporal association that follows the solar cycle is preserved, while any local correlation that might exist within a small temporal window is removed.

In order to compare the model fit for the real data with those achieved on bootstrapped

data sets, the log-likelihood ratio statistic [25]

$$LR = -\ln \left[\frac{\mathcal{L}(\mathbb{T} | H_0)}{\mathcal{L}(\mathbb{T} | H_A)} \right] = \ell(\mathbb{T} | H_A) - \ell(\mathbb{T} | H_0) \quad (1.7)$$

is utilized. In our case, the null hypothesis H_0 represents the restricted uniform model for \mathbb{T} , while the alternative hypothesis H_A represents the model fit under the Gaussian-uniform mixture model for the density of \mathbb{T} as described in Section 1.5.2. Note that the factor of two that is traditionally included in the likelihood ratio statistic has been omitted.

To create the resampled data sets, the full time spectrum was divided into blocks. The total length of the time spectrum for the data was 2,499,840 minutes, and so the spectrum was divided into $K=499$ blocks of length $L=5000$ minutes, where the last piece of the spectrum which was too short to form a complete block was removed from both the real analysis and the bootstrapped analysis.

Next, for each of the $K=499$ blocks of time, attach the solar flares that occurred inside that block of time. Then attach the CMEs that occurred either inside that block or within $\pm W$ minutes of the edges of that block. It is critical that the reader understand this detail of the procedure. By attaching those CMEs that fall within W minutes of the edge of the block, we ensure that any CME and flare which originally occurred within W minutes of each other will have the potential to be included in our bootstrapped sample. Note that CME events that fall within $\pm W$ minutes of the breakpoint between two blocks will be attached to both blocks.

We constructed $R_1=500$ bootstrapped data sets under the following procedure:

- For iteration $r = 1$ to R_1 :
 1. For $k = 1$ to K ,
 - (a) Select a block $B1$ at random with replacement from the set of K blocks.
 - (b) Condition on the number of flare events in block $B1$, and select a block $B2$ at random with replacement from the set of K blocks such that $B1$ and $B2$ contain the same number (or a similar number) of flare events.
 - (c) Overlay the two blocks $B1$ and $B2$ on top of each other so that the start points and end points are the same.

- (d) Compute the relative onset times $\tilde{\tau}_D$ for each CME event in block B_2 with respect to each flare event in block B_1 .
2. Combine all relative onset times $\tilde{\tau}_D$ from those K resampled blocks into a final collection $\tilde{\mathbb{T}}$.
3. Fit a Gaussian-uniform mixture model to the relative time points $\tilde{\mathbb{T}}$.
4. Compute the likelihood-ratio test statistic for the resampled data.

A diagram of this bootstrap procedure is presented in Figure 1.5. In practice, the degree of agreement between the number of flare events in blocks B_1 and B_2 at Step (1b) was variable and selected such that the difference between the number of flares was as small as possible while still allowing at least ten choices of block B_2 for any block B_1 . Also note that Step (1b) ensures that the long-term correlation between the two processes are preserved. That is, flares that occur during high-activity periods are paired with CMEs that also occurred during periods with high flaring activity, and vice versa.

The reason that we have chosen to measure the contributions for blocks separately rather than to create a complete realigned dataset is because our method of resampling avoids the complication that can occur at the edges when temporal blocks that did not occur together are realigned next to each other. In order to understand why this is necessary, consider an alternate scenario in which only those CMEs which occur inside of a given block are attached to the block. Suppose the above Step (1a) and Step (1b) remain the same, but then the collapsed blocks from Step (1c) are realigned side-by-side to form a single temporal spectrum and the relative times points are determined from this realigned spectrum. If we then viewed the histogram of the relative time points τ_D from this realigned spectrum we would find a tendency for the frequency in the histogram to be reduced at the edges of the temporal window $[-W, W]$. This is because we would have put blocks from high-intensity periods next to blocks from low-intensity periods, and therefore we would find an artificially reduced probability for a CME to occur as we move farther and farther away from a specified flare event in the realigned spectrum. Our block-of-blocks bootstrap technique [17] aggregates the individual contributions from each block in order to avoid this problem all together.

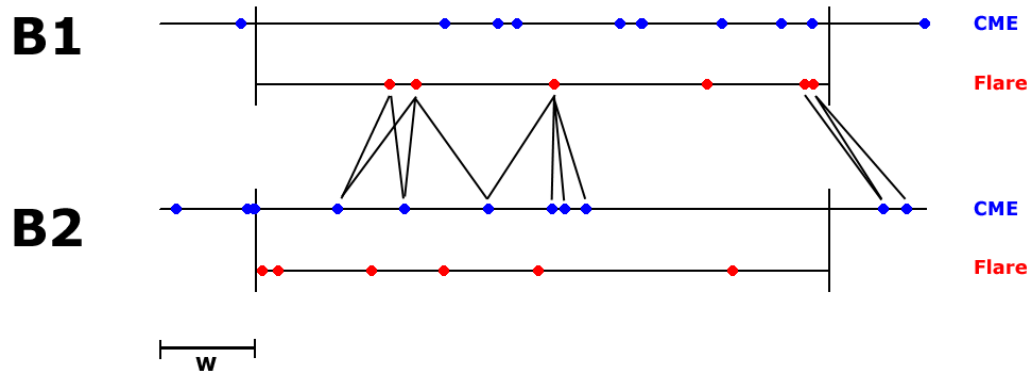


Figure 1.5: Diagram of the bootstrap procedure for evaluating the significance of the local temporal correlation. In this toy example, Block $B1$ has been selected at random from the set of K blocks. In this case, Block $B1$ contains six flare events and therefore Block $B2$ was selected at random from the set of all blocks in the set K that contain six flare events. Block $B1$ and Block $B2$ are then overlaid, and the relative onset times τ_D are computed for the CMEs in Block $B2$ relative to the flares in Block $B1$. Note that each block contains all flare events that occur inside the block, and all CME events that occur either inside the block or within $\pm W$ minutes of the block. Compare this procedure to the bootstrap method presented in Figure 1.6 which is used to form bootstrapped confidence intervals for the fitted parameter estimates.

The log-likelihood of the real data under the strictly uniform null model is

$$\ell(\mathbb{T} | H_0) = \ln \left[\left(\frac{1}{2W} \right)^{|\mathbb{T}|} \right] \quad (1.8)$$

$$= -|\mathbb{T}| \cdot \ln(2W) \quad (1.9)$$

$$= -17365 \cdot \ln(1000) \quad (1.10)$$

$$= -119891 \quad (1.11)$$

and so the observed value of the log-likelihood ratio statistic under the Gaussian-uniform mixture model is

$$LR = \ell(\mathbb{T} | H_A) - \ell(\mathbb{T} | H_0) \quad (1.12)$$

$$= -119780 + 119891 \quad (1.13)$$

$$= 111 \quad (1.14)$$

This observed value for the real data far exceeds the maximum value of the log-likelihood ratio statistic observed in the collection of $R_1=500$ bootstrapped data sets, which was $LR=6$. Therefore, the p -value of the bootstrapped significance test is $p<0.002$. Note that our p -value is based on the fact that we have only constructed $R_1=500$ bootstrapped samples. We would expect the actual p -value to be extremely small because the discrepancy between the observed statistic ($LR=111$) and the maximum observed statistic for the bootstrapped data sets ($LR=6$) is so large. Our analysis indicates a highly significant local temporal correlation between the onset time of a solar flare and the onset time of a CME.

1.5.5 Estimating the Uncertainty in the Parameter Estimates

The two fitted models provide a description of the nature of the correlation between the two processes, and therefore it is useful to have a measure of the uncertainty in each of the parameter estimates.

To this end, a second bootstrap approach [7] was implemented with the aim of preserving both the local temporal correlation and the overall trends due to the solar cycle. The

main difference between this bootstrap procedure and the bootstrap procedure described in Section 1.5.5 is that CME and flare events that occurred in the same block remain attached – there is no re-matching of flare events from one block with CME events from another block. Rather, we sample whole blocks from the set of $K=499$ blocks. This is necessary because the goal is to create resampled data sets that resemble our real data as closely as possible, so that we can feel confident that parameters estimated on those resampled data sets are comparable to those measured on the real data.

We constructed $R_2=500$ bootstrapped data sets under the following procedure:

- For iteration $r = 1$ to R_2 :
 1. For block $k = 1$ to K ,
 - (a) Select a block $B1$ at random with replacement.
 - (b) Compute the relative onset times $\tilde{\tau}_D$ for each CME onset time relative to each flare onset time within block $B1$.
 2. Combine all relative onset times $\tilde{\tau}_D$ from those K resampled blocks into a final collection $\tilde{\mathbb{T}}$.
 3. Fit a Gaussian-uniform mixture model to the relative time points $\tilde{\mathbb{T}}$, and record the fitted parameter estimates.

A diagram of this bootstrap procedure is presented in Figure 1.6.

Parameters were estimated for each of the $R_2=500$ resampled data sets, and confidence intervals for the parameter estimates were formed using the basic bootstrap confidence limits described in equation 2.10 of Davison and Hinkley [7]. The resulting 95% confidence intervals are presented in Table 1.2 along with the original maximum likelihood estimates that were computed on the real data. Note that the confidence intervals for the mixture parameter γ does not include 1.0, which would represent a purely uniform model. This provides additional evidence that a local temporal correlation exists between the two processes. It should also be noted that histograms of the bootstrapped estimates of the three parameters were all symmetric in appearance.

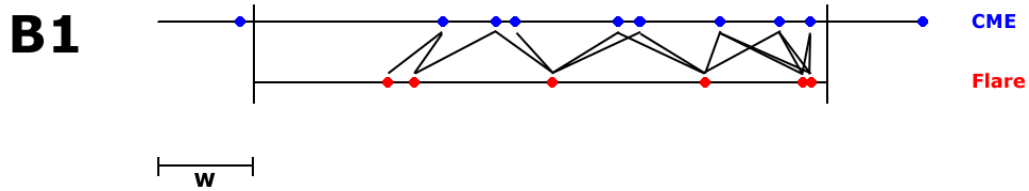


Figure 1.6: Diagram of the bootstrap procedure for constructing confidence intervals for the parameter estimates. In this toy example, Block $B1$ has been selected at random from the set of K blocks. The relative times τ_D are computed for the CMEs onset times relative to the flare onset times. Compare this procedure to the bootstrap method shown in Figure 1.5, and note that this approach preserves any local temporal correlation that might exist between the two processes. For ease of interpretation, Block $B1$ represents the same Block $B1$ shown in Figure 1.5.

Parameter	MLE	95% Confidence Interval
γ	0.965	[0.959, 0.971]
μ	-22	[-27, -17]
σ	22.0	[18.0, 25.0]

Table 1.2: Maximum likelihood estimates and bootstrapped 95% confidence intervals for model parameters under the Gaussian-Uniform mixture model for the distribution of \mathbb{T} . The maximum likelihood estimates were computed on the real data, and confidence intervals were formed using the basic bootstrap confidence limits described in Davison and Hinkley [7]. Confidence intervals are based on $R_2=500$ resampled data sets that were created using the bootstrap procedure described in Section 1.5.5.

1.6 Summary

We have found a highly statistically significant local temporal correlation between the onset times of coronal mass ejections and the onset times of solar flares. There is an increased probability for a CME to occur in the minutes preceding the onset of a solar flare. Under a Gaussian-uniform mixture model, the maximum likelihood estimates for the mean and standard deviation of the spike of increased probability are $\hat{\mu} = -22$ minutes and $\hat{\sigma} = 22.0$ minutes respectively. Bootstrapped 95% confidence intervals for the parameter estimates are $\mu \in [-26, -17]$ and $\sigma \in [19.0, 26.0]$.

Our analysis provides a statistical basis to conclude that solar flares and coronal mass ejections are indeed correlated on a local time scale. This evidence should inform the current debate in the astronomical community, and will likely provide insight for astronomers who are working to better understand the processes by which magnetic energy is released from the solar corona. It is also likely that this analysis could lead to improvements in current methodologies for the detection and prediction of both solar flare and coronal mass ejection events, as it provides a statistical basis for the inclusion of data from each type of process as a covariate when attempting to predict the occurrence of the other.

However, we would like to make it clear that although it can be tempting to view the location of the bump in density for CME events relative to flare events as an indication of causation (eg. CMEs are more likely to occur right before a flare, therefore CMEs *cause* flares), it is very important to keep in mind that this result may simply point to an underlying mechanism that is driving both types of solar events. We make no assertion about the underlying mechanism that is driving either process except to say that these events are clearly correlated on a local time scale.

Finally, the novel statistical approach presented in this chapter provides a technique for both detecting and assessing the nature of a local temporal correlation between two processes in the presence of an inhomogeneous long-term rate function to which the two processes are tied.

CHAPTER 2

ESTIMATING THE QUASAR LUMINOSITY FUNCTION

2.1 Introduction

The Quasar Luminosity Function (QLF) is the spatial density of quasars (per Mpc^3 per mag) as a function of absolute optical magnitude M and redshift z . The data are quasars detected in the Sloan Digital Sky Survey, and major complicating factors in the analysis are data truncation that occurs when quasars are too faint or too far away to be detected from Earth, and sampling bias due to the fact that quasars in the data set were not selected uniformly at random from the population of all quasars in the universe.

Following the work of Richards et al. [24], this paper presents a flexible parametric model of quasars as a non-homogeneous Poisson process over the $M \times z$ space. Our contribution to this model is two-fold: first, to allow the introduction of higher-order terms as warranted by the data, and second, to further extend the model in order to better detect and describe the $M \times z$ dependency in the data. Parameters are estimated using the method of maximum likelihood, and model selection is performed using the Bayesian Information Criterion.

2.1.1 Background

Astronomers today theorize that quasars are created when gases and materials form an accretion disk and reach luminous speeds as they fall into a supermassive black hole. Quasars can emit visible, UV, and gamma light, and some are also known to emit radio waves. Quasars are among the brightest objects in space and are therefore detectable even at extreme distances from Earth. Figure 2.1 shows images of a quasar captured by the NASA Hubble Space Telescope.

When emissions from a very distant quasar are detected on Earth, what is observed represents the state of the quasar at the time of emission. Quasars therefore present us with

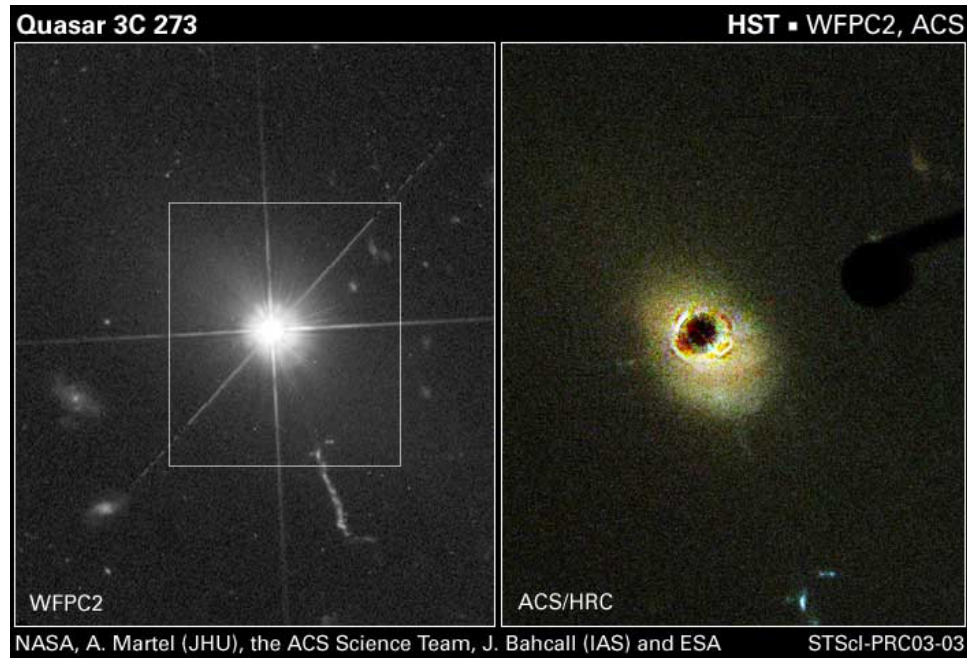


Figure 2.1: Two images of nearby quasar 3C 273 captured by the NASA Hubble Space Telescope. The image on the right was created using NASA’s Advanced Camera for Surveys, which uses a coronagraph to block the light from the central quasar so that the surrounding galaxy becomes visible.

a window into the past through which we may learn about how the state of the universe has changed over time. Changes in the distribution of quasar luminosity at different redshift distances from Earth would imply that there has been a change in the population of quasars over time.

In astronomy, a luminosity function is a description of the number of objects per cubic megaparsec per mag (a unit of luminosity) as a function of luminosity and redshift. Astronomers are interested in the estimation of luminosity functions for many different types of space objects, and the estimation of luminosity functions is frequently complicated by data truncation that occurs when objects are too faint or too far away to be detected.

This paper refers to two forms of optical magnitude: apparent optical magnitude (m) and absolute optical magnitude (M). Apparent optical magnitude is the brightness of an object (measured in mags) when it is viewed from earth. Absolute optical magnitude M is the amount of optical light that an object emits in space. Apparent magnitude m is affected

by the distance of the object from earth, as closer objects will appear brighter when viewed from Earth. Absolute magnitude is measured by first collecting the apparent magnitude of an object, and then adjusting for the distance of that object from Earth in order to determine its inherent luminosity.

The distance of a space object from Earth can be measured using redshift. Astronomers measure redshift by performing a spectrum analysis of the light emitted by the object. Common elements such as hydrogen produce a spike in the spectrum at a known frequency. The spikes becomes doppler shifted because the observer and the space object are moving relative to each other, and this shift is related to distance through the expansion of the universe. Astronomers make note of this shift of the spikes in the recorded spectrum and back-transform to produce the object's estimated distance from Earth. If an object is recorded as having redshift z , then under a simple cosmological model, the size of the universe at the time that the light was emitted was $\left(\frac{1}{1+z}\right)$ of the size universe today. For example, if a quasar has redshift $z = 1$, then we are looking at emissions from that quasar that occurred when the universe was half the size it is today.

2.1.2 Scientific Goals

The quasar luminosity function (QLF) is a description of the number of quasars per Mpc^3 per mag as a function of absolute optical magnitude M and redshift z . The first goal of this research was to provide a more accurate description of the QLF, using data available from the Sloan Digital Sky Survey, Data Release III. As of the date of this research, Richards et al. [24] presents the most complete estimation of this function. Their analysis assumes a particular functional form and then estimates model parameters via maximum likelihood analysis. This analysis takes the model a step further, allowing for the introduction of higher-order terms as warranted by the data.

The second goal was to detect whether there is a significant interaction effect between absolute optical magnitude and redshift. In the vast majority of papers analyzing these data, M and z are treated as separable (or independent). Fan et al. [9] presents a theoretical argument as to why this should be the case and is frequently cited by researchers as a basis for this assumption. This thesis presents an analysis of the quasar data that suggests that

there is in fact an interaction between absolute magnitude and redshift, and a description of the nature of this interaction is provided.

2.1.3 Overview of Methods

The locations of observed quasars in $M \times z$ space can be thought of as a two-dimensional point process. In this paper, quasar counts are modeled as a Poisson process [16] with non-homogeneous rate surface measured in quasar counts per cubic megaparsec (Mpc^3) per mag.

In an earlier paper, Richards et al. [23] describe simulation methods used to estimate observational probabilities ψ on an $m \times z$ grid. Our analysis treats these values as known and fixed. Absolute magnitude values for each quasar are K corrected [12], and redshift values are converted into cubic megaparsecs by the computation of the infinitesimal volume differential γ measured in Mpc^3 of a spherical shell about the Earth at a given redshift z .

Initially, an exponential model with polynomial functions of M and z in the exponent is constructed, and parameters are estimated by the method of maximum likelihood under a series of models with polynomials of varying degree. Model selection is performed using the Bayesian Information Criterion [27], and the selected model is presented as a description of the Poisson [19] intensity surface under the assumption that absolute optical magnitude and redshift are separable. Next, an interaction between absolute optical magnitude M and redshift z is incorporated into the model. Similar model selection techniques reveal that this term is indeed warranted by the data, suggesting that the marginal distribution of absolute magnitude for quasars is dependent on redshift, and therefore dependent on the evolutionary epoch from which the quasar was observed.

2.2 The Data

2.2.1 The Sloan Digital Sky Survey

The full data consists of 47,877 quasars from the Sloan Digital Sky Survey catalog, Data Release III. The Sloan Digital Sky Survey, or SDSS, is a large-scale astronomical survey

of one quarter of the sky. Quasars in the SDSS catalog are marked as *extended* and *non-extended*. Extended quasars are those that exist nearby or reside within a galaxy such that portions of the host galaxy are visible around the quasar. Although other quantities such as the spherical location and radio emission of each quasar are included in the SDSS quasar catalog, only the redshift and optical magnitude were utilized in this study.

2.2.2 Additional Astronomical Considerations

In this paper, apparent magnitude m will refer to the apparent optical i-band (infrared) magnitude [23], and absolute magnitude M will refer to the K corrected absolute optical i-band magnitude. The transformation from apparent magnitude m into K corrected absolute magnitude M is

$$M = m - \text{DM}(z) - \text{K}(z) \quad (2.1)$$

where $\text{DM}(z)$ is the distance modulus and $\text{K}(z)$ is the K correction evaluated at redshift z .

Hogg et al. [12] provide a detailed description of the K correction function, and a plot of the function is presented in Figure 2.2. These corrections are estimated empirically and therefore are likely to include some error, however estimates of these errors are not available. Any errors in the specification of the K correction function could have a significant impact on the results of our analysis, and therefore all results that follow should be taken in light of this information.

The distance modulus transformation is a non-linear function of redshift z . A detailed description of the science involved can be found in Carroll [3], and a plot of the function is provided in Figure 2.3.

In order to compute quasar density in units of physical volume, the redshift values can be transformed into a spherical volume differential about the Earth by assuming a simple cosmological model. This function is denoted by $\gamma(z)$, and a plot of this function is provided in Figure 2.4.

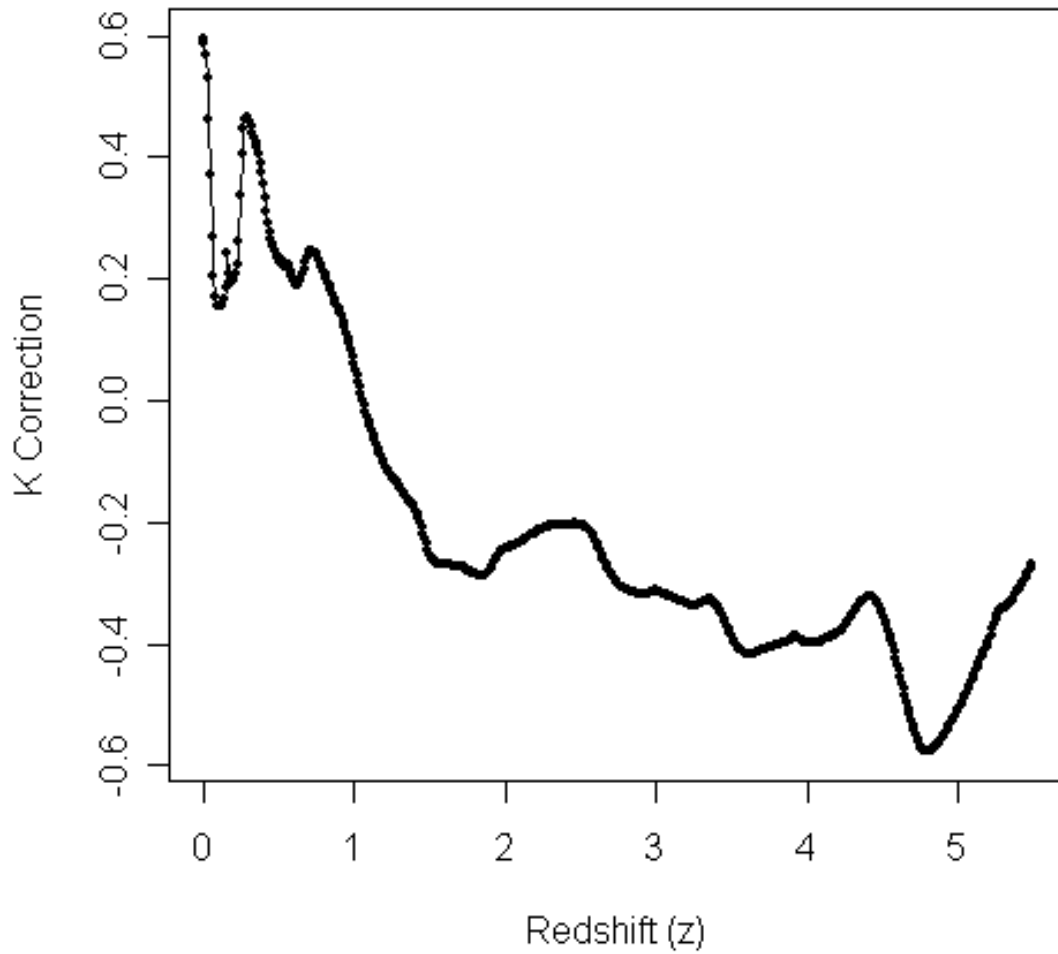


Figure 2.2: The K correction as a function of redshift z as described in Hogg et al. [12]. Note that this correction function is based largely on empirical evidence, and systematic errors in this function could introduce substantial bias and error into our resulting fitted models.

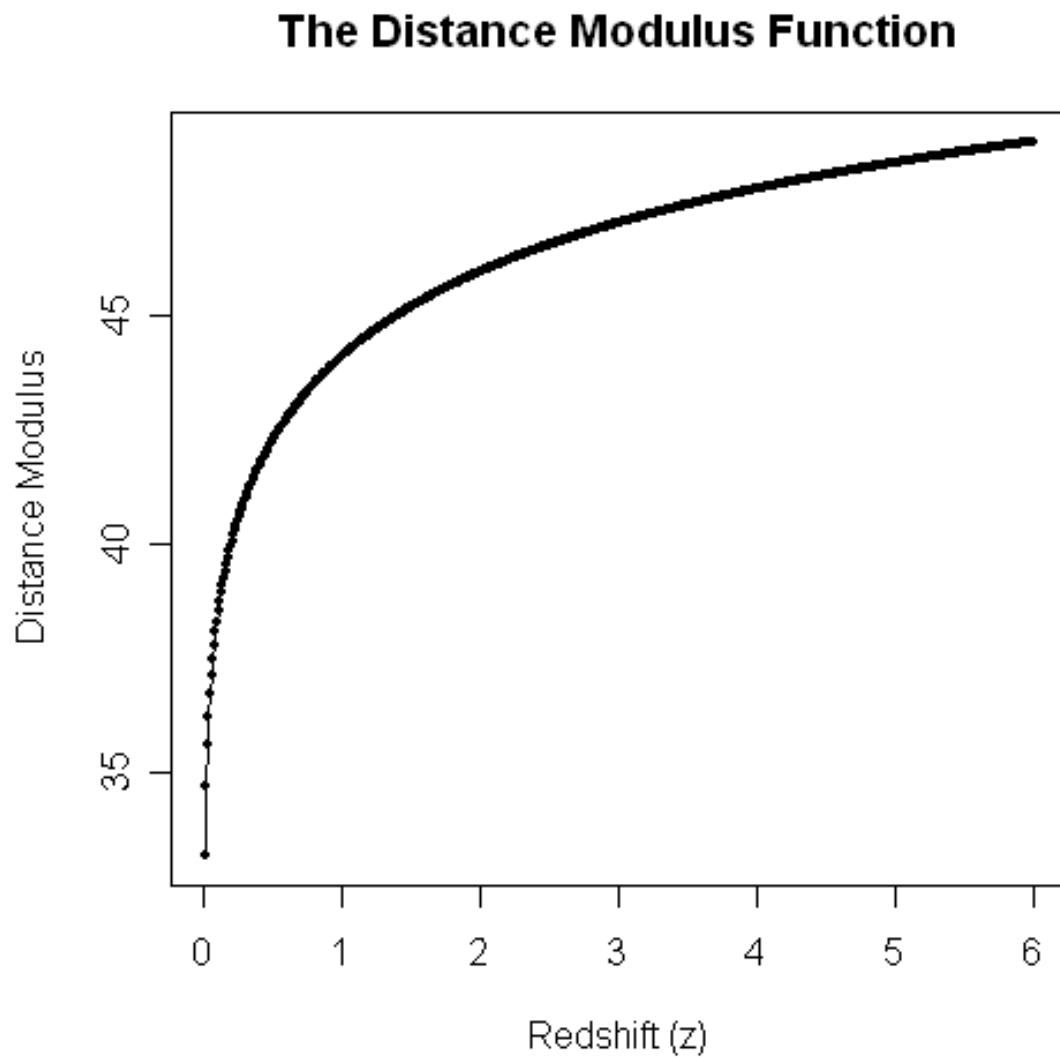


Figure 2.3: The distance modulus DM as a function of redshift z as described in Carroll [3].

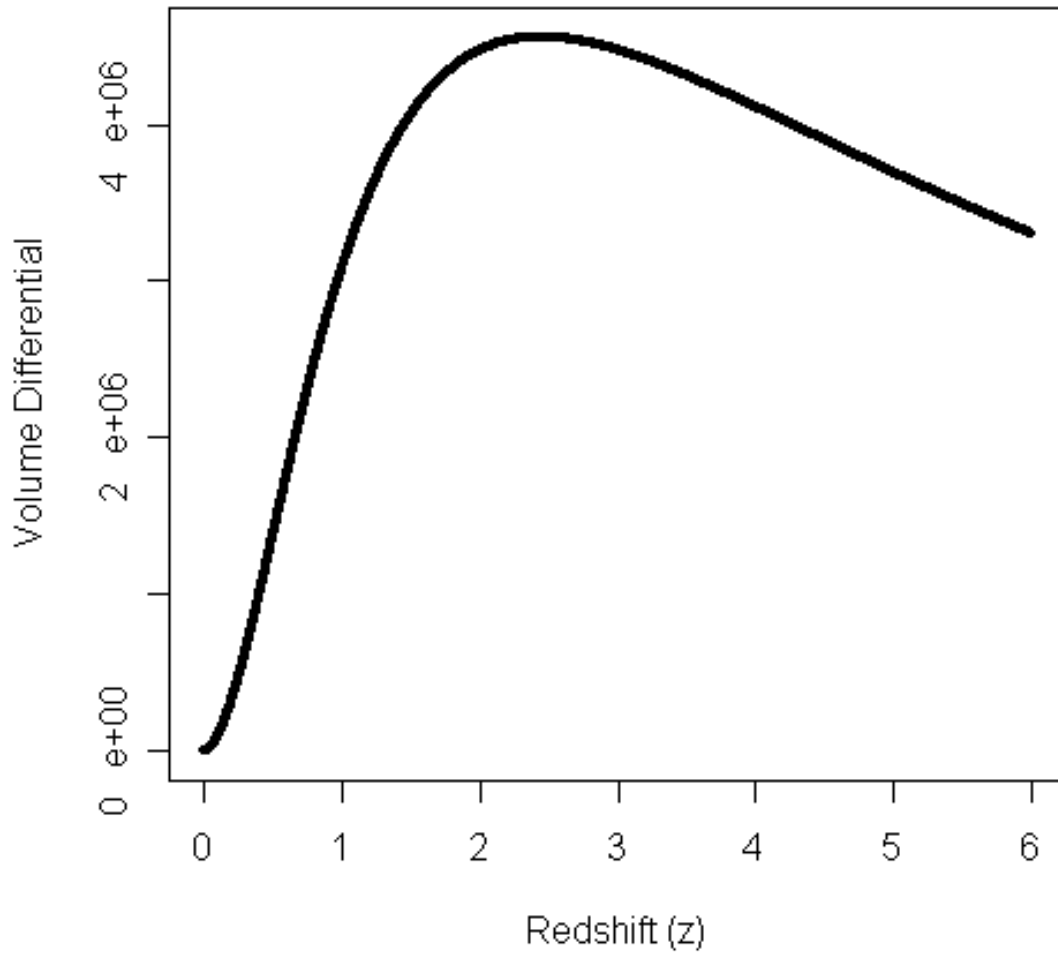


Figure 2.4: The volume differential γ as a function of redshift z assuming a simple cosmological model.

2.2.3 Observational Probabilities

A major complicating factor in the estimation of the quasar luminosity function is that observed quasars were not selected uniformly at random from the population of all detectable quasars in the universe. Rather, the selection probabilities for each quasar are dependent on that quasar's apparent brightness and its redshift distance from Earth.

Richards et al. [23] describe simulation methods used to obtain estimates of the observational probabilities ψ on the $m \times z$ grid. Each value in the table represents the probability that a quasar with the given redshift and apparent optical magnitude would be included in the data. These probabilities are adjusted for the fact SDSS only surveys one quarter of the sky, so that an observational probability of $\psi=1$ means that we would expect to observe all quasars with the given redshift and optical magnitude that are located within the quarter of the sky surveyed.

Observational probabilities are affected by whether a particular quasar is extended or non-extended, and therefore Richards et al. [23] have created two separate tables of the estimated probabilities. Our analysis is performed using only the non-extended quasars in the catalog, and observational probabilities are assumed to be flat across each grid square, with grid squares centered at the indices provided in the table. Our analysis treats the observational probabilities as known quantities because error estimates for the observational probabilities are not available.

Heat maps of the estimated observational probabilities on the apparent magnitude scale and the absolute magnitude scale are given in Figure 2.5, and Figure 2.6 respectively. Note the vertical dark band of low observational probability near redshift $z=2.75$. Many quasars at this redshift have an apparent optical magnitude that makes them difficult to distinguish from stars in the initial filtering process, and therefore these quasars have a low probability of inclusion in the catalog. Also note the black rectangular area in the upper left of Figure 2.5. This block is due to the fact that most quasars with redshift $z < 3.0$ were detected using a filter with an apparent magnitude threshold of $m \leq 19.1$, while most quasars with redshift $z \geq 3.0$ were detected using a filter with an apparent magnitude threshold of $m \leq 20.2$.

The preliminary analysis in Section 2.3.2 reveals that there may be some problems with the estimates of the observational probabilities near redshift $z=2.75$.

2.2.4 Final Data for Analysis

The final data set examined in this analysis consists of 14,113 quasars from the Sloan Digital Sky Survey, Data Release III. All extended quasars were removed from the data, as were any quasars with apparent optical magnitudes $m < 14.95$ or $m > 20.25$. In addition, any quasars that fell within $m \times z$ regions with an estimated observational probability of zero were removed from the data. It should be noted that the existence of such quasars proves that there are errors in these estimated probabilities.

Figure 2.7 gives a scatterplot of the quasar data that was used in this analysis with redshift plotted against apparent magnitude, while Figure 2.8 shows redshift plotted against absolute magnitude.

2.3 Methods

2.3.1 A Poisson Process Model for the Quasar Luminosity Function

Let the spatial density of quasars be modeled as a Poisson process in the space of absolute magnitude M and redshift z with a non-homogeneous rate surface $\lambda(M, z)$ measured in quasar counts per Mpc^3 per mag.

Let $\Phi(M, z; \theta)$ be a generic functional form to describe the idealized quasar luminosity function with model parameter θ . Let $\psi_M(M, z)$ be the observational probability for a quasar with luminosity M at redshift z , where the subscript M is used to indicate that this is the transformed function that takes absolute magnitude as its first argument rather than apparent magnitude. If we assume that ψ is known, the resulting model for the observed quasar counts is a Poisson process with rate function

$$\lambda_\theta(M, z) = \Phi_\theta(M, z) \psi_M(M, z) . \quad (2.2)$$

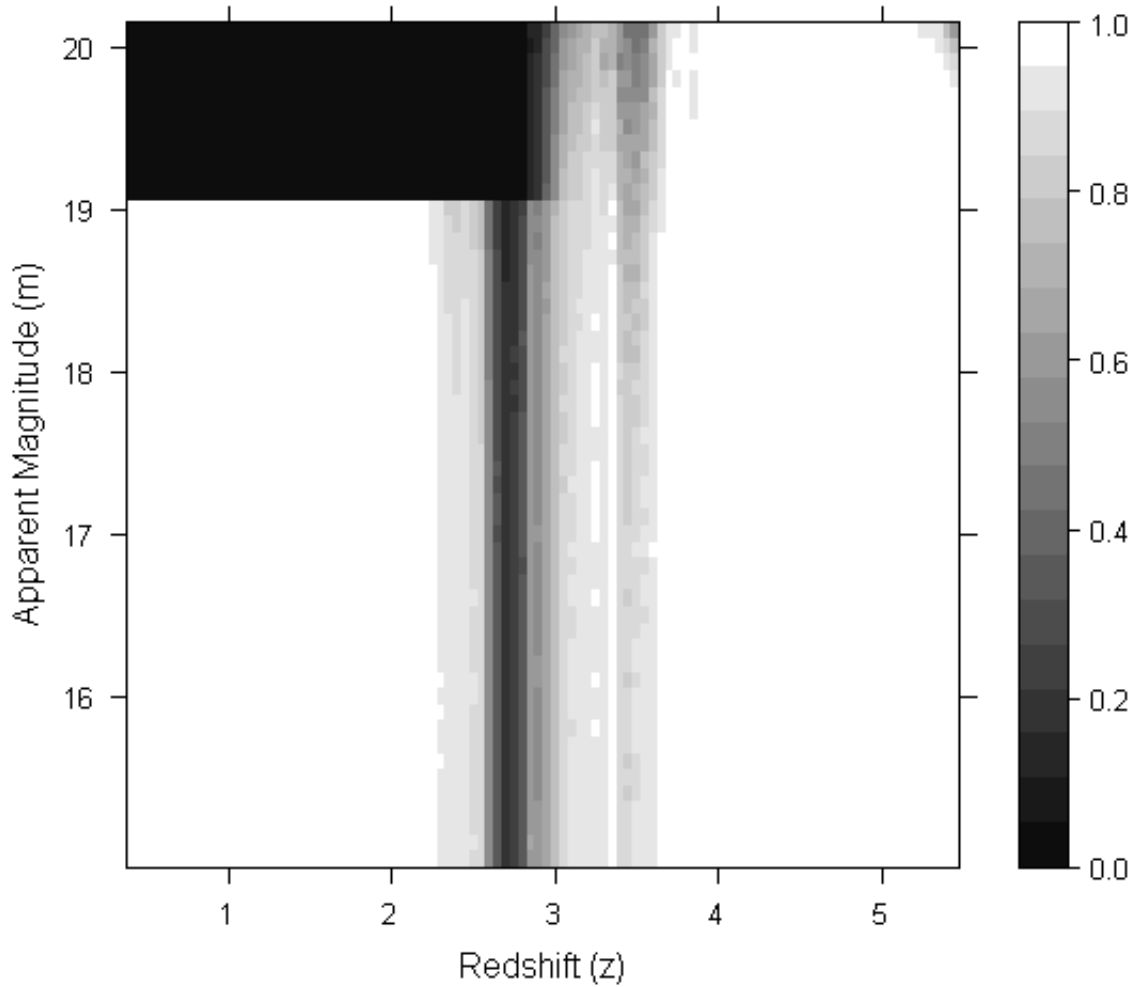


Figure 2.5: Heat map of the estimates of the observational probabilities on the apparent magnitude scale. White represents a high probability that a quasar with the given redshift and apparent magnitude will be included in the SDSS catalog, while black represents a low probability. These estimates were obtained via simulation by Richards et al. [23].

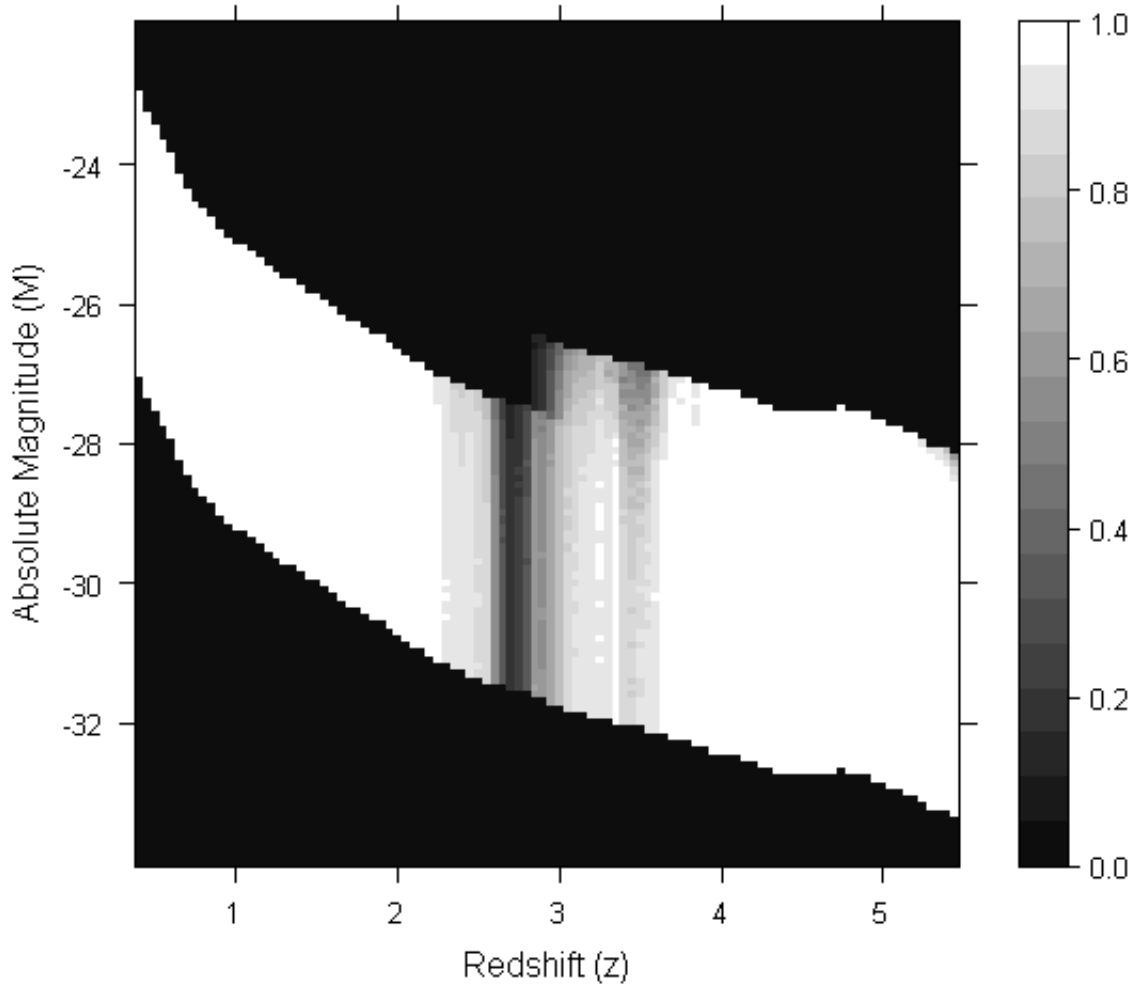


Figure 2.6: Heat map of the estimates of the observational probabilities on the absolute magnitude scale. For the purposes of visualization, this plot presents the observational probabilities on a grid, where ψ has been evaluated at the center point of each grid square. White represents a high probability that a quasar with the given redshift and absolute magnitude will be included in the SDSS catalog, while black represents a low probability. Note that transforming to the absolute magnitude scale changes the shape of the region with non-zero observational probability. The large black regions in the upper right and bottom left of this heat map represent those regions beyond the vertical limits in Figure 2.5.

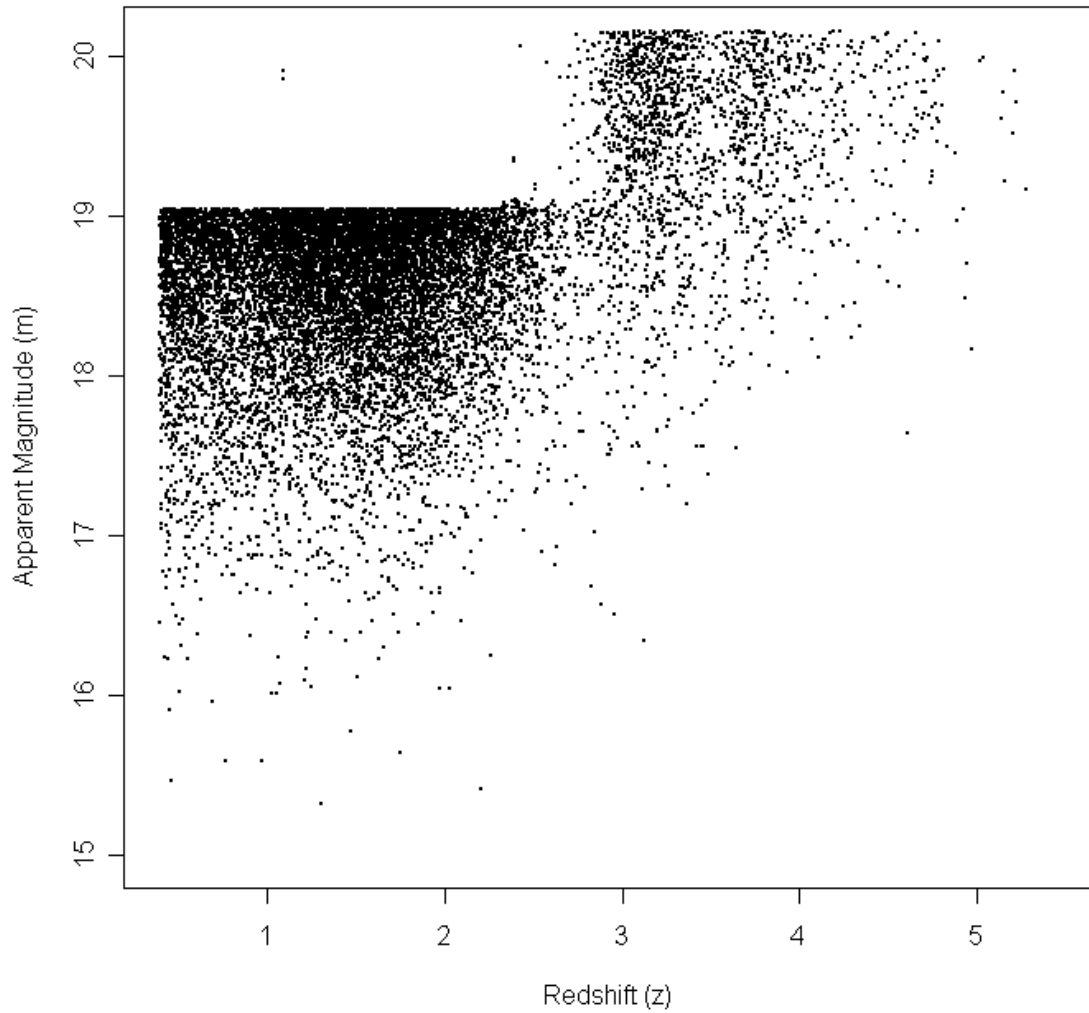


Figure 2.7: Scatterplot of SDSS quasars on the apparent magnitude scale. The change in the horizontal boundary at the top is due to the use of two different filters in the selection procedure during data collection. The thinner area visible near redshift $z=2.75$ is due to the low observational probability in that region.

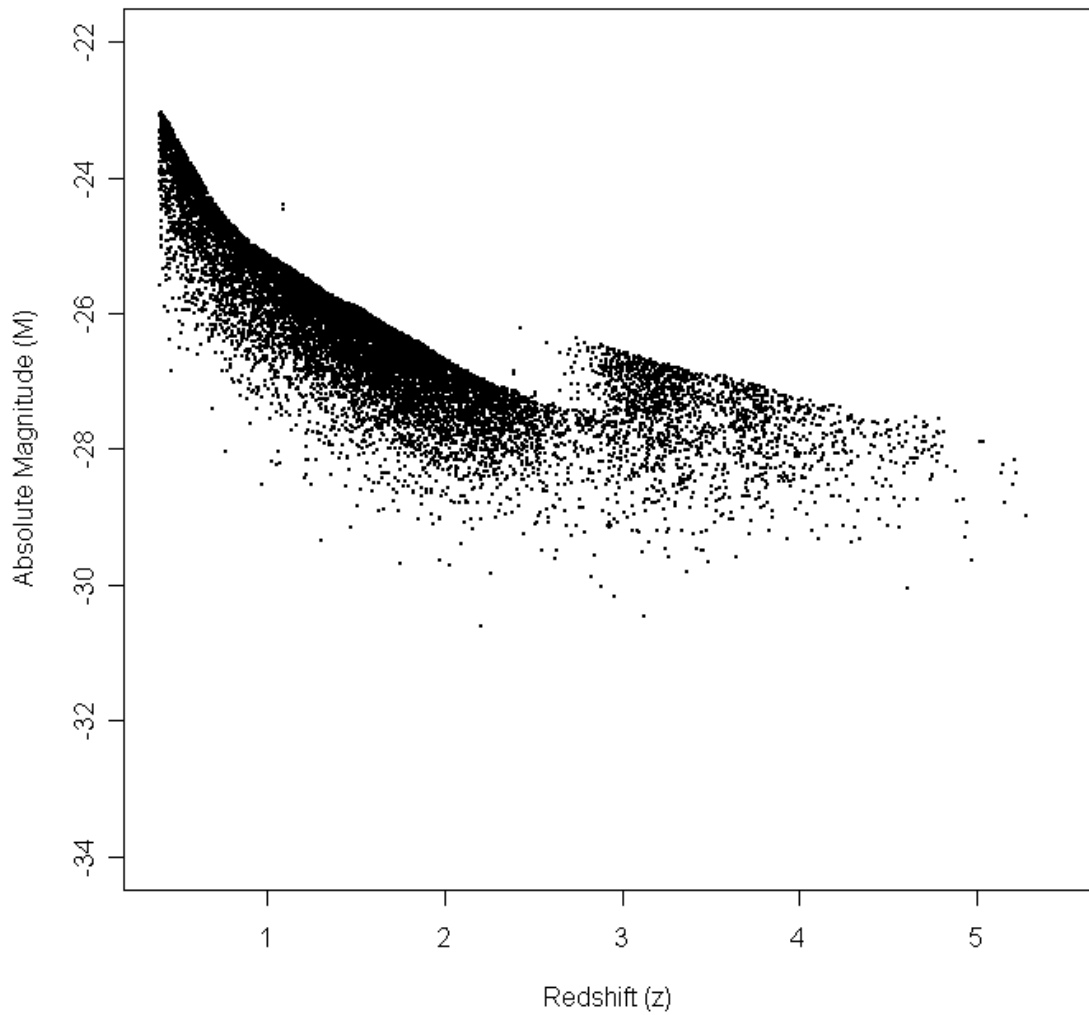


Figure 2.8: Scatterplot of SDSS quasars on the absolute magnitude scale. The change in the horizontal boundary at the top is due to the use of two different filters in the selection procedure during data collection. The thinner area visible near redshift $z=2.75$ is due to the low observational probability in that region.

The observational probability estimates ψ obtained by Richards et al. [24] are defined on a grid in the space of apparent luminosity m and redshift. The transformation from the space of absolute magnitude into the space of apparent luminosity is

$$(M, z) = (m - \text{DM}(z) - \text{K}(z), z) \quad (2.3)$$

where $\text{DM}(z)$ and $\text{K}(z)$ are the distance modulus [3] and K correction [12] at redshift z . Section A.1 describes the mathematical details of the transformed Poisson process.

Let $|\lambda_{\theta}|$ denote the integral of the Poisson process rate surface $\lambda_{\theta}(M, z)$ over the support region R . Let \mathbb{X} denote the observed quasar data with sample size N . Then the likelihood for a specific parameterization of the idealized functional form Φ is

$$\mathcal{L}(\theta|\mathbb{X}) = \frac{e^{-|\lambda_{\theta}|} |\lambda_{\theta}|^N}{N!} \cdot \prod_{i=1}^N \frac{\lambda_{\theta}(M_i, z_i)}{|\lambda_{\theta}|} \quad (2.4)$$

$$= \frac{e^{-|\lambda_{\theta}|}}{N!} \cdot \prod_{i=1}^N \lambda_{\theta}(M_i, z_i) . \quad (2.5)$$

and the log-likelihood is

$$l(\theta|\mathbb{X}) = \ln \mathcal{L}(\theta|\mathbb{X}) = -|\lambda_{\theta}| + \sum_{i=1}^N \ln(\lambda_{\theta}(M_i, z_i)) - \ln(N!) . \quad (2.6)$$

Let $\gamma(z) dz$ be the infinitesimal volume differential (measured in cubic megaparsecs) of a spherical shell about the Earth at a given redshift z . Then

$$|\lambda_{\theta}| = \int_z \int_M \lambda_{\theta}(M, z) \gamma(z) dM dz \quad (2.7)$$

$$= \int_z \int_M \Phi_{\theta}(M, z) \psi_M(M, z) \gamma(z) dM dz \quad (2.8)$$

and

$$\lambda_{\boldsymbol{\theta}}(M_i, z_i) = \Phi_{\boldsymbol{\theta}}(M_i, z_i) \psi_M(M_i, z_i), \quad (2.9)$$

which gives

$$\begin{aligned} l(\boldsymbol{\theta}|\mathbb{X}) &= - \int_z \int_M \Phi_{\boldsymbol{\theta}}(M, z) \psi_M(M, z) \gamma(z) dM dz + \\ &\quad + \sum_{i=1}^N \ln(\Phi_{\boldsymbol{\theta}}(M_i, z_i) \psi_M(M_i, z_i)) - \ln(N!) \end{aligned} \quad (2.10)$$

After removing all terms that do not depend on the model parameter $\boldsymbol{\theta}$, the maximum likelihood estimates for $\boldsymbol{\theta}$ under a fixed pair of model complexity levels (A, B) is obtained by maximizing the function $h(\boldsymbol{\theta}|\mathbb{X})$ with respect to $\boldsymbol{\theta}$, where

$$h(\boldsymbol{\theta}|\mathbb{X}) = - \int_z \int_M \Phi_{\boldsymbol{\theta}}(M, z) \psi_M(M, z) \gamma(z) dM dz + \sum_{i=1}^N \ln[\Phi_{\boldsymbol{\theta}}(M_i, z_i)] \quad (2.11)$$

2.3.2 Estimation of Poisson Intensities on a Grid

To better understand the data, we assume that quasars follow a Poisson process with a flat intensity surface over each grid square, and that the Poisson rate parameter in each grid square is independent of all other grid squares.

For a given grid square S , let μ be the constant Poisson rate parameter such that

$$\Phi(M, z) = \mu \quad (2.12)$$

within that square. From Equation (2.11),

$$h(\boldsymbol{\mu}|\mathbb{X}) = - \int_{z \in S} \int_{M \in S} \mu \cdot \psi_M(M, z) \cdot \gamma(z) dM dz + \sum_{i=1}^n \ln \mu \quad (2.13)$$

$$= -\mu \int_{z \in S} \int_{M \in S} \psi_M(M, z) \cdot \gamma(z) dM dz + n \cdot \ln \mu \quad (2.14)$$

where n is the number of quasars observed in that grid square. In order to determine the maximum likelihood estimate $\hat{\mu}$ for quasar rate within a given grid square, we differentiate to get

$$\frac{\partial h}{\partial \mu} = - \int_{z \in S} \int_{M \in S} \psi_M(M, z) \cdot \gamma(z) dM dz + \frac{n}{\mu}. \quad (2.15)$$

Setting the partial derivative equal to zero and solving for μ gives

$$\hat{\mu} = \frac{n}{\int_{z \in S} \int_{M \in S} \psi_M(M, z) \cdot \gamma(z) dM dz} \quad (2.16)$$

The integral in Equation (2.16) was approximated using the adaptive numerical integration function from the **adapt** package in R, and a level map of the resulting MLE estimates of $\hat{\lambda}$ for each grid square is shown in Figure 2.9. The resolution of the grid was set to 0.05 on the redshift scale and 0.1 on the absolute magnitude scale. Note that there is an anomaly near the break between the two filters near redshift $z=2.75$, and also an anomaly above the thrust of the data near redshift $z=1.0$. These fluctuations are likely due to a misspecification of the observational probabilities in those region, as opposed to an actual physical change in the quasar rate. For this reason, it is important that any analysis conducted on the data be examined carefully to ensure that the results are not highly influenced by these portions of the data.

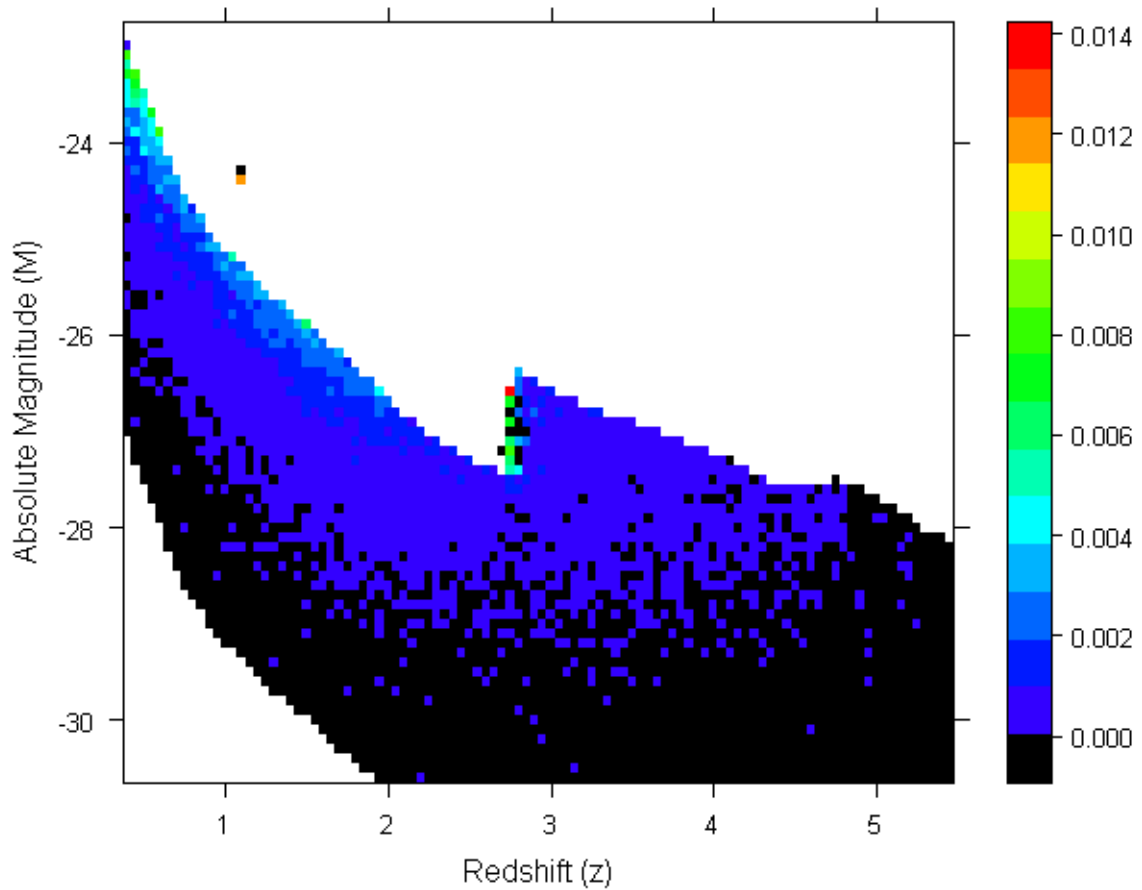


Figure 2.9: Preliminary analysis: Maximum likelihood estimates of quasar rate $\hat{\mu}$ on a grid, where we assume that quasars follow a Poisson process with a flat intensity surface over each grid square, and that the rate in each grid square is independent of all other grid squares. This analysis reveals that there may be problems with the estimated observational probabilities in the region near redshift $z=2.75$, as well as in the small blip that appears above the thrust of the data near redshift $z=1.0$. The portions of the graph that correspond to physical regions with zero observational probability have been filled with white. A black grid square represents a region with non-zero observational probability that has an estimated quasar rate of zero because there were no quasars observed in that region.

2.3.3 A Flexible Parameterization for the Quasar Luminosity

Restricted Model with No Interaction Term

A flexible parameterization of the idealized functional form for the quasar luminosity function with model parameter $\boldsymbol{\theta} = (\mu, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is

$$\Phi_{\boldsymbol{\theta}}(M, z) = 10^{\mu + f_1(M; \boldsymbol{\alpha}) + f_2(z; \boldsymbol{\beta})} \quad (2.17)$$

where

$$f_1(M; \boldsymbol{\alpha}) = \sum_{i=1}^A \alpha_i \cdot (M - M_0)^i, \quad (2.18)$$

$$f_2(z; \boldsymbol{\beta}) = \sum_{j=1}^B \beta_j \cdot \zeta(z)^j, \quad (2.19)$$

and

$$\zeta(z) = \log_{10} \left(\frac{1+z}{1+z_0} \right). \quad (2.20)$$

Note that in this functional form the terms for luminosity and redshift are separable. The expression $\zeta(z)$ has a scientific interpretation related to evolution, which is described in Wolf et al. [33]. Following Richards et al. [24], we use $M_0 = -26$ mags and redshift $z_0 = 2.45$. The flexibility of A and B in the limits of the summations in f_1 and f_2 allow for the introduction of higher-order terms in absolute magnitude and redshift if warranted by the data.

Our model selection algorithm begins by fitting all possible combinations of models with $A \in \{1, 2, 3, 4\}$ and $B \in \{1, 2, \dots, 8\}$. The Bayesian Information Criterion [27]

$$BIC(\hat{\boldsymbol{\theta}}) = -2 l(\hat{\boldsymbol{\theta}}|\mathbb{X}) + k \ln N \quad (2.21)$$

is then used to determine which of the fitted models is preferred. Here, $k = 1+A+B$ is the dimensionality of the model parameter θ .

An Extended Model with an Interaction Term

In order to test whether there might be an interaction between redshift and absolute magnitude, an interaction term $f_3(M, z)$ can be added into the exponent of the original model:

$$\Phi_{\theta}(M, z) = 10^{\mu + f_1(M; \alpha) + f_2(z; \beta) + f_3(M, z; \delta)} \quad (2.22)$$

where

$$f_1(M; \alpha) = \sum_{i=1}^A \alpha_i \cdot (M - M_0)^i, \quad (2.23)$$

$$f_2(z; \beta) = \sum_{j=1}^B \beta_j \cdot \zeta(z)^j, \quad (2.24)$$

$$f_3(M, z; \delta) = \sum_{k=1}^D \delta_k \cdot ((M - M_0) - \zeta(z))^{k+1}, \quad (2.25)$$

and

$$\zeta(z) = \log_{10} \left(\frac{1+z}{1+z_0} \right). \quad (2.26)$$

Note that the case $D=1$ has a quadratic term in $f_3(\cdot)$, while $D=2$ includes a quadratic and cubic term, and so on. We will define $D=0$ to represent the original restricted model that does not include an interaction term.

Under this extended model, we begin by fitting all possible combinations of models with $A \in \{1, 2, 3, 4\}$, $B \in \{1, 2, \dots, 8\}$, and $D \in \{0, 1, 2, 3\}$. Once again, the Bayesian Information Criterion is used to determine which of the fitted models is preferred, and in this case the dimensionality of the model parameter θ is $k = 1+A+B+D$.

2.3.4 Maximizing the Likelihood

Maximizing the function $h(\boldsymbol{\theta}|\mathbb{X})$ in Equation (2.11) directly requires the repeated computation of a double integral over the entire $M \times z$ support. To make this feasible, a few computational tricks were employed in order to collapse $h(\boldsymbol{\theta}|\mathbb{X})$ down to a simple calculation based on quantities that are computed only once.

The goal is to maximize the function $h(\boldsymbol{\theta}|\mathbb{X})$ with respect to $\boldsymbol{\theta}$, where

$$h(\boldsymbol{\theta}|\mathbb{X}) = - \int_z \int_M \Phi_{\boldsymbol{\theta}}(M, z) \psi_M(M, z) \gamma(z) dM dz + \sum_{i=1}^N \ln [\Phi_{\boldsymbol{\theta}}(M_i, z_i)] \quad (2.27)$$

$$= \text{Model Part} + \text{Data Part} \quad (2.28)$$

The Model Part of the Likelihood

The observational probabilities ψ are considered to be zero outside the range $m \in [15.0, 20.1]$ and $z \in [0.4, 5.4]$, and the Jacobian of the transformation between apparent and absolute magnitude is one (see Section A.1). Therefore, the Model Part in (2.28) can be rewritten in terms of apparent magnitude, so that

$$\text{Model Part} = - \int_z \int_M \Phi_{\boldsymbol{\theta}}(M, z) \psi_M(M, z) \gamma(z) dM dz \quad (2.29)$$

$$= - \int_z \int_m \Phi_{\boldsymbol{\theta}}^*(m, z) \psi(m, z) \gamma(z) dm dz \quad (2.30)$$

$$= - \sum_{r=1}^R \left(\psi_r \cdot \iint_{(m,z) \in r} \Phi_{\boldsymbol{\theta}}^*(m, z) \gamma(z) dm dz \right) \quad (2.31)$$

where $r=1, \dots, R$ represent the grid squares as defined by the table of observational probabilities, and $\Phi_{\boldsymbol{\theta}}^*(m, z)$ is the idealized functional form Φ on the apparent magnitude scale.

The integral can then be approximated by a 3×3 -point gaussian-quadrature summation within each grid square, such that

$$\begin{aligned}
 \text{Model Part} &\approx - \sum_{r=1}^R \left(\psi_r \cdot \int_{m \in r} \int_{z \in r} \Phi_{\theta}^*(m, z) \gamma(z) dm dz \right) \quad (2.32) \\
 &\approx - \sum_{r=1}^R \left(\psi_r \cdot \sum_{i=1}^3 \sum_{j=1}^3 [w_i \cdot w_j \cdot \right. \\
 &\quad \left. \cdot \left(\frac{W_m}{2} \right) \cdot \left(\frac{W_z}{2} \right) \Phi_{\theta}^*(m_i, z_j) \cdot \gamma(z_j)] \right) \quad (2.33)
 \end{aligned}$$

where the three-point gaussian quadrature nodes and weights w_i are as described by Abramowitz and Stegun [1]. The widths of the grid squares were $W_m=0.1$ mags and $W_z=0.05$ redshifts.

If we let P be the set of all gaussian quadrature evaluation points over the entire $m \times z$ support, and let $w_p = w_i \cdot w_j$ for a given evaluation point p , then we can write

$$\text{Model Part} \approx - \sum_{p \in P} \left[\psi_p \cdot w_p \cdot \gamma(z_p) \cdot \left(\frac{W_m}{2} \right) \cdot \left(\frac{W_z}{2} \right) \cdot \Phi_{\theta}^*(m_p, z_p) \right] \quad (2.34)$$

Next, let

$$v_p = - \psi_p \cdot w_p \cdot \gamma(z_p) \cdot \left(\frac{W_m}{2} \right) \cdot \left(\frac{W_z}{2} \right) \quad (2.35)$$

for each point p in the set of evaluation points P . Then under the general case of the idealized functional form that includes an interaction term, we can write

$$\text{Model Part} \approx \sum_{p \in P} v_p 10^{\mu} + f_1(M_p; \alpha) + f_2(z_p; \beta) + f_3(M_p, z_p; \delta) \quad (2.36)$$

Thus, the Model Part can be expressed as

$$\text{Model Part} \approx \sum_{p \in P} v_p \cdot 10^{\vec{u}_p \cdot \vec{\theta}} \quad (2.37)$$

where the model vector \vec{u}_p in the exponent is expressed as

$$\vec{u}_p = \begin{bmatrix} 1 \\ (M_p - M_0) \\ (M_p - M_0)^2 \\ \vdots \\ (M_p - M_0)^A \\ \zeta(z_p) \\ (\zeta(z_p))^2 \\ \vdots \\ (\zeta(z_p))^B \\ ((M_p - M_0) - \zeta(z_p))^2 \\ ((M_p - M_0) - \zeta(z_p))^3 \\ \vdots \\ ((M_p - M_0) - \zeta(z_p))^{D+1} \end{bmatrix} \quad (2.38)$$

and $\vec{\theta}$ is given by

$$\vec{\theta} = (\mu, \alpha_1, \dots, \alpha_A, \beta_1, \dots, \beta_B, \delta_1, \dots, \delta_D) \quad (2.39)$$

A desirable feature of writing the Model Part of the likelihood in this format is that the model parameter $\vec{\theta}$ has been isolated. Note that the values v_1, \dots, v_P and each of the components of the \vec{u}_p vector are independent of $\vec{\theta}$. Therefore, these values need only be computed once.

The Data Part of the Likelihood

For the Data Part in (2.28) we have

$$\text{Data Part} = \sum_{i=1}^N \ln [\Phi_{\boldsymbol{\theta}}(M_i, z_i)] \quad (2.40)$$

$$= \sum_{i=1}^N \ln [10^{\mu} + f_1(M_i; \boldsymbol{\alpha}) + f_2(z_i; \boldsymbol{\beta}) + f_3(M_i, z_i; \boldsymbol{\delta})] \quad (2.41)$$

$$= \ln(10) \cdot \left[N\mu + \sum_{i=1}^N f_1(M_i; \boldsymbol{\alpha}) + \sum_{i=1}^N f_2(z_i; \boldsymbol{\beta}) + \sum_{i=1}^N f_3(M_i, z_i; \boldsymbol{\delta}) \right] \quad (2.42)$$

Substituting in the equations for f_1 , f_2 , and f_3 gives

$$\begin{aligned} \text{Data Part} &= \ln(10) \cdot N\mu \\ &+ \ln(10) \cdot \sum_{i=1}^N \sum_{j=1}^A \alpha_j (M_i - M_0)^j \\ &+ \ln(10) \cdot \sum_{i=1}^N \sum_{j=1}^B \beta_j \zeta(z_i)^j \\ &+ \ln(10) \cdot \sum_{i=1}^N \sum_{j=1}^D \delta_j ((M_i - M_0) - \zeta(z_i))^{j+1} \end{aligned} \quad (2.43)$$

Next we can isolate the model parameters by writing

$$\begin{aligned}
 \text{Data Part} &= \mu \cdot (N \ln(10)) \\
 &+ \sum_{j=1}^A \alpha_j \cdot \left[\ln(10) \cdot \sum_{i=1}^N (M_i - M_0)^j \right] \\
 &+ \sum_{j=1}^B \beta_j \cdot \left[\ln(10) \cdot \sum_{i=1}^N \zeta(z_i)^j \right] \\
 &+ \sum_{j=1}^D \delta_j \cdot \left[\ln(10) \cdot \sum_{i=1}^N ((M_i - M_0) - \zeta(z_i))^{j+1} \right] \quad (2.44)
 \end{aligned}$$

Thus, the Data Part can be expressed as

$$\text{Data Part} = \vec{x} \cdot \vec{\theta} \quad (2.45)$$

where the data vector \vec{x} is

$$\vec{x} = \begin{bmatrix} N \ln(10) \\ \sum_{i=1}^N (M_i - M_0) \\ \sum_{i=1}^N (M_i - M_0)^2 \\ \vdots \\ \sum_{i=1}^N (M_i - M_0)^A \\ \sum_{i=1}^N \zeta(z_i) \\ \sum_{i=1}^N (\zeta(z_i))^2 \\ \vdots \\ \sum_{i=1}^N (\zeta(z_i))^B \\ \sum_{i=1}^N ((M_i - M_0) - \zeta(z_i))^2 \\ \sum_{i=1}^N ((M_i - M_0) - \zeta(z_i))^3 \\ \vdots \\ \sum_{i=1}^N ((M_i - M_0) - \zeta(z_i))^{D+1} \end{bmatrix} \quad (2.46)$$

and the parameter vector $\vec{\theta}$ is as described in Equation 2.39. The vector \vec{x} does not depend on the parameter vector $\vec{\theta}$, and therefore we need only compute \vec{x} once for each data set.

Evaluating the Likelihood

By the two computational reductions described above, each evaluation of the log-likelihood during the maximization process can be computed by evaluating

$$\begin{aligned}
 h(\boldsymbol{\theta}|\mathbb{X}) &= - \int_z \int_M \Phi_{\boldsymbol{\theta}}(M, z) \psi_M(M, z) \gamma(z) dM dz + \sum_{i=1}^N \ln [\Phi_{\boldsymbol{\theta}}(M_i, z_i)] \\
 &= \text{Model Part} + \text{Data Part} \\
 &\approx \sum_{p \in P} v_p \cdot 10^{\vec{u}_p \cdot \vec{\theta}} + \vec{x} \cdot \vec{\theta} \tag{2.47}
 \end{aligned}$$

where v_p and \vec{u}_p are pre-computed once for each evaluation point P , and \vec{x} is pre-computed once for each data set. Once the pre-computation step is complete, this reduction makes it possible to obtain the maximum likelihood fitted density for a given data set in less than one minute of computational time using the R statistical package on a Linux machine with a Pentium 4 processor and one gigabyte of RAM.

2.3.5 Fitted Models for the Quasar Luminosity Function

Fitted Results under the No-Interaction Model

Table 2.1 shows the complexity levels and log-likelihoods for the top 15 models sorted by the BIC selection criteria in the model without an interaction term. All 15 models have high-degree polynomials in the redshift term, and the first seven prefer at least a quadratic polynomial in the absolute magnitude term f_1 .

Figure 2.10 presents a heat map of the exponent of the fitted idealized functional form $\Phi_{\boldsymbol{\theta}}(M, z)$ under the selected model. The selected model has $(A, B)=(2, 6)$ with

$$\begin{aligned}
 \mu &= -2.16361 \\
 \alpha &= (0.72135, -0.01992) \\
 \beta &= (-0.59212, -26.84461, 1.71612, 186.3441, 199.45025, 171.66021)
 \end{aligned}$$

A	B	$\ell(X)$	BIC
2	6	-225454.0	450994.0
4	8	-225437.2	450998.6
4	4	-225456.8	450999.6
2	7	-225452.9	451001.3
3	5	-225461.1	451008.2
2	8	-225452.3	451009.7
3	6	-225464.8	451025.1
1	7	-225470.1	451026.2
3	7	-225464.9	451034.9
2	4	-225484.5	451035.9
1	5	-225485.9	451038.7
3	4	-225482.1	451040.6
1	8	-225472.7	451040.9
1	6	-225482.8	451042.0
2	5	-225488.7	451053.8

Table 2.1: BIC table for complete data under the no-interaction model. The selected model is printed in bold text, and the next best 15 models are sorted by their BIC value.

In order to assess the model fit, deviance residuals [6] were calculated by comparing the observed and expected quasar counts for each grid square on an $M \times z$ grid. The Poisson rate parameter $\hat{\mu}$ in a given grid square was determined by evaluating the integral in Equation (2.8) over that region. Let y be the observed quasar count in that square. We first compute the deviance under the fitted model

$$D = -2 \ln[p(y|\hat{\mu})] = -2 [-\hat{\mu} + y \ln \hat{\mu} - \ln(y!)] \quad (2.48)$$

and the deviance under the saturated model with Poisson rate parameter equal to the observed count

$$D_{sat} = -2 [-y + y \ln y - \ln(y!)] . \quad (2.49)$$

Note that the deviance under the saturated model is undefined when the observed count is

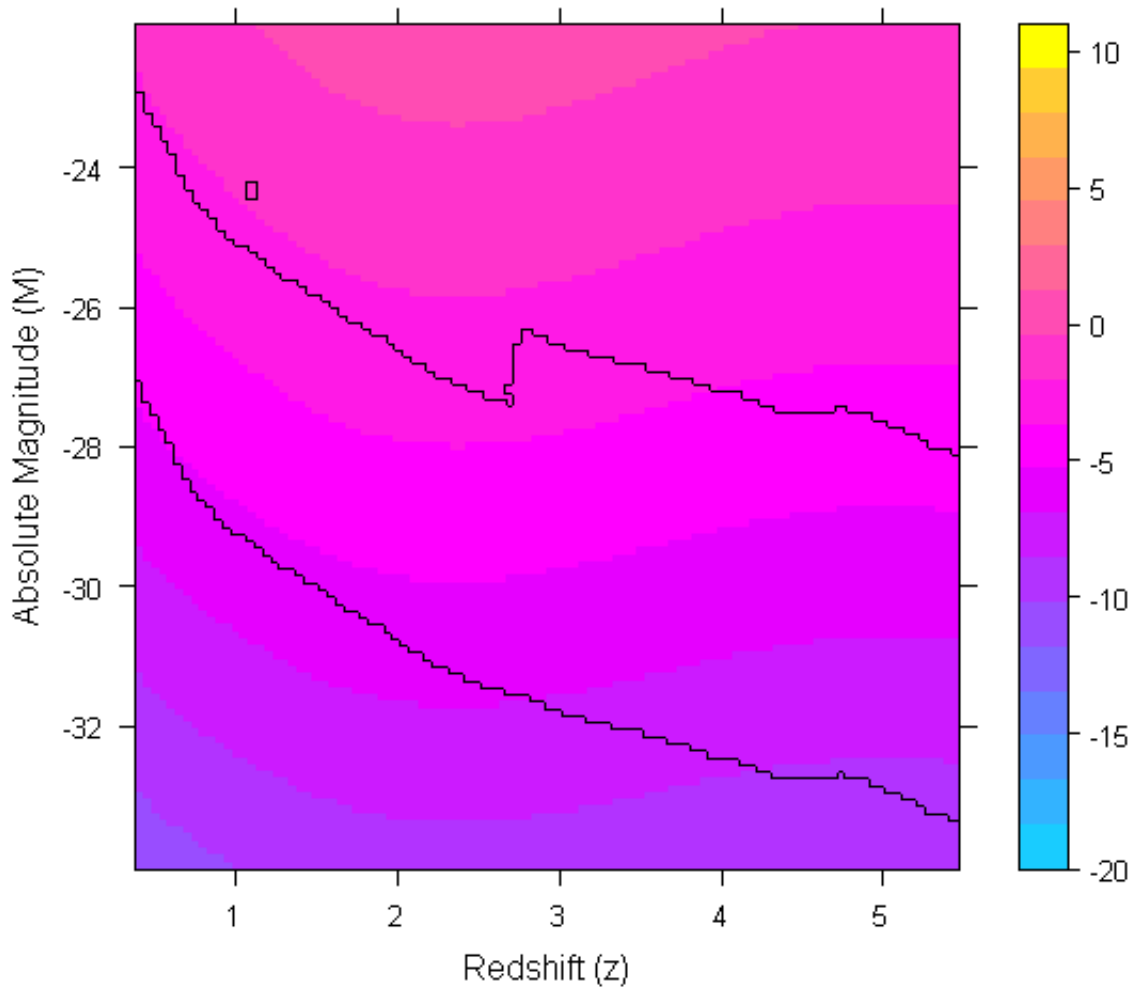


Figure 2.10: Heat map of fitted density under the no-interaction model. The density shown is the exponent of the fitted idealized functional form Φ , and the fitted model was constructed under the complete set of quasar data. The region of $M \times z$ space with non-zero observational probability has been marked with a bold outline. The fitted density outside this band is an extrapolation beyond the region where quasar data are available.

A	B	D	ℓ	$\Delta\ell$	BIC	Δ BIC
3	5	2	-225269.5	184.5	450644.1	-349.5
3	6	1	-225265.8	188.2	450636.7	-356.9
2	5	2	-225270.6	183.3	450636.8	-356.8
2	4	2	-225277.1	176.9	450640.2	-353.4
3	7	1	-225263.5	190.5	450641.7	-351.9
4	6	1	-225264.0	189.9	450642.7	-350.8
3	5	2	-225269.5	184.5	450644.1	-349.5
3	6	2	-225264.9	189.1	450644.4	-349.2
3	8	1	-225261.2	192.8	450646.6	-347.0
4	4	1	-225276.6	177.4	450648.7	-344.8
3	4	2	-225277.1	176.9	450649.6	-343.9
2	8	2	-225263.1	190.8	450650.5	-343.1
2	3	2	-225287.3	166.7	450651.0	-342.5
3	7	2	-225263.8	190.1	450651.9	-341.7
3	3	1	-225287.9	166.0	450652.3	-341.3

Table 2.2: BIC table for complete data under the model which allows for an interaction between absolute magnitude M and redshift z . The selected model is printed in bold text, and the next best 15 models are sorted by their BIC value. The differences $\Delta\ell$ and Δ BIC represent the differences when compared to the top model in the case where no interaction term is allowed.

zero. The deviance residual is

$$r_d = \text{sign}(y - \hat{\mu}) \sqrt{D - D_{sat}} . \quad (2.50)$$

A heat map of the deviance residuals for this model are plotted in Figure 2.11.

Fitted Results under the Extended Model

Table 2.2 shows the complexity levels and log-likelihoods for the top 15 models sorted by the BIC selection criteria in the extended case that allows for an interaction between absolute magnitude and redshift if warranted by the data. Note that all 15 models include the interaction term (that is, the degree $D+1$ of the polynomial in the interaction term is quadratic or better). In fact, the first 37 models sorted by BIC include this interaction term.

Figure 2.12 presents a heat map of the exponent of the fitted idealized functional form

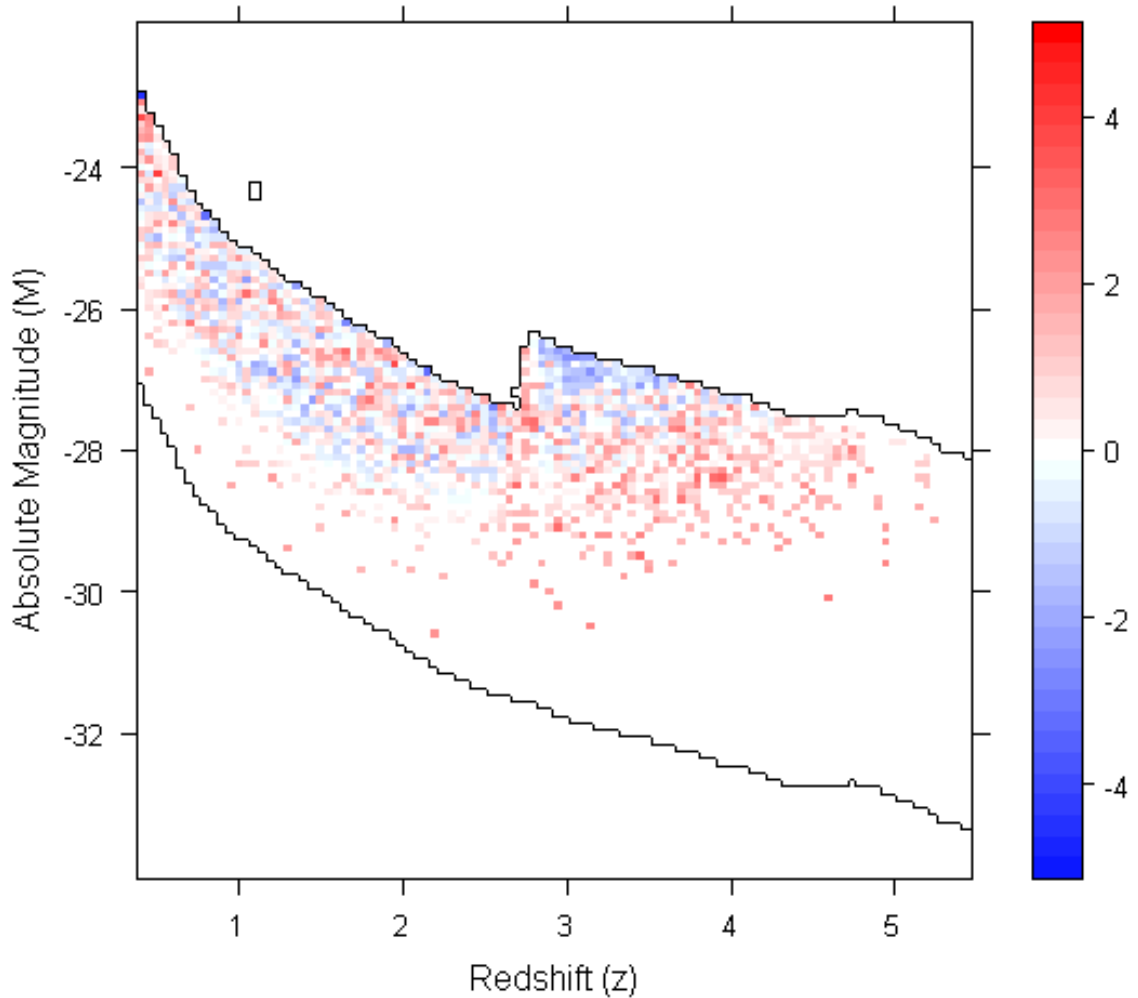


Figure 2.11: Deviance residuals r_d for fitted density under the no-interaction model. Red represents an area with more quasars observed than would be expected under the fitted model, and blue represents an area with fewer observed quasars than would be expected under the fitted model. The resolution of the grid was 0.1 on the absolute magnitude scale and 0.05 on the redshift scale, and the region of $M \times z$ space with non-zero observational probability has been marked with a bold outline.

$\Phi_{\theta}(M, z)$ under the selected model. This model has $(A, B, D)=(3, 5, 1)$ with

$$\begin{aligned}\mu &= -2.5436 \\ \alpha &= (0.25323, -1.16399, 0.00796) \\ \beta &= (-3.47642, -29.81521, -3.48286, 91.29804, 137.44952) \\ \delta &= (1.03913)\end{aligned}$$

Note that the log-likelihood of the data under this extended model is $\ell(X) = -225269.48$, which gives $\Delta\ell=184.49$ when compared to the selected model in the case where no interaction term is allowed.

Figure 2.13 shows the fitted density on the \log_{10} scale for the second best model by BIC. Note the similarity to the fitted density under the top selected model. This consistency provides some evidence that the resulting model is a good fit to the available data.

Deviance residuals r_d for the selected model were calculated in the same manner described in Section 2.3.5, and a plot of the residuals is shown in Figure 2.14. Note that the blue area which appeared at the top edge of the data near redshift $z=3$ in Figure 2.11 is no longer present here, which suggests that this extended model provides a better fit to the data in that region.

2.3.6 Fitted Models for Partitioned Quasar Data

The analysis in Section 2.3.2 suggests that there may be some problems with the estimates of the observational probabilities near redshift $z=2.75$. Therefore it is important to ensure that the interaction effect that was detected during the model fitting process is not simply the result of a misspecification of the observational probabilities.

One concern is that two separate filters were utilized in the detection of the quasars, and so there may in fact be some discrepancies between quasars detected in the low- and high-redshift regions. Another issue is that there is significantly less data available in the high redshift regions, so it may be desirable to test the analysis in those regions where we have more information to make sure that the interaction is indeed real.

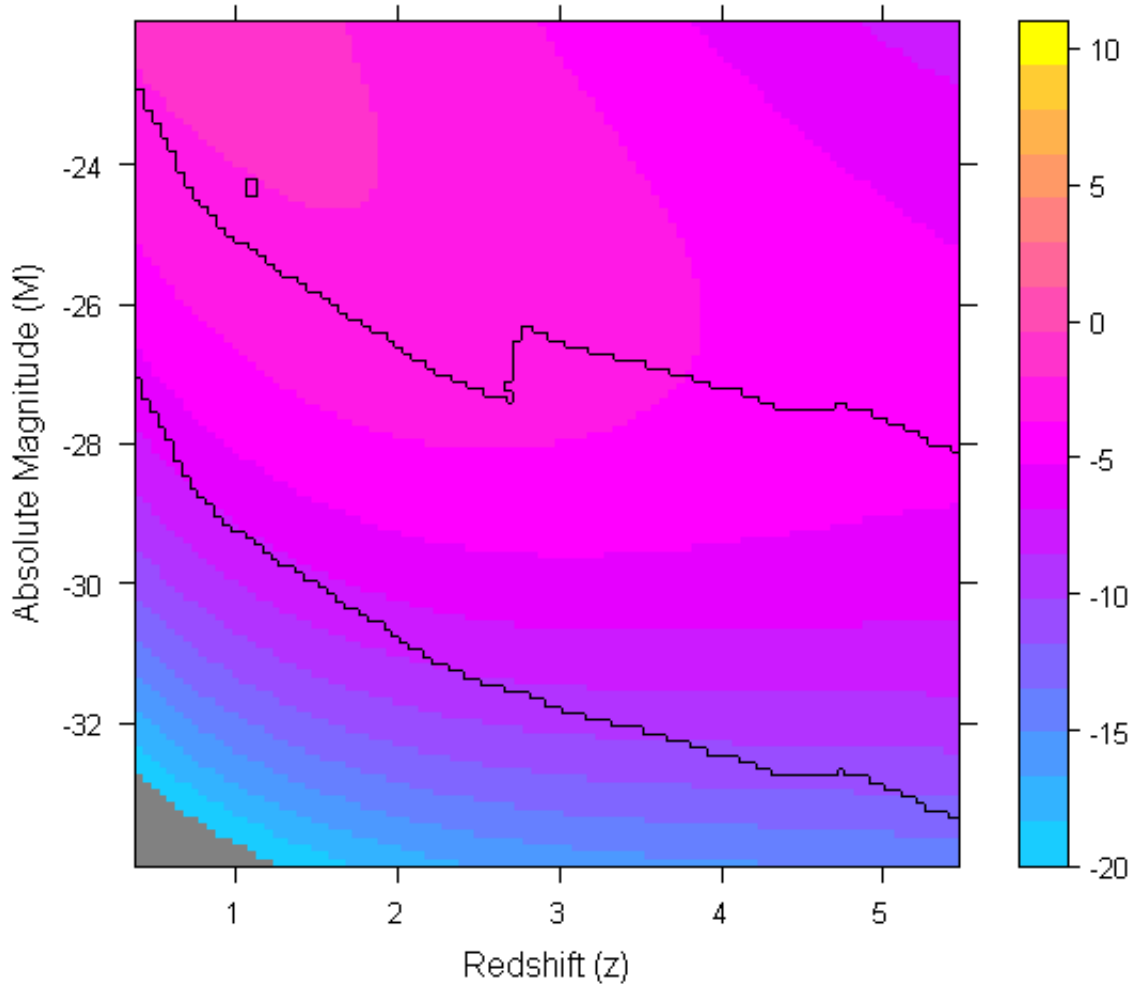


Figure 2.12: Heat map of fitted density under the model that allows for an interaction between absolute magnitude M and redshift z . The density shown is the exponent of the fitted idealized functional form Φ , and the fitted model was constructed under the complete set of quasar data. The region of $M \times z$ space with non-zero observational probability has been marked with a bold outline. The fitted density outside this band is an extrapolation beyond the region where quasar data are available. The gray region at the lower left corner indicates fitted values that were below -20 quasars per mag per megaparsec on the \log_{10} scale, and were therefore outside the range of color used in the graph.

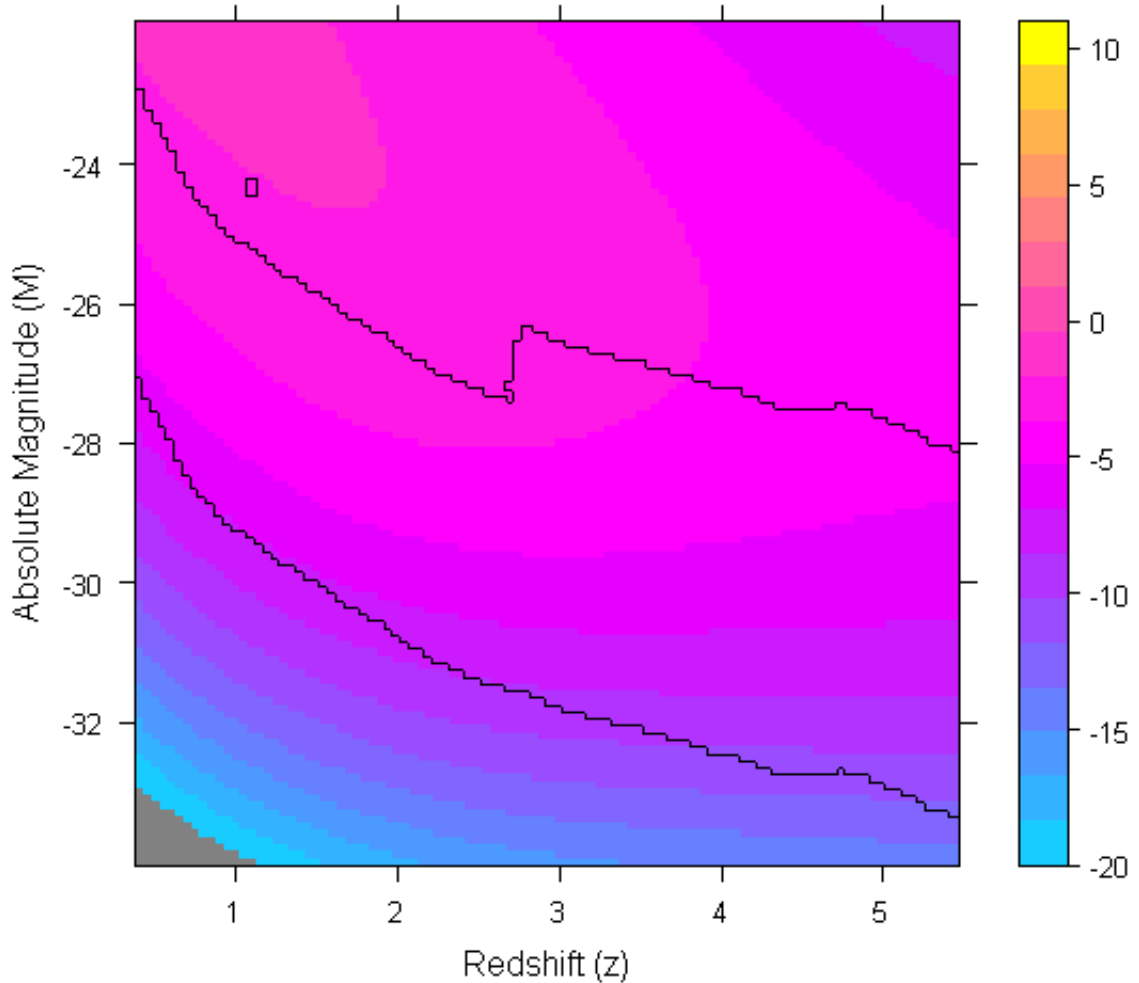


Figure 2.13: Heat map of second-best fitted density under the model that allows for an interaction between absolute magnitude M and redshift z . This was the second model selected by BIC, and it is presented here to demonstrate the similarity to the density in the top-selected model. This provides evidence that the fitted model is consistent with the data. Once again, the density shown is the exponent of the fitted idealized functional form Φ , and the fitted model was constructed under the complete set of quasar data. The region of $M \times z$ space with non-zero observational probability has been marked with a bold outline. The fitted density outside this band is an extrapolation beyond the region where quasar data are available. The gray region at the lower left corner indicates fitted values that were below -20 quasars per mag per megaparsec on the \log_{10} scale, and were therefore outside the range of color used in the graph.

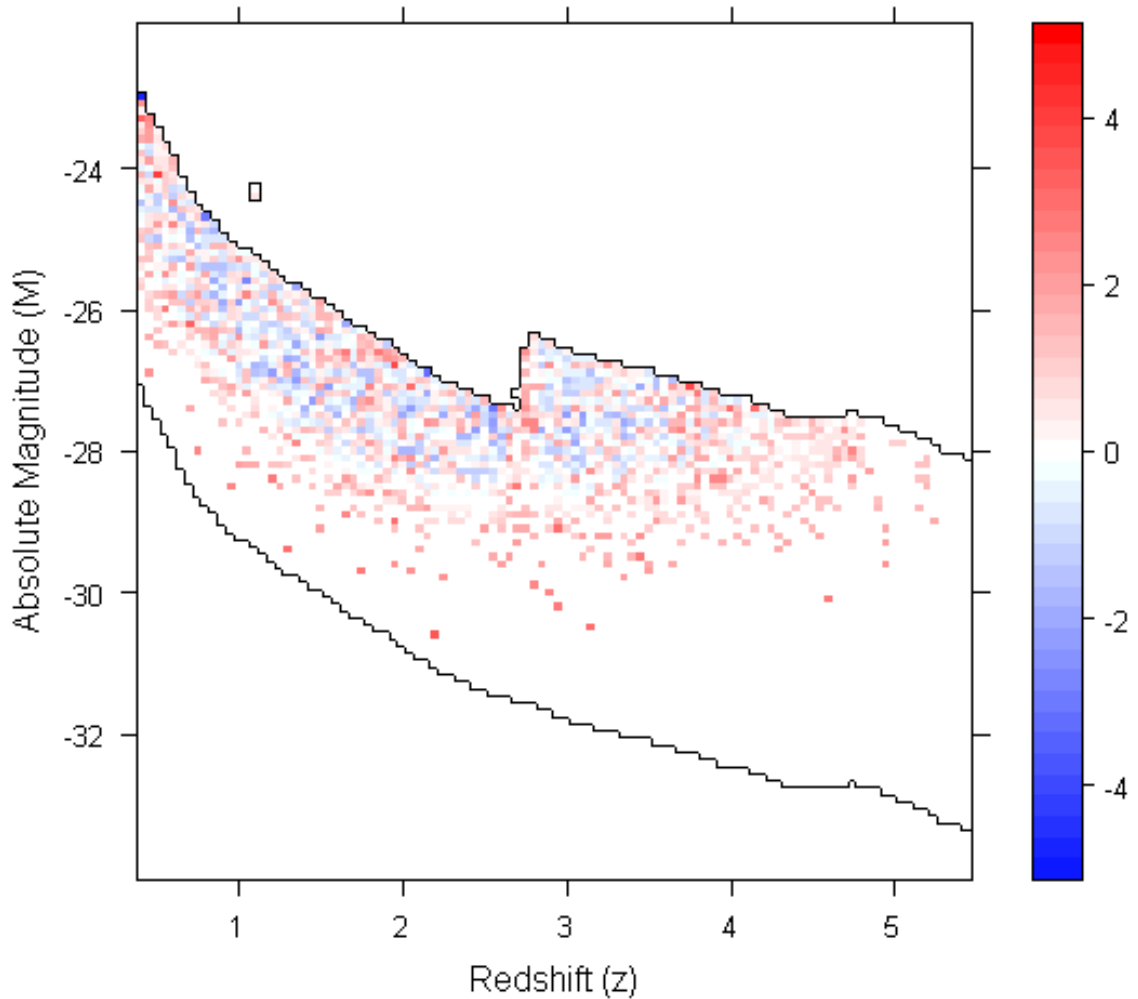


Figure 2.14: Deviance residuals r_d for fitted density under the model that allows for an interaction between absolute magnitude M and redshift z . Note that the blue area that appeared at the top edge of the data near redshift $z=3$ in Figure 2.11 is no longer present, which suggests that this model provides a better fit to the data in that region. In this figure, red represents an area with more quasars observed than would be expected under the fitted model, and blue represents an area with fewer observed quasars than would be expected under the fitted model. The resolution of the grid was 0.1 on the absolute magnitude scale and 0.05 on the redshift scale, and the region of $M \times z$ space with non-zero observational probability has been marked with a bold outline.

In order to test whether the inclusion of the interaction term was simply an artifact of errors in the specification of observational probabilities, the preceding analysis was repeated after partitioning the quasar data into two groups by redshift. The low-redshift group consists of $N_1=11,572$ quasars with redshift $z<2.5$ and the high-redshift group consists of $N_2=1,500$ quasars with redshift $z>3.0$. The subset of 1,041 quasars with redshift $z\in[2.5, 3.0]$ were omitted from this secondary analysis.

Figure 2.15 presents a side-by-side heat map of the exponent of the fitted idealized functional form $\Phi_{\theta}(M, z)$ for the two groups of data under the extended case that allows for an interaction term between absolute magnitude and redshift.

There are two key observations to be made here. First, the fitted model for the low-redshift quasars appears very similar to the selected model under the complete data set shown in Figure 2.12. The low-redshift region contains the vast majority of observed quasars, so it is encouraging to see that the detection of an interaction was not simply the result of an anomaly in the less frequent high-redshift quasar data. Second, note that the shape of the two densities shown side-by-side in Figure 2.15 appear well-matched for the region of $M \times z$ space with non-zero observational probability. Both of these observations provide evidence that the interaction detected in the data is real, and not simply an artifact of any misspecification in the observational probabilities.

The fitted model for the low-redshift group has $(A, B, D)=(2, 4, 2)$ with

$$\begin{aligned}\mu &= -2.49387 \\ \alpha &= (0.38457, -0.84892) \\ \beta &= (-6.73875, -79.94602, -230.27451, -268.53626) \\ \delta &= (0.74396, 0.00738)\end{aligned}$$

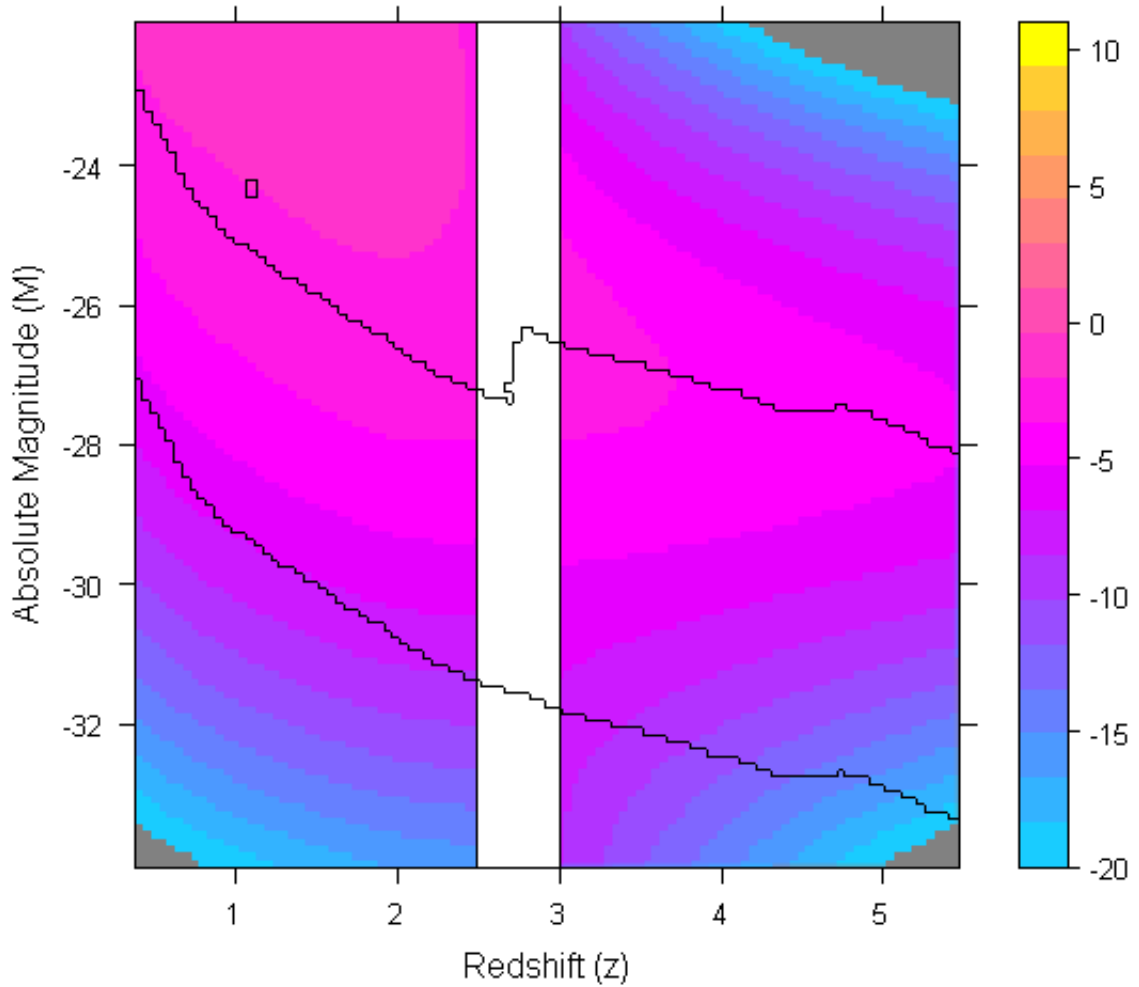


Figure 2.15: Side-by-side heat map of fitted densities for partitioned quasar data under the model that allows for an interaction between absolute magnitude M and redshift z . The model fit on the left side was obtained using only those quasars with redshift $z < 2.5$, while the model fit on the right side was obtained using only those quasars with redshift $z > 3.0$. Note that there is a large degree of agreement between the two densities. The densities shown are the exponential functions for the fitted idealized functional form Φ . The region of $M \times z$ space with non-zero observational probability has been marked with a bold outline. The fitted density outside this band is an extrapolation beyond the region where quasar data are available. The band of quasar data with redshift $z \in [2.5, 3.0]$ has been omitted from this analysis, and this region is indicated by vertical bars. The gray regions at the corners indicate fitted values that were below -20 quasars per mag per megaparsec on the \log_{10} scale, and were therefore outside the range of color used in the graph.

A	B	D	log-likelihood	BIC
2	4	2	-186581.3	373247.9
3	5	1	-186580.0	373254.8
2	5	2	-186580.2	373255.1
3	6	1	-186579.4	373263.1
2	6	2	-186579.5	373263.2
4	4	2	-186580.7	373265.7
3	4	2	-186587.4	373269.7
2	6	3	-186578.7	373271.2
3	7	1	-186579.1	373272.0
4	5	2	-186579.2	373272.2
3	6	2	-186579.5	373272.7
4	7	1	-186578.2	373279.6
4	6	2	-186578.5	373280.2
2	6	1	-186592.6	373279.9
2	8	2	-186579.0	373281.3

Table 2.3: BIC table for low-redshift group in the partitioned analysis. This table presents the top 15 models selected by BIC under the model which allows for an interaction between absolute magnitude M and redshift z . Each model was fit using only the low-redshift quasar data with $z \leq 2.5$.

while the fitted model for the high-redshift group has $(A, B, D) = (3, 2, 2)$ with

$$\begin{aligned} \mu &= -2.32728 \\ \alpha &= (0.50114, -5.54883, -0.68895) \\ \beta &= (-10.31918, -31.42667) \\ \delta &= (5.38225, 0.66555) \end{aligned}$$

Table 2.3 and Table 2.4 show the complexity levels and log-likelihoods for the top 15 models for each group of quasars sorted by BIC selection criteria.

A	B	D	log-likelihood	BIC
3	2	2	-33085.2	66229.0
2	2	1	-33095.4	66234.7
2	2	2	-33095.4	66241.9
2	3	1	-33095.4	66242.0
3	2	1	-33095.4	66242.0
3	4	2	-33085.1	66243.3
2	2	0	-33104.9	66246.4
2	3	2	-33095.4	66249.2
2	4	1	-33095.4	66249.2
3	5	2	-33084.9	66250.3
2	3	0	-33104.7	66253.2
3	2	0	-33104.9	66253.6
1	2	1	-33109.0	66254.6
2	5	1	-33095.4	66256.5
2	4	2	-33095.4	66256.6

Table 2.4: BIC table for high-redshift group in the partitioned analysis. This table presents the top 15 models selected by BIC under the model which allows for an interaction between absolute magnitude M and redshift z . Each model was fit using only the high-redshift quasar data with $z \geq 3.0$.

2.4 Simulation Studies

2.4.1 Technique for Simulating Quasar Data

On the apparent magnitude scale, quasars can be generated under a rectangular grid of observational probabilities ψ , where ψ is a constant over each grid square R . That is, for each $m \times z$ grid square R ,

1. Simulate the quasar count for the region R under model parameter θ by generating a Poisson random variable N_R with rate $|\lambda_\theta|_R$ where

$$|\lambda_\theta|_R = \iint_R \gamma(z) \Phi_\theta(M, z) \psi(m, z) dm dz \quad (2.51)$$

$$= \psi_R \iint_R \gamma(z) \Phi_\theta(M, z) dm dz \quad (2.52)$$

since $\psi(m, z)$ is constant over each $m \times z$ grid square R .

2. Generate N_R pairs (m_i, z_i) from within R by rejection sampling.

To construct the simulated data sets, the integral in step one was computed using two dimensional 3-point gaussian quadrature approximation within each grid square, and the maximum value for the rejection sampling in step two was numerically approximated using a Newton-type non-linear maximization algorithm from the **stats** package in R.

2.4.2 Performance of Model Fitting Methodology on Quasar Data

Simulated under Randomly Generated Models

For this study, we began by generating random models under the no-interaction model of the idealized functional form. We then simulated quasar data under each of the randomly generated models, and then tested the performance of our model selection methodology on simulated data.

The rule used to generate the random models was as follows: First, select A and B uniformly at random from the support set of integer values; Next, generate each element of the parameter vectors α and β by sampling uniformly on the set $[-10.0, 10.0]$; last, select the parameter μ to be a constant such that the expected number of simulated quasars is 15,000 so as to approximate the size of the quasar data.

The performance of the model fitting methodology was examined on 25 simulated quasar data sets, and scatterplots for two typical simulated quasar data sets are presented in Figure 2.16 and Figure 2.17. Two summary tables for this study are shown in Table 2.5 and Table 2.6. Note that the model fitting methodology seems to do a reasonably good job of discerning the underlying true models and does not show a tendency to prefer overly simplistic or complex models.

2.4.3 Performance of Model Fitting Methodology on Quasar Data Simulated under SDSS Fitted Model

In this simulation study, ten data sets were simulated under the selected model from Section 2.3.5 which was under the case which does not allow for an interaction between absolute magnitude and redshift. Scatterplots of two of the resulting data sets are shown in Figure 2.18 and Figure 2.19. These scatterplots can be visually compared to the plot of the real data in Figure 2.8. Fitted model complexity terms for all ten simulated data sets are shown in Table 2.7.

This simulation study seems to suggest that the BIC penalty term may be too lenient, as this model selection method seems to prefer overly complex models.

2.5 Summary

We have presented a flexible parametric model for quasars as a two-dimensional inhomogeneous poisson process that allows for the introduction of higher-order terms and an interaction term between absolute magnitude and redshift if warranted by the data.

Our findings indicate that there is some evidence of an interaction between redshift and absolute magnitude present in the data. This is in contrast to the popular scientific theory

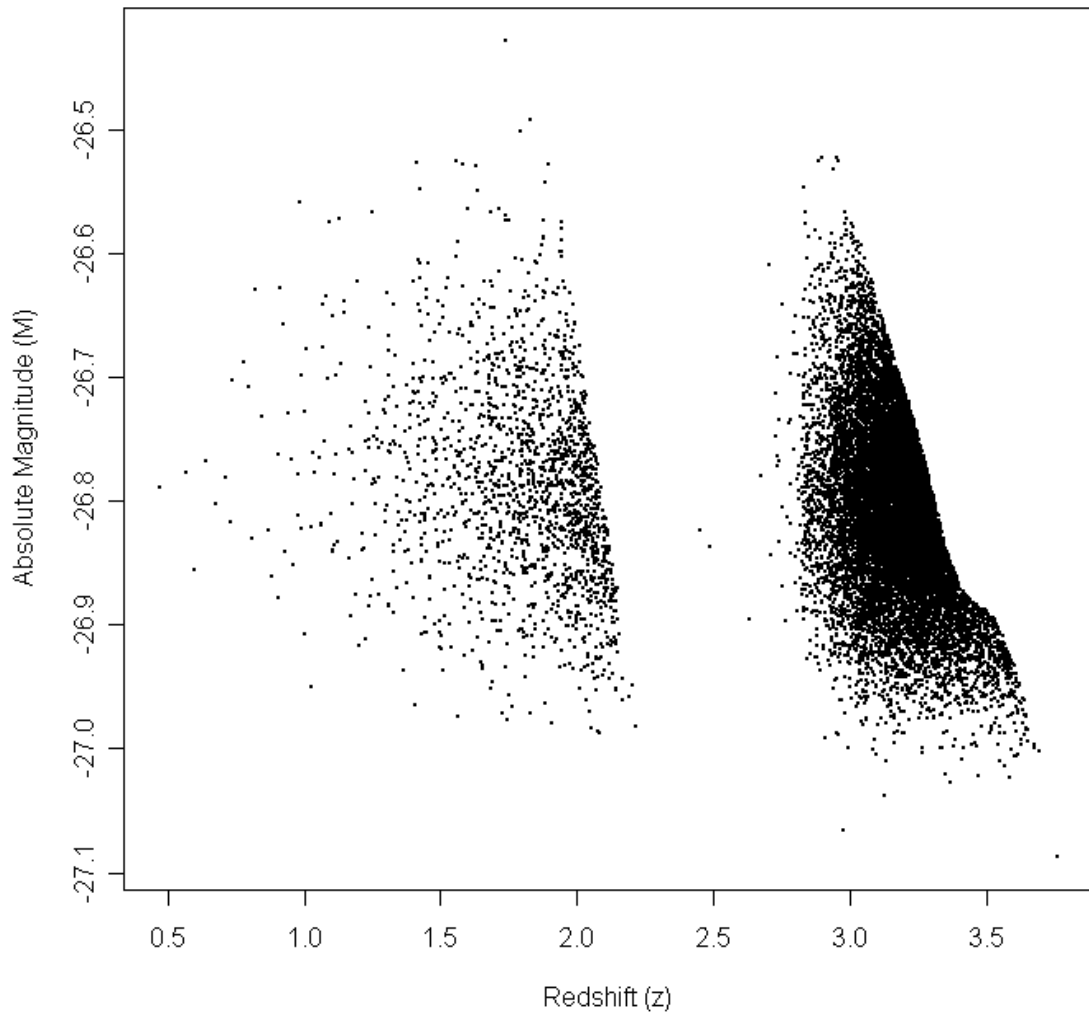


Figure 2.16: Scatterplot of simulated quasar dataset #1 under a randomly generated model.

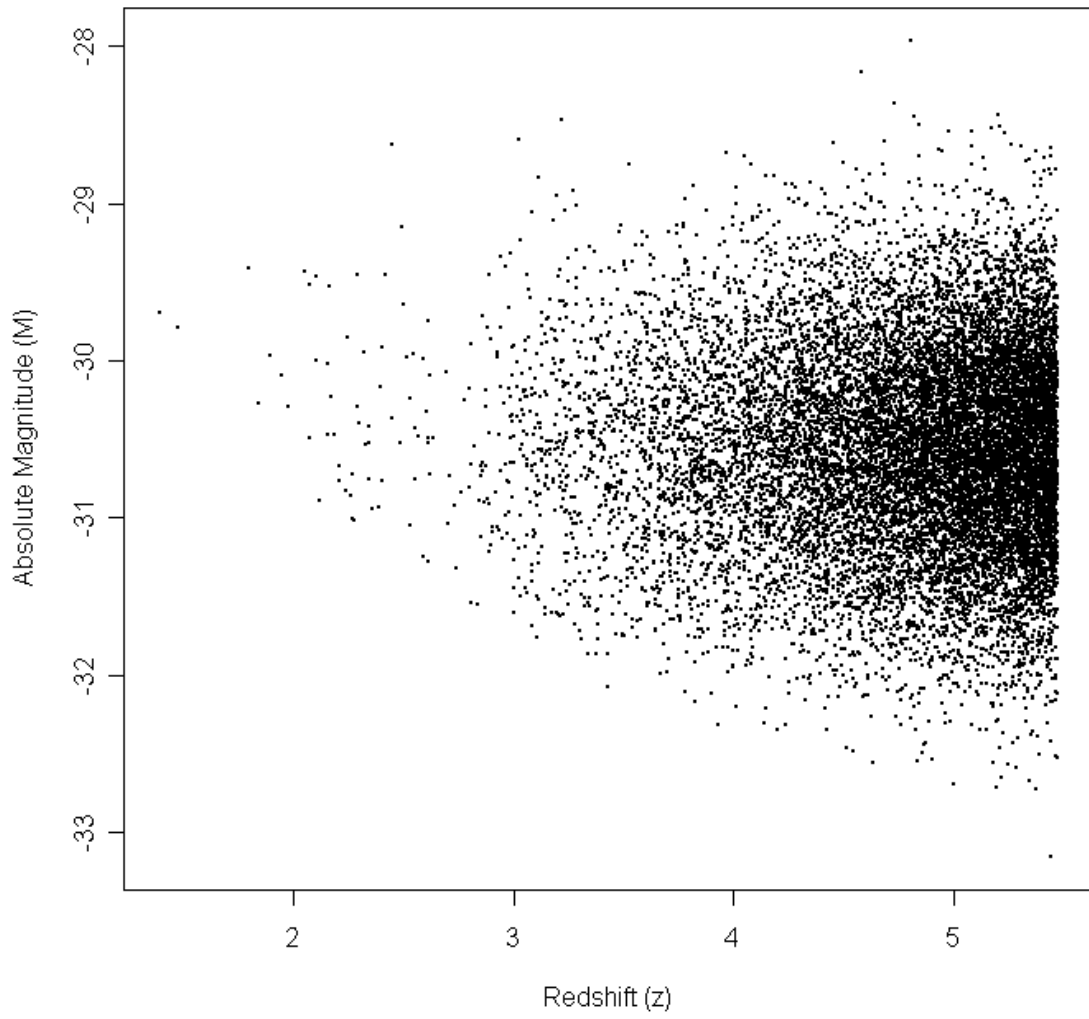


Figure 2.17: Scatterplot of simulated quasar dataset #2 under a randomly generated model.

Simulation #	true A	fitted \hat{A}	true B	fitted \hat{B}
1	6	4	4	5
2	2	2	3	3
3	8	2	5	5
4	2	2	1	6
5	3	2	7	1
6	3	2	4	6
7	7	6	3	4
8	8	2	3	6
9	1	1	8	7
10	1	1	4	6
11	4	4	1	4
12	8	4	5	7
13	1	2	8	3
14	2	2	5	6
15	4	4	5	2
16	6	4	6	7
17	1	2	6	1
18	7	6	6	5
19	6	4	4	4
20	3	2	8	5
21	1	1	5	7
22	5	2	7	7
23	6	4	3	3
24	4	4	1	4
25	1	1	1	8

Table 2.5: Simulation Study: Fitted model complexities for simulated quasar data under randomly generated models. This table presents the true model complexity for each randomly generated model, along with the fitted values for the model complexities. A summary of these results is shown in Table 2.6.

Complexity Parameter	true < fitted	true = fitted	true > fitted	total
A	2	10	13	25
B	13	5	7	25

Table 2.6: Simulation Study: Comparison of true model complexities to fitted model complexities for simulated quasar data under randomly generated models. This table is a summary of the results in Table 2.5. The counts represent the number of simulations out of 25 total which had the given relationship between the true model complexity parameter and the fitted model complexity parameter. This simulation study does not appear to indicate a tendency for BIC to over-specify or under-specify the fitted models.

Simulation #	A	B
1	4	7
2	2	8
3	2	7
4	3	7
5	3	7
6	3	8
7	2	7
8	2	8
9	3	7
10	3	7
true	2	6

Table 2.7: Simulation Study: Fitted model complexities for simulated quasar data under the fitted model for the SDSS quasar data in the case that does not allow for an interaction term, as described in Section 2.3.5. This table presents the model complexities selected by BIC under each of the simulated quasar data sets. The true model complexity parameters used to simulate the quasar data are listed at the bottom.

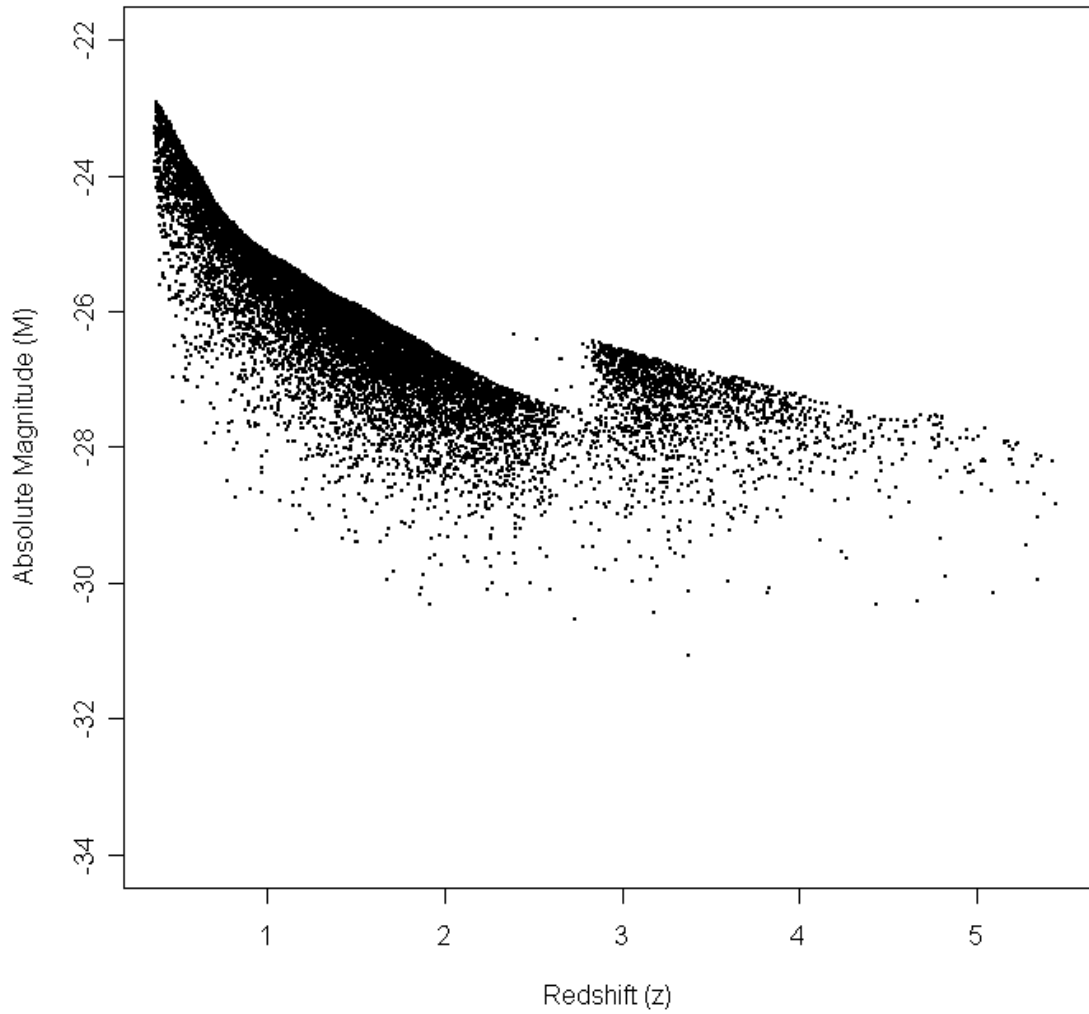


Figure 2.18: Scatterplot of simulated quasar dataset #1 under the fitted model for the SDSS quasar data in the case that does not allow for an interaction term, as described in Section 2.3.5.

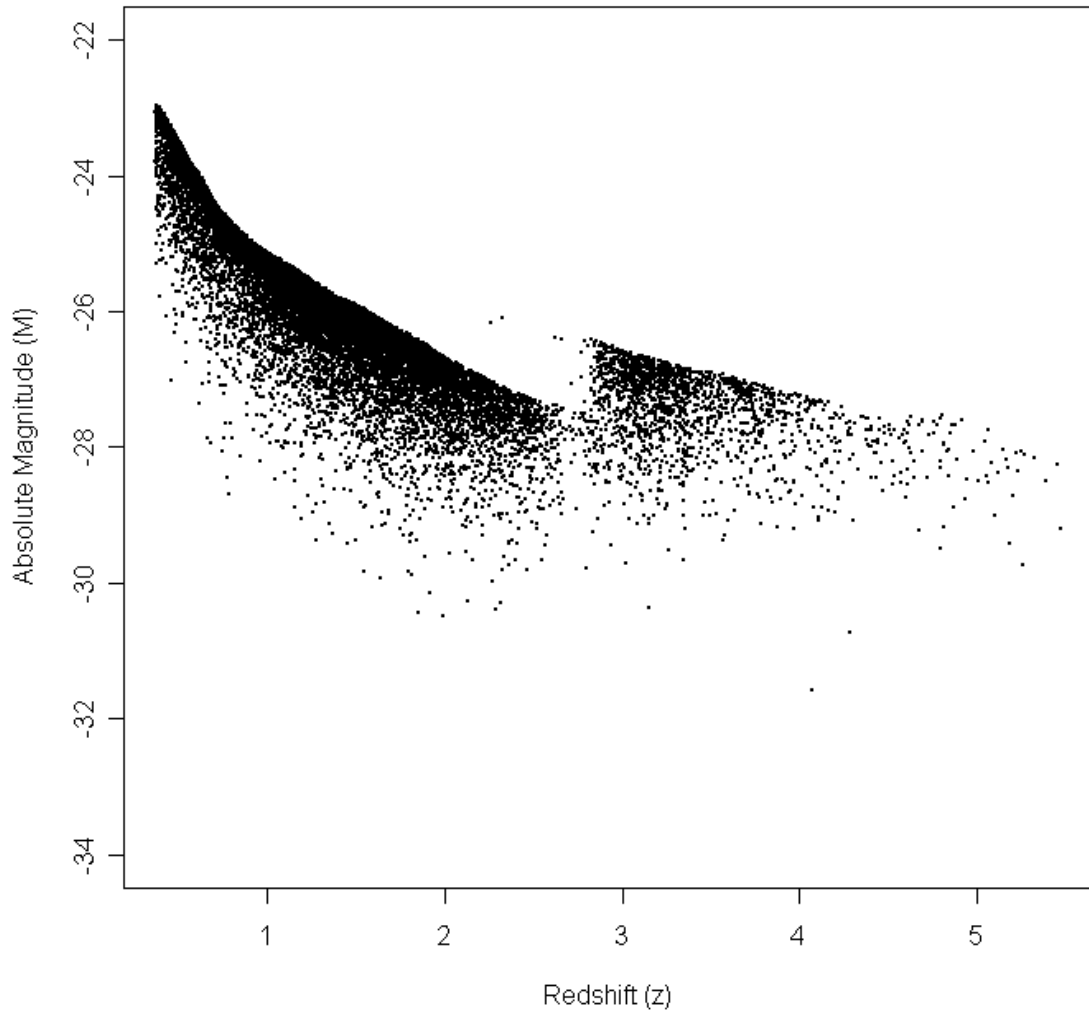


Figure 2.19: Scatterplot of simulated quasar dataset #2 under the fitted model for the SDSS quasar data in the case that does not allow for an interaction term, as described in Section 2.3.5.

proposed by Fan et al. [9] which suggests that absolute magnitude and redshift should be treated as separable components in the quasar luminosity function. However, it is important to keep in mind that unknown errors in the K correction function or the estimates of the observational probabilities could have a significant impact on the results of our analysis. We are unable to test for the robustness of our methods against these types of errors, and therefore we must be cautious about making any strong claims as to the true nature of the quasar luminosity function.

The presence of an interaction effect would suggest that new scientific theories are needed in order to explain the way in which the distribution of quasars in our universe has changed over time. We hope that our findings will encourage astronomers to pursue this research further, and perhaps inspire new theories about quasars and the evolution of our universe.

The statistical contribution of this work is the development of a flexible parametric framework for model selection that can be used to model point process data. Our method is generalizable in that it allows for the inclusion of an observational probability surface which can be applied whenever there are known selection biases in the data, and we have provided computational tools that make maximum likelihood fitting under this model framework computationally feasible even for large data sets.

APPENDIX A

MATHEMATICAL DETAILS

A.1 Poisson Process on a Transformed Space

Let X be a Poisson process on \mathbb{R} with rate function $\lambda(\cdot)$. Then for any interval $[a, b] \in \mathbb{R}$, the number of points in the interval is Poisson, with disjoint intervals independent and

$$N_X[a, b] \sim Poi \left(\int_a^b \lambda(x) dx \right). \quad (\text{A.1})$$

Let Y be the Poisson Process defined by the rate function $\lambda(\cdot)$ on a transformed space, namely $h(\mathbb{R})$, where $h: \mathbb{R} \rightarrow \mathbb{R}$. If h is assumed to be monotonic, then for any interval $[a, b] \in \mathbb{R}$, the new process will still place a Poisson number of points, with disjoint intervals being independent.

For an interval $[a, b] \in \mathbb{R}$, define the following:

$$N_X([a, b]) = \text{count under the original process} \quad (\text{A.2})$$

and

$$N_Y([a, b]) = \text{count under the transformed process} \quad (\text{A.3})$$

Then,

$$N_Y([a, b]) = N_X([h^{-1}(a), h^{-1}(b)]) \quad (\text{A.4})$$

which implies that

$$N_Y([a, b]) \sim Poi \left(\int_{h^{-1}(a)}^{h^{-1}(b)} \lambda(x) dx \right). \quad (\text{A.5})$$

So in order to obtain the rate function of Y , a function $y \mapsto \rho(y)$ is required which obeys:

$$\int_a^b \rho(y) dy = \int_{h^{-1}(a)}^{h^{-1}(b)} \lambda(x) dx . \quad (\text{A.6})$$

The Change of Variable Formula states that if h is regular on H , then for $a, b \in \text{image}(H)$,

$$\int_a^b \lambda(x) dx = \int_{h^{-1}(a)}^{h^{-1}(b)} \lambda(h(y)) |h'(y)| du \quad (\text{A.7})$$

Note that the order of the endpoints for the integral depends on h . Now, expressing the change of variables formula under the transformation h^{-1} gives

$$\int_a^b \lambda(x) dx = \int_{h(a)}^{h(b)} \lambda(h^{-1}(y)) |(h^{-1})'(y)| dy \quad (\text{A.8})$$

$$\Rightarrow \int_{h^{-1}(a)}^{h^{-1}(b)} \lambda(x) dx = \int_a^b \lambda(h^{-1}(y)) |(h^{-1})'(y)| dy \quad (\text{A.9})$$

Thus from Equation (A.6) we have

$$\rho(u) = \lambda(h^{-1}(y)) \cdot |(h^{-1})'(y)| , \quad (\text{A.10})$$

and therefore

$$N_Y([a, b]) \sim Poi \left(\int_a^b \lambda(h^{-1}(y)) \cdot |(h^{-1})'(y)| dy \right) , \quad (\text{A.11})$$

and a similar computation can be performed in the two-dimensional case where the absolute value becomes a Jacobian term.

For the quasar research presented in this paper, we have

$$h(M, z) = (M + \mathbf{DM}(z) + \mathbf{K}(z) , z) \quad (\text{A.12})$$

as the transformation from absolute magnitude into apparent magnitude, so that

$$h^{-1}(m, z) = (m - \text{DM}(z) - \text{K}(z) , z) \quad (\text{A.13})$$

is the corresponding function to go from apparent magnitude into absolute magnitude. For a region $R_m = h(R_M)$ in $m \times z$ space, the two-dimensional version of Equation (A.11) gives

$$N_m(R_m) \sim \text{Poi} \left(\iint_{R_m} \gamma(z) \cdot \lambda(h^{-1}(m, z)) \cdot \left| \frac{\partial h^{-1}}{\partial(m, z)} \right| dm dz \right). \quad (\text{A.14})$$

But,

$$h^{-1}(m, z) = (h_1^{-1}(m, z), h_2^{-1}(m, z)) = (m - \text{DM}(z) - \text{K}(z), z) \quad (\text{A.15})$$

which implies that

$$\left| \frac{\partial h^{-1}}{\partial(m, z)} \right| = \begin{vmatrix} \frac{\partial h_1^{-1}}{\partial m} & \frac{\partial h_1^{-1}}{\partial z} \\ \frac{\partial h_2^{-1}}{\partial m} & \frac{\partial h_2^{-1}}{\partial z} \end{vmatrix} = \begin{vmatrix} 1 & \left(-\frac{\partial \text{DM}}{\partial z} - \frac{\partial \text{K}}{\partial z} \right) \\ 0 & 1 \end{vmatrix} = 1 - 0 = 1 \quad (\text{A.16})$$

And therefore,

$$N_m(R_m) \sim Poi \left(\iint_{R_m} \gamma(z) \cdot \lambda(h^{-1}(m, z)) \cdot |(h^{-1})'(m, z)| \, dm \, dz \right) \quad (\text{A.17})$$

$$\sim Poi \left(\iint_{R_m} \gamma(z) \cdot \lambda(h^{-1}(m, z)) \cdot 1 \, dm \, dz \right) \quad (\text{A.18})$$

$$\sim Poi \left(\iint_{R_m} \gamma(z) \cdot \Phi(h^{-1}(m, z)) \cdot \psi_M(h^{-1}(m, z)) \, dm \, dz \right) \quad (\text{A.19})$$

$$\sim Poi \left(\iint_{R_m} \gamma(z) \cdot \Phi(h^{-1}(m, z)) \cdot \psi_m(m, z) \, dm \, dz \right) \quad (\text{A.20})$$

$$\begin{aligned} \sim Poi \left(\iint_{R_m} \gamma(z) \cdot \Phi(m - \mathbf{DM}(z) - \mathbf{K}(z), z) \cdot \right. \\ \left. \cdot \psi_m(m, z) \, dm \, dz \right) \end{aligned} \quad (\text{A.21})$$

APPENDIX B
ADDITIONAL TABLES

timeStart	timeEnd	timePeak	location	class	number
4/2/1999 8:06	4/2/1999 8:29	4/2/1999 8:21		M	1.1
4/3/1999 13:39	4/3/1999 13:47	4/3/1999 13:44	S27E05	C	1.4
4/3/1999 14:52	4/3/1999 15:09	4/3/1999 15:01	N11E63	C	1.9
4/3/1999 20:00	4/3/1999 20:26	4/3/1999 20:17	S26E02	C	1.6
4/3/1999 20:44	4/3/1999 20:58	4/3/1999 20:50	N18E74	C	3.1
4/3/1999 22:16	4/3/1999 22:23	4/3/1999 22:20	S26E01	C	2.8
4/3/1999 22:56	4/3/1999 23:19	4/3/1999 23:10	N29E81	M	4.3
4/4/1999 5:15	4/4/1999 5:30	4/4/1999 5:25	N18E72	M	5.4
4/4/1999 21:47	4/4/1999 21:54	4/4/1999 21:50		C	2.3
4/5/1999 8:28	4/5/1999 8:33	4/5/1999 8:31	N17E57	C	1.2
4/5/1999 19:09	4/5/1999 19:16	4/5/1999 19:14	N18E52	C	2.6
4/5/1999 23:01	4/5/1999 23:15	4/5/1999 23:09	S28W26	C	3.3
4/6/1999 6:52	4/6/1999 7:22	4/6/1999 7:05	N22E47	C	3.5
4/6/1999 23:35	4/7/1999 1:15	4/7/1999 0:03		C	1.6
4/7/1999 4:32	4/7/1999 4:37	4/7/1999 4:35		C	1.1
4/7/1999 5:33	4/7/1999 5:39	4/7/1999 5:36		C	1.2
4/7/1999 9:20	4/7/1999 9:27	4/7/1999 9:24	N20E34	C	1.3
4/8/1999 5:11	4/8/1999 5:25	4/8/1999 5:19	N23E23	C	2.1
4/8/1999 7:39	4/8/1999 7:54	4/8/1999 7:48	N23E22	C	2.2
4/8/1999 8:10	4/8/1999 8:21	4/8/1999 8:16	N22E21	C	1.1

Table B.1: Example GOES soft X-ray flare data

firstObs	onset	centPA	width	linSpd	measPA
4/1/1999 11:54	4/1/1999 10:52	56	6	713	59
4/1/1999 14:30	4/1/1999 13:30	67	15	658	74
4/1/1999 19:31	4/1/1999 18:21	25	49	529	34
4/2/1999 1:31	4/1/1999 23:37	91	74	294	78
4/2/1999 6:06	4/2/1999 5:23	288	26	748	293
4/2/1999 7:31	4/2/1999 5:50	291	69	302	305
4/2/1999 8:30	4/2/1999 7:37	90	191	1068	79
4/2/1999 13:31	4/2/1999 11:57	244	20	326	244
4/2/1999 23:30	4/2/1999 22:13	162	133	447	158
4/3/1999 15:06	4/3/1999 13:43	228	88	375	234
4/3/1999 21:30	4/3/1999 20:10	92	58	797	79
4/3/1999 23:47	4/3/1999 22:38	74	156	923	38
4/4/1999 1:54	4/4/1999 1:27	171	87	307	164
4/4/1999 4:30	4/4/1999 3:23	119	70	671	115
4/4/1999 4:30	4/4/1999 3:23	273	132	451	299
4/4/1999 5:30	4/4/1999 4:51	65	173	1203	73
4/4/1999 14:06	4/4/1999 12:26	357	5	297	1
4/4/1999 19:31	4/4/1999 18:43	80	6	647	85
4/4/1999 22:30	4/4/1999 21:07	79	29	789	89
4/5/1999 1:54	4/5/1999 0:38	258	53	471	243

Table B.2: Example Catholic University CDAW coronal mass ejection data

Redshift z	Apparent Magnitude m	Extended
2.240	18.18	0
0.989	18.99	0
0.768	18.89	0
1.411	19.02	0
0.453	18.91	0
0.404	18.45	0
1.924	18.85	0
0.530	18.75	0
0.399	18.90	0
1.790	18.01	0
1.272	18.92	0
1.272	18.74	0
1.823	18.44	0
3.675	19.47	0
0.901	18.98	0
1.568	18.89	0
0.792	18.88	0
1.148	18.99	0
0.858	17.73	0
3.686	19.35	0

Table B.3: Example quasar data from the SDSS Data Release III

Redshift z	Distance Modulus $DM(z)$
0.01	33.175
0.02	34.697
0.03	35.594
0.04	36.234
0.05	36.735
0.06	37.146
0.07	37.496
0.08	37.801
0.09	38.072
0.10	38.315
0.11	38.537
0.12	38.740
0.13	38.928
0.14	39.102
0.15	39.266
0.16	39.420
0.17	39.565
0.18	39.702
0.19	39.832
0.20	39.956

Table B.4: Example entries in the distance modulus DM table

Redshift z	K Correction $K(z)$
0.00	0.596
0.01	0.587
0.02	0.569
0.03	0.531
0.04	0.462
0.05	0.372
0.06	0.268
0.07	0.203
0.08	0.170
0.09	0.157
0.10	0.156
0.11	0.155
0.12	0.154
0.13	0.160
0.14	0.167
0.15	0.184
0.16	0.241
0.17	0.208
0.18	0.193
0.19	0.192

Table B.5: Example entries in the K correction table

Redshift z	Volume Differential $\gamma(z)$
0.0000	0.000000
0.0001	0.090617
0.0002	0.362437
0.0003	0.815410
0.0004	1.449488
0.0005	2.264621
0.0006	3.260760
0.0007	4.437857
0.0008	5.795863
0.0009	7.334728
0.0010	9.054404
0.0011	10.954842
0.0012	13.035993
0.0013	15.297808
0.0014	17.740238
0.0015	20.363234
0.0016	23.166747
0.0017	26.150728
0.0018	29.315129
0.0019	32.659900

Table B.6: Example entries in the volume differential $\gamma(z)$ table

Redshift z	Apparent Magnitude m	Observational Probability $\psi(m, z)$
0.00	15.00	1.000
0.05	15.00	1.000
0.10	15.00	1.000
0.15	15.00	1.000
0.20	15.00	1.000
0.25	15.00	1.000
0.30	15.00	1.000
0.35	15.00	1.000
0.40	15.00	1.000
0.45	15.00	1.000
0.50	15.00	1.000
0.55	15.00	1.000
0.60	15.00	1.000
0.65	15.00	0.980
0.70	15.00	0.995
0.75	15.00	1.000
0.80	15.00	1.000
0.85	15.00	1.000
0.90	15.00	1.000
0.95	15.00	1.000

Table B.7: Example entries in the observation probabilities $\psi(m, z)$ table

APPENDIX C
INDEX OF SYMBOLIC NOTATION

Notation	Definition
τ_D	Onset time of a CME relative to the onset time of a flare
\mathbb{T}	Collection of relative times τ_D
W	Window half-width
γ	Mixture parameter
μ	Center of the bump
σ	Spread of the bump
LR	log-likelihood ratio statistic
M, P	Parameters for the semi-greedy grid search
K	Number of blocks that the temporal spectrum is divided into
L	Length of a block in minutes
R_1, R_2	Number of bootstrapped data sets constructed

Table C.1: Symbolic Notation for Chapter 1

Notation	Definition
z	Redshift
m	Apparent i-band magnitude
M	Absolute i-band Magnitude (Luminosity)
$DM(z)$	Distance modulus
$K(z)$	K correction
$\gamma(z)$	Spherical shell volume differential
$\psi(m, z)$	Observational probability
$\Phi(M, z)$	Idealized quasar luminosity function
A	Complexity parameter for absolute magnitude
B	Complexity parameter for redshift
D	Complexity parameter for interaction term
α	Parameter vector for absolute magnitude
β	Parameter vector for redshift
δ	Parameter vector for interaction term
r_d	Deviance residual

Table C.2: Symbolic Notation for Chapter 2

REFERENCES

- [1] M. Abramowitz and I.A. Stegun, 1972, "Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables", Dover, New York
- [2] A. Baddeley & J. Moller, 2000, "Non- and Semi-parametric estimation of interaction in inhomogeneous point patterns", *Statistica Neerlandica*. 54:329-50
- [3] S. Carroll, 2001, "The Cosmological Constant", *Living Reviews in Relativity* IV
- [4] D.R. Cox, 1972, "Multivariate Point Processes", *Proceedings of the Sixth Berkeley Symposium of Mathematical Statistics and Probability*, 3:401-445
- [5] Daley & Vere-Jones, 1988, "An Introduction to the Theory of Point Processes", Springer-Verlag, New York
- [6] A.C. Davison, A. Gigli, 1989, "Deviance Residuals and Normal Scores Plots", *Biometrika*, 76:2,211-221
- [7] A.C. Davison, D.V. Hinkley, 1997, "Bootstrap Methods and their Application", Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge
- [8] R.A. DeVore, 1996, "Some remarks on greedy algorithms", *Advances in Computational Mathematics*, 5:173-187
- [9] Fan et al., 2001, "High-Redshift Quasars Found in Sloan Digital Sky Survey Commissioning Data. IV.", *The Astronomical Journal*, 121:54-65
- [10] J.T. Gosling, 1993, "The Solar Flare Myth", *Journal of Geophysical Research*, 98:18937-18949
- [11] R.A. Harrison, 1986, "Solar coronal mass ejections and flares", *Astronomy and Astrophysics*, 162:283-291
- [12] Hogg, Ivan, Blanton, and Eisenstein, 2002, "The K Correction", *ArXiv Astrophysics e-prints*, astro-ph/0210394
- [13] R.A. Howard et al., 1985, "Coronal Mass Ejections: 1979-1981", *Journal of Geophysical Research*, 90:8173-8192
- [14] H.S. Hudson, 1991, "Solar flares, microflares, nanoflares, and coronal heating", *Solar Physics*, 133:357-369

- [15] S.W. Kahler, 1992, "Solar Flares and Coronal Mass Ejections", *Annual Review of Astronomy and Astrophysics*, 30:113-41
- [16] J.F.C. Kingman, 1993, "Poisson Processes", Clarendon Press, Oxford
- [17] H.R. Kunsch, 1989, "The jackknife and the bootstrap for general stationary observations", *The Annals of Statistics*, 17:1217-1241
- [18] van Lieshout & Baddeley, 1999, "Indices of Dependence between types in multivariate point patterns", *Scandinavian Journal of Statistics*, 26:511-532
- [19] J.K. Lindsey, 2004, "Introduction to Applied Statistics: A Modelling Approach", Oxford University Press, Oxford
- [20] J.M. Loh & M.L. Stein, 2004, "Bootstrapping a Spatial Point Process", *Statistica Sinica*, 14:69-101
- [21] J. Moller, 1998, "Log Gaussian Cox Processes", *Scandinavian Journal of Statistics*, 25:451-482
- [22] J. Moller, 2004, "Statistical Inference and Simulation for Spatial Point Processes", Chapman and Hall, London
- [23] Richards et al., 2002, "Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Quasar Sample", *The Astronomical Journal*, 123:2945-2975
- [24] Richards et al., 2006, "The SDSS Quasar Survey: Quasar Luminosity Function from Data Release Three", *The Astronomical Journal*, 131:2766-2787
- [25] G.G. Roussas, 1997, "A Course in Mathematical Statistics", Academic Press, London
- [26] J.D. Scargle, 2003, "Adaptive Piecewise-constant modelling of signals in Multidimensional Spaces", PHYSTAT2003, SLAC, Stanford
- [27] G. Schwarz, 1978, "Estimating the dimension of a model", *Annals of Statistics*, 6:461-464
- [28] Wheatland, Sturrock, and McTiernan, 1998, "The Waiting-Time Distribution of Solar Flare Hard X-Ray Bursts", *The Astrophysical Journal*, 509:448-455
- [29] M.S. Wheatland, 2001, "The local poisson hypothesis for solar flares", *Astrophysical Journal Letters*
- [30] M.S. Wheatland, 2001, "Rates of Flaring in Individual Active Regions", *Solar Physics*, 203:87-106
- [31] M.S. Wheatland, 2002, "The Coronal Mass Ejection Waiting-Time Distribution", *Solar Physics*, 214:361-373

- [32] M.S. Wheatland, 2004, "A Bayesian Approach to Solar Flare Prediction", *The Astrophysical Journal*, 609:1134-1139
- [33] Wolf et al., 2003, "The evolution of faint AGN between $z \sim 1$ and $z \sim 5$ from the COMBO-17 survey", *Astronomy & Astrophysics*, 408:499-514
- [34] Chin-Teh Yeh, 2005, "Waiting time distribution of coronal mass ejections", *Chinese Journal of Astronomy & Astrophysics*, 5:2:193-197