

Point Process Models for Astronomy: Quasars, Coronal Mass Ejections, and Solar Flares

Angela B. Hugeback, Google, Inc.

Joint work with:

Marc A. Coram, Stanford University

C. Alex Young, NASA GSFC

Thesis research supported by:

Department of Statistics, University of Chicago

University of Washington Department of Statistics Seminar Series

October 22, 2007

Project 1:

“A Local Temporal Correlation Between
Solar Flares and Coronal Mass Ejections”

Contributing Astronomers:

Dr. C. Alex Young, NASA GSFC

Dr. Jeffrey Love, U.S. Geological Survey

Solar Flares and Coronal Mass Ejections

Solar Flares

- Enormous explosions on surface of Sun
- Duration: a few minutes.

Coronal Mass Ejections (CMEs)

- Tremendous bubbles of plasma from the solar corona
- Duration: several hours

Why study them?

- Settle a longstanding debate (correlation/sequence)
- Potential to result in geomagnetic storms
- Learn more about solar energy release

GOES solar flare catalog

- 6,506 soft X-ray flares
- Observed between 4/1/1999 and 12/31/2003
- Variables:
 - onset time
 - peak time
 - end time
 - location on Sun

Solar Flares and Coronal Mass Ejections

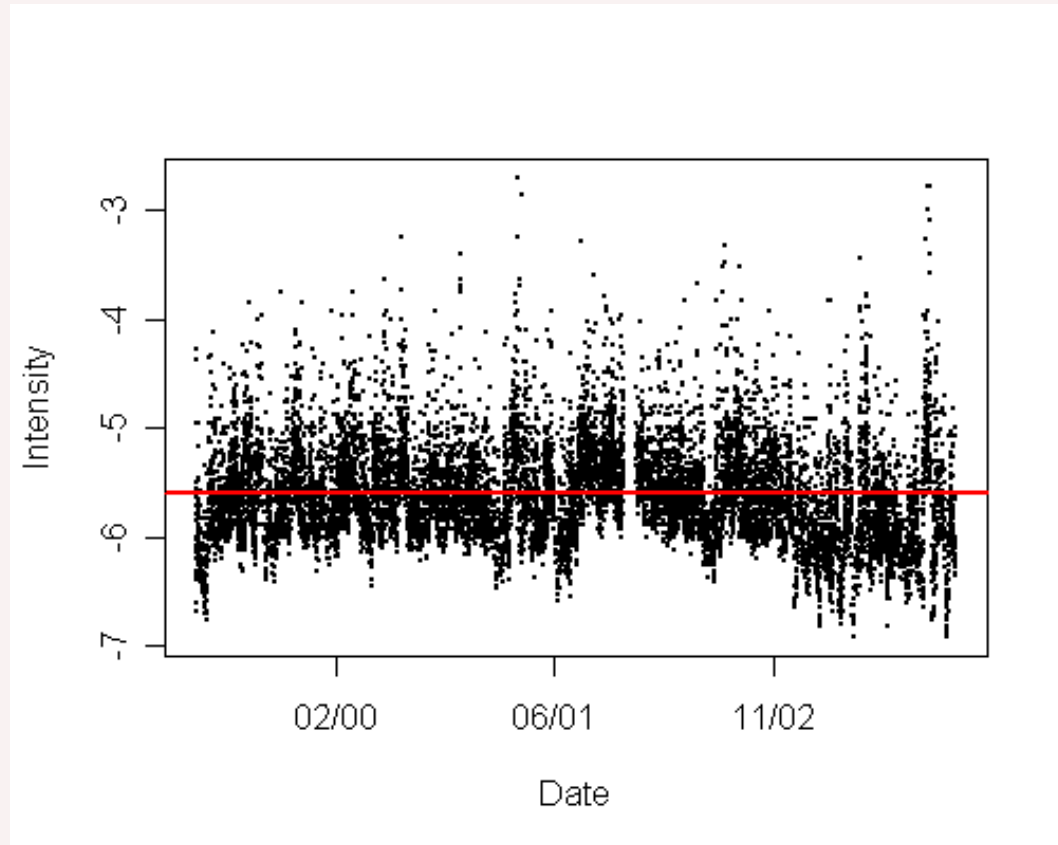


Figure 1: GOES Solar Flare Catalog

Catholic University CDAW CME Catalog

- 6,604 CME events
- Observed between 4/1/1999 and 12/31/2003
- Variables:
 - onset time
 - time of first observation
 - velocity
 - positional angle
 - angular width
- Measurement issues!

Solar Flares and Coronal Mass Ejections

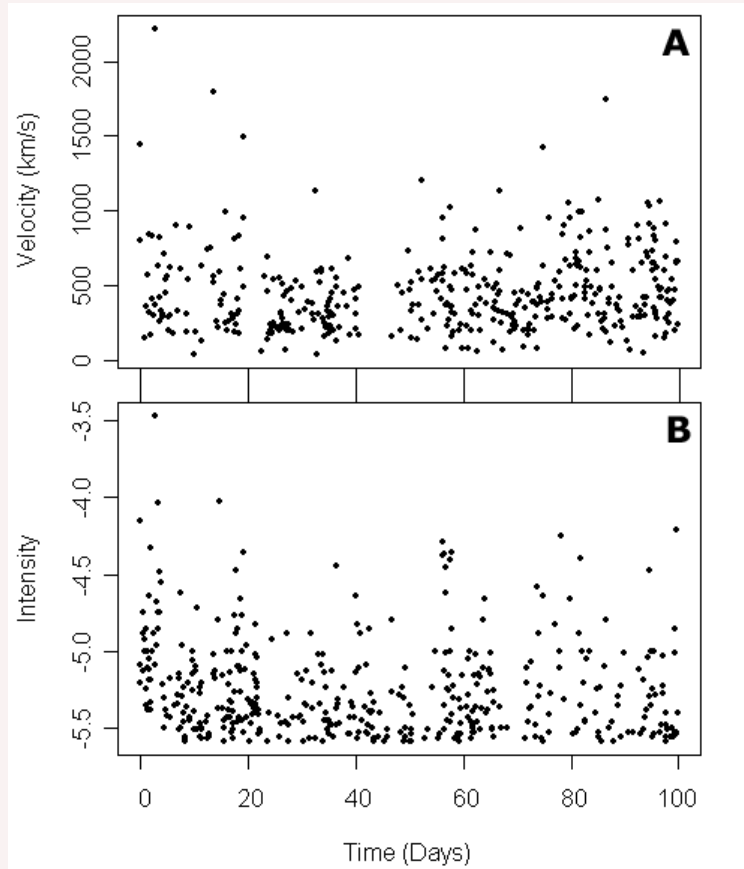


Figure 2: Typical 100 Day Interval A: CDAW CME events B: GOES solar flare events

Astronomy Literature Review

- No previous work on superimposed/marked point process
- Numerous articles study isolated cases
- Wheatland (2001, 2001, 2002)
 - Local behavior of solar flares and CMEs as Poisson processes
 - Obscuration effect for small solar flares after large ones
- Wheatland (2004)
 - Piecewise-constant Poisson process provides a good model for solar flaring activity associated with a particular region.

Solar Flares and Coronal Mass Ejections

Preliminary Analysis:

- Sliding window estimates of the two intensity functions

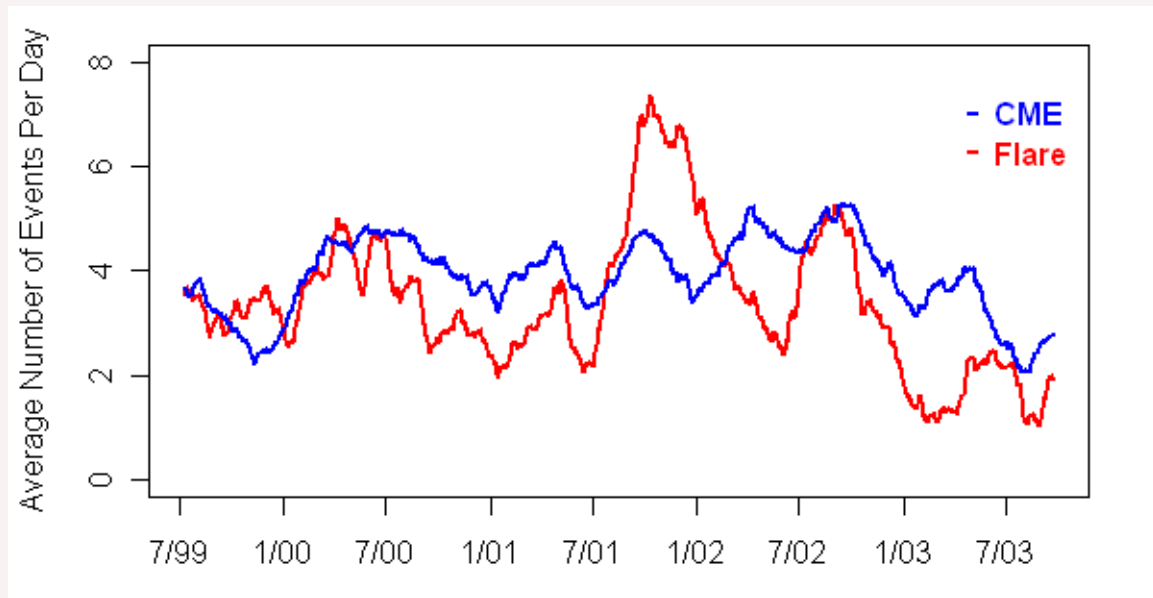


Figure 3: Sliding window estimates. Average number of events per day (100 day bandwidth)

Objective:

- Q: Does correlation exist on a much smaller time scale?
- Q: Does the occurrence of a single solar flare event make the occurrence of a CME more or less likely in the surrounding minutes or hours, and vice versa.

Statistical Goal:

Assess significance and nature of local correlation after adjusting for correlation due to inhomogeneous long-term trends.

- Cox, *Multivariate point processes*, 1972
- Møller, *Statistical inference and simulation for spatial point processes*, 2004
- Baddeley & Møller, *Non- and semi-parametric estimation of interaction in inhomogeneous point patterns*, 2000
- van Lieshout & Baddeley, *Indices of dependence between types in multivariate point patterns*, 1999

Approach:

1. Compute relative onset times

$$\tau_D = \tau_C - \tau_F \quad (1)$$

2. Create collection $\mathbb{T} = \{ \tau_D \mid \tau_D \in [-W, W] \}$

3. Examine distribution of \mathbb{T}

- Note: Not independent observations!

Solar Flares and Coronal Mass Ejections

Drum roll...

Solar Flares and Coronal Mass Ejections

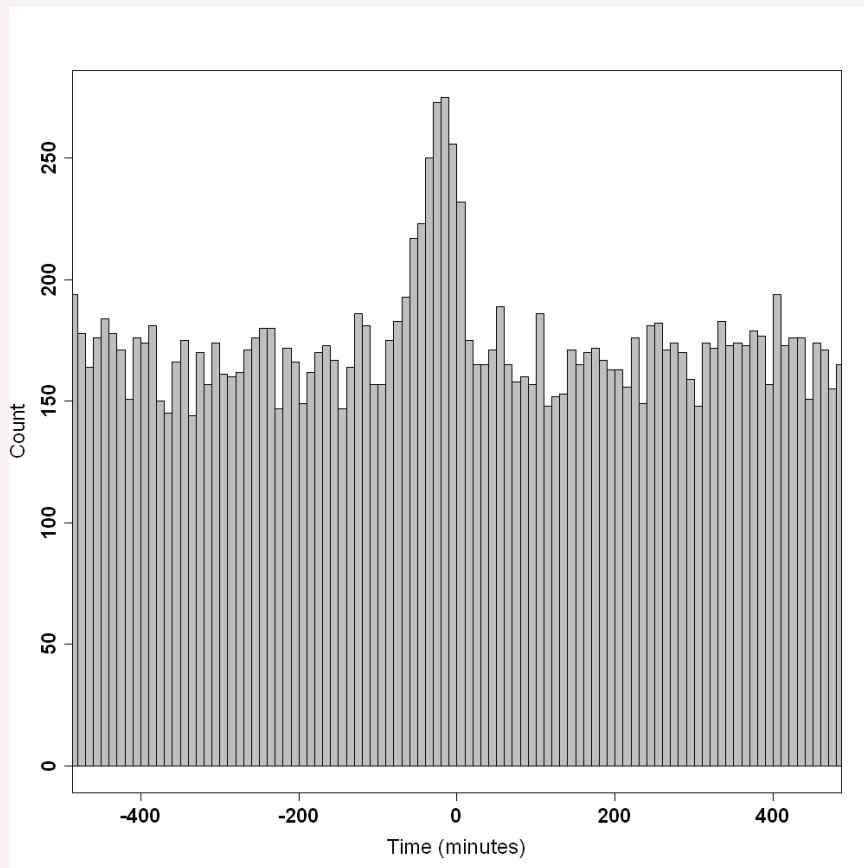


Figure 4: Histogram of relative onset times T . $W = 500$ min ($8\frac{1}{3}$ hours), Bin width = 10 min

Two models for the distribution of \mathbb{T} ($t \in [-W, W]$)

1. Piecewise-Uniform Model

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

2. Gaussian-Uniform Mixture Model

$$\mathbf{f}(t; \gamma, \mu, \sigma) = \frac{\gamma}{2W} + \frac{1-\gamma}{I} \cdot \mathbf{g}(t; \mu, \sigma) \quad (3)$$

Model Fitting Approach:

- Method of maximum likelihood
- Semi-greedy grid search algorithm

Solar Flares and Coronal Mass Ejections

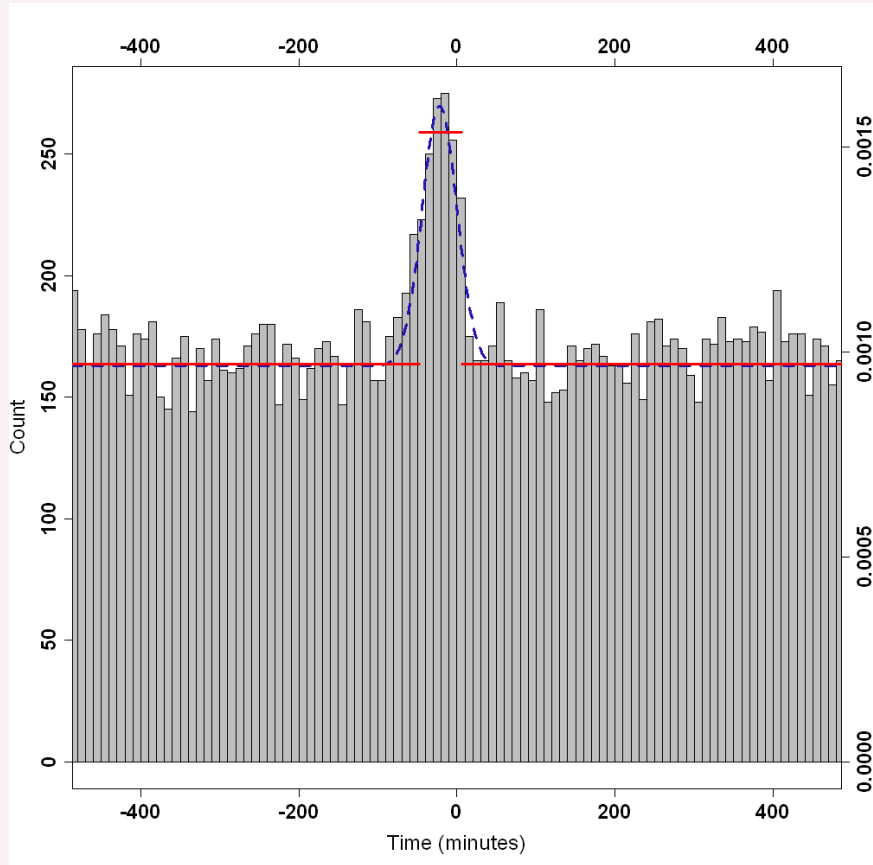


Figure 5: Histogram of relative times T . Red: Piecwise-Uniform Model, Blue: Gaussian-Uniform Mixture Model

Solar Flares and Coronal Mass Ejections

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

Table 1: Maximum likelihood estimates for the distribution of \mathbb{T} .

Next Steps:

1. Compute confidence intervals for parameter estimates

- Approach:

Create bootstrapped samples that resemble the observed data as closely as possible, preserving any long-term and local correlation.

2. Assess statistical significance of local temporal correlation

- Approach:

Create bootstrapped samples that mirror the long-term correlation as closely as possible without incorporating any local dependence.

Block-of-Blocks Bootstrap Approach

(Kunsch 1989, Loh & Stein 2004)

1. Partition time into K blocks of length L

- $K=499$ blocks
- $L=5000$ minutes

2. Resample blocks and collect contributions to the statistic,

where contributions are relative occurrence times:

$$\tau_D = \tau_C - \tau_F$$

Solar Flares and Coronal Mass Ejections

Example Block:

- Flare: completely within the block
- CME: within block edges plus $W=500$ minutes

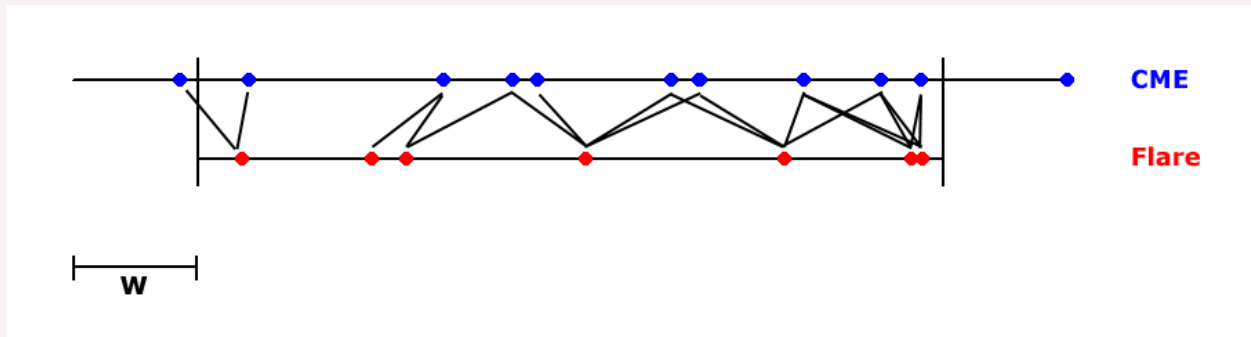


Figure 6: Diagram for Block-of-Blocks Bootstrap Approach

Bootstrap Confidence Intervals for Model Parameters

- To create each bootstrapped sample,
 1. Resample blocks uniformly at random with replacement
 2. Collect all relative onset times $\tilde{\tau}_D$ into collection $\tilde{\mathbb{T}}$.
 3. Fit Gaussian-uniform mixture model to the collection $\tilde{\mathbb{T}}$.
- Compute “basic bootstrap confidence limits” (Davison, 1997)
- Form 95% confidence intervals for model parameters

Bootstrap Confidence Intervals for Model Parameters

Parameter	MLE	95% confidence interval
γ	0.965	[0.959, 0.971]
μ	-22.0	[-27, -17] ←
σ	22.0	[18.0, 25.0] ←

Table 2: Bootstrap 95% confidence intervals based on Gaussian-Uniform mixture model for 500 bootstrapped samples.

Assessing Significance

- Need a test statistic:

Log-likelihood ratio statistic (Roussas, 1997)

$$LR = -\ln \left[\frac{\mathcal{L}(\mathbb{T} | H_0)}{\mathcal{L}(\mathbb{T} | H_A)} \right] = \ell(\mathbb{T} | H_A) - \ell(\mathbb{T} | H_0) \quad (4)$$

- H_0 : \mathbb{T} is uniformly distributed
- H_A : \mathbb{T} follows the fitted Gaussian-uniform mixture model

Assessing Significance (cont'd)

- **The Goal:**
Preserve long-term correlation while removing any local correlation.

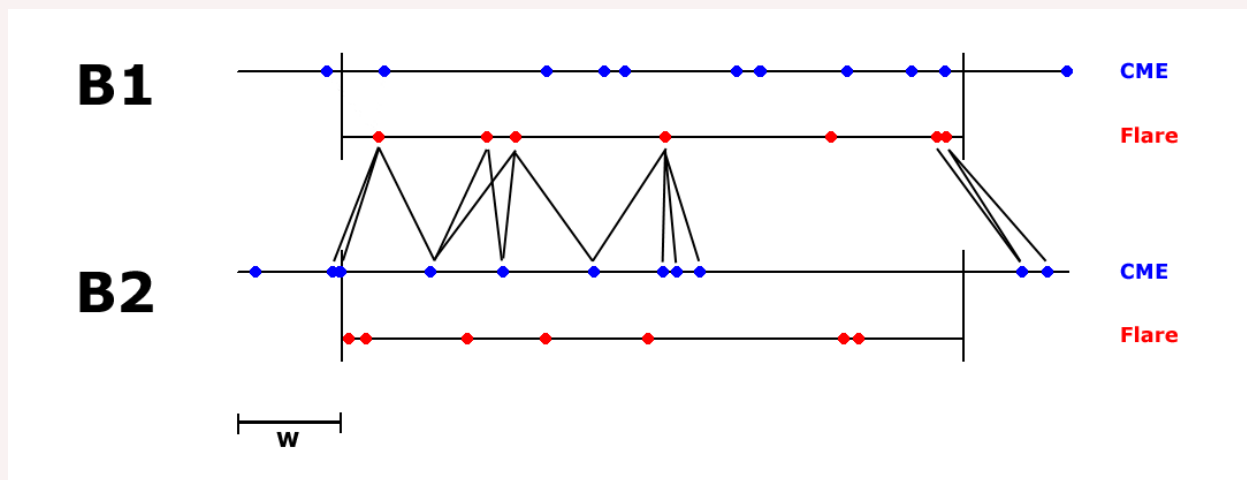


Figure 7: Diagram for Block-of-Blocks Bootstrap Approach for Assessing Significance

Assessing Significance (cont'd)

- For each bootstrapped sample:
 1. Perform the resampling: For $k = 1$ to K ,
 - Select block $B1$ at random with replacement from set of K blocks.
 - Condition on the number of flare events in block $B1$, and select 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.
 - Compute the relative onset times $\tilde{\tau}_D$ for each CME event in block $B2$ with respect to each flare event in block $B1$.
 2. Combine all relative onset times $\tilde{\tau}_D$ from those K resampled blocks into a final collection $\tilde{\mathbb{T}}$.
 3. Compute the log-likelihood ratio statistic (LR)

Result:

- Observed statistic:

$$LR=111$$

Maximum observed statistic in 500 bootstrapped samples:

$$LR=6$$

- p-value $\ll 0.002$
 \Rightarrow highly statistically significant local temporal correlation

Summary of Findings:

- Detected a statistically significant local temporal correlation between the solar flare and CME processes, after adjusting for long-term trends in the two intensity functions.
- Given that a solar flare has occurred, there is a spike of increased probability that a CME will have occurred immediately preceding that flare.
- Under the Gaussian-uniform mixture model,
 $\hat{\mu} = -22.0 \text{ minutes}$ $[-26.0, -17.0]$
 $\hat{\sigma} = 22.0 \text{ minutes}$ $[19.0, 26.0]$

Scientific Implications:

- Provides clear evidence of local dependence, which will inform current debate in solar physics community
- May allow improvements in current methodologies for the detection and prediction of both solar flare and coronal mass ejection events (covariates)

Statistical Contribution:

- A novel statistical approach for both detecting and assessing the nature of a local temporal correlation in a marked point process in the presence of an inhomogeneous long-term rate function to which the two processes are tied.

Future Research Goals:

- Build descriptive models for the solar processes
- Incorporate location/size data into these models
- Ideas:
 - Description of proportional variation over solar cycle
 - Models for particular active regions
 - Models for event size (conditional on last event)
 - Models for waiting time to next event
 - Two-state, non-homogeneous Markov chain with decay in dependence on previous event as time lag increases

Solar Flares and Coronal Mass Ejections

References: (and sources of inspiration)

- A. Baddeley & J. Møller, 2000, “Non- and Semi-parametric estimation of interaction in inhomogeneous point patterns”, *Statistica Neerlandica*. 54:329-50
- D.R. Cox, 1972, “Multivariate Point Processes”, *Proceedings of the Sixth Berkeley Symposium of Mathematical Statistics and Probability*, 3:401-445
- Daley & Vere-Jones, 1988, “An Introduction to the Theory of Point Processes”, Springer-Verlag, New York
- A.C. Davison, D.V. Hinkley, 1997, “Bootstrap Methods and their Application”, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge
- J.T. Gosling, 1993, “The Solar Flare Myth”, *Journal of Geophysical Research*, 98:18937-18949
- R.A. Harrison, 1986, “Solar coronal mass ejections and flares”, *Astronomy and Astrophysics*, 162:283-291
- R.A. Howard et al., 1985, “Coronal Mass Ejections: 1979-1981”, *Journal of Geophysical Research*, 90:8173-8192
- H.S. Hudson, 1991, “Solar flares, microflares, nanoflares, and coronal heating”, *Solar Physics*, 133:357-369

Solar Flares and Coronal Mass Ejections

- S.W. Kahler, 1992, “Solar Flares and Coronal Mass Ejections”, *Annual Review of Astronomy and Astrophysics*, 30:113-41
- J.F.C. Kingman, 1993, “Poisson Processes”, Clarendon Press, Oxford
- van Lieshout & Baddeley, 1999, “Indices of Dependence between types in multivariate point patterns”, *Scandinavian Journal of Statistics*, 26:511-532
- loh2004 J.M. Loh & M.L. Stein, 2004, “Bootstrapping a Spatial Point Process”, *Statistica Sinica*, 14:69-101
- J. Moller, 1998, “Log Gaussian Cox Processes”, *Scandinavian Journal of Statistics*, 25:451-482
- J. Moller, 2004, “Statistical Inference and Simulation for Spatial Point Processes”, Chapman and Hall, London
- Wheatland, Sturrock, and McTiernan, 1998, “The Waiting-Time Distribution of Solar Flare Hard X-Ray Bursts”, *The Astrophysical Journal*, 509:448-455
- M.S. Wheatland, 2001, “The local poisson hypothesis for solar flares”, *Astrophysical Journal Letters*
- M.S. Wheatland, 2001, “Rates of Flaring in Individual Active Regions”, *Solar Physics*, 203:87-106

Solar Flares and Coronal Mass Ejections

- M.S. Wheatland, 2002, “The Coronal Mass Ejection Waiting-Time Distribution”, *Solar Physics*, 214:361-373
- M.S. Wheatland, 2004, “A Bayesian Approach to Solar Flare Prediction”, *The Astrophysical Journal*, 609:1134-1139
- Chin-Teh Yeh, 2005, “Waiting time distribution of coronal mass ejections”, *Chinese Journal of Astronomy & Astrophysics*, 5:2:193-197

Project 2:

“A Flexible Model for the Quasar Luminosity Function
as an Inhomogeneous Poisson Process”

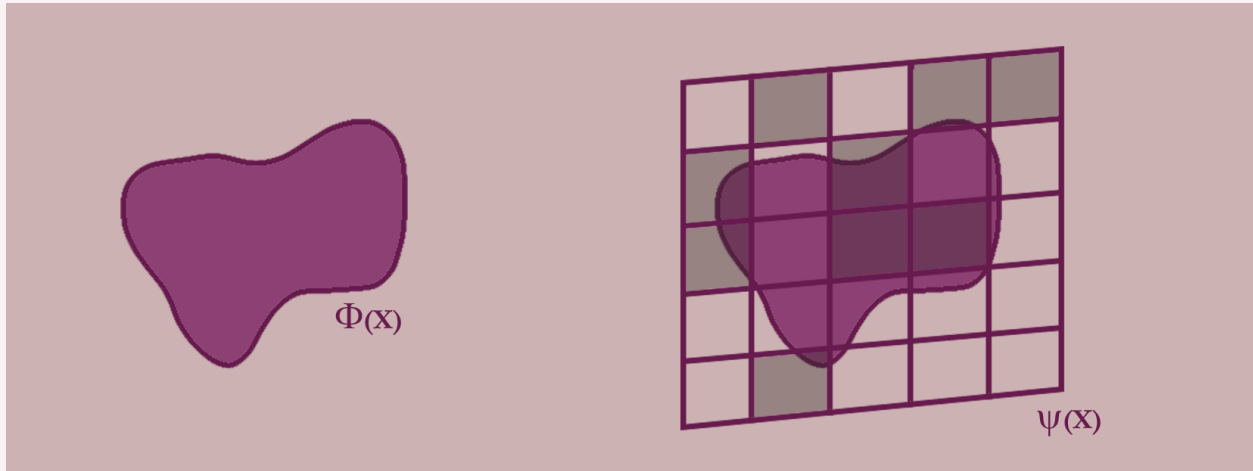
Contributing Astronomers:

Dr. Sebastian Jester, Max Planck Institute for Astronomy

Dr. Jasper Wall, University of British Columbia

Dr. Don Lamb, University of Chicago

The Quasar Luminosity Function



- $\Phi(\mathbf{X})$: unknown density (intensity surface)
- $\psi(\mathbf{X})$: “known” observational probabilities
- Data : observed point process on space X
- General question: How to estimate $\Phi(\mathbf{X})$?
- Motivation: Estimate the Quasar Luminosity Function

Statistical Contributions:

- Developed a flexible parametric framework for model selection that can be used to model point process data
 - Exponential model for Poisson intensity surface
 - Iteratively add polynomial terms into the exponent and compute MLE for model parameters (coefficients)
 - Model Selection:
Bayesian Information Criterion (Schwarz 1978)
- Method is generalizable in that it allows an observational probability surface if known selection biases exist in the data
- Computational tricks/tools make maximum likelihood fitting computationally feasible even for large data sets

The Quasar Luminosity Function

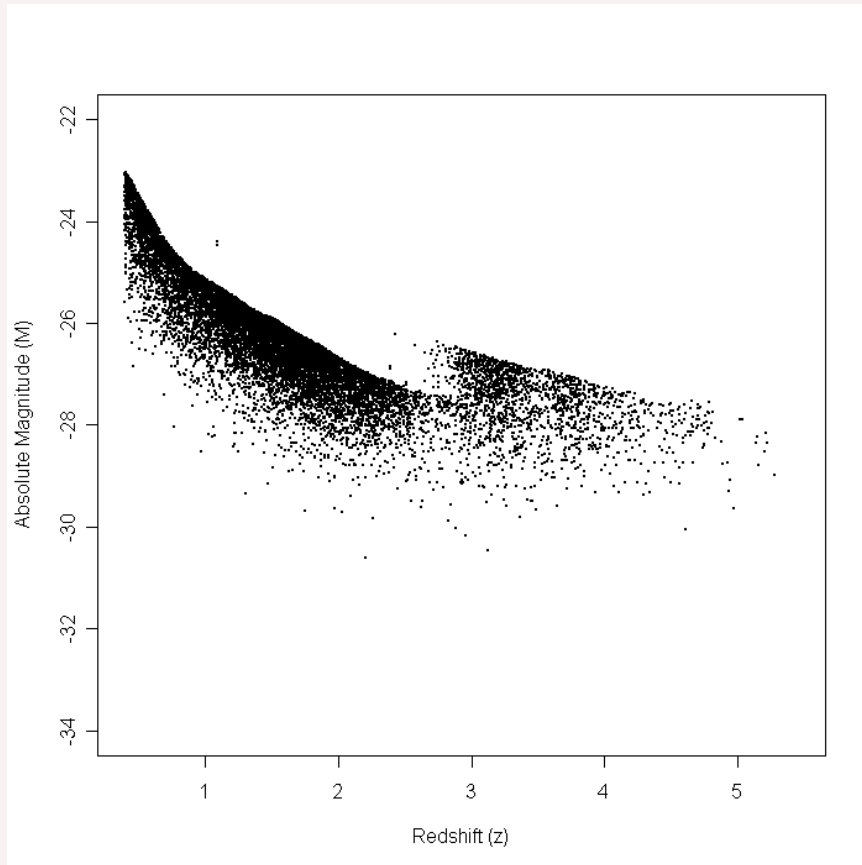


Figure 8: 14,113 quasars from SDSS DR3 ($M \times z$ space)

The Quasar Luminosity Function

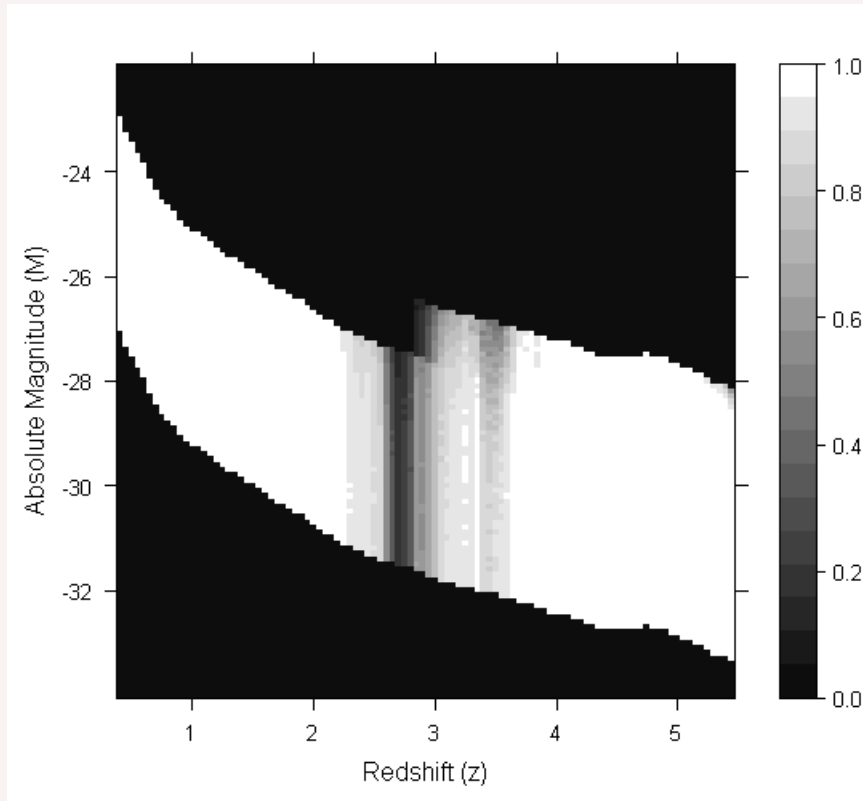


Figure 9: Observational probabilities, $\psi_M(M, z)$

The Quasar Luminosity Function

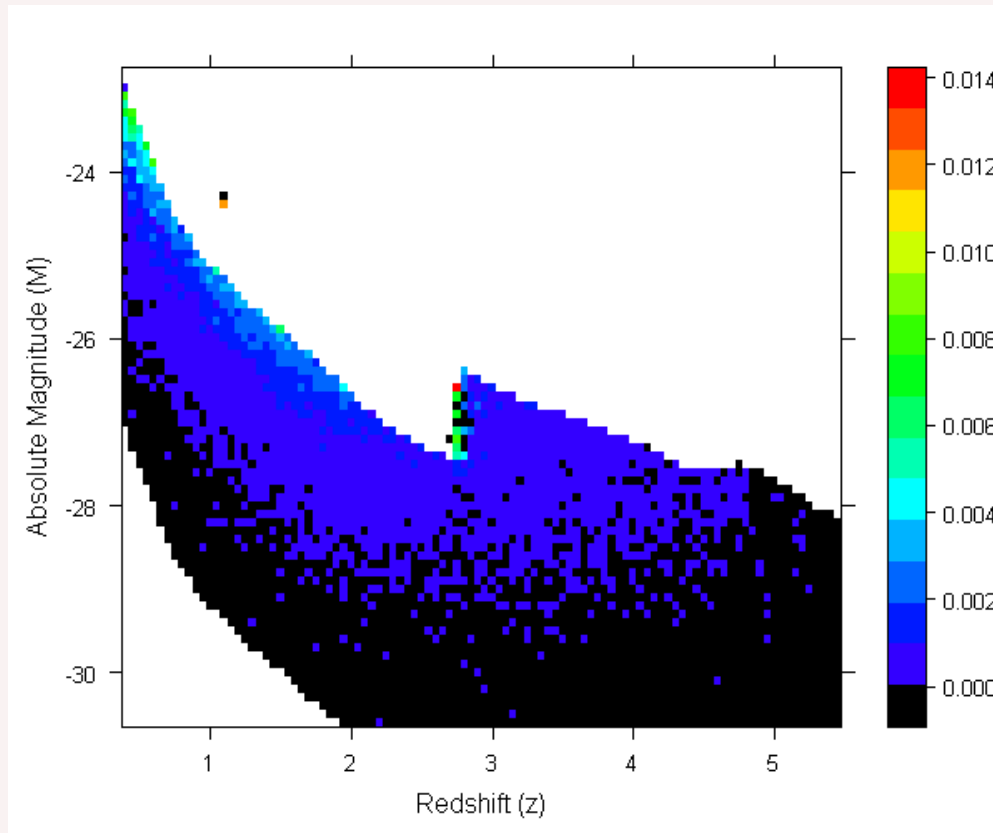


Figure 10: Estimates of Poisson Intensities on a Grid

The Quasar Luminosity Function

Separable Model Framework:

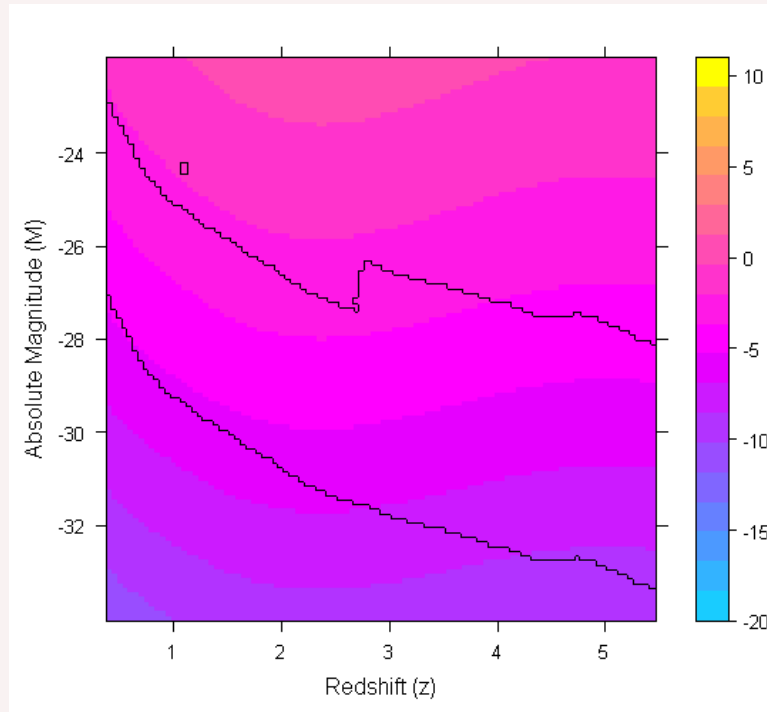


Figure 11: Exponent of idealized intensity function Φ for selected model under separable model framework

The Quasar Luminosity Function

Interaction Model Framework:

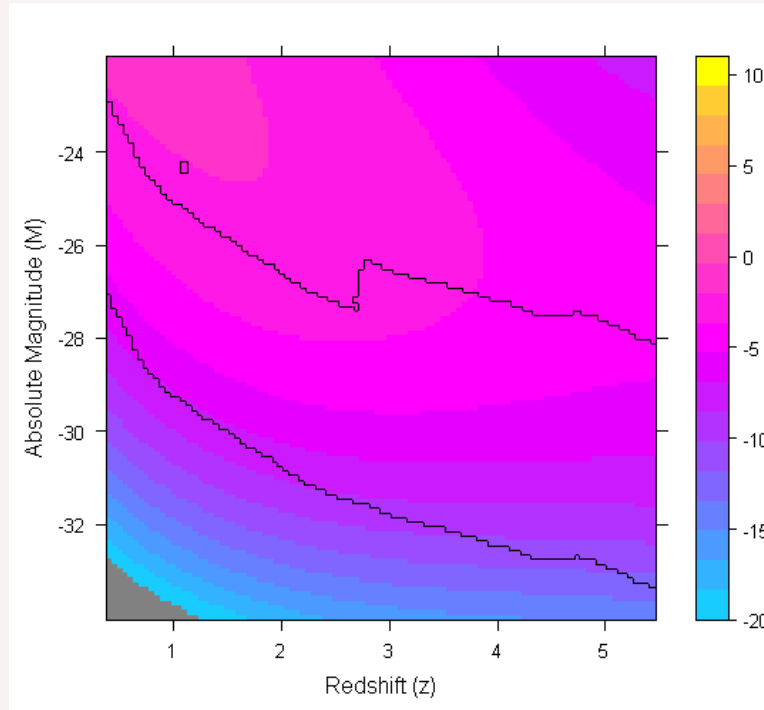


Figure 12: Exponent of idealized intensity function Φ for selected model under interaction model framework

The Quasar Luminosity Function

Interaction Model Framework: 2^{nd} best model

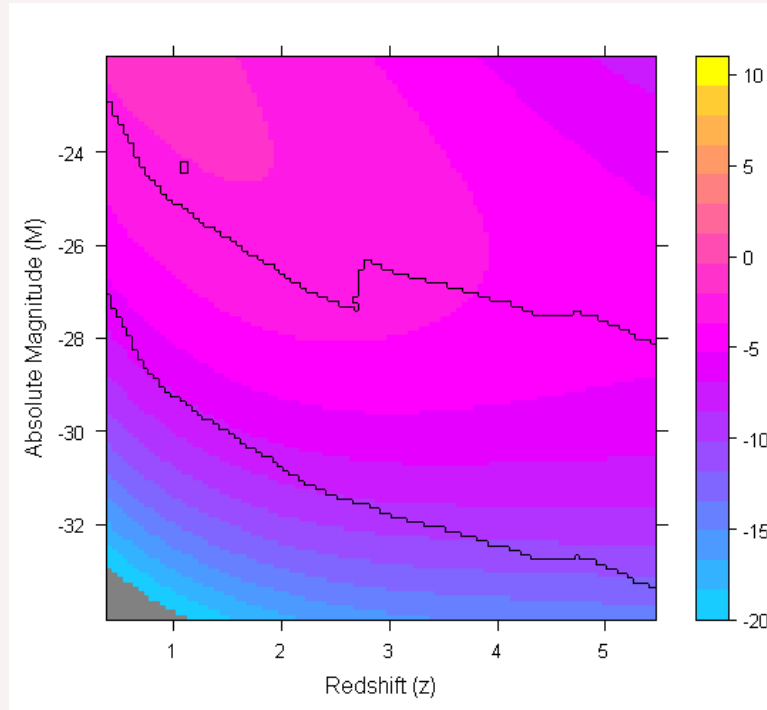


Figure 13: Exponent of idealized intensity function Φ for 2^{nd} best selected model under interaction model framework

The Quasar Luminosity Function

Partitioned Analysis:

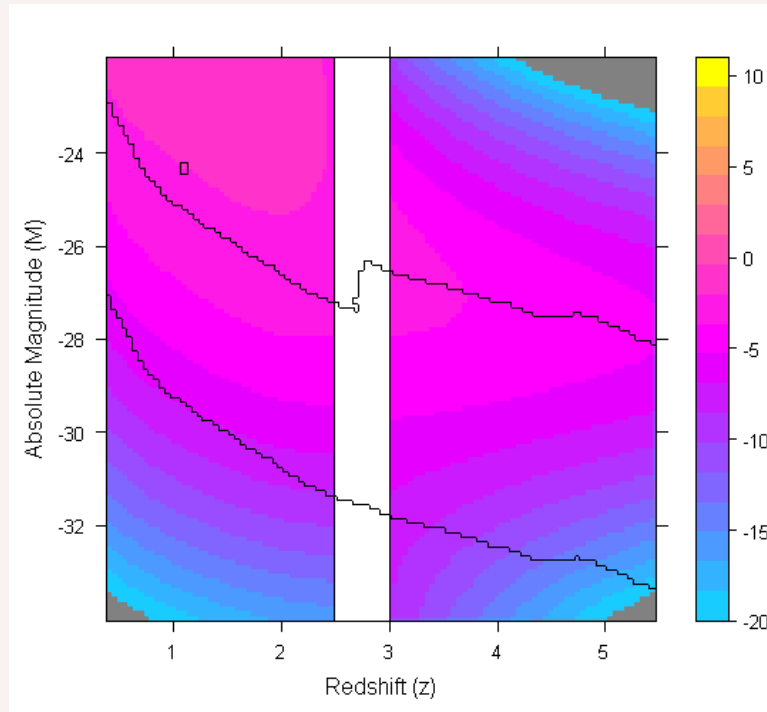


Figure 14: Exponent of idealized intensity function Φ for selected models in the the low- and high- redshift groups

Observations:

- Fitted model for the low-redshift group appears very similar to the selected model for full data
- The two side-by-side densities appear well-matched
- More plausible behavior in the limit for high-redshift group

Conclusion 1:

Evidence of $M \times z$ interaction

- This is in contrast to popular scientific theory.
- Apparent interaction observed in the full data is also present in the low-redshift region where the vast amount of data lies and where observational probabilities are high.
- Further research into this interaction is warranted.

Conclusion 2:

Problematic issues with the data

- Appear to be discrepancies in the mid-redshift data.
- Prior model of the QLF by Richards et al. (2006) achieved better behavior in the high redshift region due to convenient choice of parametric form.
- Unknown errors in traditional astronomical correction functions or the estimates of the observational probabilities could have a significant impact on the results of our analysis (unable to test for robustness).

The Quasar Luminosity Function

References: (and sources of inspiration)

- S. Carroll, 2001, “The Cosmological Constant”, Living Reviews in Relativity IV
- A.C. Davison, A. Gigli, 1989, “Deviance Residuals and Normal Scores Plots”, *Biometrika*, 76:2,211-221
- Fan et al., 2001, “High-Redshift Quasars Found in Sloan Digital Sky Survey Commissioning Data. IV.”, *The Astronomical Journal*, 121:54-65
- Hogg, Ivan, Blanton, and Eisenstein, 2002, “The K Correction”, ArXiv Astrophysics e-prints, astro-ph/0210394
- J. Moller, 2004, “Statistical Inference and Simulation for Spatial Point Processes”, Chapman and Hall, London
- Richards et al., 2002, “Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Quasar Sample”, *The Astronomical Journal*, 123:2945-2975
- Richards et al., 2006, “The SDSS Quasar Survey: Quasar Luminosity Function from Data Release Three”, *The Astronomical Journal*, 131:2766-2787
- G. Schwarz, 1978, “Estimating the dimension of a model”, *Annals of Statistics*, 6:461-464
- 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