Astrostatistics:

Overview and Highlights

Eric D. Feigelson

Penn State University

Center for Astrostatistics

PHYSTAT-Gamma 2022 Workshop

Outline

I Role & history of statistics in astronomy

II Highlight: Variability in gamma-ray sources

What is astronomy?

Astronomy is the observational study of matter beyond Earth: planets in the Solar System, stars in the Milky Way Galaxy, galaxies in the Universe, and diffuse matter between these concentrations.

Astrophysics is the study of the intrinsic nature of astronomical bodies and the processes by which they interact and evolve. This is an indirect, inferential intellectual effort based on the assumption that physics – gravity, electromagnetism, quantum mechanics, etc – apply universally to distant cosmic phenomena.

What is statistics? *(No consensus !!)*

- "… briefly, and in its most concrete form, the object of statistical methods is the reduction of data"

  (R. A. Fisher, 1922)

- "Statistics is the mathematical body of science that pertains to the collection, analysis, interpretation or explanation, and presentation of data."

  (Wikipedia, 2014)

- "Statistics is the study of the collection, analysis, interpretation, presentation and organization of data."

  (Wikipedia, 2015)

- "A statistical inference carries us from observations to conclusions about the populations sampled"

  (D. R. Cox, 1958)

Does statistics relate to scientific models?

The pessimists …

"Essentially, all models are wrong, but some are useful."

(Box & Draper 1987)

"There is no need for these hypotheses to be true, or even to be at all like the truth; rather … they should yield calculations which agree with observations" (Osiander's Preface to Copernicus' *De Revolutionibus*, quoted by C. R. Rao in *Statistics and Truth*)

"The object [of *statistical* inference] is to provide ideas and methods for the critical analysis and, as far as feasible, the interpretation of empirical data ... The extremely challenging issues of *scientific* inference may be regarded as those of synthesising very different kinds of conclusions if possible into a coherent whole or theory ... The use, if any, in the process of simple *quantitative* notions of probability and their numerical assessment is unclear."

(D. R. Cox, 2006)

The positivists …

"The goal of science is to unlock nature's secrets. … Our understanding comes through the development of theoretical models which are capable of explaining the existing observations as well as making testable predictions. …

"Fortunately, a variety of sophisticated mathematical and computational approaches have been developed to help us through this interface, these go under the general heading of statistical inference."

(P. C. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences,* 2005)

Recommended steps in the statistical analysis of scientific data

The application of statistics can reliably quantify information embedded in scientific data and help adjudicate the relevance of theoretical models. But this is not a straightforward, mechanical enterprise. It requires:

- exploration of the data
- careful statement of the scientific problem
- model formulation in mathematical form
- choice of statistical method(s)
- calculation of statistical quantities ← *easiest step with R*
- judicious scientific evaluation of the results

Astronomers often do not adequately pursue each step

Why astrostatistics is difficult !

- Modern statistics is vast in its scope and methodology. It is difficult to find what may be useful (jargon problem!), and there are usually several ways to proceed. Very confusing.

- Some statistical procedures are based on mathematical proofs while others are not. It is perilous to violate mathematical truths! Some issues are debated among statisticians, or have no known solution.

- Scientific inferences should not depend on arbitrary choices in methodology & variable scale. Start with nonparametric & scale-invariant methods. Try multiple methods.

- It can be difficult to interpret the meaning of a statistical result with respect to the scientific goal. Statistics is only a tool towards understanding nature from incomplete information.

We should be knowledgeable in our use of statistics and judicious in its interpretation

Astronomy & Statistics: A glorious past

For most of western history,
the astronomers were the statisticians!

Ancient Greeks to today

Best estimate of the length of a year from discrepant data?

- Middle of range: Hipparcos (4th century B.C.)
- Observe only once! (medieval)
- Mean: Brahe (16th c), Galileo (17th c), Simpson (18th c)
- Median with bootstrap (21th c)

19th century

Discrepant observations of planets/moons/comets used to estimate orbital parameters using Newtonian celestial mechanics

- Legendre, Laplace & Gauss develop least-squares regression and normal error theory (~1800-1820)
- Prominent astronomers contribute to least-squares theory (~1850-1900)

The lost century of astrostatistics….

In the late-19th and 20th centuries, statistics moved towards human sciences (demography, economics, psychology, medicine, politics) and industrial applications (agriculture, mining, manufacturing).

During this time, astronomy recognized the power of modern physics: electromagnetism, thermodynamics, quantum mechanics, relativity. Astronomy & physics were wedded into astrophysics.

Thus, astronomers and statisticians substantially broke contact; e.g. the curriculum of astronomers heavily involved physics but little statistics. Statisticians today know little modern astronomy.

The state of astrostatistics today
(not so good but rapidly improving)

Many astronomical studies are confined to a narrow suite of familiar statistical methods:

- Fourier transform for temporal analysis (Fourier 1807)
- Least squares regression (Legendre 1805, Pearson 1901)
- Kolmogorov-Smirnov goodness-of-fit test (Kolmogorov, 1933)
- Principal components analysis for tables (Hotelling 1936)

Even traditional methods are sometimes misused!

- *Kolmogorov-Smirnov test has three limitations*
- *Likelihood ratio test can't be used for parameters near zero*
- *Bayesian priors should not be improper*

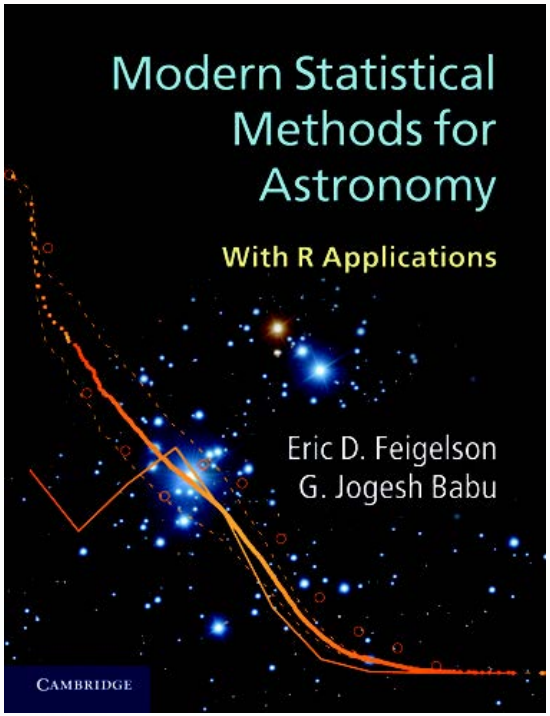
https://asaip.psu.edu/Articles/beware-the-kolmogorov-smirnov-test/
Protassov et al. 2002
Tak et al. 2018

Under-utilized methodology from the 20th century:

- modeling (MLE, EM Algorithm, BIC, bootstrap)
- multivariate classification (LDA, SVM, CART, RFs)
- time series (autoregressive models, state space models)
- spatial point processes (Ripley's K, kriging)
- nondetections (survival analysis)
- image analysis (computer vision methods, False Detection Rate)
- statistical computing (R)

Advertisement …

Modern Statistical Methods for Astronomy with R Applications
E. D. Feigelson & G. J. Babu,
Cambridge Univ Press, 2012



Winner 2012 PROSE Award for Best Astronomy & Cosmology Book

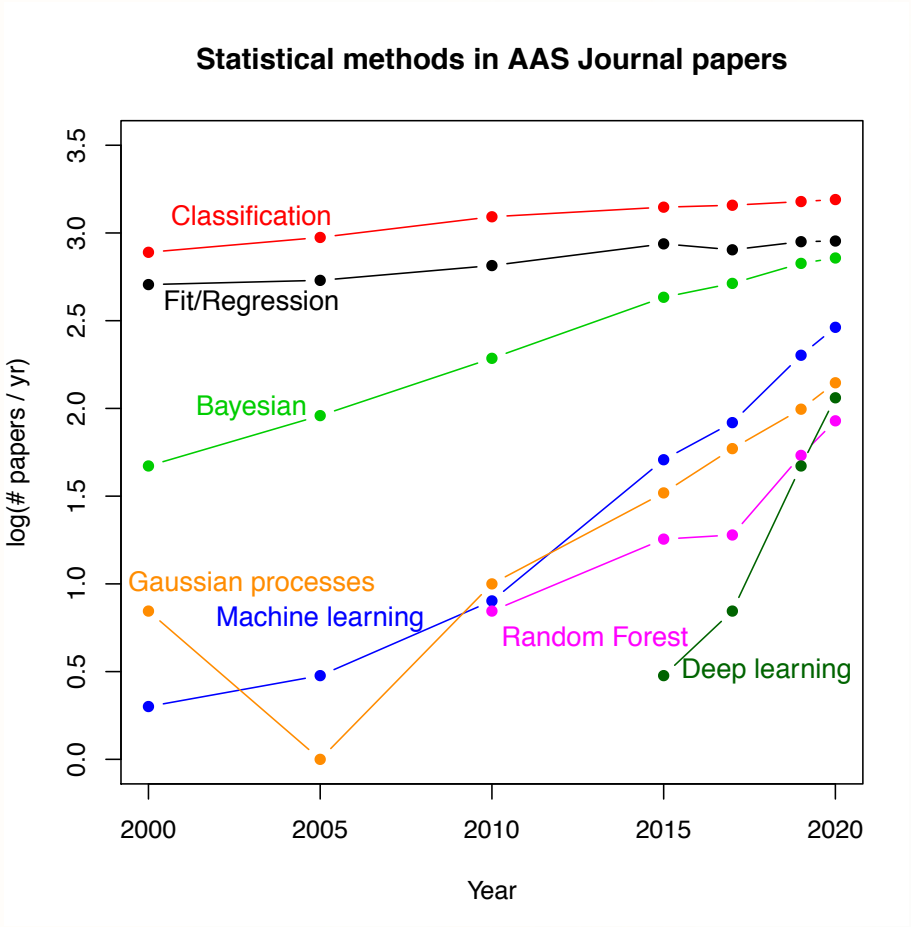
Astrostatistics is difficult: it involves many fields of statistics

Cosmology	⟷	**Statistics**
Galaxy clustering	⟷	Spatial point processes, clustering
Galaxy morphology	⟷	Regression, mixture models
Galaxy luminosity fn	⟷	Gamma distribution
Power law relationships	⟷	Pareto distribution
Weak lensing morphology	⟷	Geostatistics, density estimation
Strong lensing morphology	⟷	Shape statistics
Strong lensing timing	⟷	Time series with lag
Faint source detection	⟷	False Discovery Rate
Multiepoch survey lightcurves	⟷	Multivariate classification
CMB spatial analysis	⟷	Markov fields, ICA, etc
ΛCDM parameters	⟷	Bayesian inference & model selection
Comparing data & simulation	⟷	Uncertainty Quantification

Recent resurgence in astrostatistics

- Improved access to statistical software: R/CRAN, Matlab & Python

- A significant fraction of papers in the astronomical literature use modern methodology and is growing exponentially



- Short training courses (Penn State has run tutorials in ~17 nations)
- Cross-disciplinary research collaborations (Harvard, CMU, Penn State, CEA-Saclay, Cornell, Imperial College London …)
- Cross-disciplinary conferences *(Statistical Challenges in Modern Astronomy 1991-2023, Astronomical Data Analysis 1991-2016, SAMSI 2006/2012/2016, Astroinformatics 2012-2020)*
- Scholarly societies:
  - Internationl Stat Institute SIGAstro
  - International Astrostatistical Assn
  - International Astro Union Commission B3
  - American Astro Soc Working Group
  - American Stat Assn Interest Group
  - LSST Info/Stat Science Collaboration
  - IEEE Astro Data Miner Task Force

Several textbooks in astrostatistics

Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica Support
Gregory, 2005

Practical Statistics for Astronomers
Wall & Jenkins, 2nd ed, 2012

Modern Statistical Methods for Astronomy with R Application,
Feigelson & Babu, 2012

Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data,
Ivecic, Connolly, VanderPlas & Gray, 2014 (2nd edition in preparation)

+ many texts written by statisticians to teach specific fields of methodology, often with R code. R has 18K packages growing ~5/day. A new text with "R" in the title has been published every ~10 days for the past decade.

A vision of astrostatistics by 2030 …

- Astronomy graduate curriculum has 1 year of statistical and computational methodology
- Some astronomers have M.S. in statistics or data science
- Astrostatistics and astroinformatics is a well-funded, cross-disciplinary research field involving a few percent of astronomers pushing the frontiers of methodology (similar to astrochemistry)
- Astronomers regularly use advanced methods coded in R.
- *Statistical Challenges in Modern Astronomy* meetings are held biannually with hundreds of participants

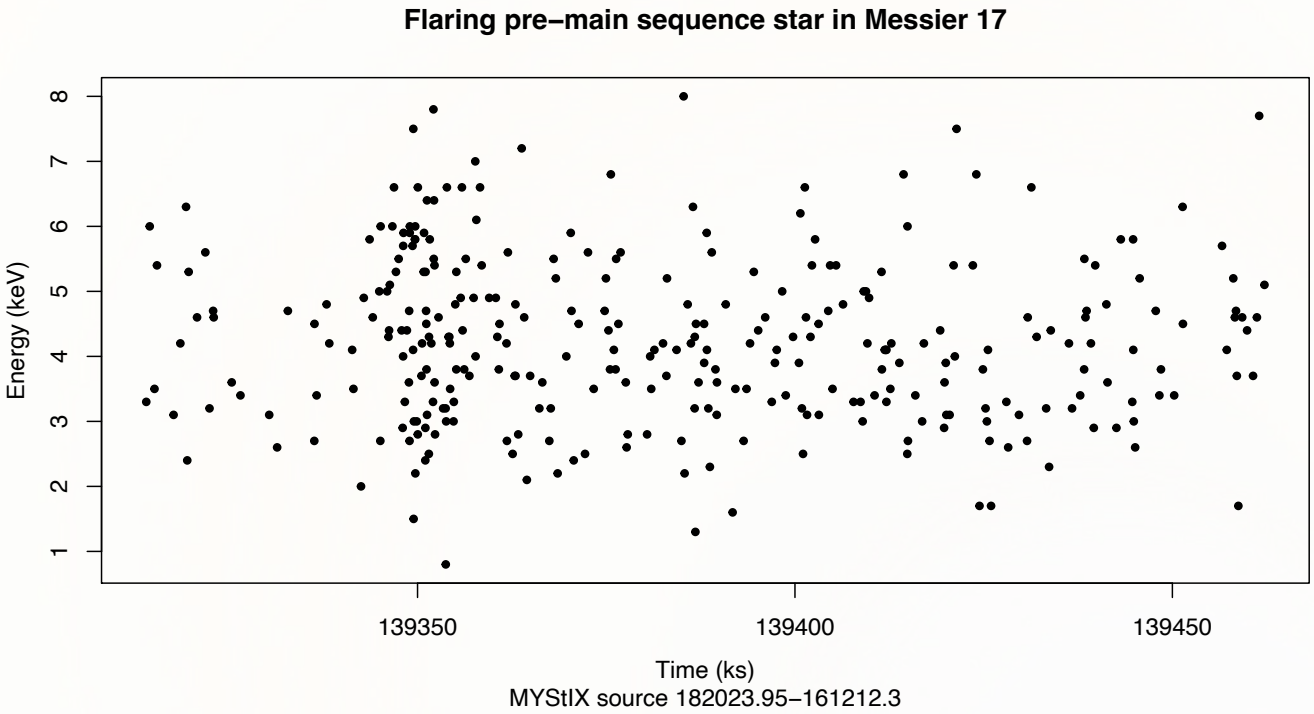
Highlight

Statistical Approaches to Detecting Variability in Low Count Sources

The Dataset

312 photons from a pre-main sequence star arriving during a 148 ks Chandra ACIS exposure of the Messier 17 star forming region.

Time	Energy
139314.06	3.3
139314.52	6.0
139315.15	3.5
139315.48	5.4
139317.68	3.1
139318.58	4.2
139319.33	6.3
139319.49	2.4
…	
…	

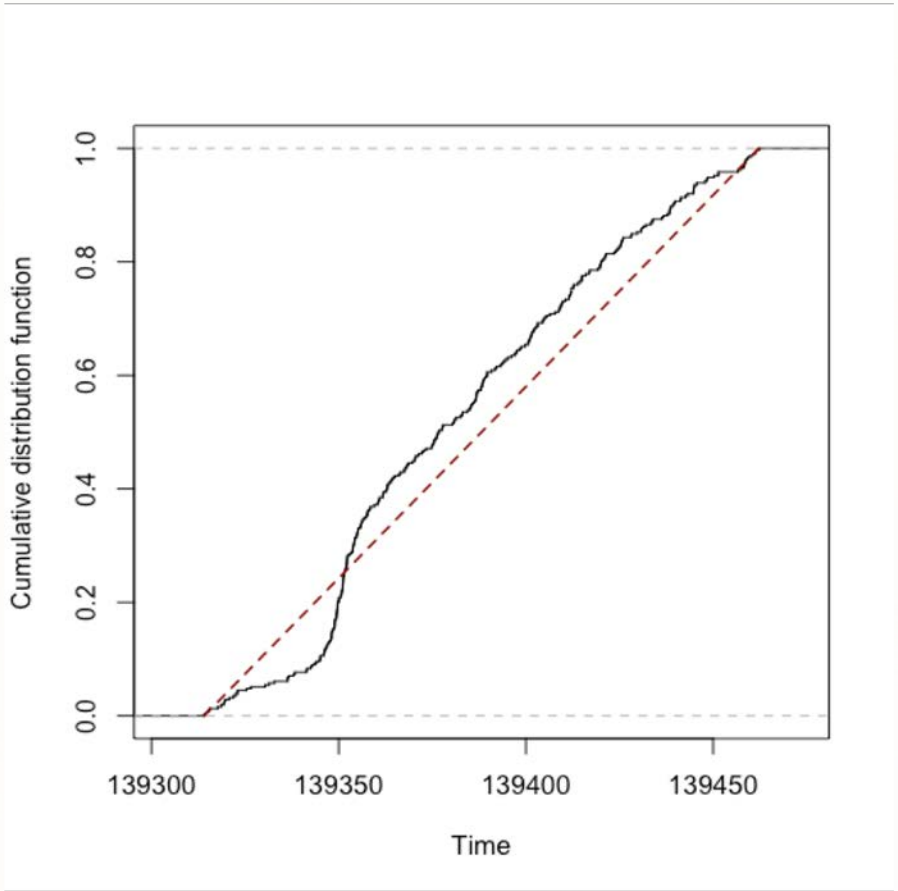


Unbinned nonparametric tests for variability

Kolmogorov–Smirnov test
P = 4%

Anderson–Darling test
P = 0.0004%

The K-S test is most sensitive to long-timescale variations, and does not pick this short-lived flare. The Anderson-Darling test is much more effective here.

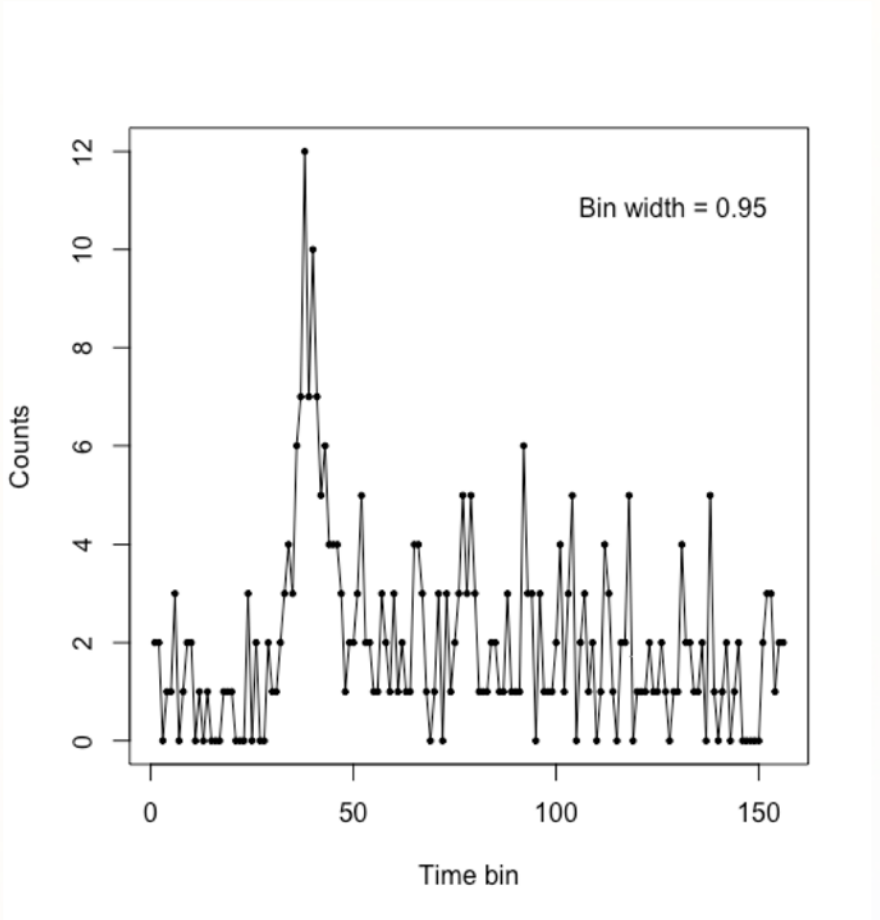


R: ecdf, ks.test
CRAN: DescTools

Binned tests for variability

The purpose of binning is not to approximate a Gaussian distribution, but rather to obtain an evenly-spaced time series of Poisson-distributed count data. Here the binwidth is chosen so the Poisson intensity is

$$\lambda = 2.0 \text{ counts/bin}$$



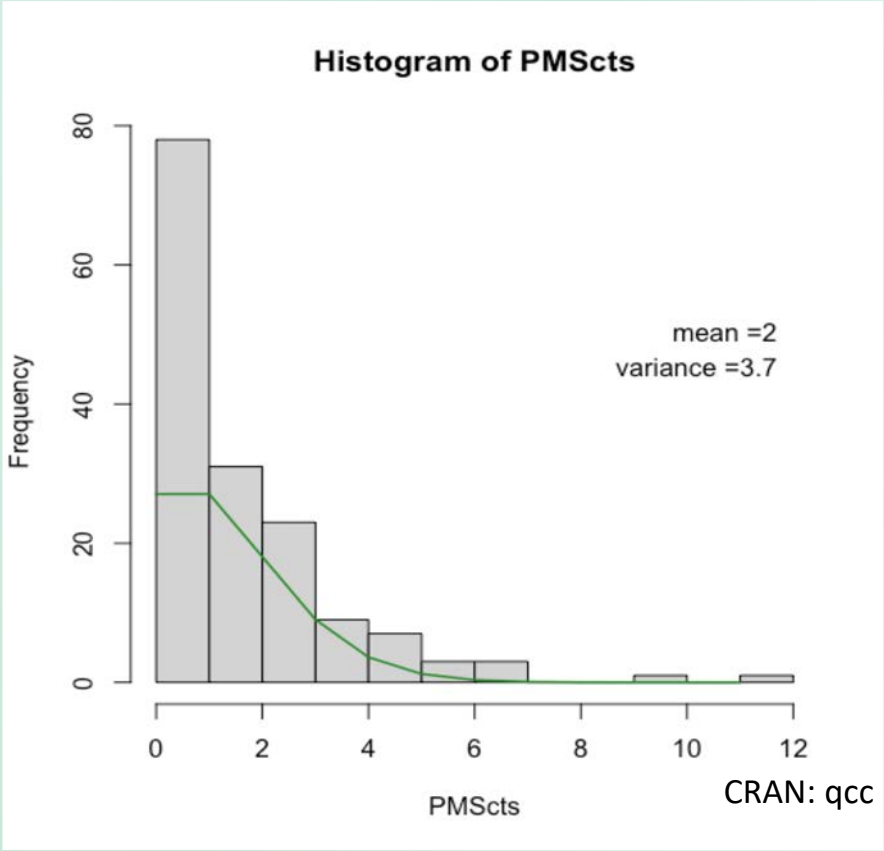
Binned tests for variability

(without temporal information)

First we apply two tests on the distribution of counts, paying no attention to the temporal behavior.

We find that the distribution is strongly ***overdispersed*** *where the variance is greater than the mean. For a homogeneous Poisson process, the variance is equal to the mean (green curve).*

Using a χ^2 *test,* $P < 0.0001\%$ *that the count distribution arises from a constant intensity Poisson process.*

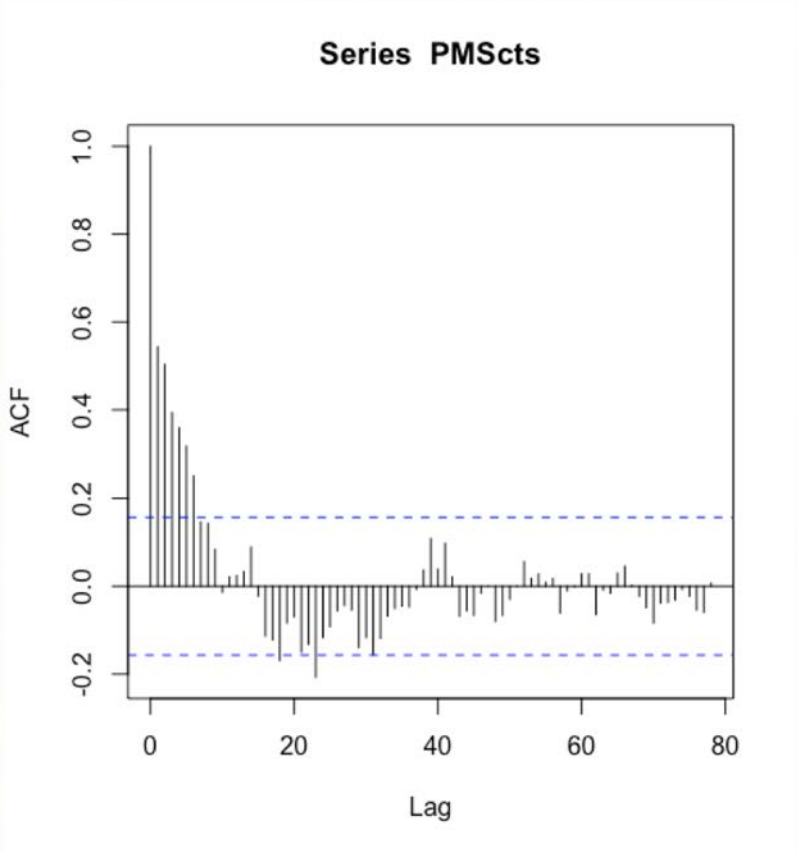


CRAN: qcc

Binned tests for variability
(with temporal information)

The nonparametric autocorrelation function shows strong autocorrelation at short lags.

The dashed lines here show 95% confidence intervals for Gaussian data. These are underestimates for Poisson data.



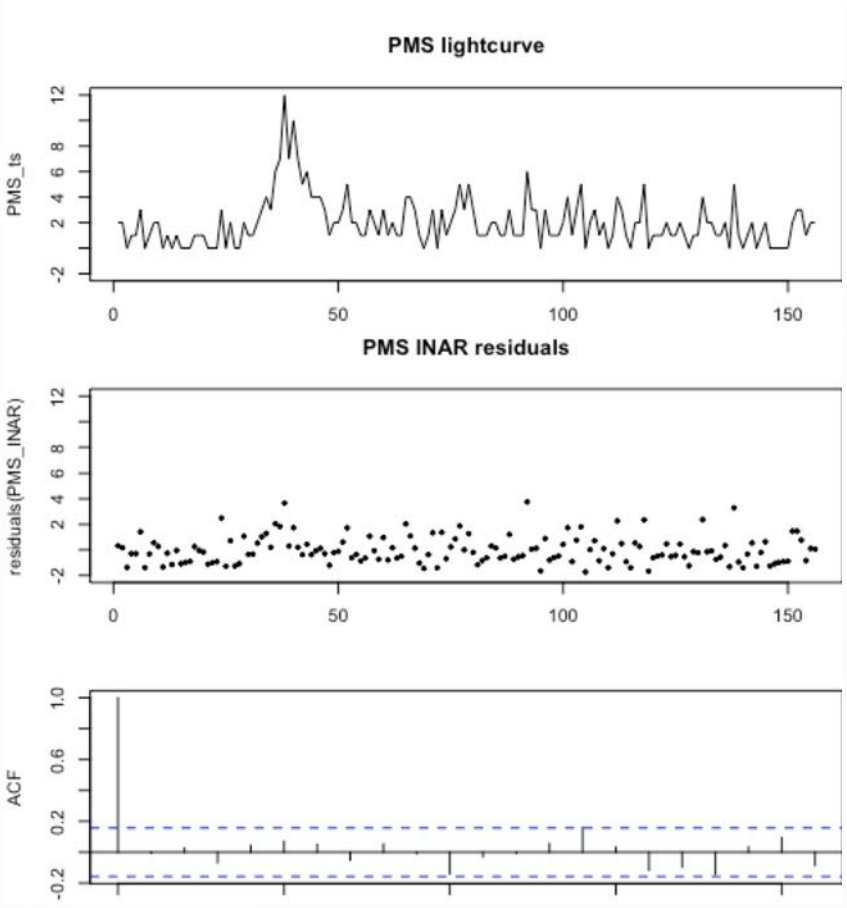
R: acf

Binned modeling of variability

If the autocorrelated behavior is stationary (present at all times in the lightcurve), then it can often be modeled with parametric autoregressive functions in the ARMA class.

Since we have count data, the appropriate models are INAR (Integer AR) or PAR (Poisson AR) models. We fit here a linear INAR(1) model.

Surprisingly, this simple linear INAR(1) removes most of the flare and the residuals show no autocorrelation. It suggests that most of the variations in the lightcurve arise from an autoregressive stochastic process, such as the 'avalanche' model of solar flare occurrences.

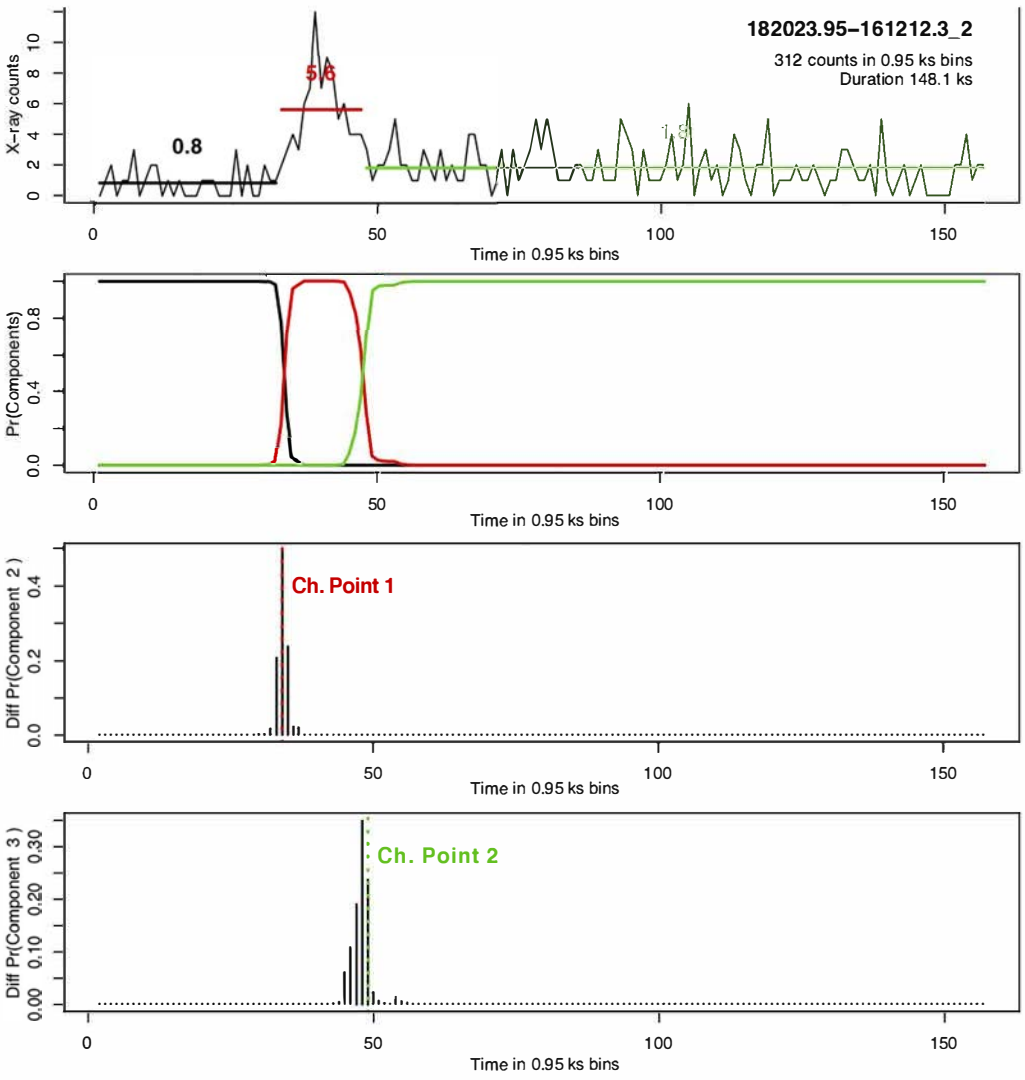


CRAN: acp

Multiple Poisson changepoint modeling

Here we assume that the behavior is deterministic with flux that jumps discontinuously between constant intensity levels. This is the statistical model of Jeff Scargle's Bayesian Blocks.

It is a considerable statistical challenge to fit this model, as the number and location of changepoints is unknown. Bayesian methods are used with different advanced computational algorithms: dynamic programming (Scargle) or latent variables (Chibb).



Multiple changepoint model for an inhomogeneous Poisson process using the latent variable algorithm of Chibb (1998).

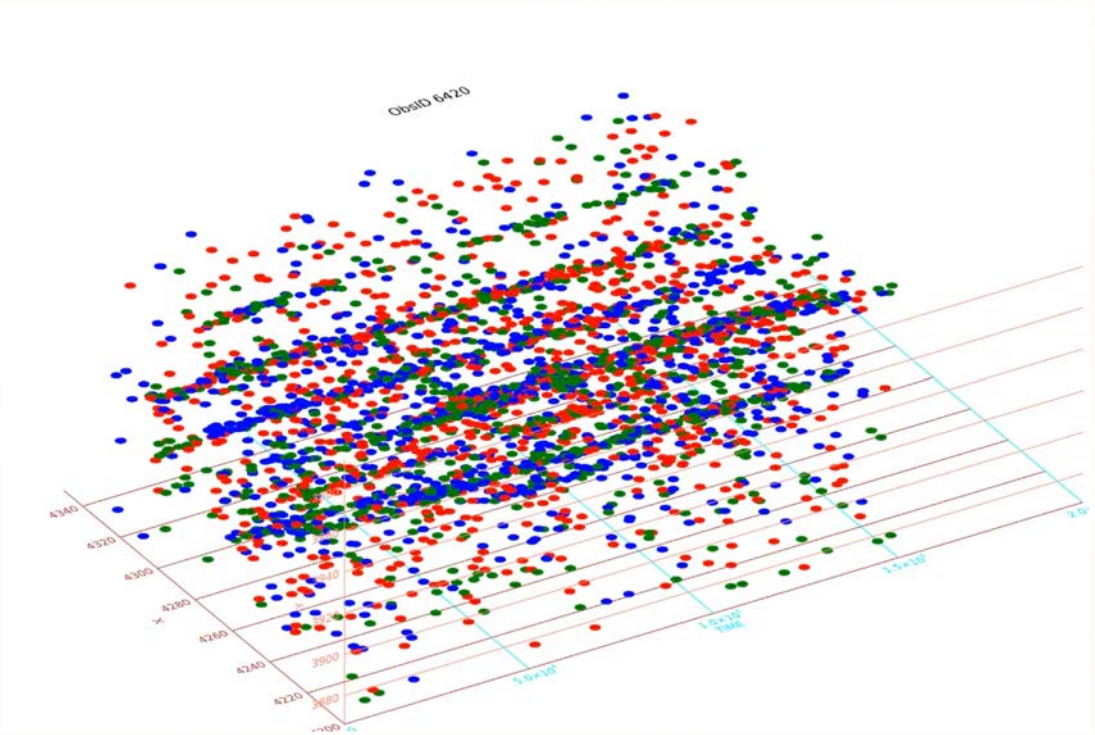
CRAN: acp, MCMCpack

Getman & Feigelson, ApJ 916:32 2021

Multivariate detection of source variability

Though typically studied as a univariate time series, Chandra data is a Poisson process in four dimensions: RA, Dec, energy & time. For crowded fields, a multivariate treatment can be effective.

Here is a movie of the vicinity of the flaring M17 pre-main sequence star. It shows the (RA, Dec, time) datacube with photon energy in color.



4D_Automark
Xu+ ApJ 2021

Change-point Detection and Image Segmentation for Time Series of Astrophysical Images

C. Xu, H. Gunther, V.Kashyap, T. Lee & A. Zezas, AJ 161:184 (2021)

Model is piecewise constant flux with multiple changepoints for a Poisson process (similar to Bayesian Blocks). Model selection based on *minimum description length (MDL)* criterion.

Algorithm is computationally intensive, requiring preprocessing and good starting points

Python code is available: *4D_Automark (*Github)

The Automark method is very general, finding changepoints in space, time and energy

Some book references

- P. Del Moral & S. Penev, *Stochastic Processes: From Applications to Theory* (2017)
- J. Kingman, *Poisson Processes* (1993)
- A. Tartakovsky, I. Nikiforov & M. Basseville, *Sequential Analysis: Hypothesis Testing and Changepoint Detection* (2014)
- J. Hilbe, *Modeling Count Data* (2014)
- O. Pons, *Estimations and Tests in Change-Point Models (2018)*
- J. Grandell, *Mixed Poisson Processes* (1997)
- R. Streit, *Poisson Point Processes: Imaging, Tracking and Sensing (2010)*

Highlight conclusions

A wide variety of statistical procedures are available to detect and characterize variability in low-count-rate X-ray sources. Most of these methods are not used by X-ray astronomers.

They can be roughly classified as: nonparametric and Poisson-based hypothesis tests for variability; tests based on binned count rate distributions; autoregressive and changepoint models based on Poisson processes; and multivariate methods.

Each method has distinctive capabilities and limitations. Some methods are from classical statistics, others are focused on Poisson processes, and others are emerging from astrostatistical innovations.

Existing software packages (e.g. CIAO, XRONOS) are completely inadequate for this problem. Codes for many methods are available in R; these are easily wrapped into Python.