Introduction to astrostatistics

Eric Feigelson (Astro & Astrophys)
&
G. Jogesh Babu (Stat)

Center for Astrostatistics
Penn State University

Astronomy & statistics: A glorious history

Hipparchus (4th c. BC): **Average via midrange of observations**

Galileo (1572): **Average via mean of observations**

Halley (1693): **Foundations of actuarial science**

Legendre (1805): **Cometary orbits via least squares regression**

Gauss (1809): **Normal distribution of errors in planetary orbits**

Quetelet (1835): **Statistics applied to human affairs**

But the fields diverged in the late 19-20th centuries,
astronomy → astrophysics (EM, QM, GR)
statistics → social sciences & industries

Do we need statistics in astronomy today?

- **Are the stars/galaxies/sources we are studying an unbiased sample of the vast underlying population?**
- **When should these objects be divided into 2/3/… classes?**
- **What is the intrinsic relationship between two properties of a class (especially with confounding variables)?**
- **Can we answer such questions in the presence of observations with measurement errors & flux limits?**

Do we need statistics in astronomy today?

- **Are the stars/galaxies/sources we are studying an unbiased sample of the vast underlying population?** **Sampling**
- **When should these objects be divided into 2/3/… classes?** **Multivariate classification**
- **What is the intrinsic relationship between two properties of a class (especially with confounding variables)?** **Multivariate regression**
- **Can we answer such questions in the presence of observations with measurement errors & flux limits?**

  Censoring, truncation & measurement errors

- **When is a blip in a spectrum, image or time series a real signal?** **Statistical inference**
- **How do we model the vast range of variable objects (extrasolar planets, BH accretion, GRBs, …)?** **Time series analysis**
- **How do we model the 2/4/6-dimensional points representing galaxies in the Universe or photons in a detector?** **Spatial point processes & image processing**
- **How do we model continuous structures (CMB fluctuations, interstellar/intergalactic media)?** **Density estimation, regression**

How often do astronomers need statistics?

(a bibliometric measure)

Of ~15,000 refereed papers annually:

1% have `*statistics*' in title or keywords

5% have `*statistics*' in abstract

10% treat variable objects

5-10% (est) analyze data tables

5-10% (est) fit parametric models

The state of astrostatistics today

The typical astronomical study uses:

- Fourier transform for temporal analysis (Fourier 1807)
- Least squares regression (Legendre 1805) & minimum χ^2 (Pearson 1901)
- Kolmogorov-Smirnov goodness-of-fit test (Kolmogorov, 1933)
- Principal components analysis for tables (Hotelling 1936)

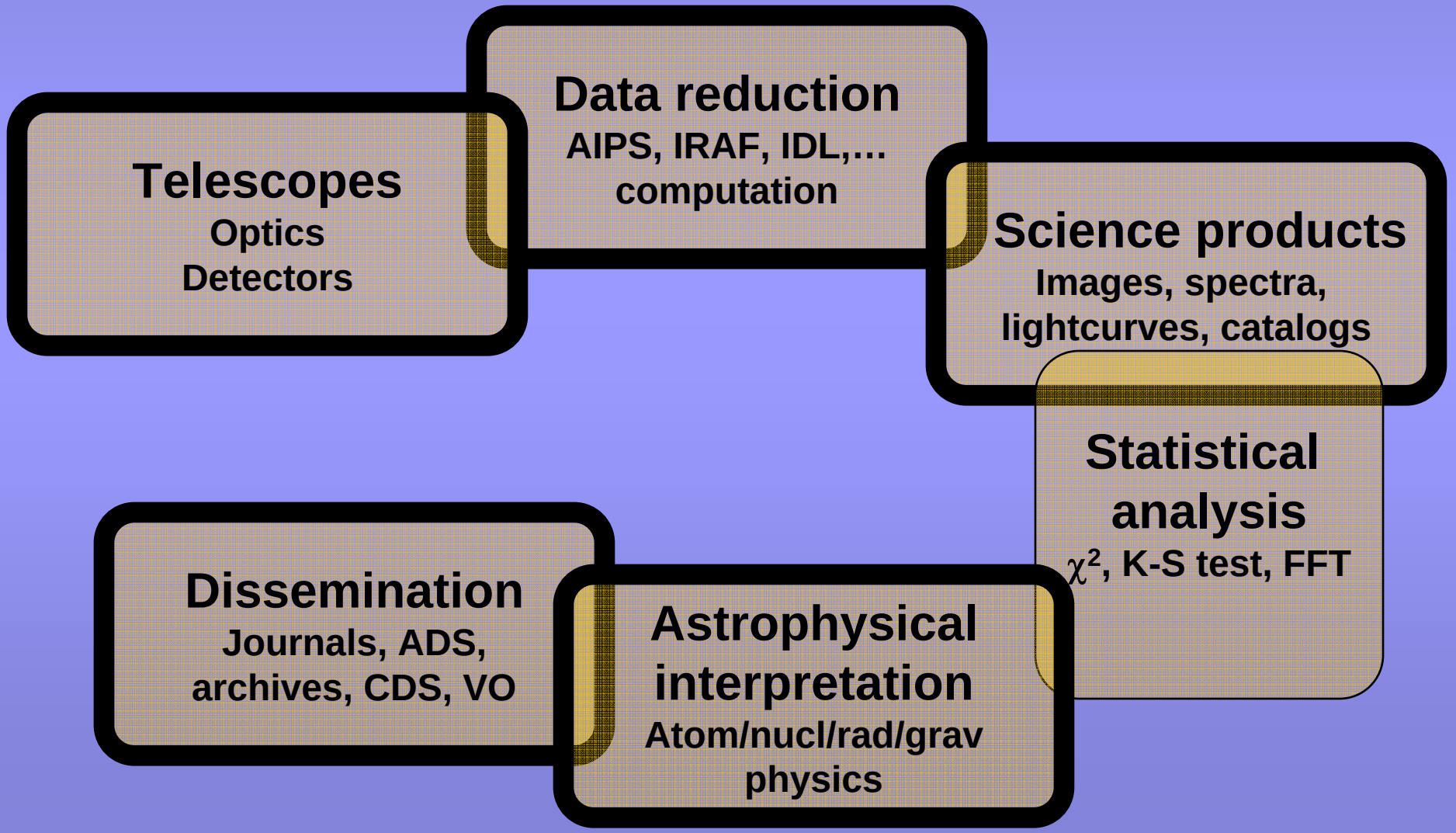
Even traditional methods are often misused:

- Six unweighted bivariate least squares fits are used interchangeably in H_o studies with wrong confidence intervals

  Feigelson & Babu ApJ 1992
- Likelihood ratio test (F test) usage typically inconsistent with asymptotic statistical theory

  Protassov et al. ApJ 2002

Statistical methodology is the weak link in astronomical research



But astrostatistics is undergoing a resurgence …

- Growing interest in some modern methods -- wavelets, MCMC, neural nets – but not in a full context of capabilities & alternatives

- Conferences & books:
  - *Statistical Challenges in Modern Astronomy* at Penn State *(1991, 1996, 2001, 2006)*
  - Data analysis meetings & monographs (e.g. by Fionn Murtagh & Jean-Luc Starck)
  - Astrostat sessions as AAS and JSM/ISI meetings

- Emerging astro-stat collaborations:
  - Harvard/Smithsonian (David van Dyk, Chandra scientists, students)
  - CMU/Pitt = PiCA (Larry Wasserman, Chris Genovese, Bob Nichol, …)
  - NASA-ARC/Stanford (Jeffrey Scargle, David Donoho)
  - Berger/Jeffreys/Loredo/Connors
  - Efron/Petrosian, Stark/GONG, …

A new imperative: Virtual Observatory

Huge, uniform, multivariate databases are emerging from specialized survey projects & telescopes:

- **10^9-object catalogs from USNO, 2MASS & SDSS opt/IR surveys**
- **10^6- galaxy redshift catalogs from 2dF & SDSS**
- **10^5-source radio/infrared/X-ray catalogs**
- **10^{3-4}-samples of well-characterized stars & galaxies with dozens of measured properties**
- **Many on-line collections of 10^2-10^6 images & spectra**
- **Planned Large-aperture Synoptic Survey Telescope will generate ~10 Pby**

The Virtual Observatory is an international effort underway to federate these distributed on-line astronomical databases.

Powerful statistical tools are needed to derive scientific insights from extracted VO datasets

Types of astronomical datasets
(and their statistical challenges)

Continuous univariate data

Examples: **Spectra, lightcurves, instrumental time series**

Issues: **Unevenly-spaced observations, heteroscedastic errors, Gaussian/Poissonian signals, periodic/ aperiodic/explosive variations**

Bivariate imaging data

Examples: **IR, visible, X-ray images**

Issues: **Gaussian/Poissonian signals, distorting optics, hierarchical structures**

Multivariate data

Examples: Catalogs & science studies

(rows: stars/galaxies/sources, cols: properties,)

Issues: Heteroscedastic errors, truncated & censored data, non-Gaussian populations

Combinations

- Interferometric data (radio, IR, gamma-ray) using Fourier transforms to give 3-dim spectro-imaging
- Interferometric gravitational wave observatories seeking weak periodic & single event signals
- SDSS obtains 10^5 spectra of galaxies & quasars
- LSST will obtain 10^{10} lightcurves

Multivariate classification of gamma ray bursts

GRBs are extremely rapid (1-100s), powerful (L~10^{51} erg), explosive events occurring at random times in normal galaxies. They probably arise from a subclass of supernova explosions or colliding binary neutron stars which produces a collimated relativistic fireball. Current detectors can see ~1/day. ~2500 papers have been written on GRBs with \$$10^{8}$-$10^{9}$ spent on their detection & study.

We consider here only the limited question of whether the bulk properties of the GRB itself suggest multiple classes of events. Properties include: location in the sky, arrival time, duration, fluence, and spectral hardness.

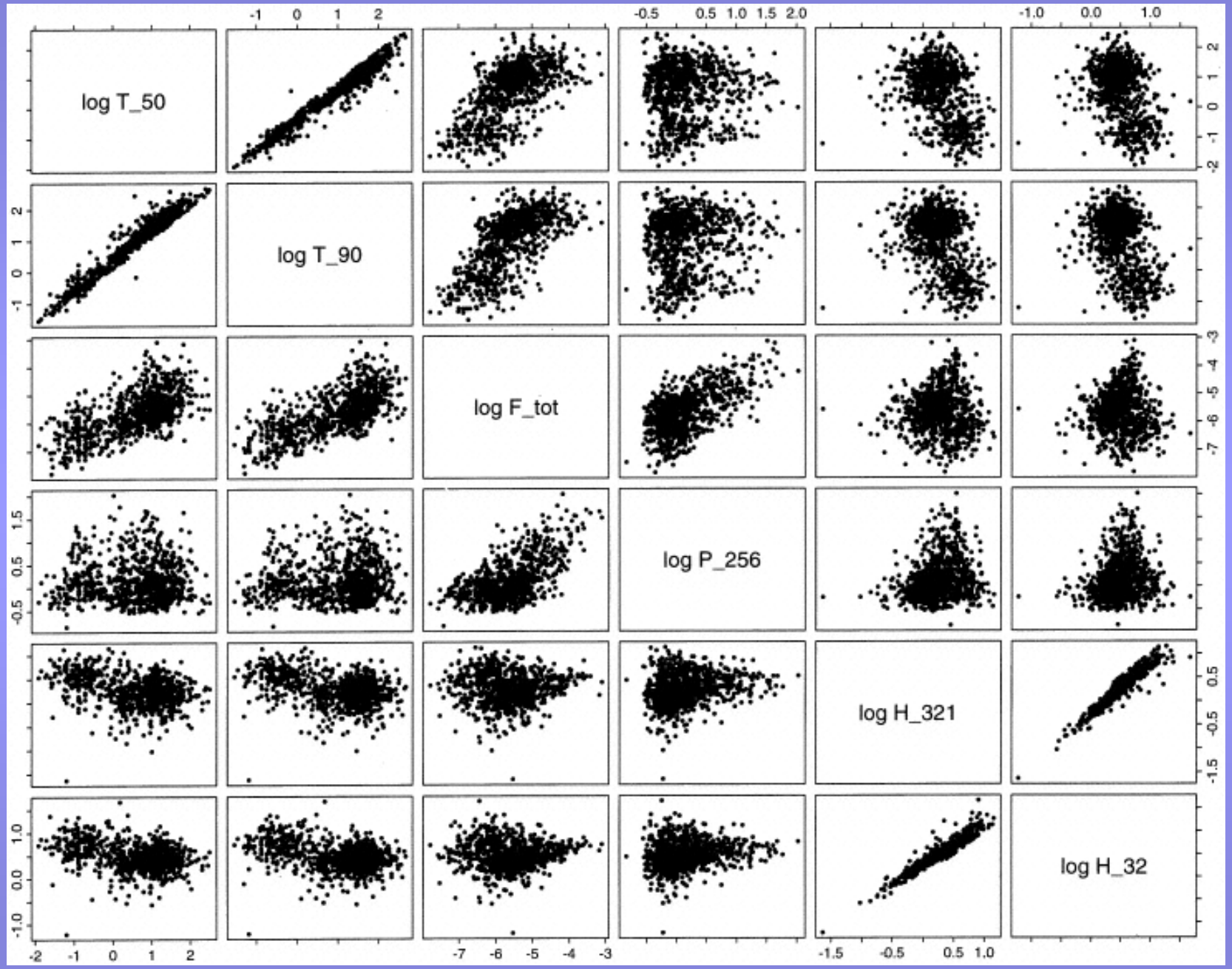
THE THIRD BATSE GAMMA-RAY BURST CATALOG

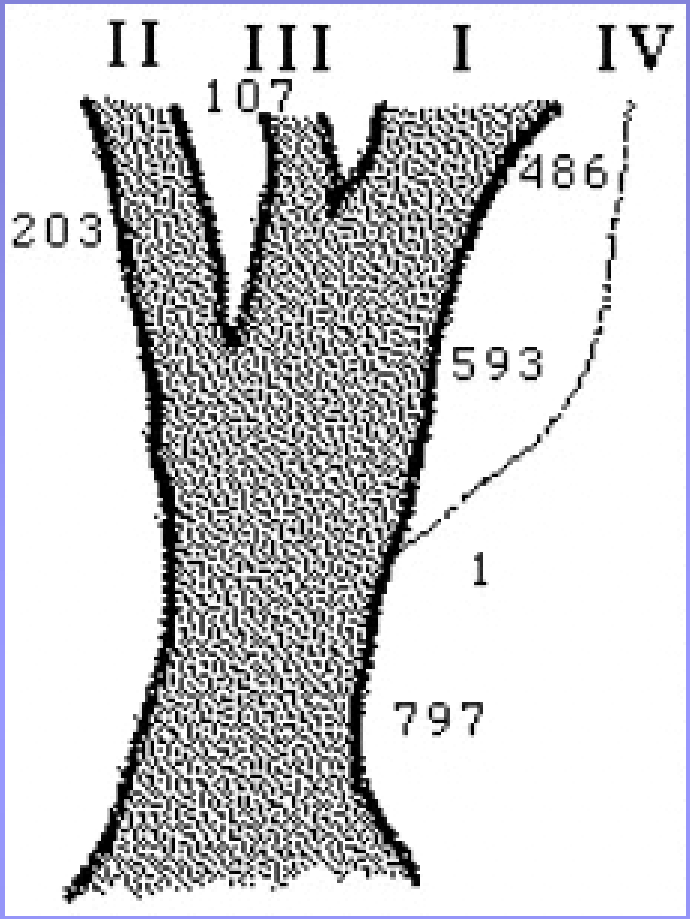
CHARLES A. MEEGAN,[1] GEOFFREY N. PENDLETON,[2] MICHAEL S. BRIGGS,[2] CHRYSSA KOUVELIOTOU,[3]
THOMAS M. KOSHUT,[2] JOHN PATRICK LESTRADE,[4] WILLIAM S. PACIESAS,[2]
MICHAEL L. MCCOLLOUGH,[5] JEROME J. BRAINERD,[2] JOHN M. HORACK,[1]
JON HAKKILA,[6] WILLIAM HENZE,[7] ROBERT D. PREECE,[2]
ROBERT S. MALLOZZI,[2] AND GERALD J. FISHMAN[8]

Received 1995 October 30; accepted 1996 February 5

Trigger Number	Burst Name	Time (TJD:s)	Time (DOY:h:m:s)	α (°)	δ (°)	l^{II} (°)	b^{II} (°)	Stat. Loc. Error (°)	C_{max}/C_{min}	C_{min}	Time Scale (ms)
105	3B 910421	8367:33243.8	111:09:14: 3.8	270.7	24.8	50.8	21.2	0.5	34.1	286	1024
107	3B 910423	8369:71684.7	113:19:54:44.7	193.5	−8.4	304.0	54.5	11.1	1.1	264	1024
108	3B 910424	8370:71006.6	114:19:43:26.6	201.3	−45.4	309.1	17.1	13.8	1.0	60	64
109	3B 910425	8371: 2265.7	115:00:37:45.7	91.3	−22.8	229.0	−19.9	1.0	13.1	286	1024
110	3B 910425B	8371:20253.3	115:05:37:33.3	335.9	25.8	85.8	−26.3	4.8	1.6	264	1024
111	3B 910426	8372:80046.7	116:22:14: 6.7	75.8	−19.5	219.8	−32.3	2.7	1.4	264	1024
114	3B 910427	8373:32720.7	117:09:05:20.7	78.9	−15.6	216.9	−28.1	9.1	1.6	264	1024

Trigger Number	Burst Name	T_{90} (s)	Peak Flux (ph cm^{-2} s^{-1})	Fluence 50–300 keV (erg cm^{-2})	Hardness Ratio	Fluence > 20 keV (erg cm^{-2})	Comments
105	3B 910421	(5.18 ± 0.18)E+0	11.86 ± 0.26	(3.37 ± 0.18)E−6	1.55 ± 0.14	(5.27 ± 0.25)E−6	A(pu),G,H
107	3B 910423	(2.09 ± 0.01)E+2	0.30 ± 0.11	(1.16 ± 0.23)E−7	1.04 ± 0.40	< 1.1E−6	U
108	3B 910424	(3.14 ± 0.59)E+0	0.44 ± 0.11	(4.69 ± 1.34)E−8	···	(1.52 ± 0.42)E−6	A(c)
109	3B 910425	(9.02 ± 0.03)E+1	3.62 ± 0.17	(2.50 ± 0.01)E−5	2.42 ± 0.02	(5.89 ± 0.09)E−5	A(cuw),B
110	3B 910425B	(4.30 ± 0.01)E+2	0.48 ± 0.11	(2.24 ± 0.15)E−6	3.80 ± 0.45	(3.09 ± 0.72)E−6	
111	3B 910426	(9.82 ± 0.23)E+1	0.65 ± 0.12	(3.91 ± 0.10)E−6	1.07 ± 0.05	(5.02 ± 0.45)E−6	
114	3B 910427	(3.81 ± 0.10)E+2	0.72 ± 0.12	(1.42 ± 0.17)E−6	1.03 ± 0.25	(2.08 ± 1.22)E−6	A(u)

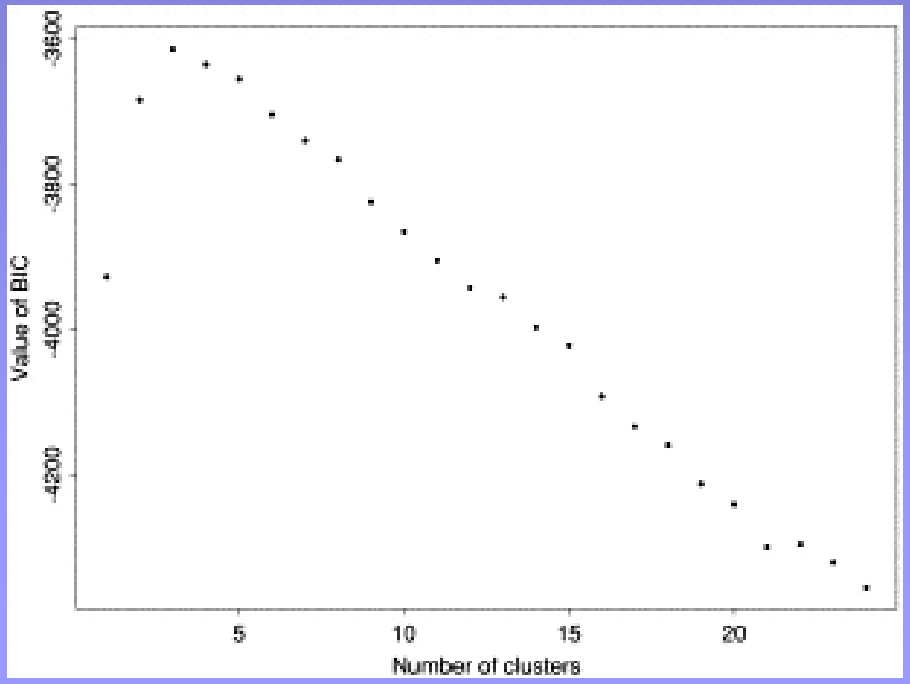

log T_50
log T_90
log F_tot
log P_256
log H_321
log H_32



Dendrogram from nonparametric agglomerative hierarchical clustering

Must choose sample, variables, unit-free (e.g. log), metric (e.g. Euclidean), merging procedure (e.g. Ward's criterion) & cluster level. The 3 cluster model is validated with MANOVA tests (e.g. Wilk's Λ and Pillai's trace) at P<<0.0001 level.

Mukherjee, Feigelson, Babu, Murtagh, Raftery, Fraley ApJ 1998



Parametric clustering using maximum-likelihood estimator (assuming multivariate normal) using EM Algorithm. Validated using Bayesian Information Criterion:

$$BIC = 2L - p \log N$$

Mukherjee et al. ApJ 1998

The three cluster model has been confirmed by several subsequent studies. The BATSE Team believes the intermediate class is due in part to biases produced by the on-board burst triggering algorithm.

Hakkila et al. ApJ 2003

Some methodological challenges for astrostatistics in the 2000s

- Simultaneous treatment of measurement errors and censoring (esp. multivariate)
- Statistical inference and visualization with very-large-N datasets too large for computer memories
- A user-friendly cookbook for construction of likelihoods & Bayesian computation of astronomical problems
- Links between astrophysical theory and wavelet coefficients (spatial & temporal)
- Rich families of time series models to treat accretion and explosive phenomena

Structural challenges for astrostatistics

Cross-training of astronomers & statisticians

New curriculum, summer workshops
Effective statistical consulting

Enthusiasm for astro-stat collaborative research

Recognition within communities & agencies
More funding (astrostat gets <0.1% of astro+stat)

Implementation software

StatCodes Web metasite (astrostatistics.psu.edu/statcodes)
Standardized in R & Inference (www.r-project.org)

Inreach & outreach

Center for Astrostatistics is designed to further these goals