Statistical Applications in the Astronomy Literature II

Jogesh Babu

Center for Astrostatistics

PennState University, USA

The likelihood ratio test (LRT) and the related F-test

Protassov et al. (2002, ApJ, 571, 545)

This paper by an astrostatistical group discusses how astronomers often misuse the likelihood ratio test (LRT) (known in X-ray astronomy as Cash's C-statistic) for comparing two models.

The standard chi-squared probabilities are incorrect to establish whether a faint additional feature (e.g. spectral line) is present within a simpler (e.g. continuum) model. Known to statisticians as *Nested models*

The LRT and the F-test, popularized in astrophysics by

- Eadie and coworkers in 1971
- Bevington in 1969
- Lampton, Margon & Bowyer in 1976
- Cash in 1979
- Avni in 1978

do not (even asymptotically) adhere to their nominal χ^2 and F-distributions in many statistical tests, thereby casting many marginal line or source detections and nondetections into doubt.

- It is common practice to use the LRT or the F-test to detect a line in a spectral model or a source above background despite the lack of certain required regularity conditions.

- In these and other settings that involve testing a hypothesis that is on the boundary of the parameter space, contrary to common practice, the nominal χ^2 distribution for the LRT or the F-distribution for the F-test should not be used.

- In this paper, the authors characterize an important class of problems in which the LRT and the F-test fail and illustrate this nonstandard behavior.

- There are numerous cases of the inappropriate use of the LRT and similar tests in the literature, bringing substantive scientific results into question.

To understand the difficulty with using these tests in this setting, begin with a formal statement of the asymptotic result that underlies the LRT.

Suppose($x_1; \ldots ; x_n$) is an independent sample (e.g., measured counts per PHA [pulse height analyzer] channel or counts per imaging pixel) from the common probability distribution $f(\mathbf{x}) = f(\mathbf{x}|\theta)$, with parameters $\theta=(\theta_1; \ldots ; \theta_p)$. We denote the

likelihood ratio statistic $T_{\mathrm{LRT}}(x) = -2 \log R(x)$, where

$$R(x) = \frac{\max \prod_{i=1}^{n} f(x_i|\theta_1^T, \ldots, \theta_q^T,\ \theta_{q+1}, \ldots, \theta_p)}{\max \prod_{i=1}^{n} f(x_i|\theta_1, \ldots, \theta_p)}, \quad (4)$$

where the maxima are found by varying the parameters. In

the numerator, the θ^T terms represent parameters that are not varied but held at their " true" values, i.e., the values assumed under the null model. Under some regularity conditions, if $(\theta_1; \ldots ; \theta_p)$ actually equals $(\theta^T_1; \ldots ; \theta^T_p)$, the distribution of the LRT statistic converges to a Chi-square distribution with q degrees of freedom as the sample size n increases without bound. The degrees of freedom is the difference between the number of free parameters specified.

STATISTICS: HANDLE WITH

- Although the LRT is a valuable statistical tool with many astrophysical applications, it is not a universal solution for model selection.

- When testing a model on the boundary of the parameter space (e.g., testing for a spectral line), the (asymptotic) distribution of the LRT statistic is unknown.

- Using this LRT and its nominal Chi-square distribution can lead to unpredictable results (e.g., false positive rates varying from 1.5% to 31.5% in the nominal 5% false positive rate test in Monte Carlo studies).

- There is no replacement for an appreciation of the subtleties involved in any statistical method.

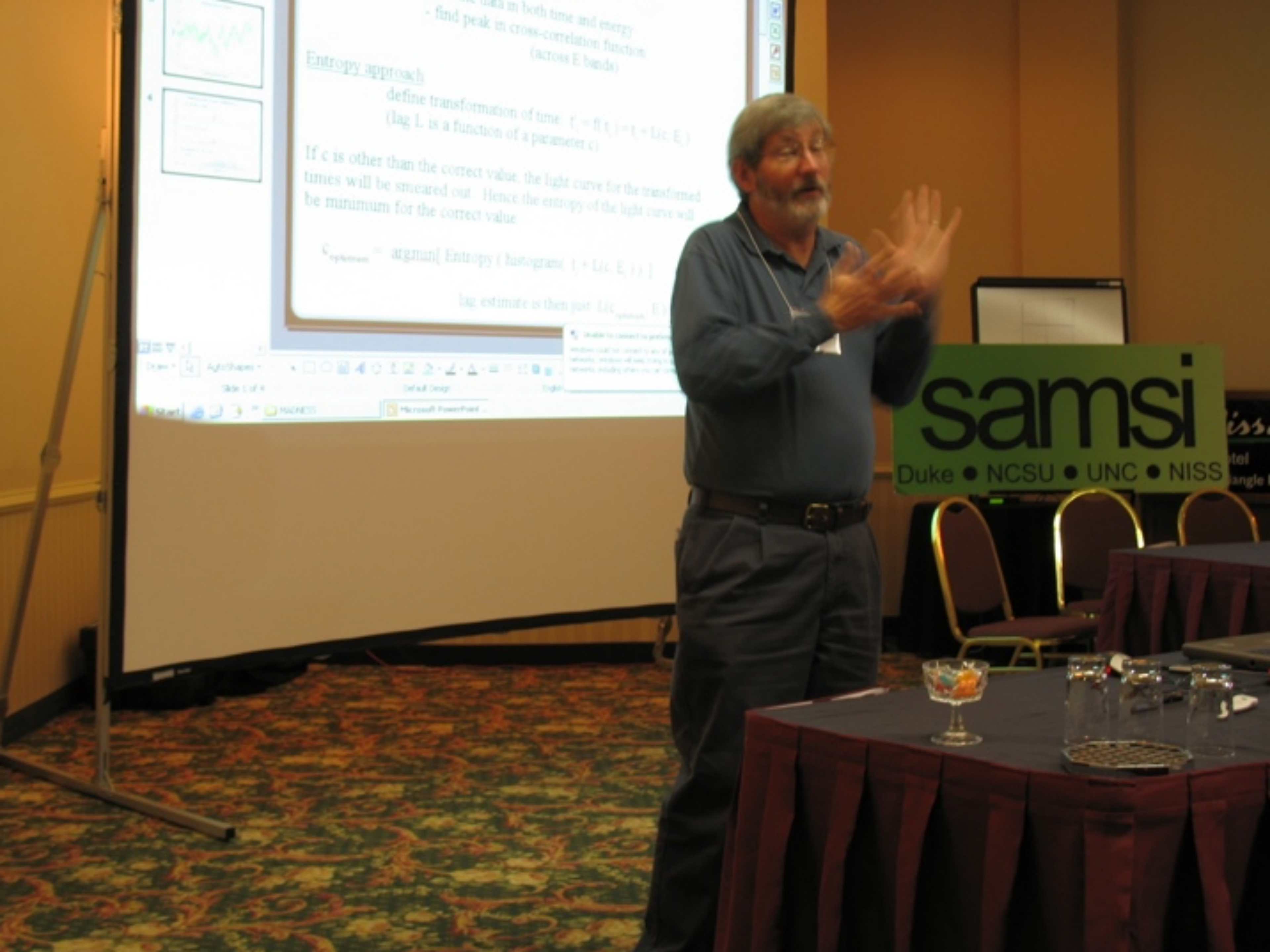
- Practitioners of statistics are forever searching for statistical ‘‘black boxes”.

- For sophisticated models that are common in spectral, spatial, as well as other applications in astrophysics, such black boxes simply do not exist.

- The highly hierarchical structures inherent in the data must be, at some level, reflected in the statistical model.

Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data

- Detection of periodic signal hidden in noise
- This paper studies the reliability and efficiency of detection with the periodogram, when the observation times are unevenly spaced.
- A modification of the classical definition of periodogram is suggested.
- With this modification, periodogram


- find peak in cross-correlation function
(across E bands)
Entropy approach
define transformation of time:
(lag L is a function of a parameter c)
If c is other than the correct value, the light curve for the transformed
times will be smeared out. Hence the entropy of the light curve will
be minimum for the correct value
samsi
Duke • NCSU • UNC • NISS

- A physical variable X is measured at a set of times t_j, the resulting time series data $\{X(t_j), j= 1,2, ..., N\}$, are sum of signal and random observational errors (noise): $X_j= X(t_j)= X_s(t_j)+ R(t_j)$
- The signal X_s is taken to be strictly periodic, $R(t_j)$ are i.i.d. normal with mean 0 & variance 1.

The Periodogram

- The basic tool of spectral analysis is the discrete Fourier transform (DFT)

$$FT_X(\omega) = \sum^{N}_{j=1} X(t_j) \exp(-i\omega t_j)$$

- The classical periodogram is

$P_X(\omega) = (1/N)\ |FT_X(\omega)|^2$

$= (1/N)[(\sum_j X(t_j) \cos \omega t_j)^2 + (\sum_j X(t_j) \sin \omega t_j)^2]$ (1)

- This expression is traditionally evaluated at $M=N/2$.

- If X contains a sinusoidal component of frequency ω_0 then P would be large at and near $\omega = \omega_0$

- If the observational times are evenly spaced, say $t_j = j$,

  $$P_X(\omega) = (1/N)\left|\Sigma_j X(t_j) \exp(-ij\,\omega)\right|^2$$

  (2)

- Proposed version

  $$P_X(\omega) = (1/2)\left[\frac{(\Sigma_j X(t_j)\cos\omega t_j)^2}{\Sigma_j \cos^2\omega t_j} + \frac{(\Sigma_j X(t_j)\sin\omega t_j)^2}{\Sigma_j \sin^2\omega t_j}\right]$$

- This does not have exponential distribution unless,

- Invariance to time translation is a useful property possessed by the classical periodogram, i.e., it is unchanged if t_j is changed to $T+t_j$ for every j.

- There are many ways to restore invariance.

$$P_X(\omega) = (1/2) \frac{[\Sigma_j X(t_j) \cos \omega(t_j-\tau)]^2}{\Sigma_j \cos^2 \omega (t_j-\tau)} + \frac{[\Sigma_j X(t_j) \sin \omega (t_j-\tau)]^2}{\Sigma_j \sin^2 \omega (t_j-\tau)},$$

where

$$\tau = (1/2\omega) \arctan \frac{\Sigma_j \sin 2\omega t_j}{\Sigma_j \cos 2\omega t_j}$$

- Unevenly-sampled signals: a general formalism of the Lomb-Scargle periodogram

R. Vio1, P. Andreani, and A. Biggs

A&A preprint doi
http://dx.doi.org10.1051/0004-6361/201014079

The authors suggest methods to handle data with non-Gaussian errors.

Hou et al. (2009 ApJ 702, 1199)

This is a nice and straightforward study of nonparametric goodness-of-fit tests in the context of a problem in extragalactic astronomy, It shows that the poorly-known Anderson-Darling test performs better than the popular Kolmogorov-Smirnov test. However, they still make an error by not noting that the tabulated probabilities are incorrect for models with parameters derived from the same dataset.

These issues were discussed in Babu & Feigelson (2006 ASPC, 351, 127).

Hinshaw et al. (2003 ApJ S, 148, 35 & Genovese et al. (2004)
http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.ss/1105714165

Here are two contrasting ways to fit functions derived from astrophysical theory to a dataset. The problem is estimation of the LambdaCDM cosmological model parameters from WMAP cosmic microwave background radiation fluctuations.

Hinshaw et al. (Appendix A) give a classical maximum likelihood estimation approach with error analysis from the Fisher Information Matrix.

Genovese et al. are statisticians who use new techniques of `semi-parametric regression' to estimate the curve (and its uncertainty) without assuming a cosmological model.

A brief discussion, of when parametric vs. semi-parametric modeling is appropriate, is given by Feigelson (2007) http://adsabs.harvard.edu/abs/2007ASPC..371..280F).

Hundreds of astronomical papers seek the slope of a power-law distribution of an astronomical properties in a sample by binning the data and computing the least-squares regression line. This is done for cosmic ray energy spectra, stellar Initial Mass Functions, interstellar molecular cloud clump functions, galaxy luminosity function, and so forth. Yet, the method gives both biased and inaccurate results, even for large samples. A very simple maximum likelihood estimator has been known since 1957 with excellent performance that does not require binning. Here are a few papers on the issue.

http://adsabs.harvard.edu/abs/1970ApJ...162..405C
http://adsabs.harvard.edu/abs/2004EPJB...41..255G
http://adsabs.harvard.edu/abs/2007EPJB...58..167B

http://en.wikipedia.org/wiki/
Binomial_proportion_confidence_interval
http://www.projecteuclid.org/DPubS?
service=UI&version=1.0&verb=Display&handle=euclid.ss/
1009213286
http://adsabs.harvard.edu/abs/2006ApJ...652..610P

Astronomers sometimes want to compute the ratios of two
Poisson-distributed quantities. This might be a signal-to-noise ratio or a hardness ratio in fields such as X-ray or gamma-ray astronomy. The ratio of two Poissons (known in statistics as the `binomial proportion problem') is surprisingly tricky; for example, the maximum likelihood estimator is unstable and chaotic. Several analytical solutions are discussed but none is obviously best. When background is subtracted from the numerator and/or denominator, the problem requires numerical