THE ANNOTATED SOURCE DETECTION

2^{nd} X-Ray Astronomy School, Berkeley Springs, WV
2002 August 21
Vinay Kashyap & Peter Freeman

Overview

Why source detection?

⋆ why is it so difficult in X-rays to know that there is a source? consider that for the same flux as at, say, 5000 Å, there are a thousand fewer photons at 5 Å; not to mention that for most objects – e.g., stars – the X-ray flux is lower than the optical by many orders of magnitude, leaving us scrounging for photons in the Poisson regime
[HST image of 47 Tuc]
[Chandra ACIS image of central portion of 47 Tuc]
⋆ dealing only with source detection, not image reconstruction, not image deconvolution, nor modeling

How to detect sources

- In Theory
  ⋆ Peter Freeman speaks on the theory of source detection:
  - [Detection Theory: the Short Form]
  - [The Five-Fold Path]
  - [Potholes on the Five-Fold Path]
  - [Classic Detection: CELLDETECT]
  - [Figure of source and background cells]
  - [New Detection: WAVDETECT]
  - [Figure of MexHat wavelet and negative annulus]
- In Practice
  [Source Detection in the Real World: Snares and Traps]

Dealing with detected sources

- Hardness Ratios
  [Hardness Ratios]
- $\log N - \log S$
  [Biases in the construction of $\log N - \log S$ curves]

Upper limits

⋆ there is a difference between the confidence range limits derived on the source strength parameter and asking the question "at what count rate would the source be detected?"
⋆ For example, consider a source cell with 5 counts and a background cell of 100 times larger area with 200 counts. Then the background estimate in the source cell is $B = 2.0 \pm 0.141$, and the estimated source strength is $S = 3 \pm 2.24$ (assuming simple root-N errors). A "3σ" upper limit could be set based on the measured error, i.e., $S < 9.7$, i.e., 12 counts in the source cell are required for a 3σ detection.

On the other hand, if a hypothesis test is performed by computing the probability that a background of 2 counts can generate $> N$ counts, then the putative 3σ limit is reached at 6.5
$\star$ Of course, note that the former method is not internally self-consistent, i.e., if 12 counts are indeed observed in the source cell, that would be still only be classified as a 2.9σ source
$\star$ The difference is due to the fact that it is not necessary to measure the flux of a source to determine its existence or lack thereof
$\star$ A Bayesian approach would consider the relative probabilities of two models, one with only the background and one with a source and a background

Source Detection in the Real World: Snares and Traps

1. The Starter Kit
   $\star$ these are basic issues without which source detection is impossible, and ones that the analysis software will work out for you, if not in canned routines, then at least in threads

   - Aspect dither
     [Capella HRC-S/LETG 0^{th}-order image, pre-aspect correction v/s post-aspect correction]
   - Bad pixels
     [Capella ACIS-S/HETG dispersed region, showing the hot pixels as points in detector coordinates, and dithered squares in sky coordinates]
   - Exposure maps
     $\star$ imperative that instrument effects are correctly accounted for, otherwise the detection algorithm will find these structures in the data and call them sources
     $\star$ this occurs because of incorrect background determination at the edges and at support structure shadows, etc. – an exposure map tells the algorithm to *expect* a differect background
     [ROSAT/PSPC exposure map]
     [ACIS-I/ACIS-S data showing the effects of including the exposure map – fewer sources at the edge – v/s not including one – lots of sources at the edges]
     $\star$ the changing energy sensitivity of the exposure maps results in the remaining few false sources along the boundaries of the BI chips

2. Here be Hippogrif
   $\star$ these are problems that are by and large taken care of by the algorithms, but they are also traditionally where source detection algorithms tend to break down, and results must be double checked

   - Point Spread Functions
     [HZ 43 at $10'$ off-axis]
   - Varying background
     [diffuse background in the Pleiades, with shadow due to high-density molecular cloud]
     $\star$ this is a wavdetect product, notice that the image has been flat-fielded, i.e., exposure map effects have been taken into account
   - Extended Sources
     [Galaxy cluster MS1137 showing the effects of an extended source on background determination]
     $\star$ not as bad as it looks, the dynamic range in the source image is in the 1000's, while in the background image it is less than $2\times$
   - Overlapping Sources
     [Sirius A and B]

3. What, me worry?
   $\star$ these are problems that the source detection algorithms generally do not consider in any detail, and also do not cause much of a problem in the general run of things, but have the potential to have a large impact in certain cases

   - Source Position and plate scale
     $\star$ at large off-axis locations and for weak sources, source position determination may be

contaminated by background photons
$\star$ when large numbers of sources have been detected in the field and identified with counterparts, it is always a good idea to double check the plate scale

- Source spectrum and ExpMaps and PSFs
[exposure map showing energy dependence, courtesy Jonathan McDowell]
$\star$ an exposure map that has the correct energy dependence will take into account the differences between the ACIS-S FI and BI chip responses, so that the algorithms will know that the change in background level across the chip boundaries is not real
- Pileup
[ACIS-S/LETG 0^{th} order image of XTE J1118+48]
- Detection Sensitivity
$\star$ Type I Errors, aka false positives, or sources that are detected even if there aren't real
- Detection Probability
$\star$ Type II Errors, aka false negatives, or real sources that are not detected because of count fluctuations
$\star$ Type I and Type II errors must be properly accounted for when computing $\log N - \log S$ curves

Hardness Ratios
(Not as simple as they seem)

Useful with large samples and when spectral fits are unfeasible

Types of hardness ratios:

- $R = \frac{S}{H}$, $\sigma_R = R\sqrt{\left(\frac{\sigma_S}{S}\right)^2 + \left(\frac{\sigma_H}{H}\right)^2}$
  $\star$ A simple ratio of counts in two passbands, Soft and Hard; range is $R \in [0, \infty]$
- $C = \log_a(S) - \log_a(H)$, $\sigma_C = \ln(a)\sqrt{\left(\frac{\sigma_S}{S}\right)^2 + \left(\frac{\sigma_H}{H}\right)^2}$
  $\star$ the equivalent of color, such as $B - V$; range is $C \in [-\infty, +\infty]$
- $HR = \frac{H-S}{H+S}$, $\sigma_{HR} = \frac{2}{(H+S)^2}\sqrt{H^2\sigma_S^2 + S^2\sigma_H^2}$
  $\star$ the fractional difference, often used in extragalactic astronomy because of the better behavior of the denominator; range is $HR \in [-1, 1]$

Caveats

- Mathematical obstinacy
  $\star$ formal error propagation is mathematically invalid for both R and HR, which result in Cauchy distributions for which a mean cannot be defined and the variance is infinite; this is not a problem that goes away with "better statistics", in fact it gets worse
  $\star$ use the log form, C, which is well behaved, as much as possible
- Poisson statistics
  $\star$ Better to use the full Poisson likelihoods to determine the errors, especially since the ratios behave very unintuitively in the Poisson regime – for instance, for small values of H, the mean value of R increases, but the mode decreases
- Background
  $\star$ Poisson formulations can correctly take background into account, but for regular error propagation just square-add the errors
- Upper limits
  $\star$ again, Poisson formulations can deal naturally with the cases where one or both of S and H are consistent with 0. DO NOT use simplistic constructs like using the 1σ limit to compute the ratios

$\log N - \log S$

Say $n(r) = n_0$, $f(L_x) = \delta(L_x - L_{x0})$, $S = \frac{L_{x0}}{4\pi r^2}$. Number of sources within r, $N(<r) = \frac{4\pi}{3} r^3 n_0$, or $N(>S) = n_0 \frac{4\pi}{3} \left(\frac{L_{x0}}{4\pi S}\right)^{\frac{3}{2}}$, i.e., $N(>S) \propto S^{-\frac{3}{2}}$. In general,

$$N(>S, l, b) = d\Omega \int_0^\infty dL_x \int_0^\infty dL'_x f(L'_x) \sigma(L_x, L'_x) \int_0^{r'} n(r, l, b) r^2 dr$$

where r' is implicitly defined by $S = \frac{L'_x}{4\pi r'^2} e^{-\tau(r')}$, and $\sigma(L_x, L'_x)$ is a function that takes into account statistical fluctuations in the intrinsic luminosity.

Biases:

- Source confusion
  ⋆ in crowded fields, detection algorithms tend to miss weaker sources near strong sources, and sometimes they merge multiple sources into a single one
- False sources (false positives)
  ⋆ false sources that arise due to fluctuations in the background will contaminate the numbers at the low flux end
- Lost sources (false negatives)
  ⋆ because of statistical fluctuations, sources of a given intrinsic strength will produce different numbers of counts at any given observation, and as the source strength becomes smaller, the chances of it not being detected increase even if nominally it is above the detection threshold
- Malmquist Bias
  ⋆ the volume in which high-luminosity sources can be detected is larger than the volume in which low-luminosity sources are detected; thus luminous objects will be overrepresented in flux-limited samples
- Faint source fluctuations
  ⋆ when there are larger numbers of low-flux sources than high-flux sources, then statistical fluctuations result in a larger number of the weaker sources deflected into higher flux regimes than stronger sources that are deflected into lower flux regimes
- Eddington Bias
  ⋆ when fluxes of sources with intensities near the detection threshold are measured, there is a tendency for the average measured flux to be higher than the true fluxes, because of the fact that fluctuations towards smaller counts will be censored out because of the detection threshold

Construction:

- Sky coverage
  ⋆ construct the area of the field corresponding to a given sensitivity and correct the measured $N(>S)$ v/s S
- Modeling
  ⋆ recent efforts have concentrated on developing a method for *modeling* $f(L_x)$ to match the observed $\log N - \log S$