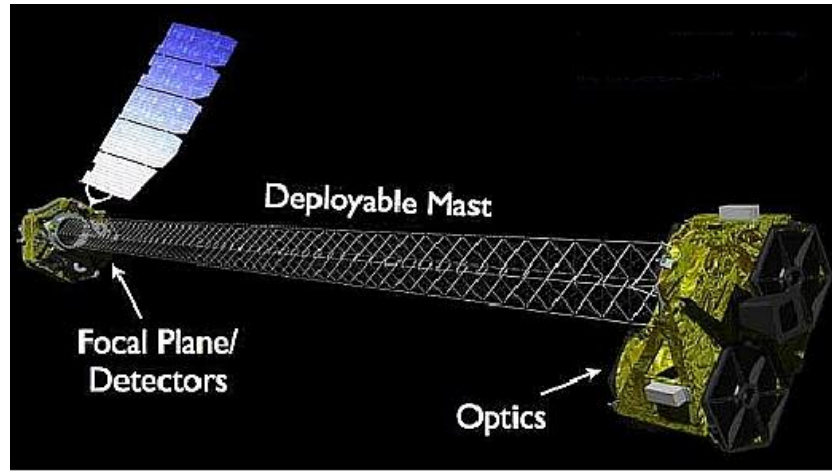
STATISTICAL METHODS FOR THE PHYSICAL SCIENCES

Week 7 tutorial: Fitting data and hypothesis testing in practice

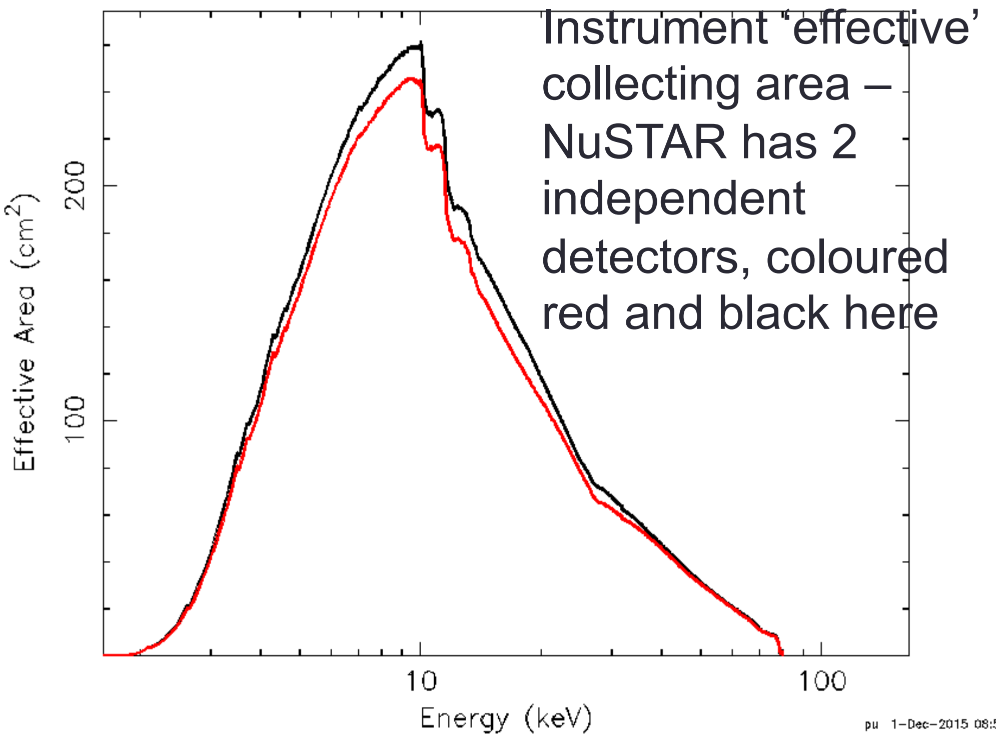
For a nice review of the likelihood ratio and important caveats for its use:
"Statistics, handle with care: detecting multiple model components with the likelihood ratio test"
by Protassov et al., 2002, The Astrophysical Journal, 571, p545-559.

NuSTAR spectra from an unusual black hole

The Nuclear Spectroscopic Telescope Array (NuSTAR) was launched in June 2012 and is the first hard (high-energy) X-ray telescope to use highly sensitive focussing (grazing incidence) optics



Swift J1753.5-0127 is a black hole accreting gas from a stellar companion, around ~10^4 light years from us. In March 2015 it entered a very unusual never-before-seen ‘low state’ where the emission in hard X-rays (>4 keV) dropped away, while at lower energies it remained bright. NuSTAR and the XMM-Newton observatory were tasked with finding out what had happened…



Instrument ‘effective’ collecting area – NuSTAR has 2 independent detectors, coloured red and black here

X-ray spectral fitting with Xspec

- Xspec is the most popular software package for fitting X-ray data ~30 years old and currently in v12.9
- Developed and supported by NASA, so it sticks around…
- Old-fashioned interface but many adaptations under the hood.
- Default fitting approach is Levenberg-Marquardt (same as curve_fit in scipy.stats), so works with chi-squared
- But other max-likelihood approaches possible and Monte Carlo methods available for trickier cases

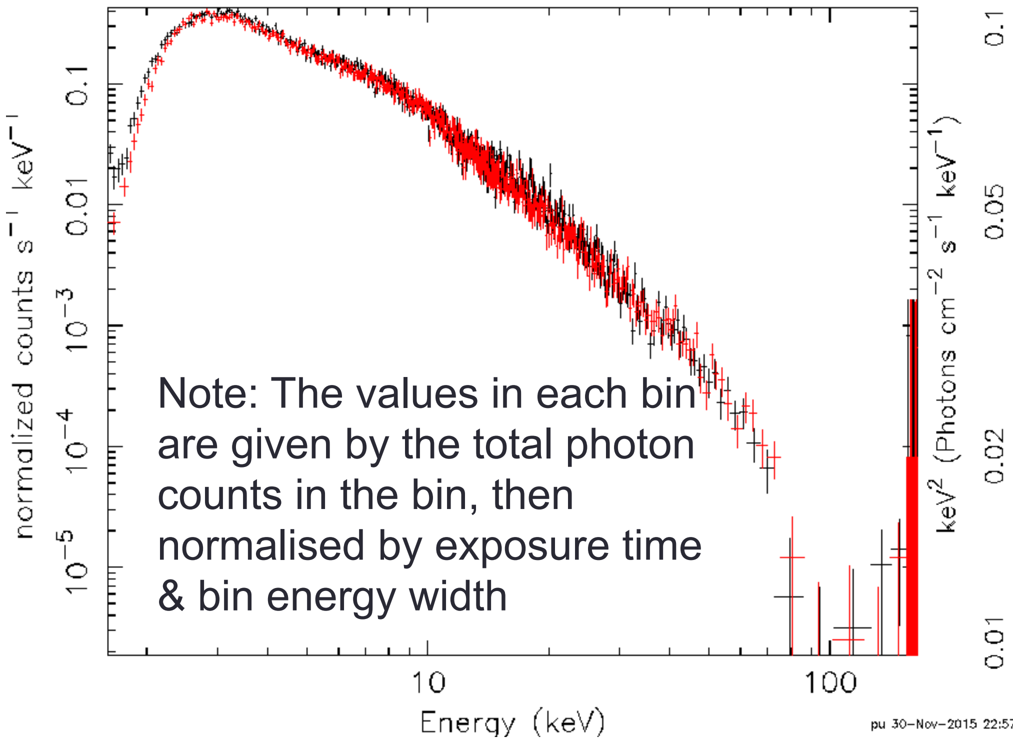
pietervlasurvey — xspec — 82×42

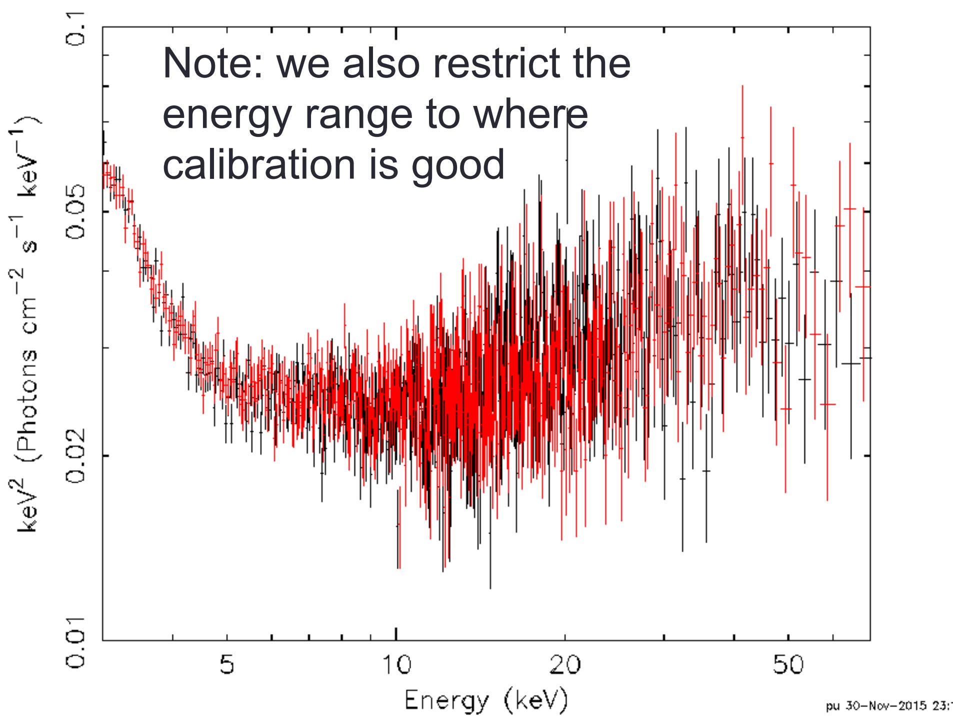
```
 Current data and model not fit yet.
XSPEC12>fit
                                   Parameters
Chi-Squared  |beta|/N    Lvl    1:PhoIndex        2:norm
1055.36      6.91216e-11  -3       2.00000     0.0267959
1044.52      154.922      -4       1.96678     0.0248310
1044.05      2848.53      -5       1.96393     0.0247427
1044.04      4.40173      -6       1.96369     0.0247297
==============================
 Variances and Principal Axes
                 1        2
 1.0458E-04| -0.9985  -0.0549
 1.1846E-08|  0.0549  -0.9985
------------------------------

========================
  Covariance Matrix
        1           2
   1.043e-04   5.729e-06
   5.729e-06   3.267e-07
------------------------

========================================================================
Model powerlaw<1> Source No.: 1   Active/On
Model Model Component  Parameter  Unit     Value
 par  comp
                            Data group: 1
   1    1   powerlaw   PhoIndex            1.96369      +/-  1.02112E-02
   2    1   powerlaw   norm                2.47297E-02  +/-  5.71590E-04
                            Data group: 2
   3    1   powerlaw   PhoIndex            1.96369      = 1
   4    1   powerlaw   norm                2.47297E-02  = 2
________________________________________________________________________


Fit statistic : Chi-Squared =         1044.04 using 834 PHA bins.

Test statistic : Chi-Squared =         1044.04 using 834 PHA bins.
 Reduced chi-squared =        1.25486 for    832 degrees of freedom
 Null hypothesis probability =   6.878935e-07
XSPEC12>plo ldata
XSPEC12>plo ratio
```

https://heasarc.gsfc.nasa.gov/xanadu/xspec/

1. Take a look at the data

Before a statistical analysis we have to follow some standard procedures to clean the data, also making a 'background spectrum' which is subtracted from all the spectra shown here. We then look at the data in xspec…



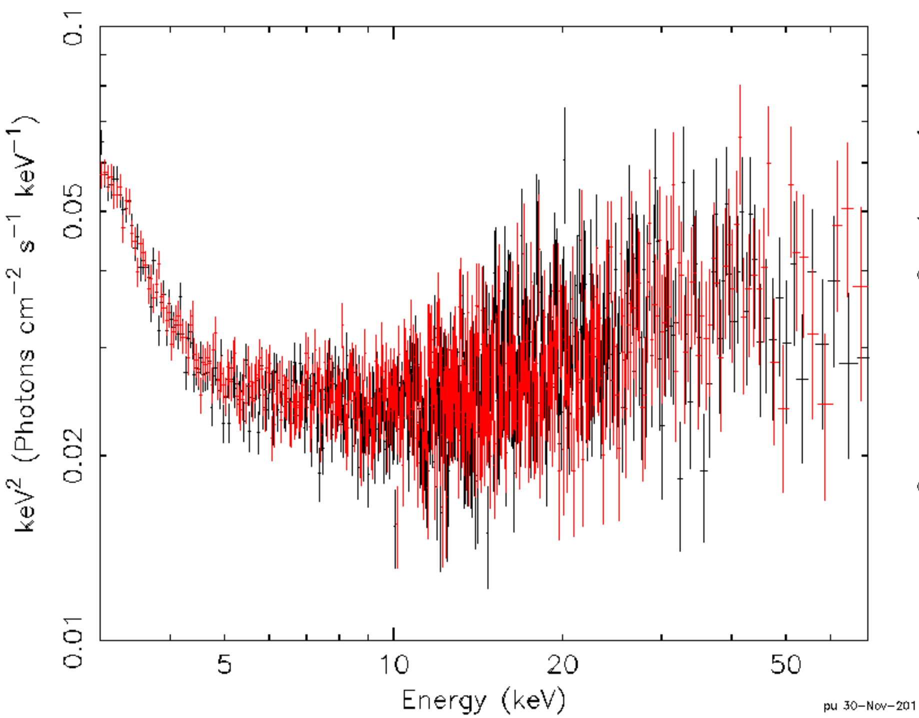


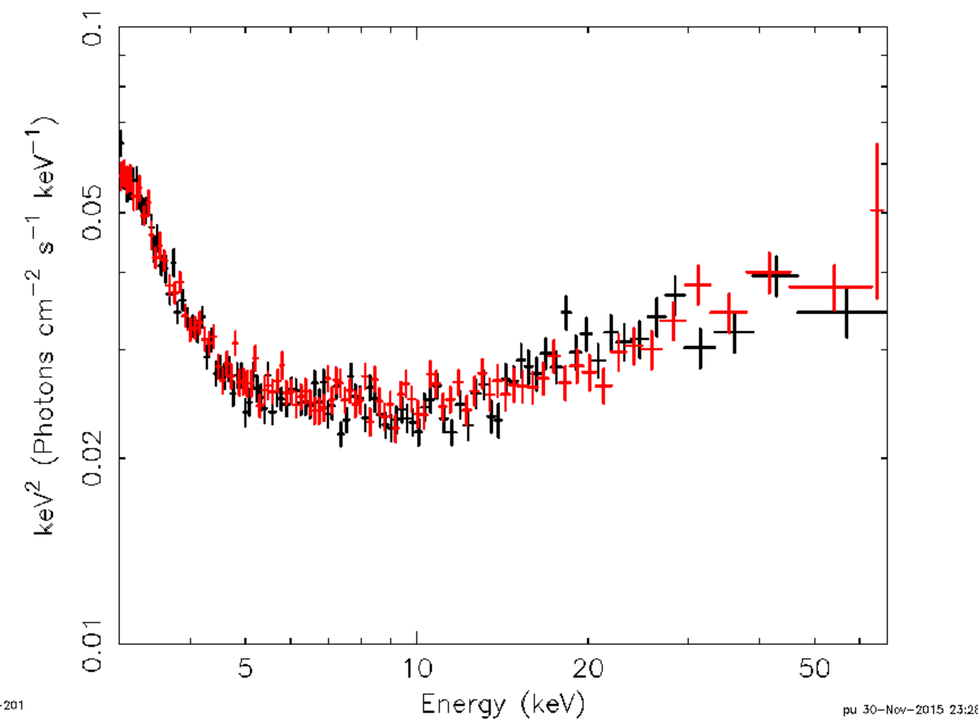
The raw spectrum does not make much physical sense since the underlying 'true' spectrum is affected by the 'response' - the effective collecting area - of the instrument.

The 'unfolded' spectrum is more physical: above 5 keV it looks power-law like. Below 5 keV there is a clear upturn.

Some words about binning

The error bars on each spectral bin are determined by Poisson counting statistics: they are sqrt(n) where n is the number of counts in each bin.





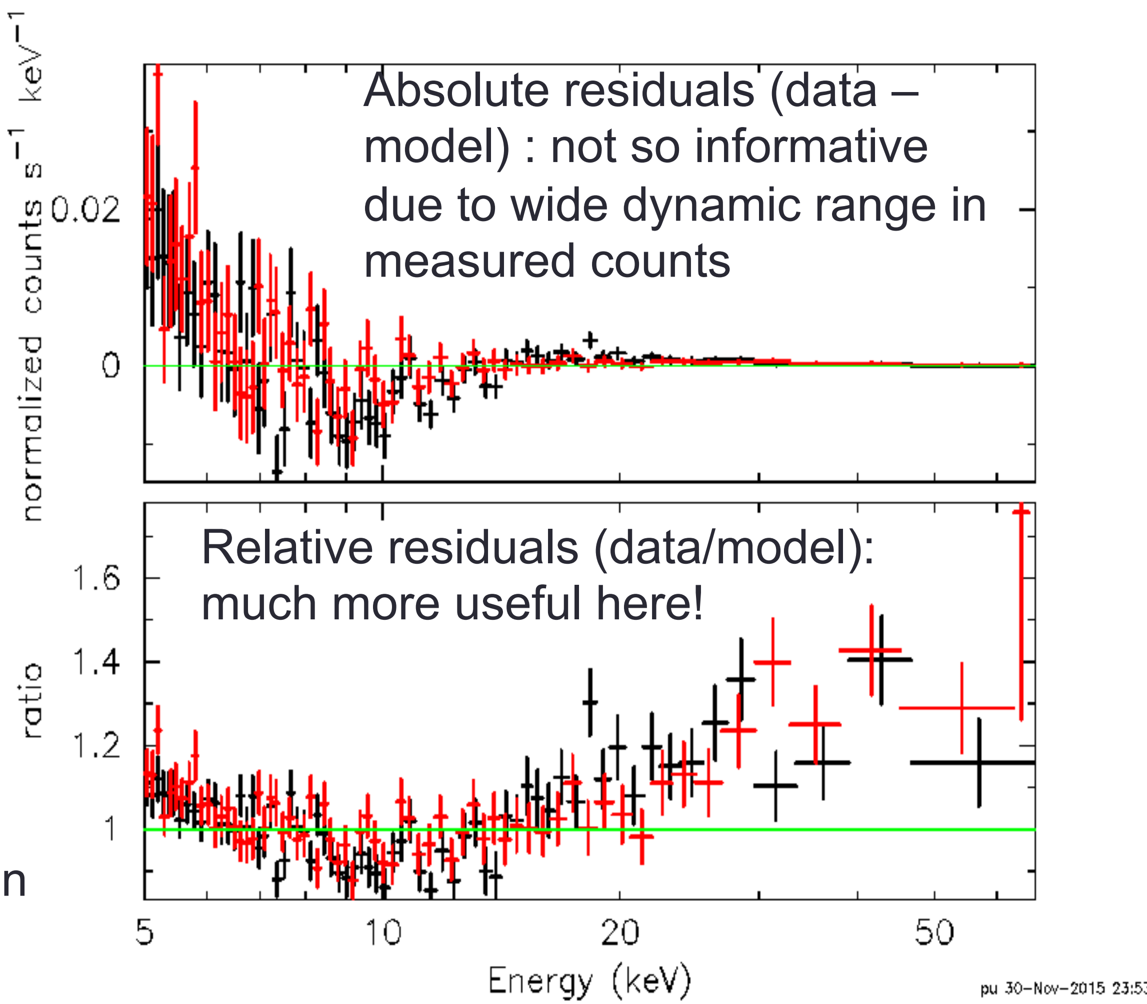
To ensure errors are ***close to normally distributed***, bins are combined so there are at least 20 photons per bin (left). The resulting spectrum is still noisy, so we can bin up even further (right). The main advantage is aesthetic and so that we can see features in our data more clearly: ***it shouldn't change the statistics.***

2. A simple fit to the data > 5 keV: power-law model

To keep things simple, we first try fitting a single power-law to the data above 5 keV. (Note: technically, xspec folds the model through the instrument response and compares directly with the raw spectrum, but the effect is the same as direct fitting to the unfolded spectrum)

Not surprisingly, this is a bad fit:
Chi-sq = 1044 for 832 d.o.f.
(834 bins and 2 free parameters)
p-value = 7e-7

To figure out where the fit has gone wrong and ensure it isn't in a stuck in a bad part of the parameter space (i.e. it still broadly fits the continuum shape) we can look at the residuals…

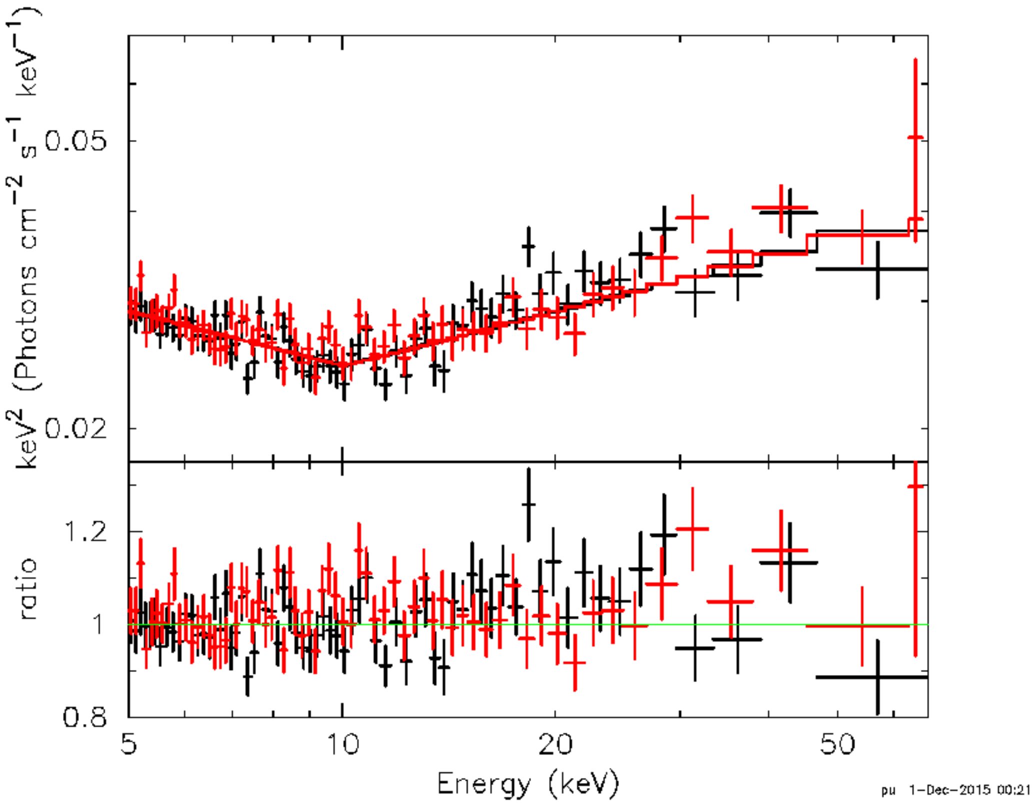


3. Try a more complex model: broken power-law

There is no point in estimating errors for a bad fit (they don't mean anything), so we move on to fitting a more suitable model: the data and residuals clearly suggest a more complex inflected shape. Try a broken power-law?

Success!! The model fits well!
Chi-sq=832.4 for 830 d.o.f.
p-value = 0.47

Also, no obvious systematic features are apparent in the residuals.



Whether we take this successful model seriously or not depends on what our aims are. Do we have a physical expectation for such a model? Are we unsure about the physics and simply want an empirical parameterisation of the spectral shape? Other models may fit equally well, and be more physically motivated.

4. Parameters and confidence intervals 1

Let's assume we will just accept the broken power-law model for now. What about the best-fitting parameter values (the MLEs) and their confidence intervals?

xspec gives the square-roots of the diagonal elements of the covariance matrix (inverse Hessian) as first-order error estimates:

Index 1 (below break): 2.25+/-0.03
Break energy: 10.05+/-0.35
Index 2 (above break): 1.75+/-0.02
Normalisation (at 1 keV): 0.0439 +/- 0.0024

Care must be taken when deciding how many significant figures to use

We can also use brute force to determine precise confidence intervals:

Recall the log likelihood-ratio approach:

$$\Lambda = -2\ln\frac{l_H(\mathbf{x})}{l_A(\mathbf{x})} = \underbrace{-2L_H(\mathbf{x}) + 2L_A(\mathbf{x})} \sim \chi^2_k$$

This is the difference in chisq (delta-chisq) between the null and the alternative

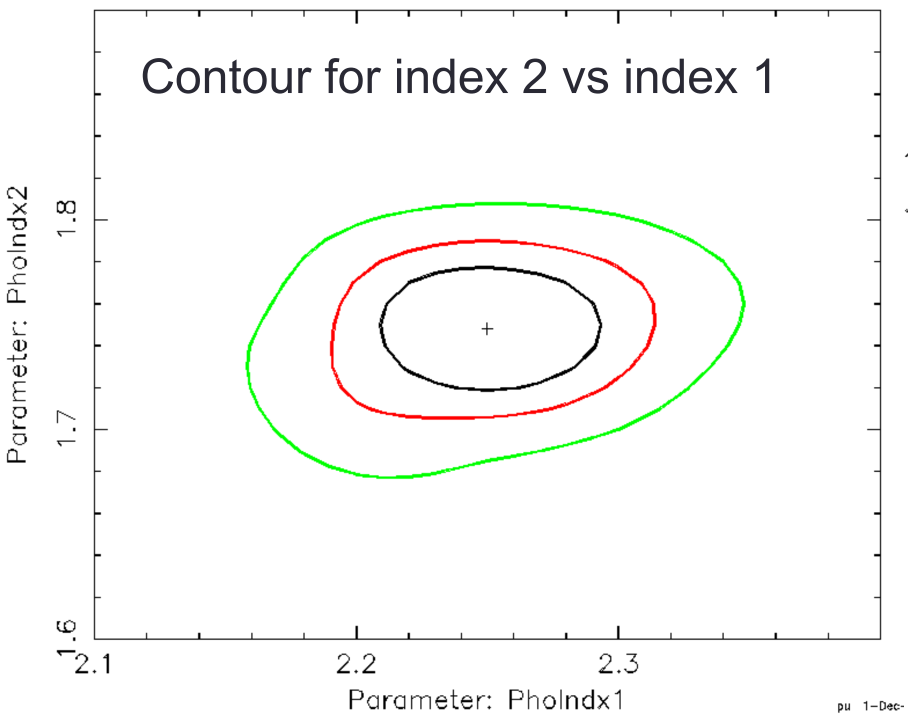
where *k* is the number of additional constrained parameters in the null hypothesis

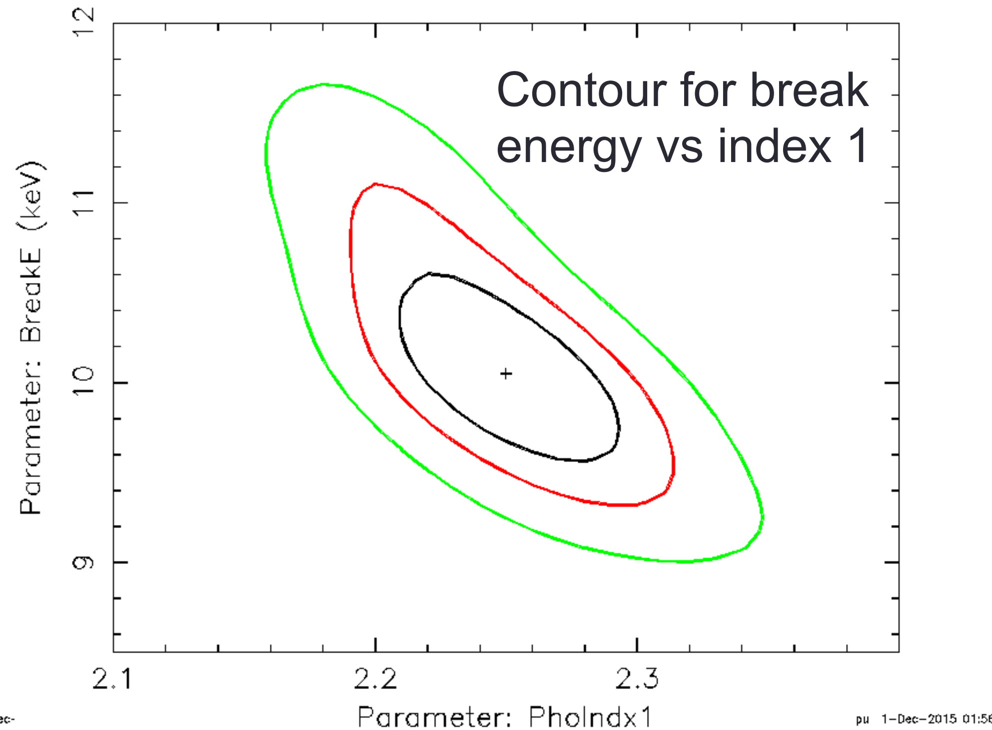
$$\int_c^\infty P(\Lambda|H)d\Lambda = \alpha$$

Thus we need to find the corresponding value of χ^2_k for a given significance α (i.e. 1 – coverage)

4. Parameters and confidence intervals 2

Our 2-parameter contours mean we are comparing two extra constraints on the null with the best-fitting chi-square (the MLEs, marked by the crosses). Thus we use χ^2_2 with contours corresponding to delta-chisq of 2.3 (68.4%), 4.61 (90%) and 9.21 (99%).





Note the covariant errors for one pair but not the other… Compare also with the single-parameter estimates (+/-0.35 for the break, 0.03 and 0.02 for index 1 and 2 respectively)

5. Hypothesis test: do both NuSTAR detectors show the same spectral shape?

Now lets go beyond using log likelihood ratio for confidence interval determination and try a simple hypothesis test. Are the 2 detectors consistent with having the same spectral shape, or could there be cross-calibration issues?

So far we have forced the parameters for models fitted to both spectra to be the same (the parameters are ***tied*** between the two spectra but are free to vary together). Now we successively free up (***untie***) parameters and see what happens:

All parameters tied: chisq = 832.4 for 830 d.o.f.
Untie break energy: chisq = 832.0 for 829 d.o.f. ➔ not signifiant
Untie index 1: chisq = 826.3 for 828 d.o.f. ➔delta-chisq = 5.7 for 1 extra free parameter ➔ p-value 0.02, ***could be significant***
Untie index 2: chisq = 825.6 for 827 d.o.f. ➔ not signifiant
Untie normalisation: chisq = 825.3 for 826 d.o.f. ➔ not signifiant

Our conclusion is that there is some (weak) evidence that the spectra are different at the lowest energies.

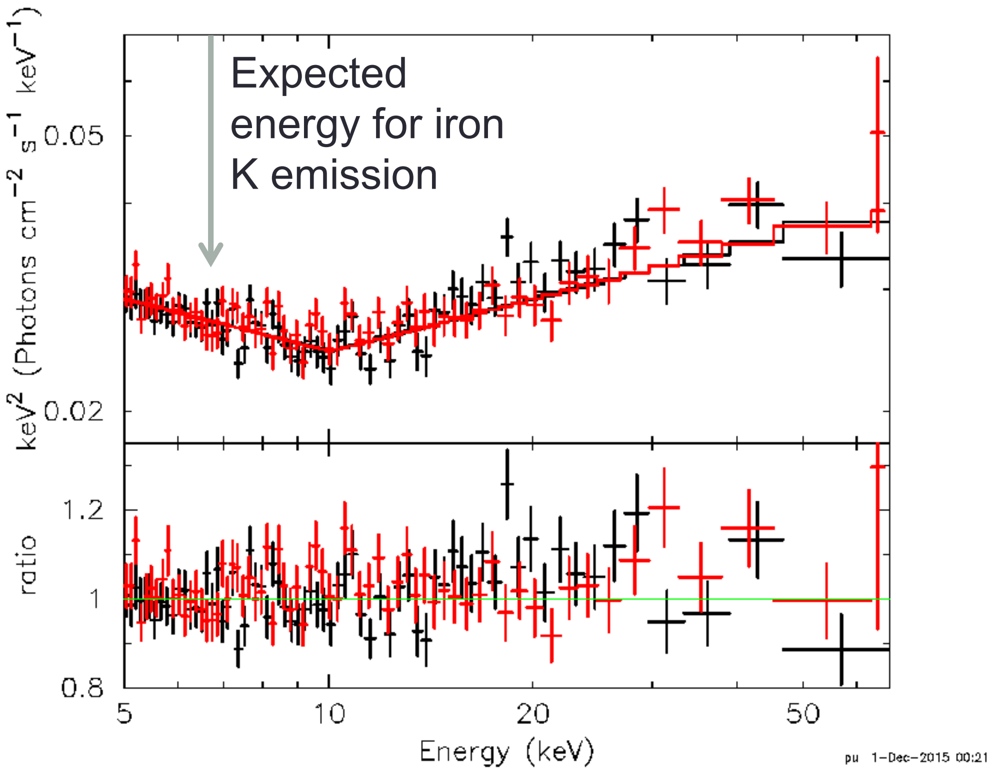
6. Searching for discrete signals: emission lines

The evidence for spectral differences between detectors was marginal, so here we continue to tie the fits together.

Most accreting black holes show detectable emission lines from Fe fluorescent emission, sometimes quite broad, sometimes narrow.

It is both surprising and intriguing that no line is apparent here. Is a line not required? What is the upper limit on line strength?

First add narrow (unresolved) line at 6.4 keV and fit: chisq = 831.4 for 829 d.o.f. (not significant, only 68.4% confidence)



We can also now vary the line normalisation to find the delta-chisq for a chosen significance and thus set an upper limit: we choose 3-sigma, delta-chisq for 1 parameter (normalisation) = 8.81 ➔ upper limit is ~1.3e-5 photons/cm^2/s, EW ~ 20 eV, very weak!

6. Searching for discrete signals: number of trials

An emission line can also be detected using a likelihood ratio hypothesis testing approach

The concerns of Protassov et al. (null is at a boundary, i.e. zero flux, can be avoided by allowing the fits to go negative – i.e. allow for absorption lines)

But unless one knows the energy of the line a priori, one must also correct the significance for the number of trials (i.e. line widths in the energy range searched over).