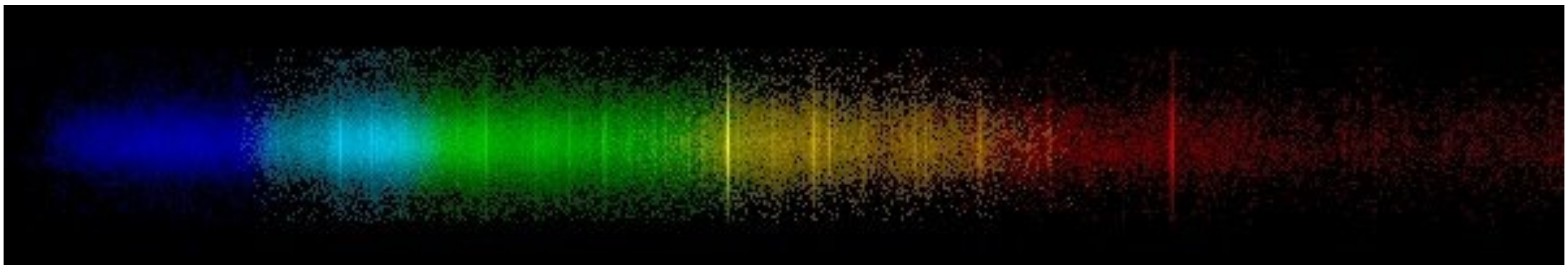
Statistical Challenges in Modeling High Resolution X-ray Spectra

Aneta Siemiginowska
Vinay Kashyap
(CfA/CXC/CHASC)

Challenges

- Global fitting
- Line detection and fitting low counts lines
- Massive model misspecification
- Building complex models

Lack of Pretty Pictures?

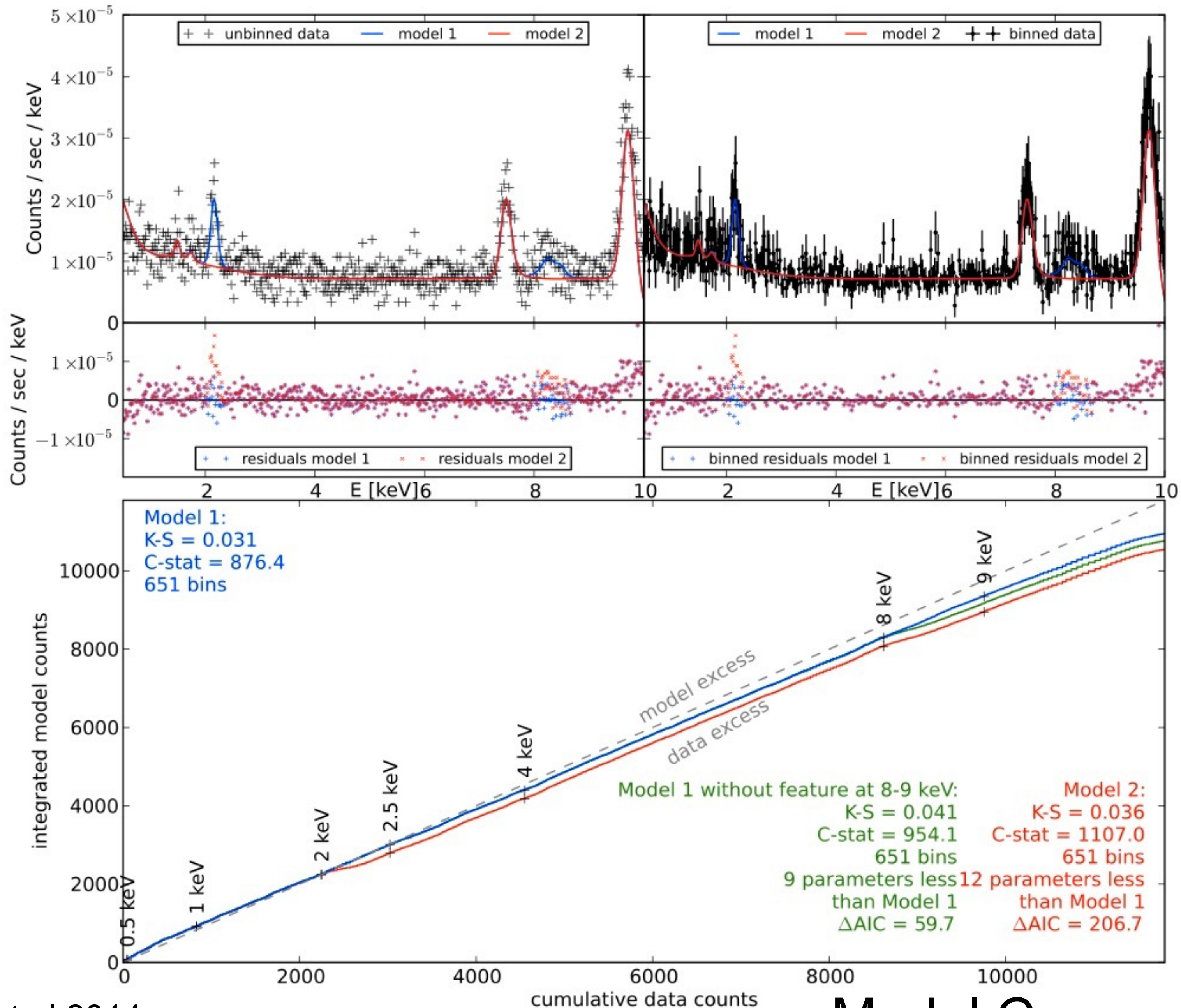


Global Fitting

- Poisson data
- Poisson Likelihood in modeling the data
  - Cash/Cstat/Wstat
  - Bias in chi-gehrels $1+\mathrm{sqrt}(d_i+0.75)$
- Background?
- Group data ?
- High S/N - Fit individual lines
- Large number of parameters
  - MCMC methods (Markov-Chain Monte Carlo)
  - Bayesian methodology

Ungrouped for Fitting

Group for visualization



Buchner et al 2014

Model Comparison

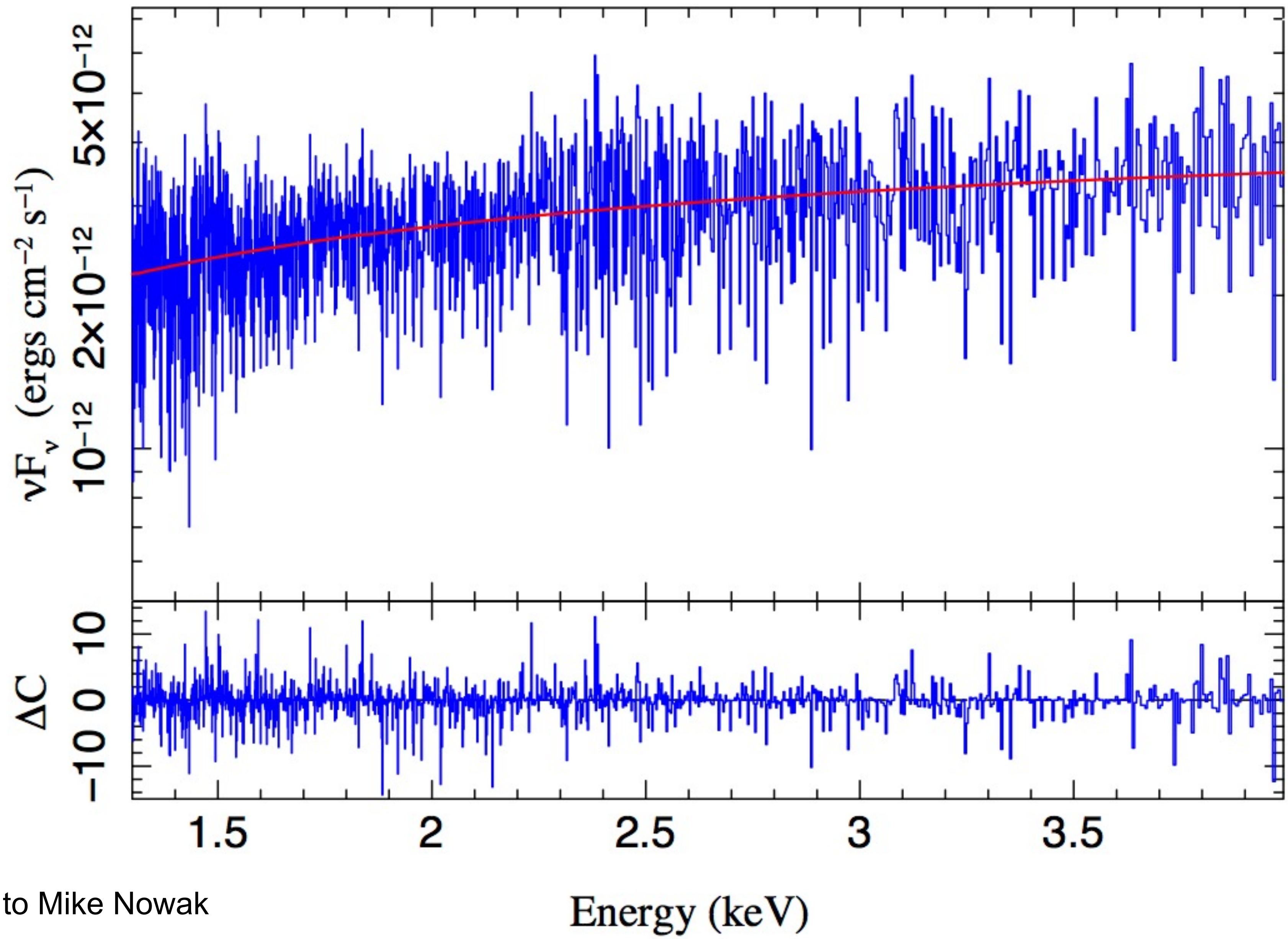
Line detection

- Critical for plasma diagnostics
  - Densities from line ratios, e.g. He-like OVI and NeIX
  - Temperature diagnostics
- Measurements of line intensities:
  - Accounting for background, continuum, instrumental line spread, effective area calibration, contamination from other lines, line profile fitting.
- Issues
  - Binning the data
  - Uncertainties
  - Line significance

Binning

- Loss of resolution
  - blending of lines
  - line disappearing
  - wrong line properties etc.
- Loss of information
  - binned lines can contribute to the continuum
  - wrong normalization and source flux

Continuum vs. Lines

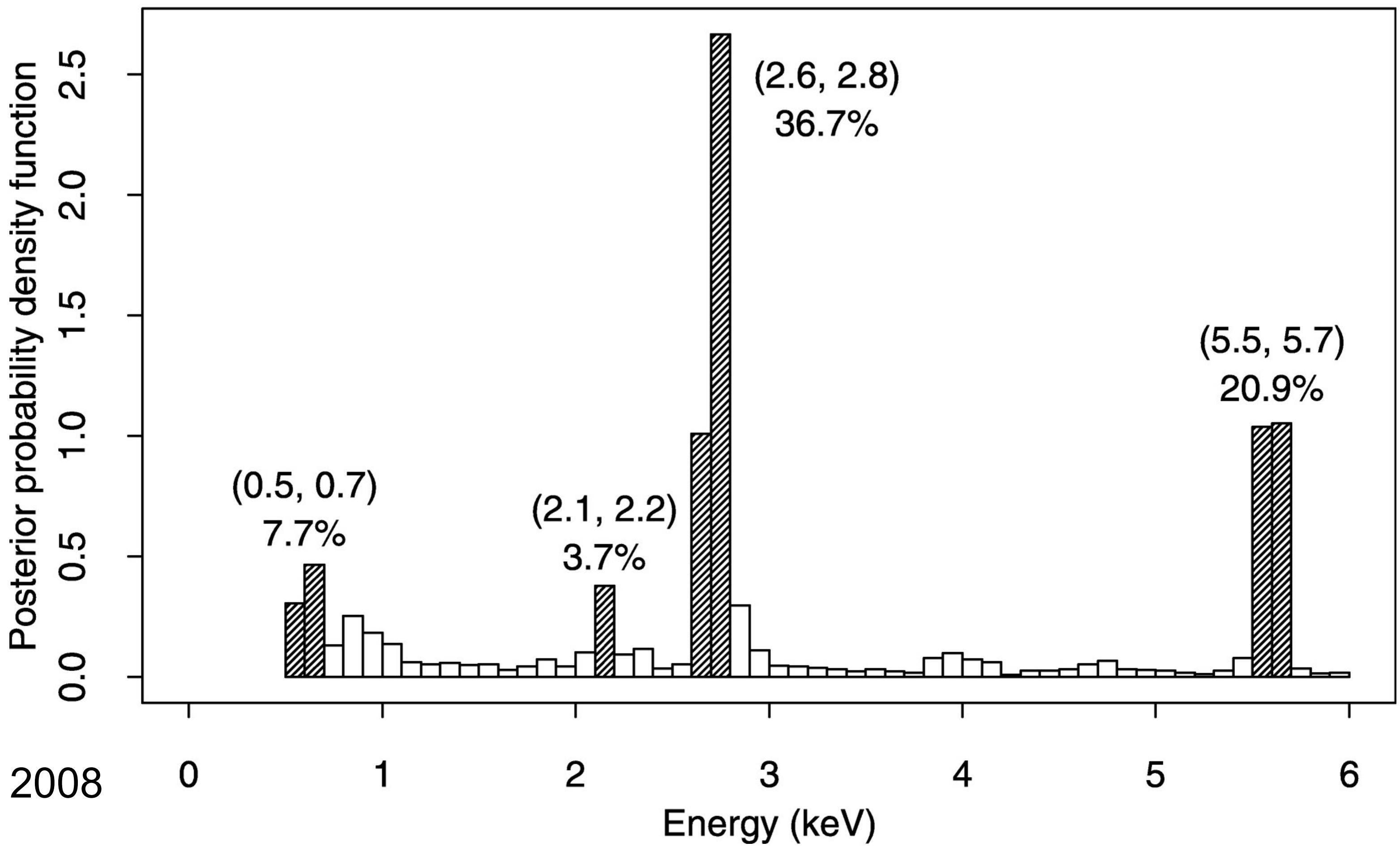


Thanks to Mike Nowak

Detecting Lines in Poisson Data

Computational challenges
Multimodal likelihood
Bayesian methods

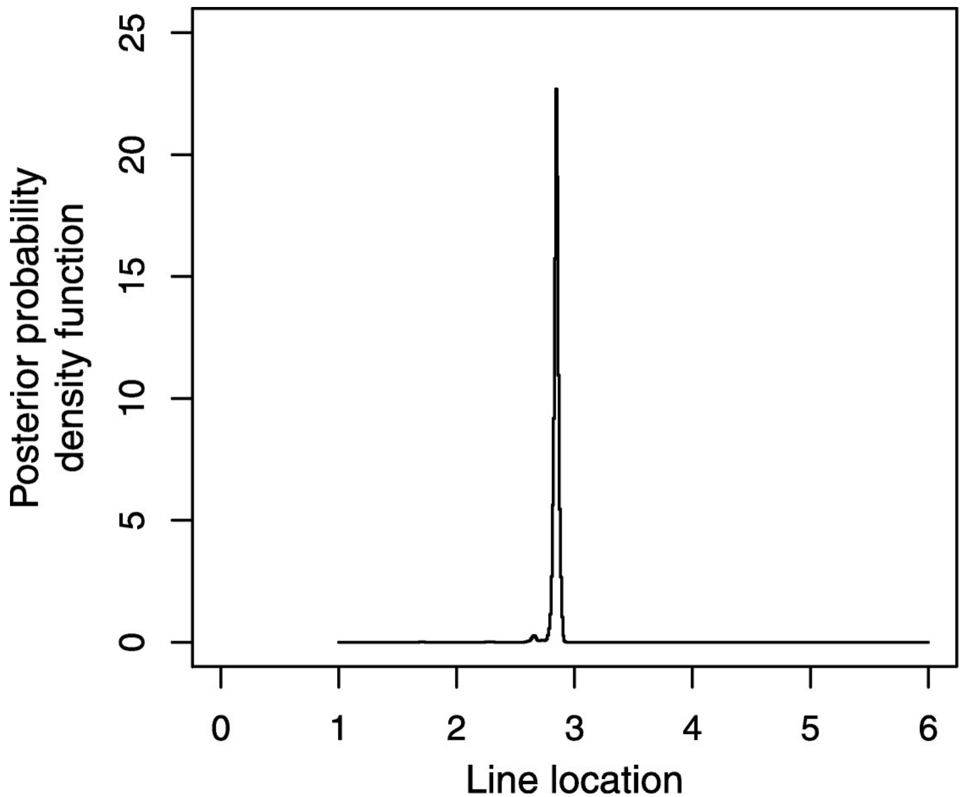
Profile posterior distribution

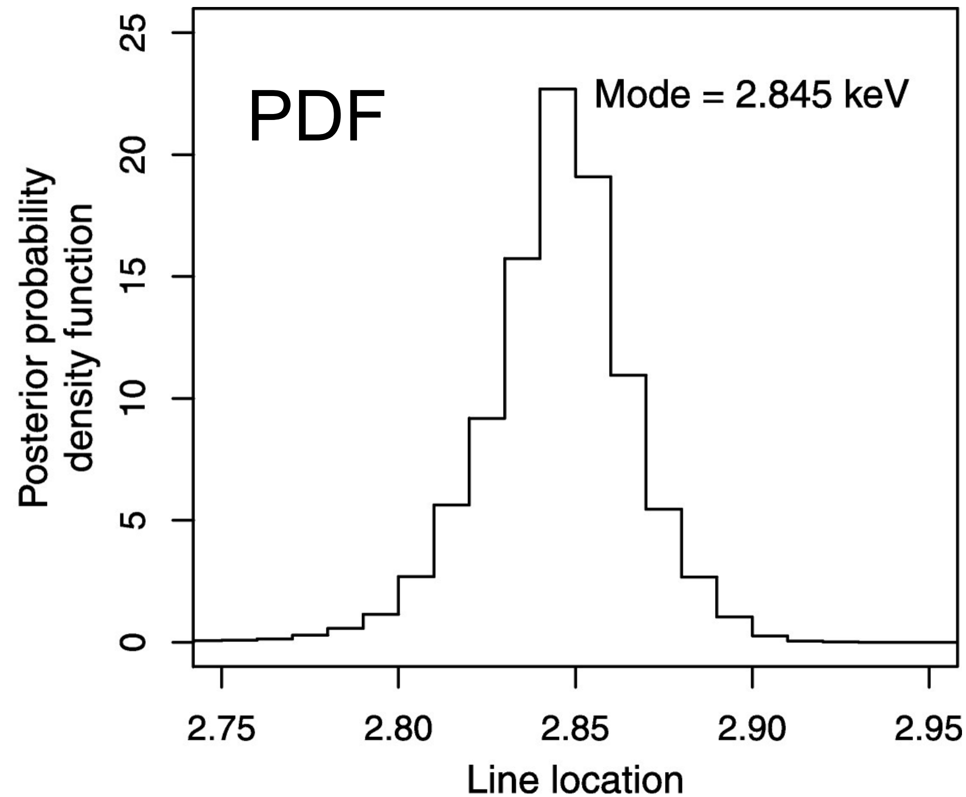


Park et al 2008

Uncertainties

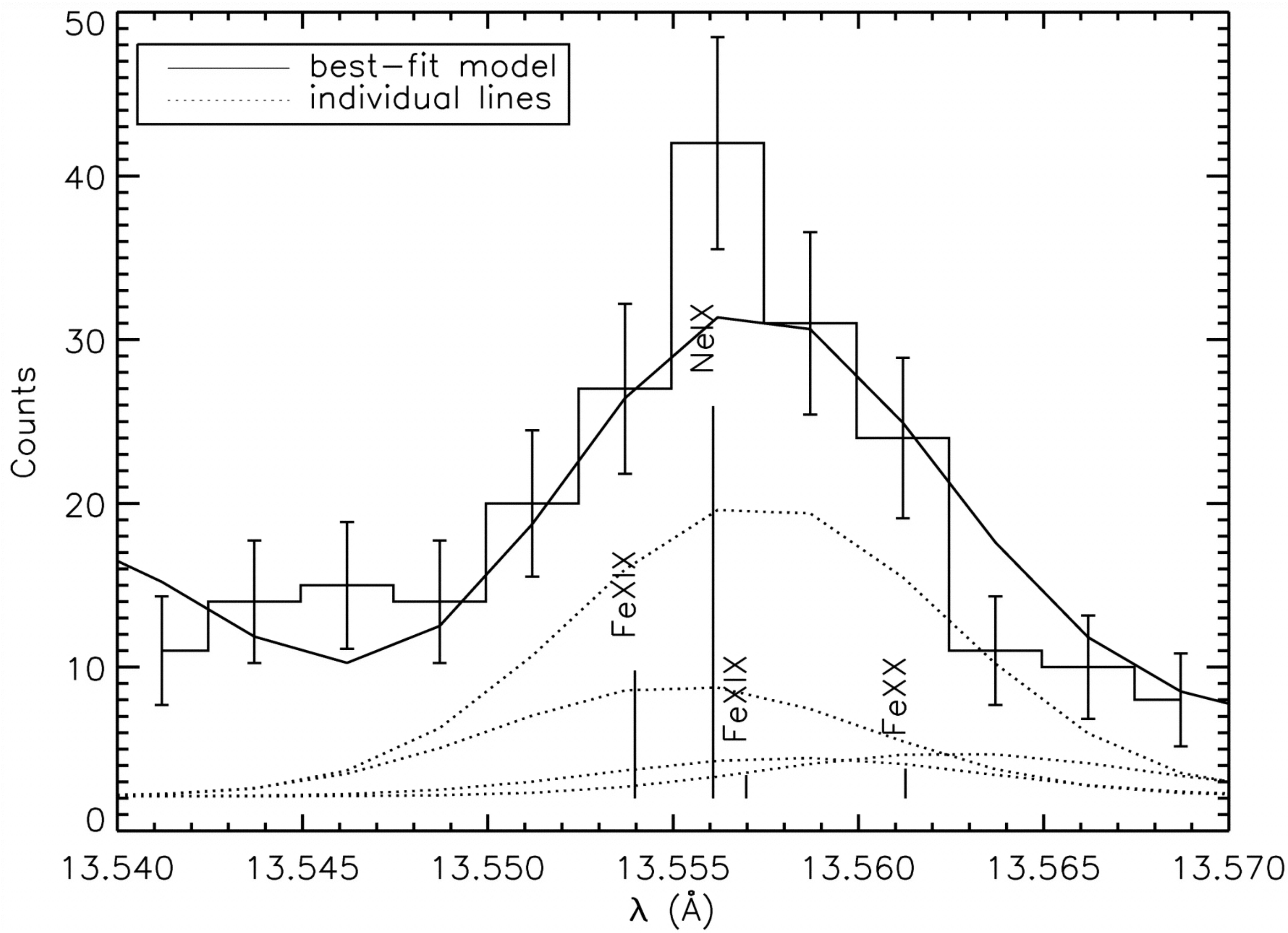
Based on MCMC samples





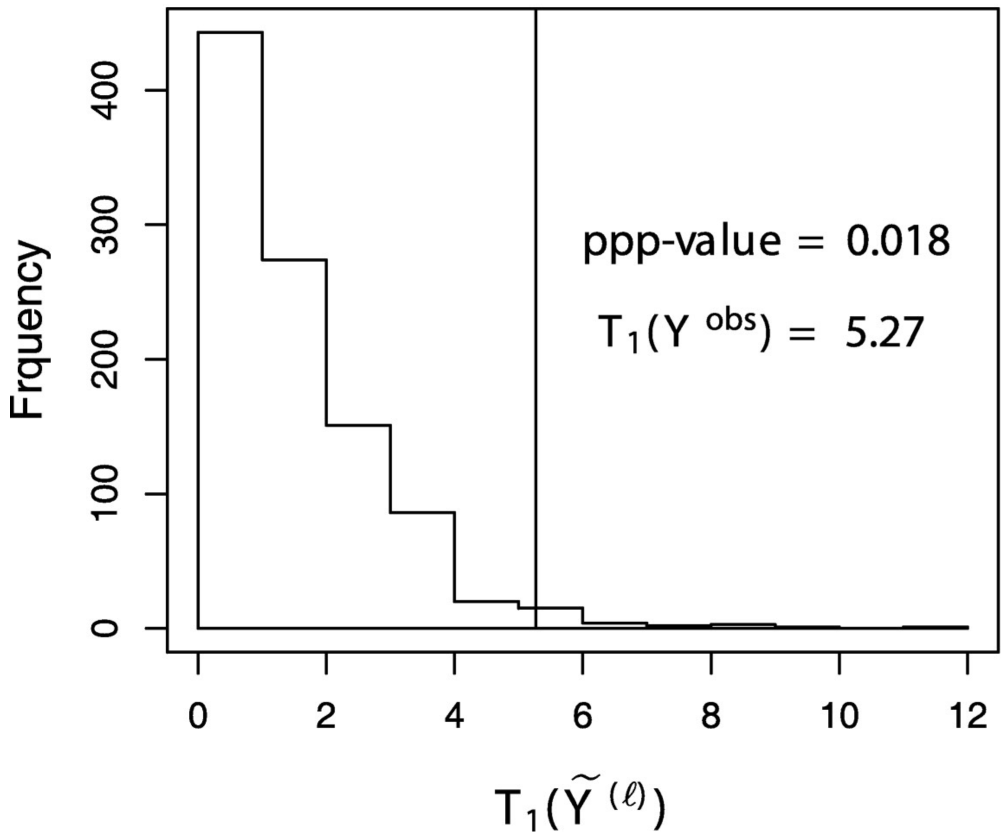
Park et al 2008

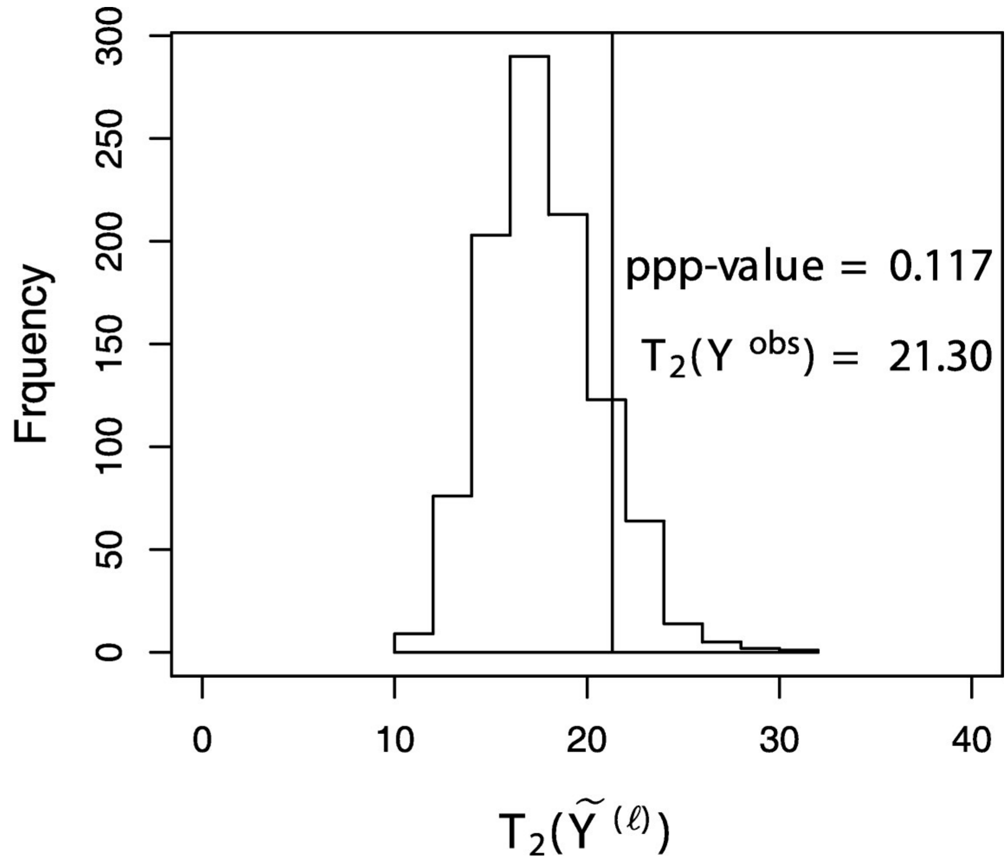
Blended Lines



Uncertainties via MCMC?

Significance of the Line - Simulations





Park et al 2008
Protassov et al 2002

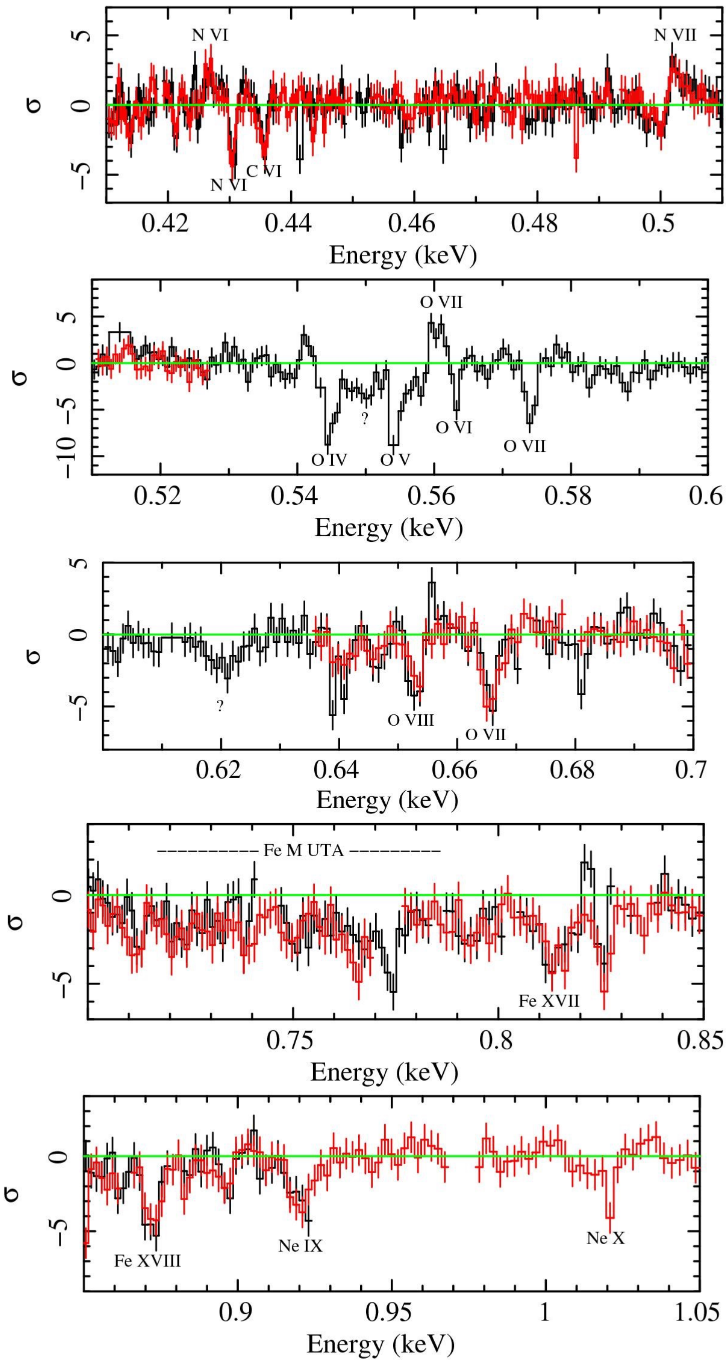
Model Misspecification

- Atomic data bases have offsets between theoretical and true line locations
- Blending of individual lines
- Problems with pseudo-continuum
- Variable abundances across the line-of-sight
- Multi-temperature plasmas require DEM analyses - subject to high frequency instabilities, requiring ad hoc regularization

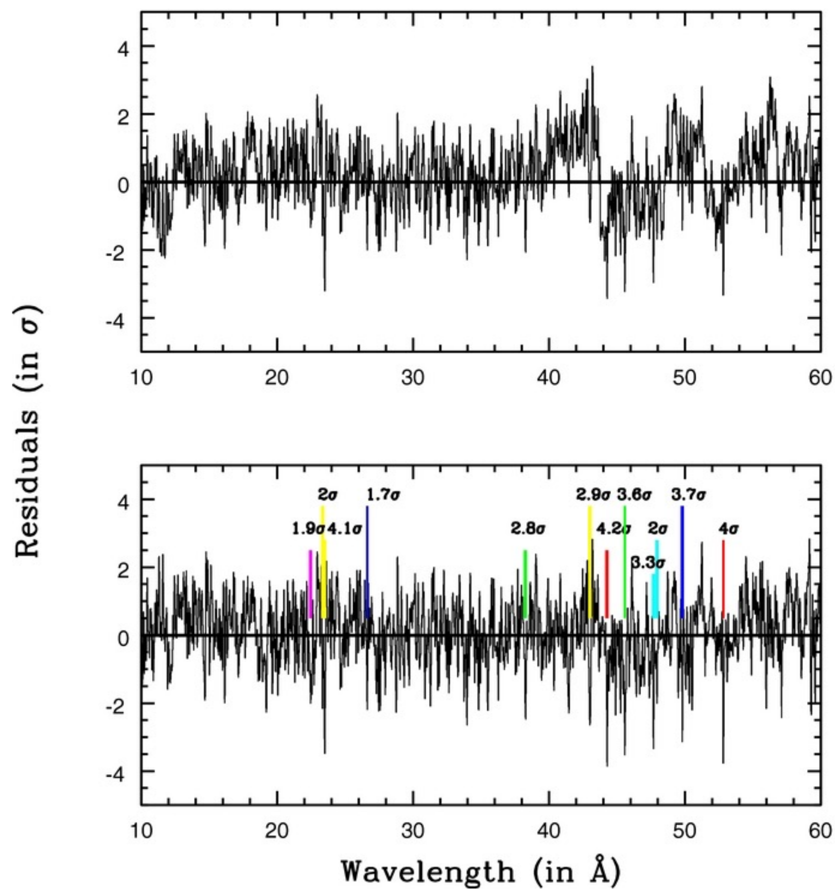
Building Complex Models: Decision Process

- First look at the residuals, use some simple tests:
    1. symmetry of distribution
    2. Cumulative sums
    3. Model the residuals and then move this model to the primary model
- Look at the parameters posterior pdf if you have them
- If you do MCMC, always double check the convergence
    - Multiple starting points
    - Lots of iterations

Look at Residuals

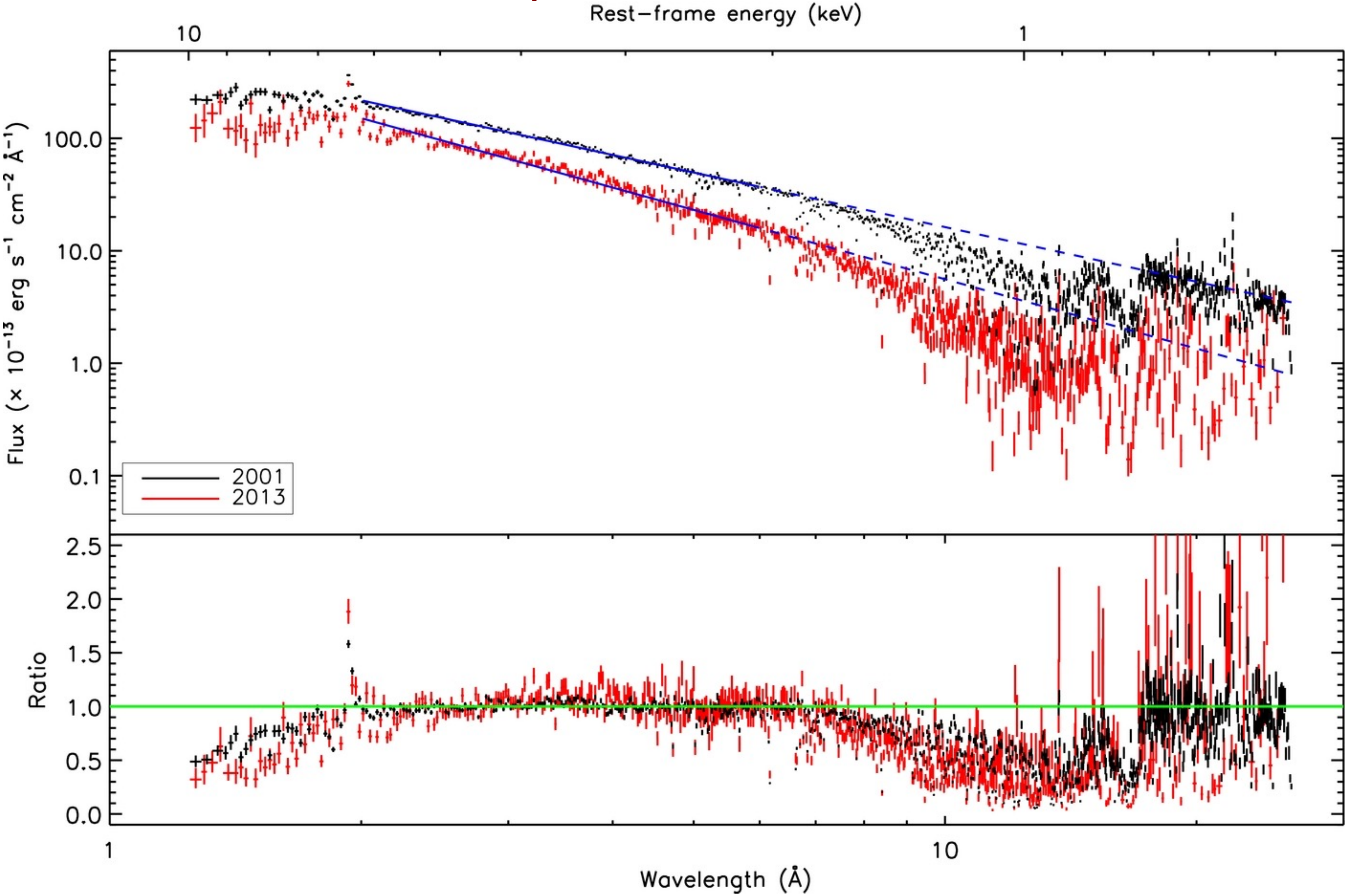


Giustini et al 2015

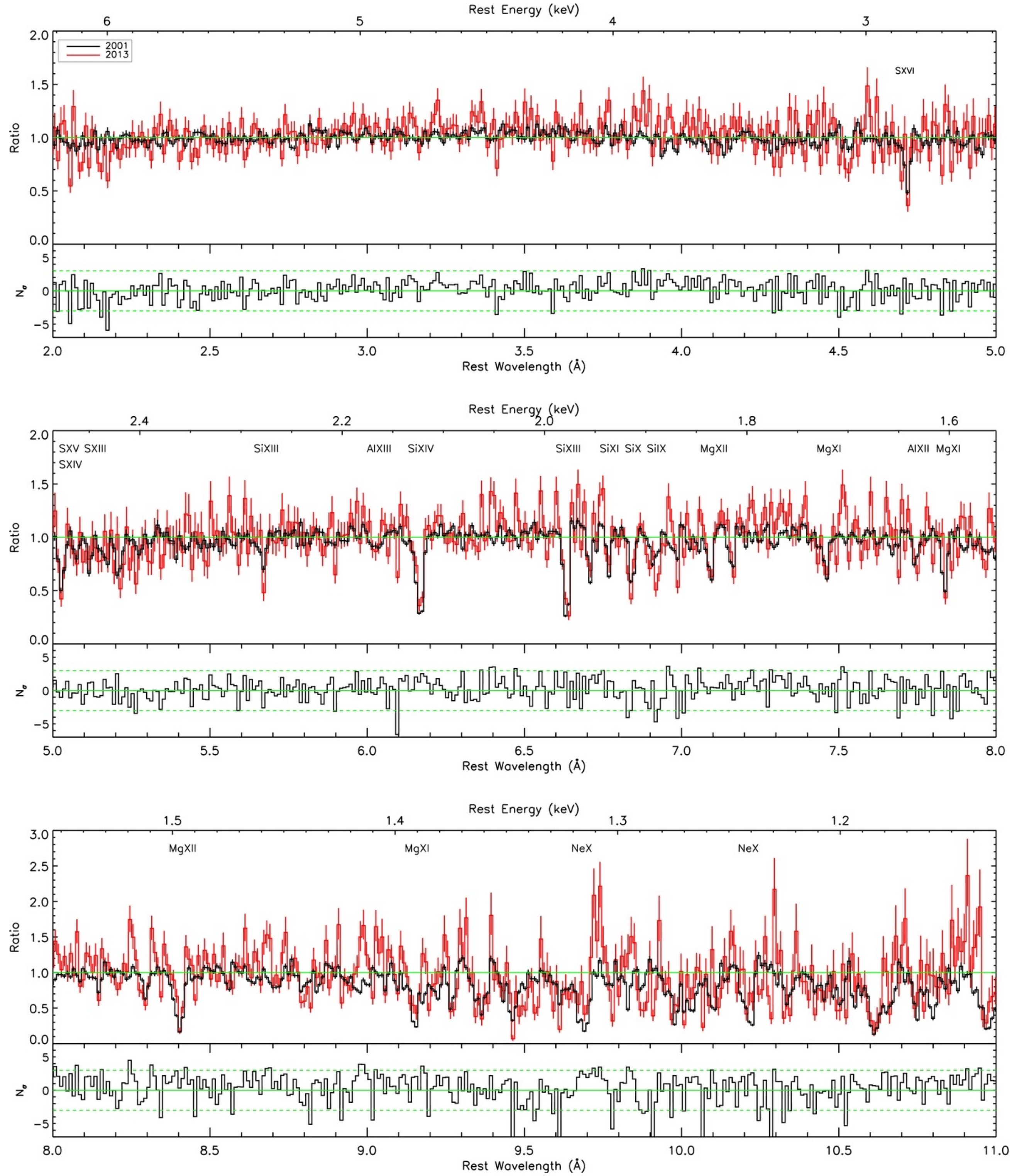


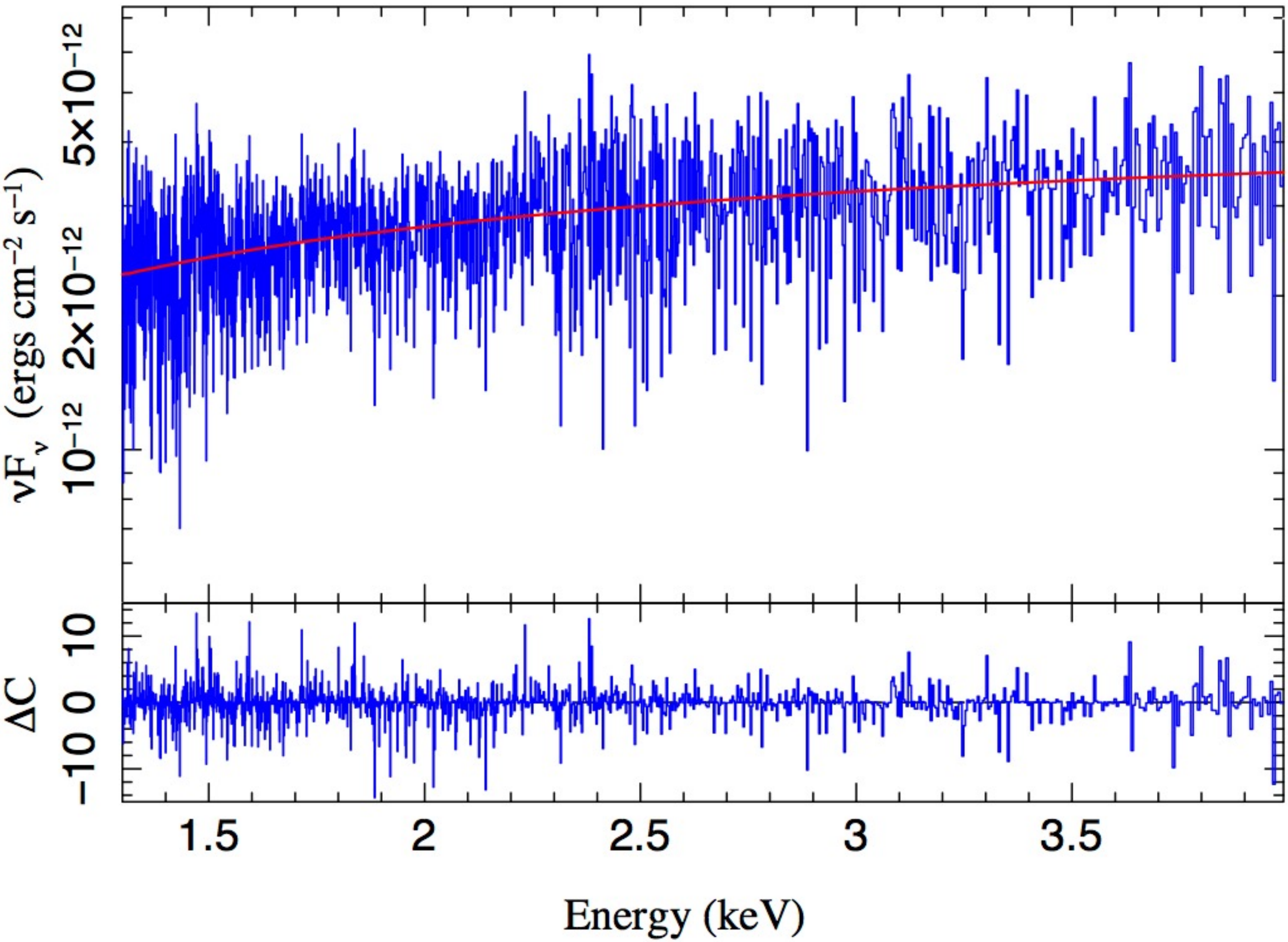
Nicastro et al 2013

NGC 3783 - variability



Scott et al 2014


Rest Energy (keV)
2001
2013
SXVI
Ratio
N_σ
Rest Wavelength (Å)
SXV SXIII
SXIV
SiXIII
AlXIII
SiXIV
SiXIII
SiXI SiX SiIX
MgXII
MgXI
AlXII MgXI
MgXII
MgXI
NeX
NeX



νF_ν (ergs cm^{-2} s^{-1})
5×10^{-12}
2×10^{-12}
10^{-12}
ΔC
10
0
−10
1.5
2
2.5
3
3.5
Energy (keV)

Building Complex Models: Decision Process

- Selecting a model - number of parameters vs. model simplicity:
  - Ockham's razor - selecting a simpler model
  - AIC/BIC/DIC/MDL - criteria as a guide
  - Bayes Factors - calculation of integrals, depending on prior

- Classical tests require that a simpler model must be nested inside a complex model
  - mixture models
  - KDE (kernel density estimates)

Classical Model Selections

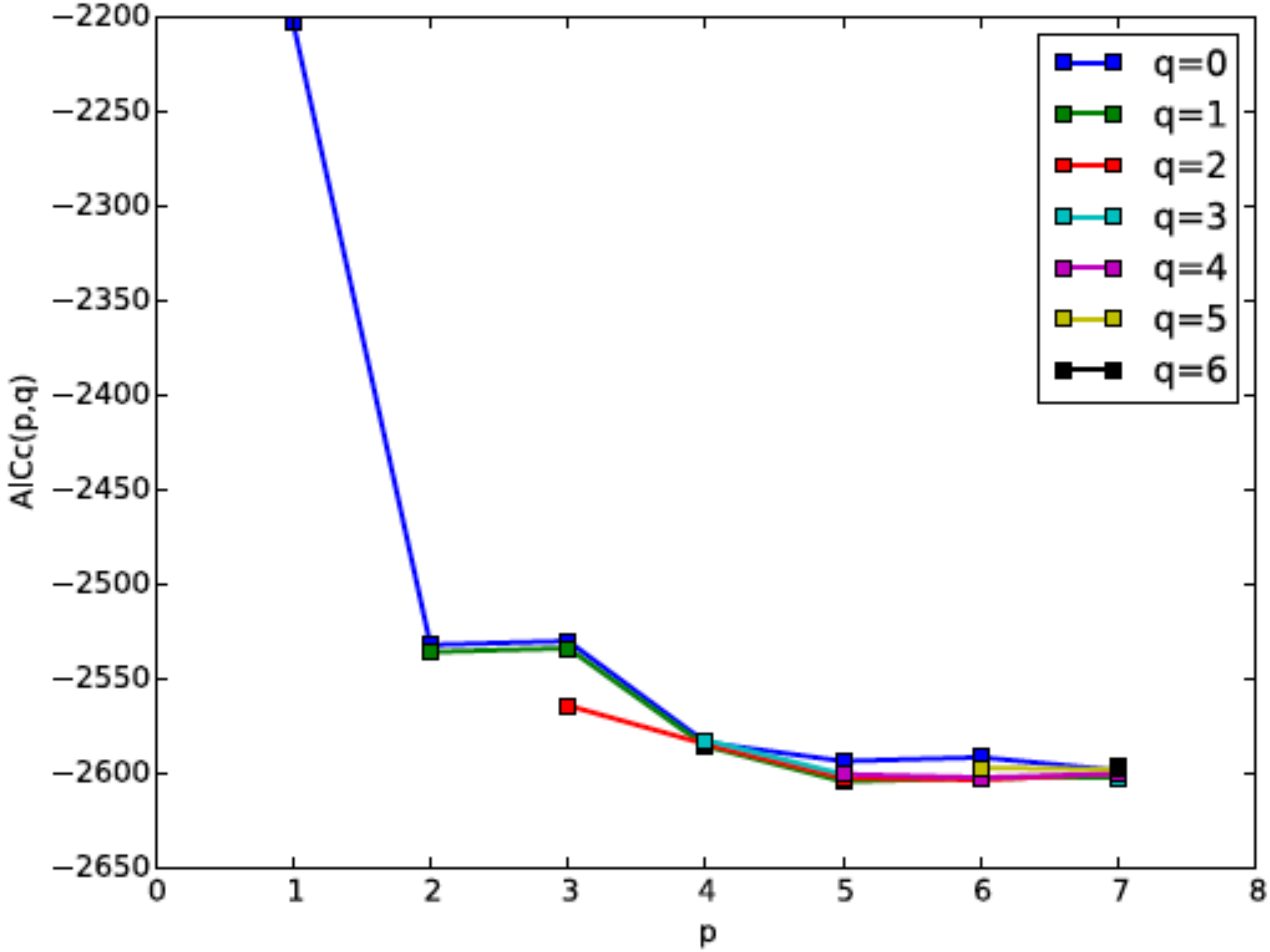
- χ^2 - goodness of fit test
- F-test
- Likelihood Ratio Tests
- AIC - Akaike Information Criterion

Given ML for a set of models. The model with the largest value provides the best description of the data. Need to incorporate number of model parameters. The model with the lowest AIC value is the best model.

$$AIC = \underbrace{-2\ln[L(M)]}_{\chi^2 \text{ - assuming Normality}} + 2k + \underbrace{\frac{2k(k+1)}{N-k-1}}_{\text{finite sample correction}}$$

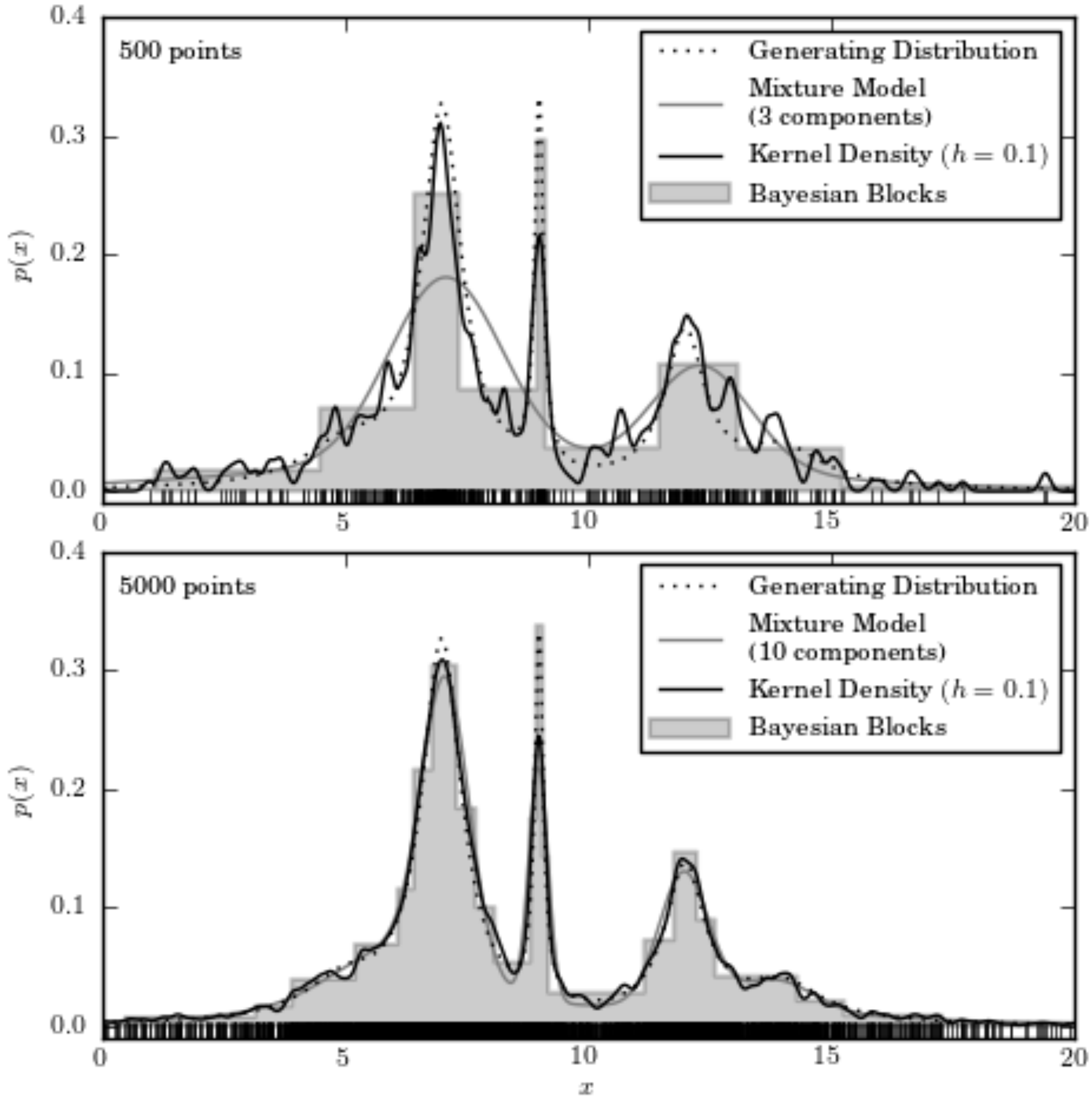
K - number of model parameters
N - number of data points

Example



Kelly et al 2014, ApJ 788,33

Mixture Models: Example



AstroML package

Summary

- Global fitting
- Line detection and fitting low counts lines
- Massive model misspecification
- Building complex models

- Statistical methods exist and can be applied
- Need good statistics software tools.

Bayesian Model Selection

- Odds Ratio $$O_{21} = \frac{p(M_2|D,I)}{p(M_1|D,I)}$$ M_2,M_1 - models$_I$

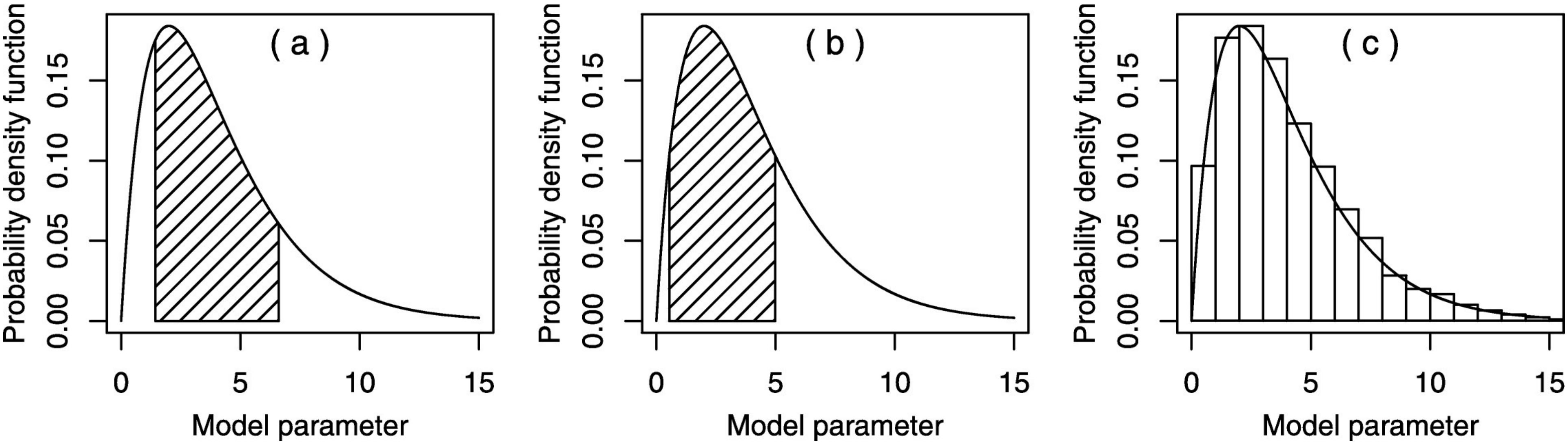
- Bayes Factors $$B_{21} = \frac{p(D|M_2,I)}{p(D|M_1,I)}$$

- BIC - Bayesian Information Criterion
$$BIC = -2\ln[L(M)] + k\ln N$$

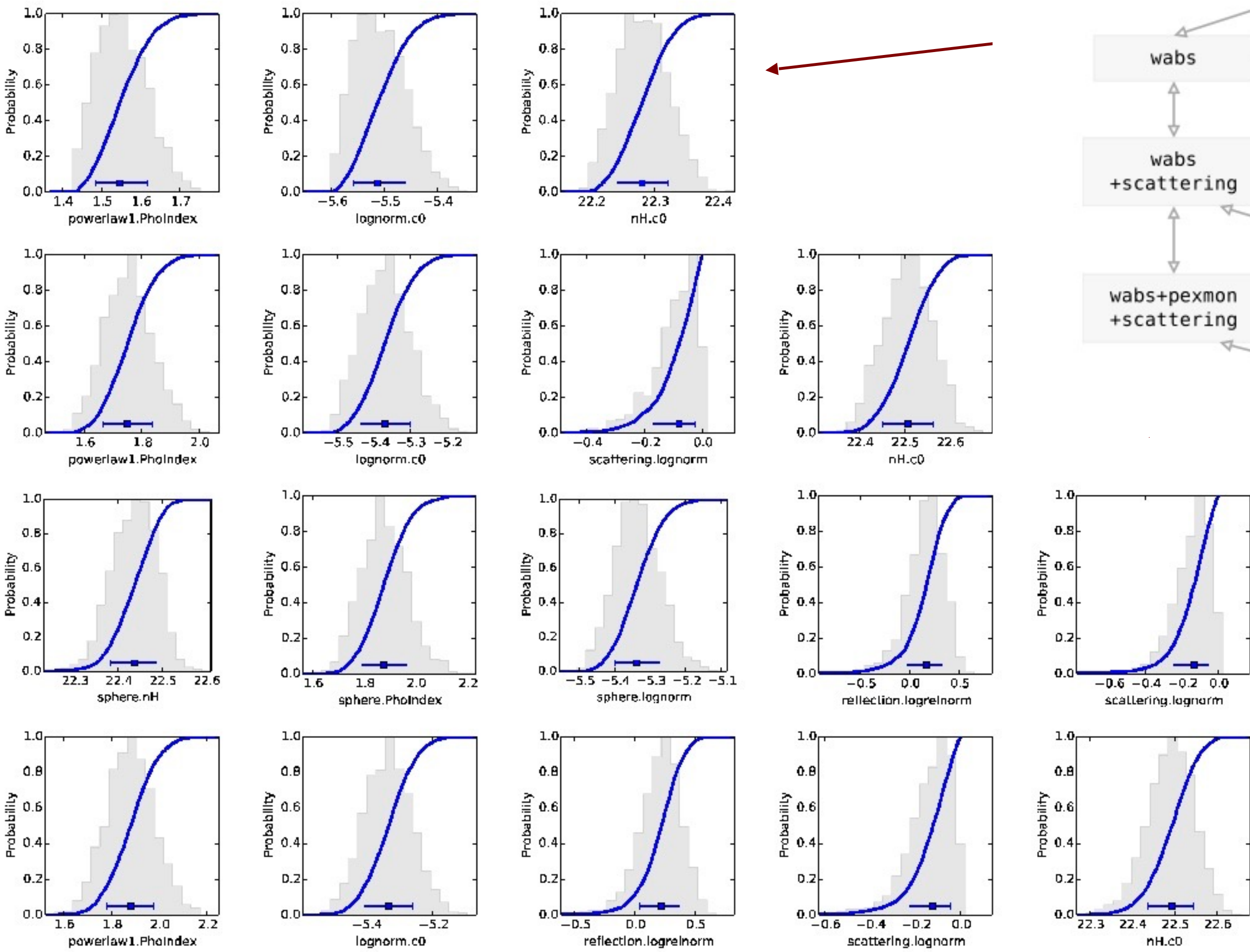
- DIC - Deviance Information Criterion
$$DIC = -2\ln p(\mathrm{y}|\theta) + 2\mathrm{p}_{\mathrm{DIC}}$$

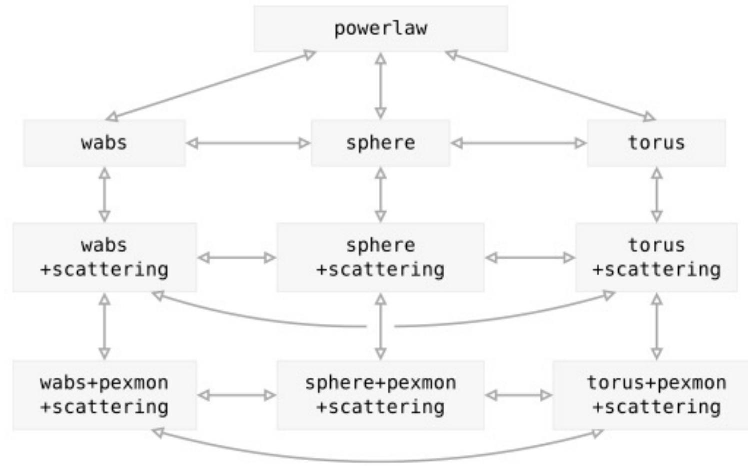
Equal tail

HPD
Highest Posterior Density



Model Summaries: Posterior Distributions





Buchner et al. 2014

Calibrate the test statistics

THE ASTROPHYSICAL JOURNAL, 571:545–559, 2002 May 20

STATISTICS, HANDLE WITH CARE: DETECTING MULTIPLE MODEL COMPONENTS WITH THE LIKELIHOOD RATIO TEST

ROSTISLAV PROTASSOV AND DAVID A. VAN DYK
Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138; protasso@stat.harvard.edu, vandyk@stat.harvard.edu

ALANNA CONNORS
Eureka Scientific, 2452 Delmer Street, Suite 100, Oakland, CA 94602-3017; connors@frances.astro.wellesley.edu

AND

VINAY L. KASHYAP AND ANETA SIEMIGINOWSKA
Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138; kashyap@head-cfa.harvard.edu, aneta@head-cfa.harvard.edu
Received 2001 June 1; accepted 2002 January 25

ABSTRACT

The likelihood ratio test (LRT) and
in 1971, Bevington in 1969, Lampton
(even asymptotically) adhere to their
astrophysics, thereby casting many
Although the above authors illustrate
can be impossible to compute the cor
use the LRT or the *F*-test to detect a l
certain required regularity conditions

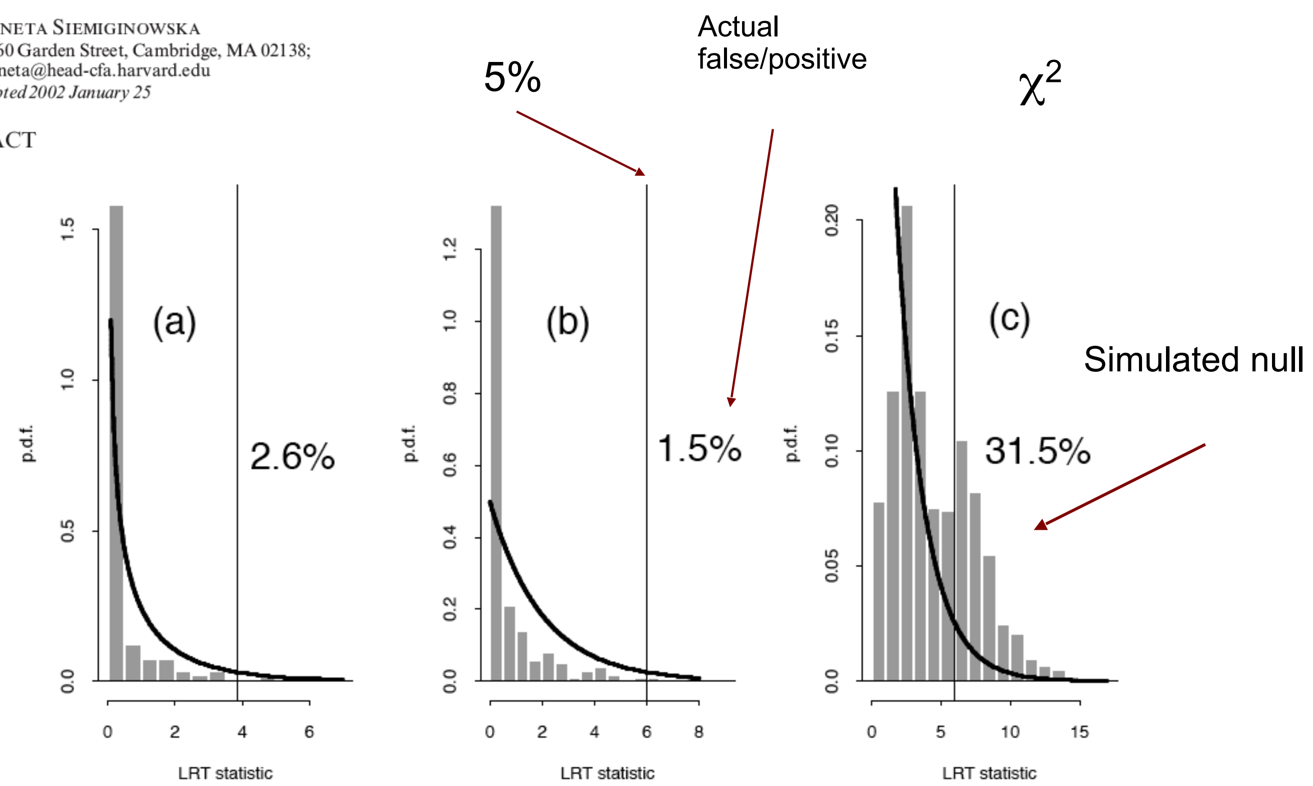
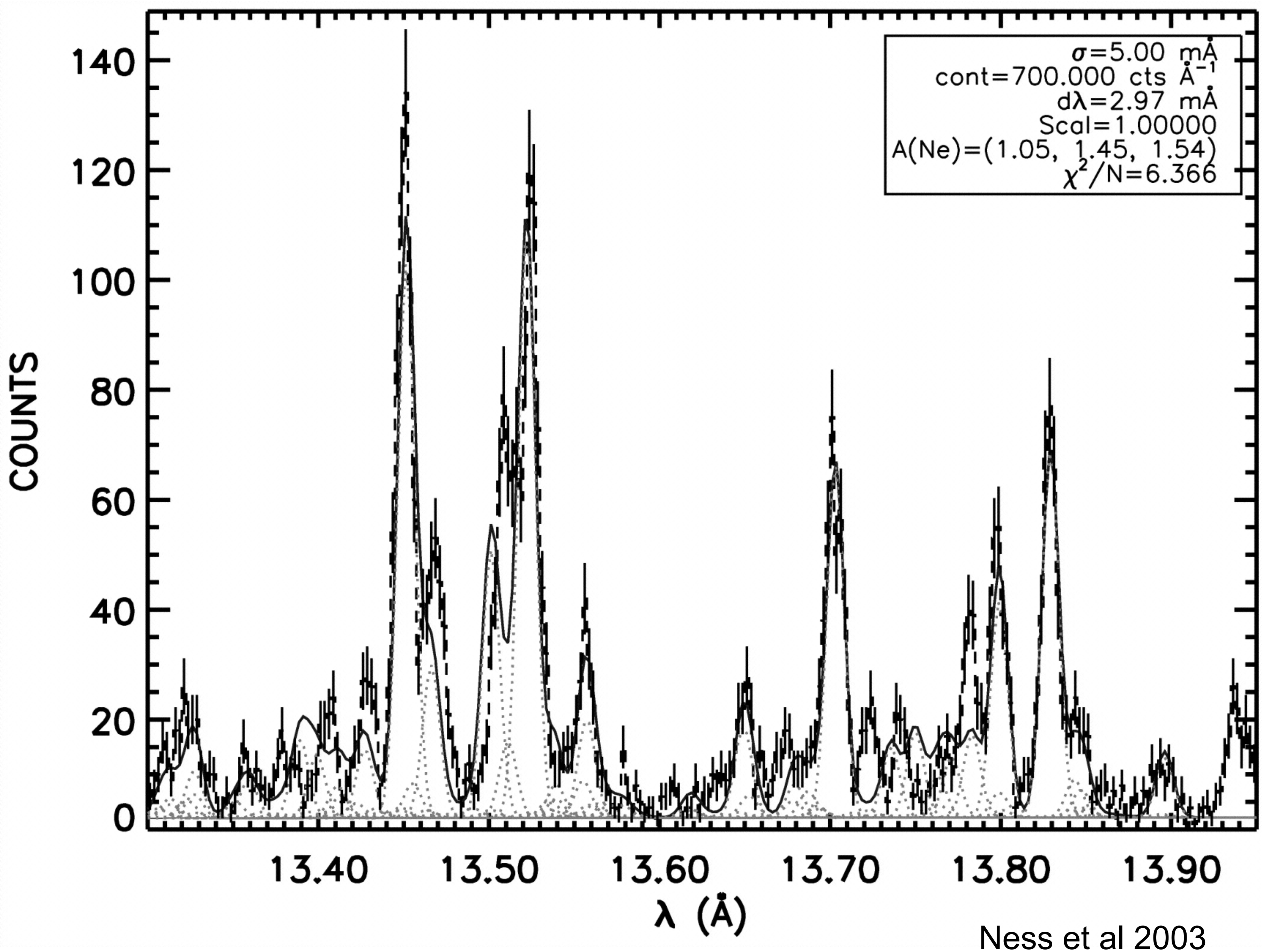


FIG. 1.—Null distribution of the LRT test statistic. The histograms illustrate the simulated null distribution of the LRT statistic in three scenarios and should be compared with nominal χ^2 distributions, which are also plotted. As detailed in § 3.2, the histograms corresponds to (*a*) testing for a narrow emission line with fixed location, (*b*) testing for a wide emission line with fitted location, and (*c*) testing for an absorption line. The vertical lines show the nominal cutoff for a test with a 5% false positive rate; note that the actual false positive rates vary greatly at 2.6%, 1.5%, and 31.5%. The label on the *y*-axis stands for the probability density function.

Multi-line model of NeIX region in HEG spectrum of Capella



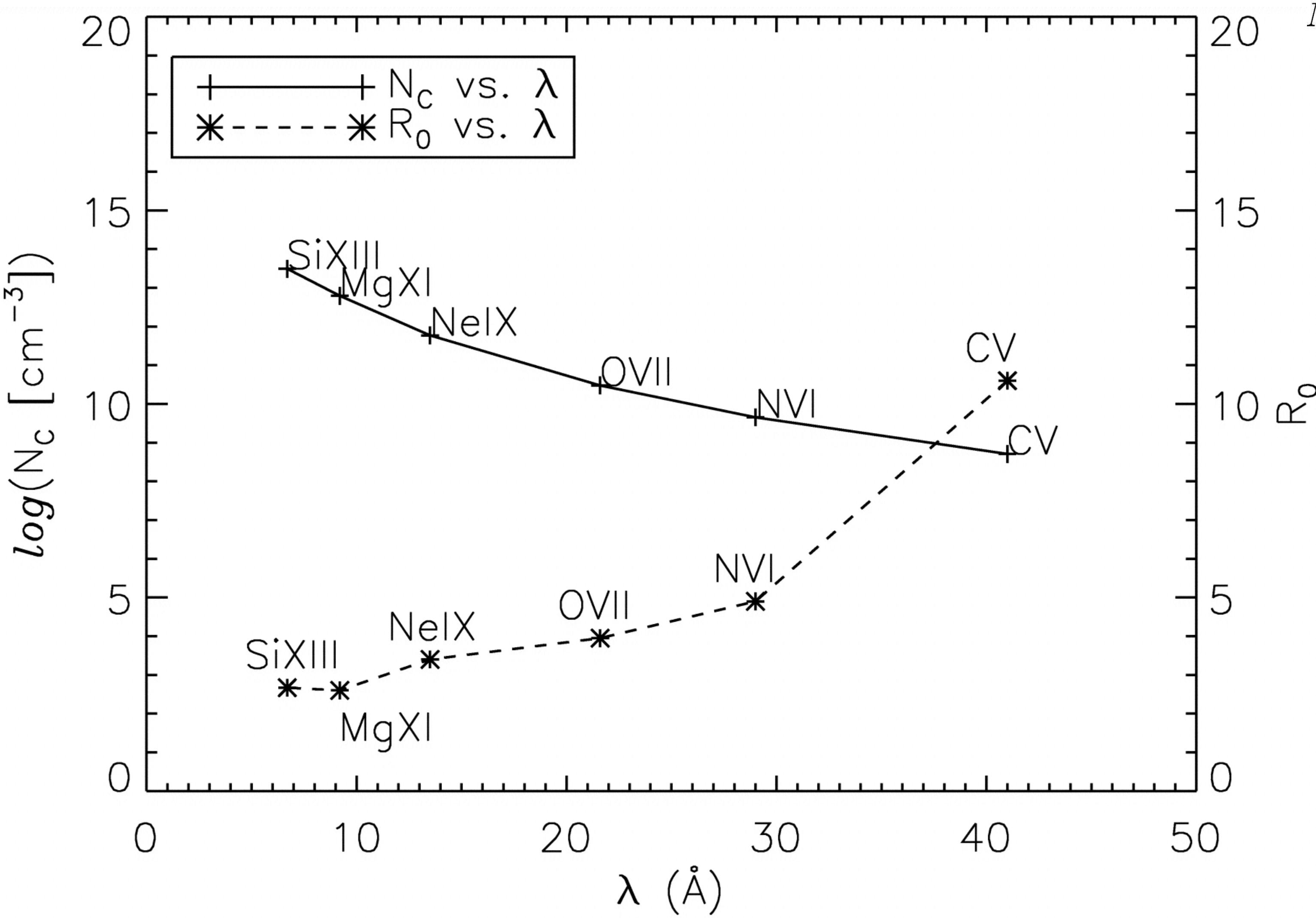
Ness et al 2003

Line Ratios as Density diagnostics

$$\frac{f}{i} = R = \frac{R_0}{1 + \frac{n_e}{N_c}}$$

R_0 – low density limit

$N_c \Rightarrow R = 1/2R_0$ – critical density



Ness et al 2003