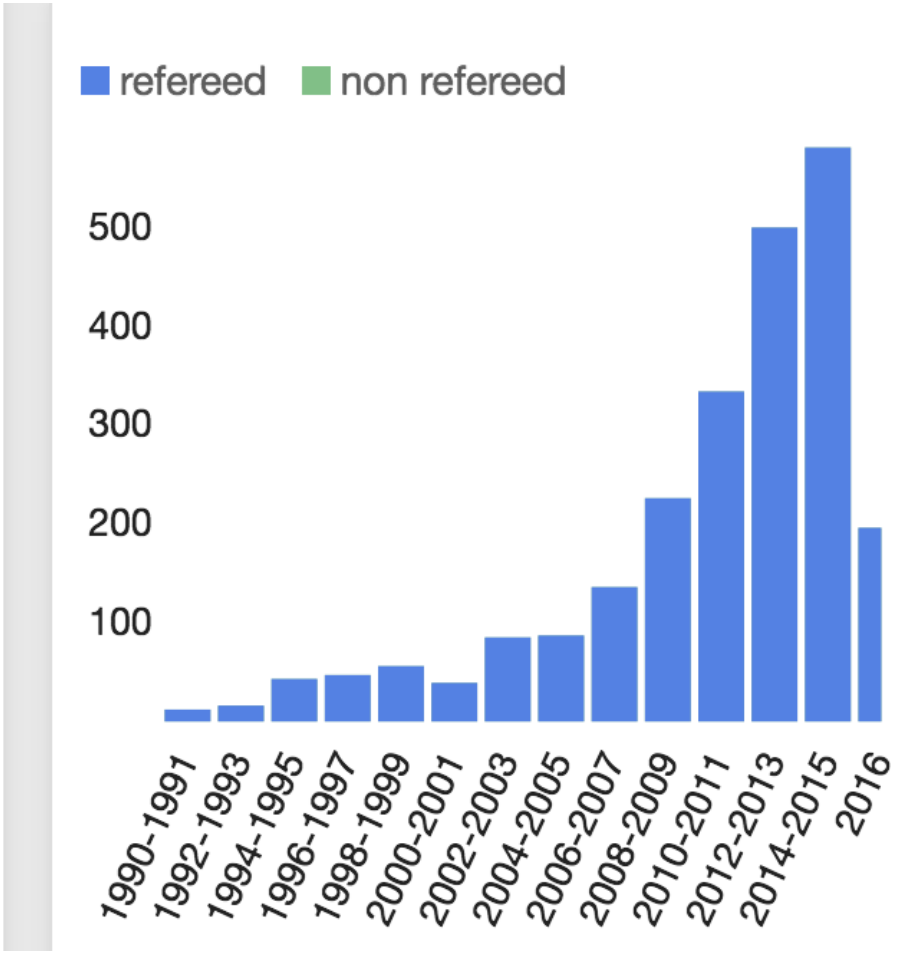
Bayesian Data Analysis

Aneta Siemiginowska
Harvard-Smithsonian Center for Astrophysics
Chandra X-ray Center

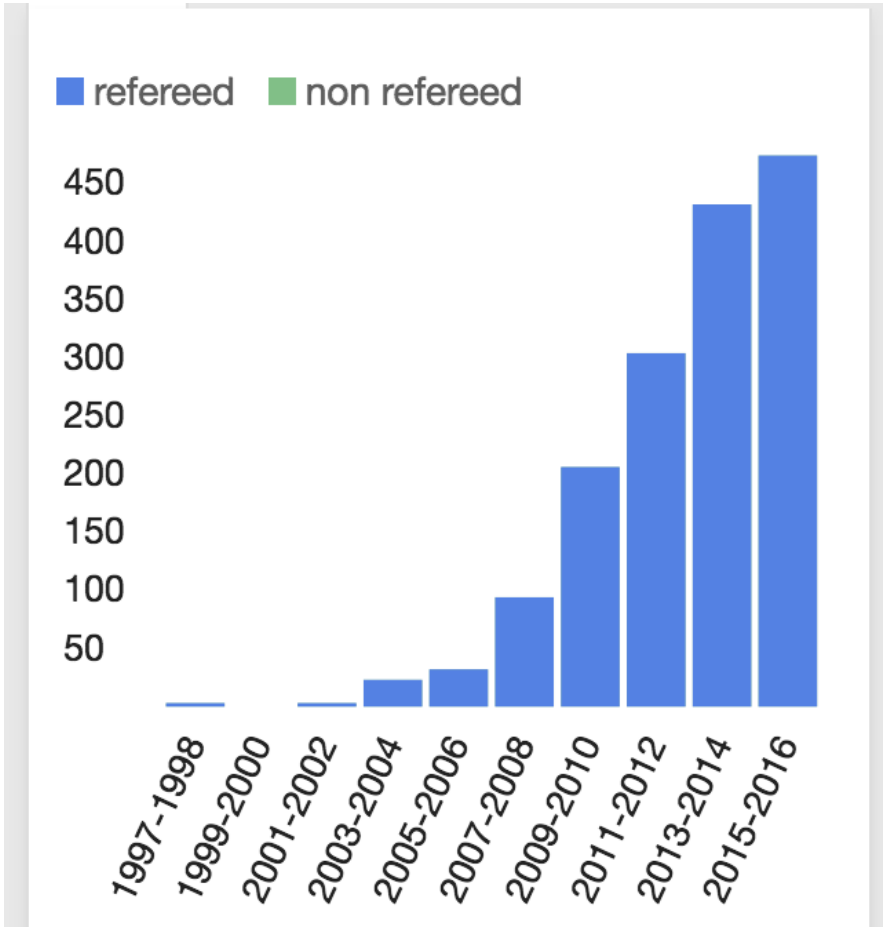
CHASC - Astrostatistics
http://hea-www.harvard.edu/AstroStat/

Publications in ApJ

Bayesian

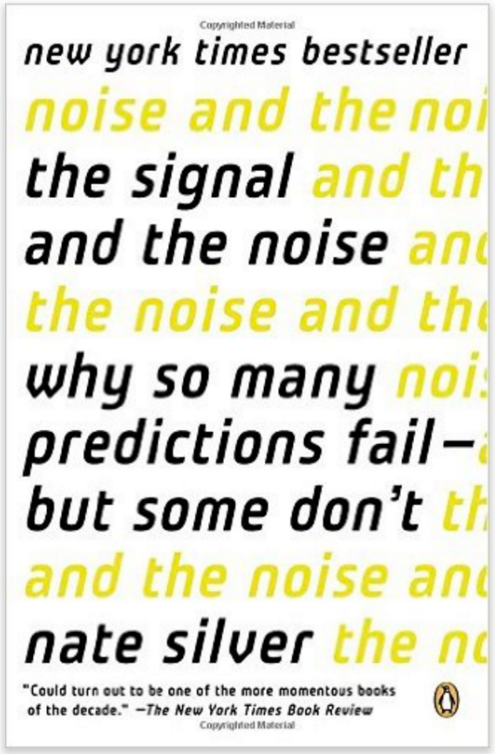


MCMC

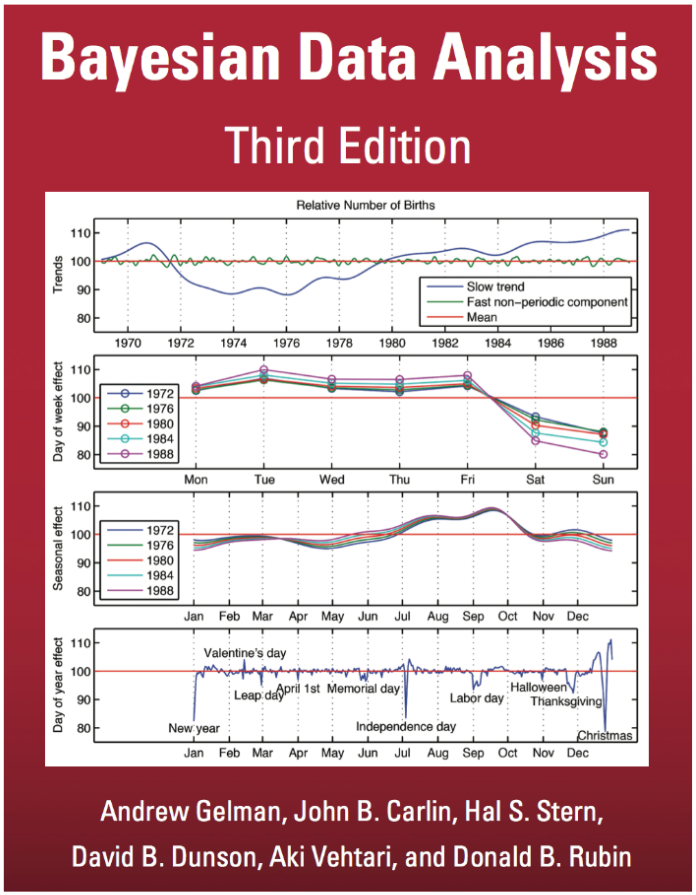


References

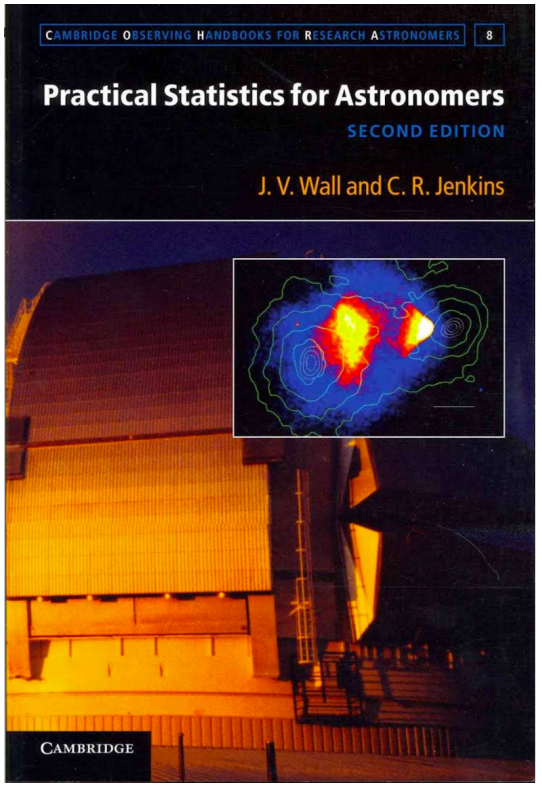
Probabilistic view of the world



Main Reference



For Astronomers



Topics

- Probability
- Distributions
- Bayes Theorem
- Likelihood
- Priors
- Statistical Models
- Bayesian Inference
- Model Checking
- Model Selection
    - Bayesian Information Criteria
    - Bayes Factors

- Components of Bayesian Analysis
- Models and Inference
- Model checking and Model selection

Probability

Probability quantifies randomness and uncertainty.

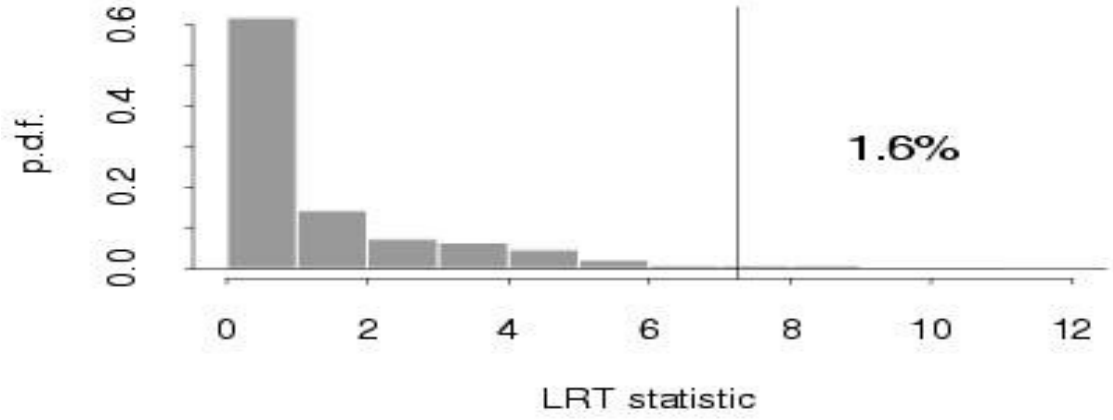
Statistics uses probability to make scientific inferences based on data.

Bayesian: Probability quantifies the degree of belief that an event will occur.

Frequentist: Probability is the relative frequency of an event occurring, in the limit of infinite trials.

Probability Distributions

- Probability is crucial in decision process:
- Example:



Limited data yields only partial idea about the line width in the spectrum. We can only assign the probability to the range of the line width roughly matching this parameter. We decide on the presence of the line by calculating the probability.

Probability Distributions

- Continuous Distribution Function “probability density function”
- Probability that the random variable x takes a value between x and x+dx

$$p(x)dx$$

- Probability that x is between x1 and x2

$$\Pr(x_1 < x < x_2) = \int_{x_1}^{x_2} p(x)dx$$

Probability Distributions

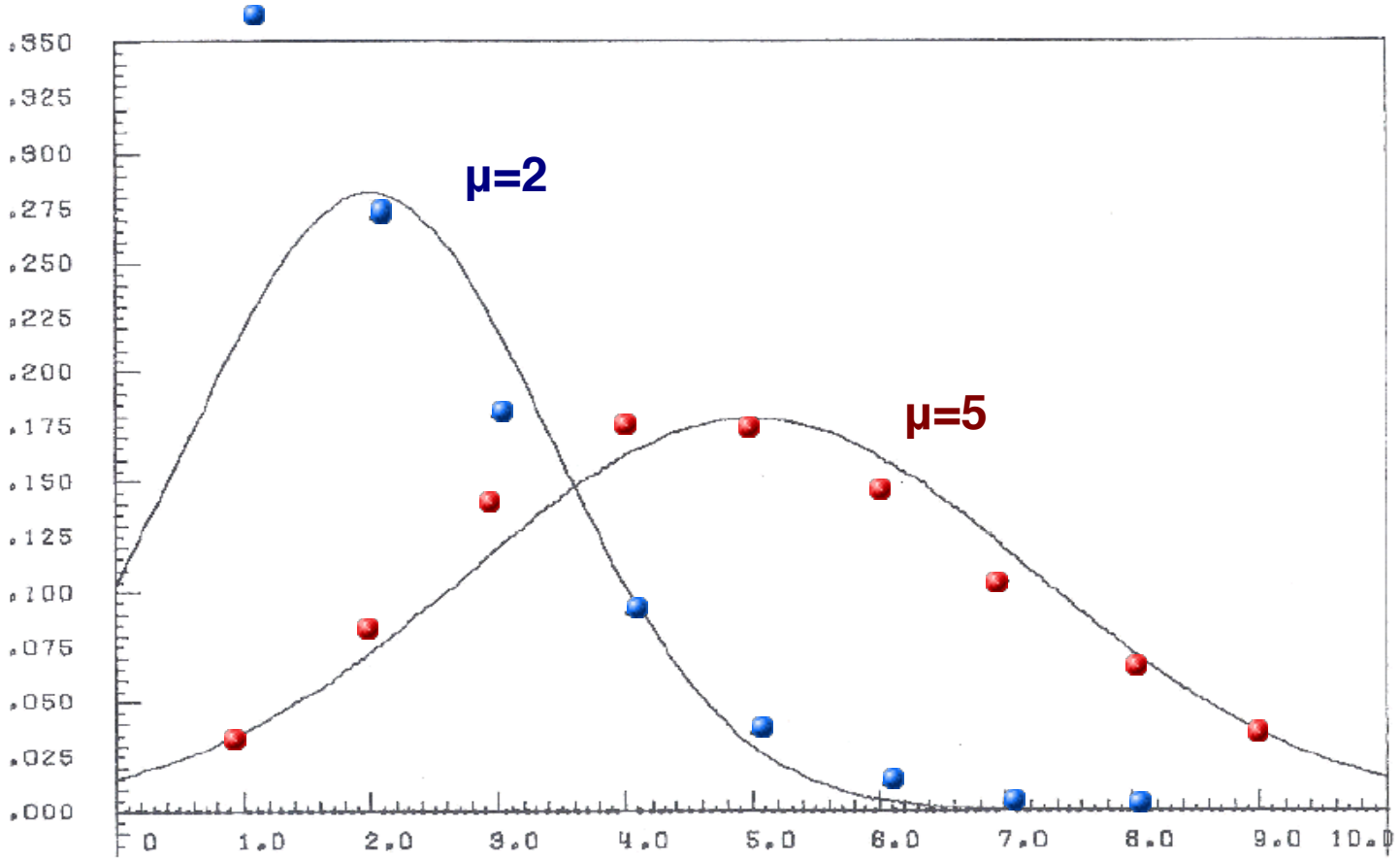
- **Joint:** "probability of x and y" => $p(x,y)$
- **Marginal:** "probability of x" => $p(x) = \int p(x,y)dy$
- **Conditional:** "probability of x at given (fixed) y"

$$p(x \mid y)$$

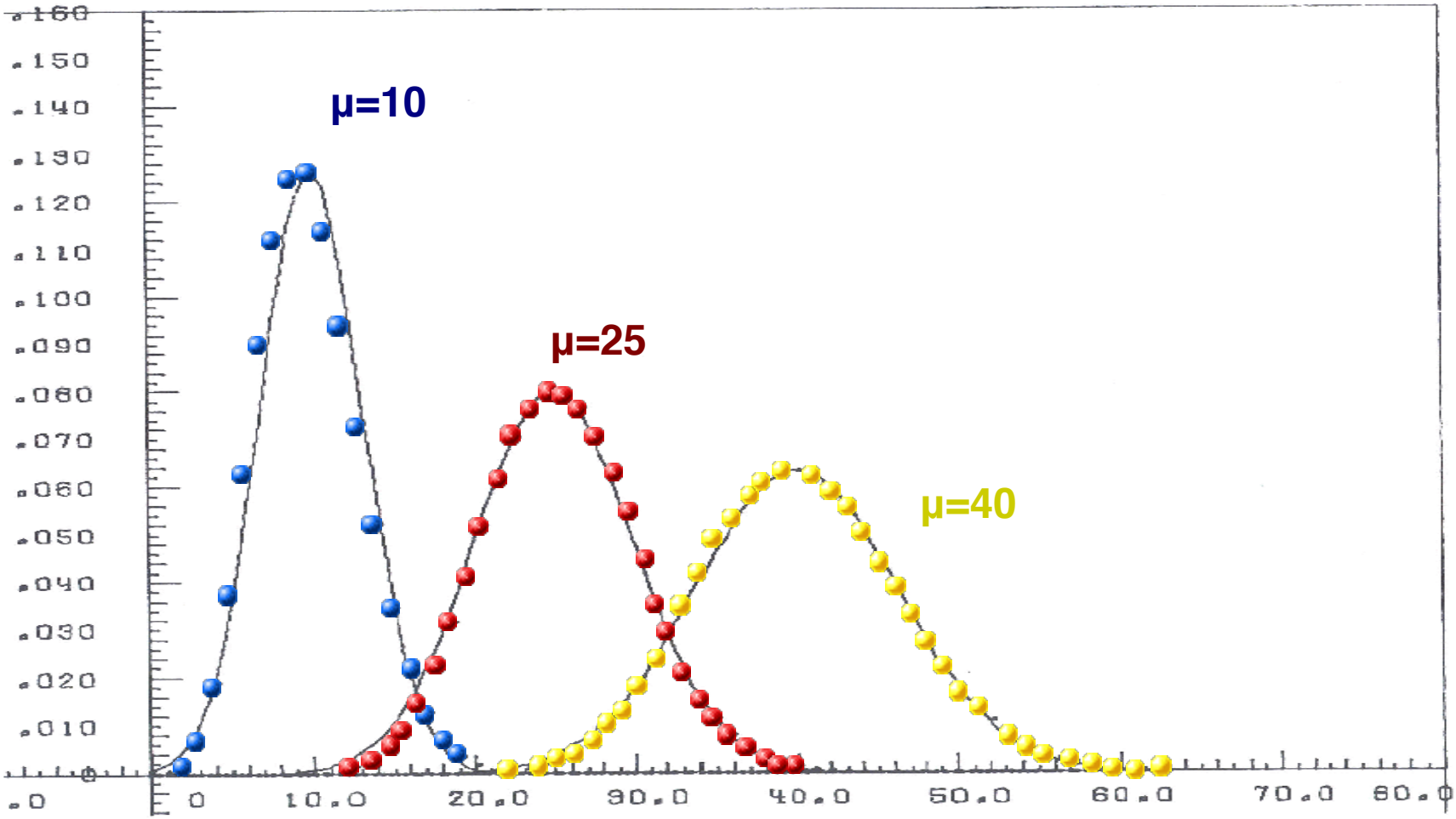
Distributions - Notations

- Binomial $\mathrm{Bin}(n,p)$
- Poisson $\mathrm{Poisson}(\lambda)$
- Normal $\mathrm{N}(\mu,\sigma^2)$
- Chi-square χ^2_ν
- t distribution $t_\nu(\mu,\sigma^2)$

Poisson vs. Gaussian Distributions – Low Number of Counts



Poisson vs. Gaussian



Bayesian Data Analysis

- Setting up a full probability model - joint probability distribution for all observable and unobservable quantities in a problem.
- Conditioning on observed data - calculating and interpreting the posterior distribution
- Evaluating fit of the model and the implications of the posterior probability distribution

Building Blocks of Bayesian Inference

- The sampling distribution, e.g. Likelihood $p(Y|\theta)$
- Prior distribution for model parameters $p(\theta)$
- Bayes Theorem - posterior distribution

$$p(\theta \mid Y) \propto p(Y \mid \theta) p(\theta)$$

Bayesian Inference

- Strength:

  (1) ability to combine information from multiple sources

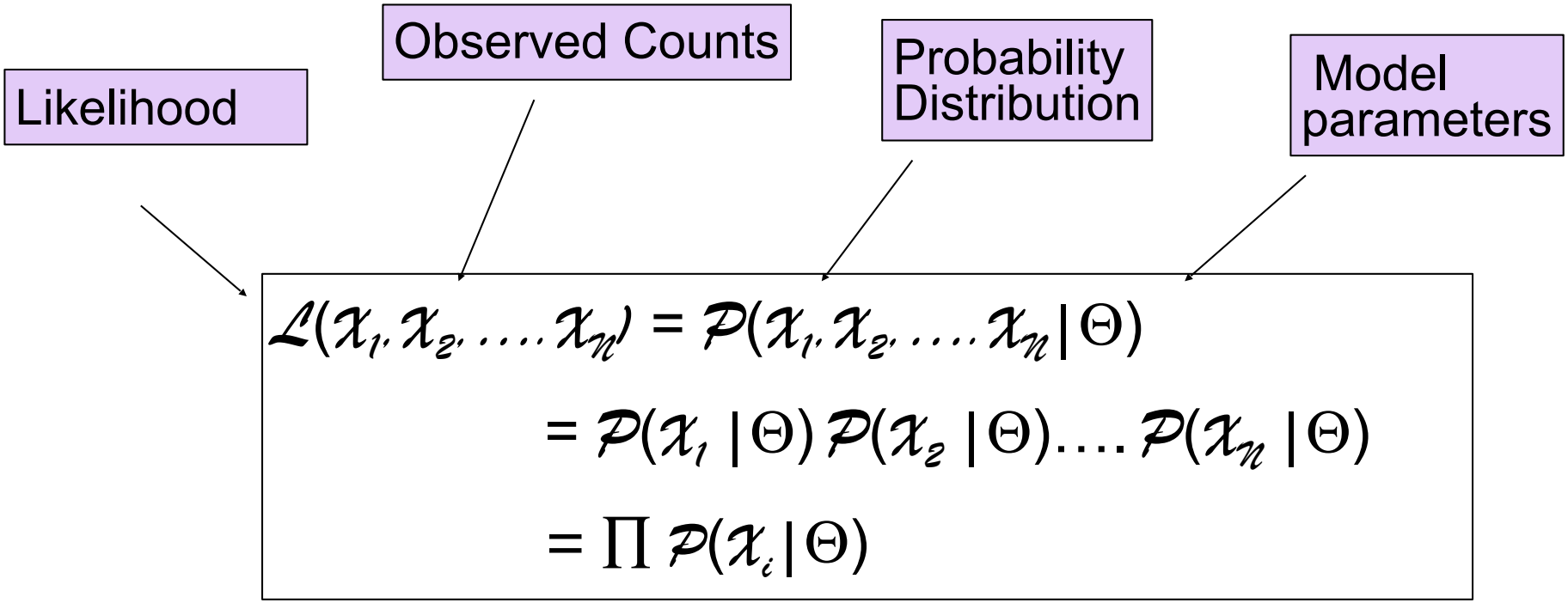
  (2) encompassing accounting of uncertainty about the unknowns in the problem

- Additional points:

  - many parameters in the model
  - hierarchical structuring of models
  - model checking
  - emphasis on inference in the form of distributions
  - the use of simulations
  - the use of probability models as tools for understanding and possibly improving techniques that may not involve a Bayesian model
  - including as much background information as possible - data can be viewed as a random sample, conditional on all the variables in the model
  - design studies of the inference is robust to model assumptions.

Likelihood of one data point

Likelihood



$\mathcal{P}$ - Poisson Probability Distribution for X-ray data

$x_1, \ldots x_n$ - X-ray data - independent

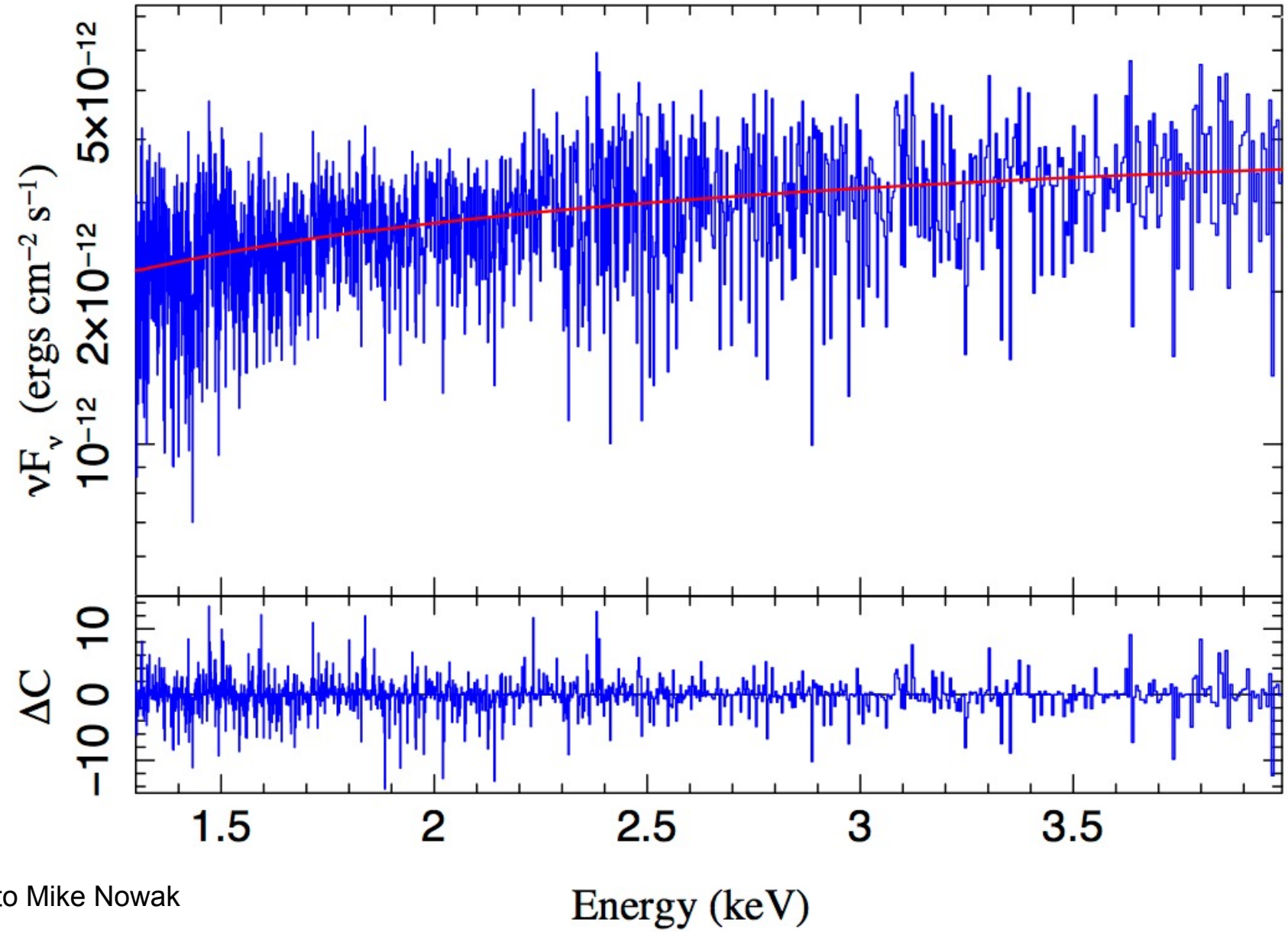
Θ - model parameters

Priors

- Non-informative and informative priors
- **Conjugate** prior distributions - the posterior and prior distributions follow the same parametric forms, for example the beta prior distribution is a *conjugate* family for binomial likelihood.

Statistical Models

Continuum vs. Lines



Thanks to Mike Nowak

Statistical Model

1. Physical source model -

   continuum plus emission lines

$$\Lambda_j(\theta) = \Delta_j f(\theta^C, E_j) + \sum_{k=1}^{K} \lambda_k \pi_j(\mu_k, \nu_k)$$

source

continuum

Lines

Calibration files - RMF/ARF in X-rays

$$C_l(\theta) = \sum_{j \epsilon \mathcal{J}} R_{lj} \Lambda_j(\theta)\, A_j \; + \theta_l^B$$

Counts

source

background

$\theta -$ model parameters

$R_{l,j} -$ redistribution matrix

$A_j -$ effective area

Likelihood and Posterior Distribution

- The likelihood for the Poisson model for photon counts C_l

parameters

$$L(Y^{obs}|\theta) \propto \prod_l C_l(\theta)^{Y_l^{obs}} \exp[-C_l(\theta)]$$

$$p(\theta|Y^{obs}) = \frac{p(\theta)L(Y^{obs}|\theta)}{\int p(\theta)L(Y^{obs}|\theta)d\theta}$$

$$\propto p(\theta)L(Y^{obs}|\theta)$$

posterior distribution

$$p(Y'|Y^{obs}) = \int L(Y'|\theta)p(\theta|Y^{obs})\mathrm{d}\theta$$

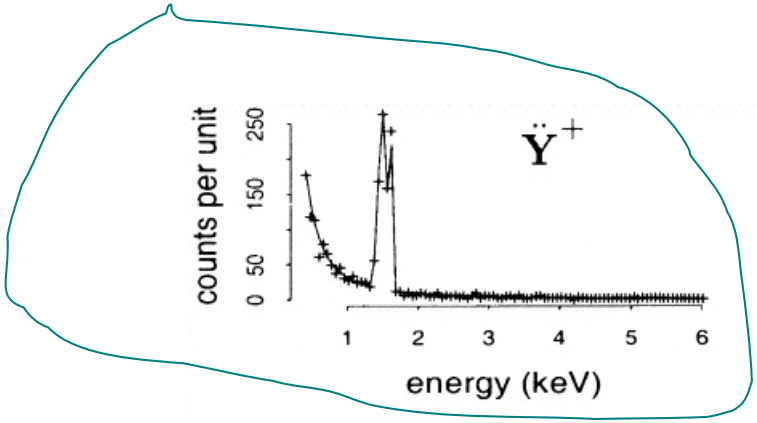
posterior predictive distribution

future

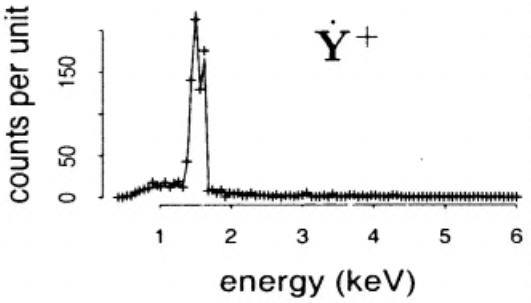
Structured Statistical Models in X-rays

Model directly the physical source and data collection, and include statistical procedure to fit the resulting structured statistical model

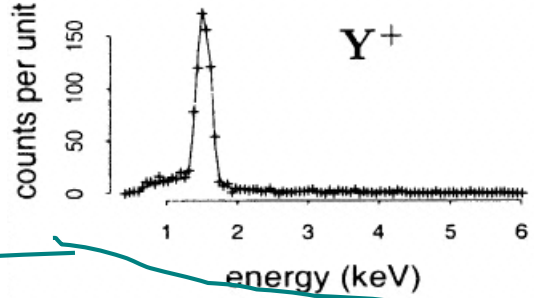
Emitted Spectrum



absorbtion and submaximal effective area

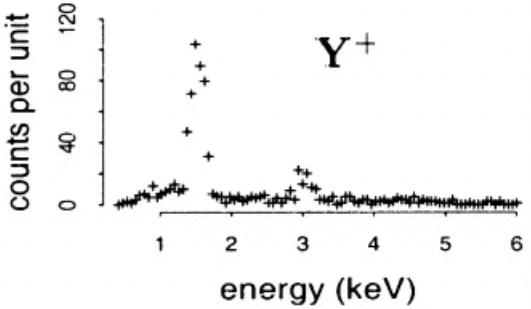


instrument response

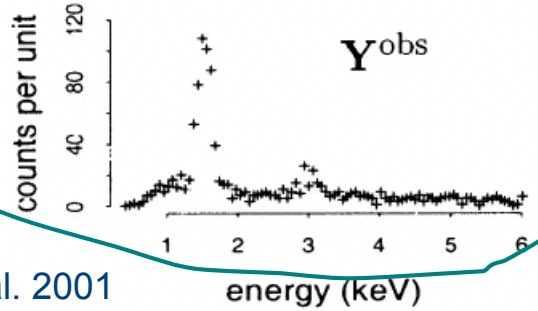


pile-up

Observed



background



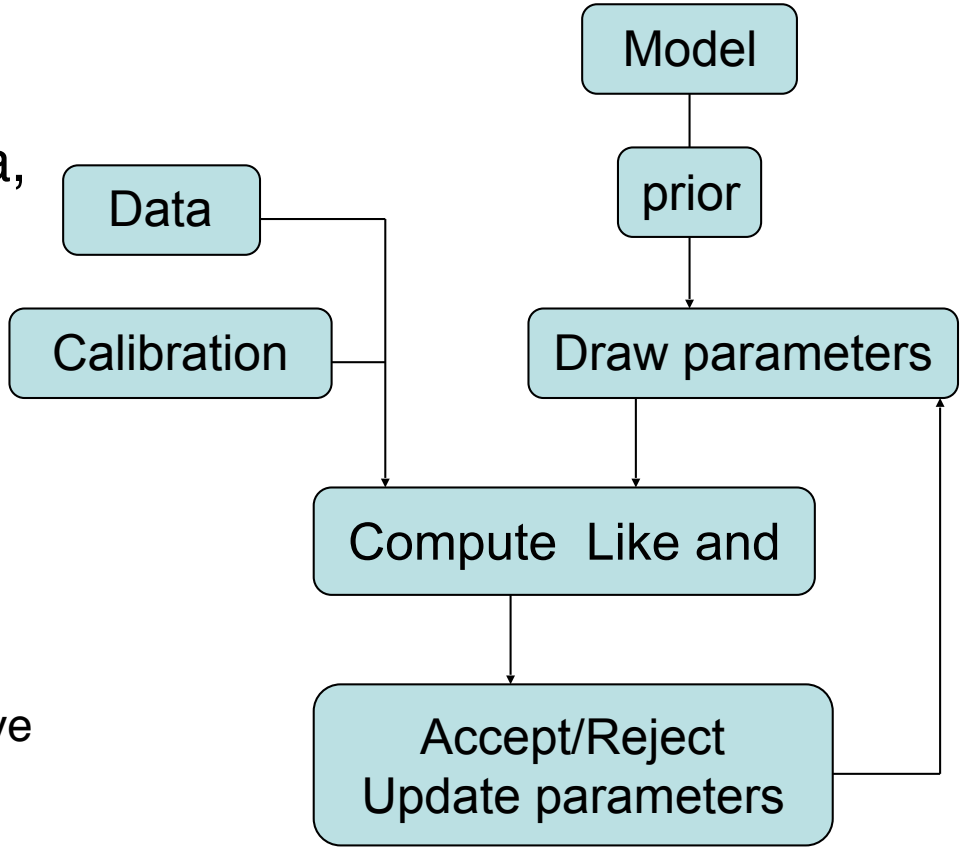
Loss of data

van Dyk et al. 2001

MCMC Simulations from the Posterior

Simulation from the posterior distribution requires careful and efficient algorithms:

Draw parameters from a "proposal distribution", compute likelihood and posterior probability of the "proposed" parameter value given the observed data, use a Metropolis-Hastings criterion to accept or reject the "proposed" values.

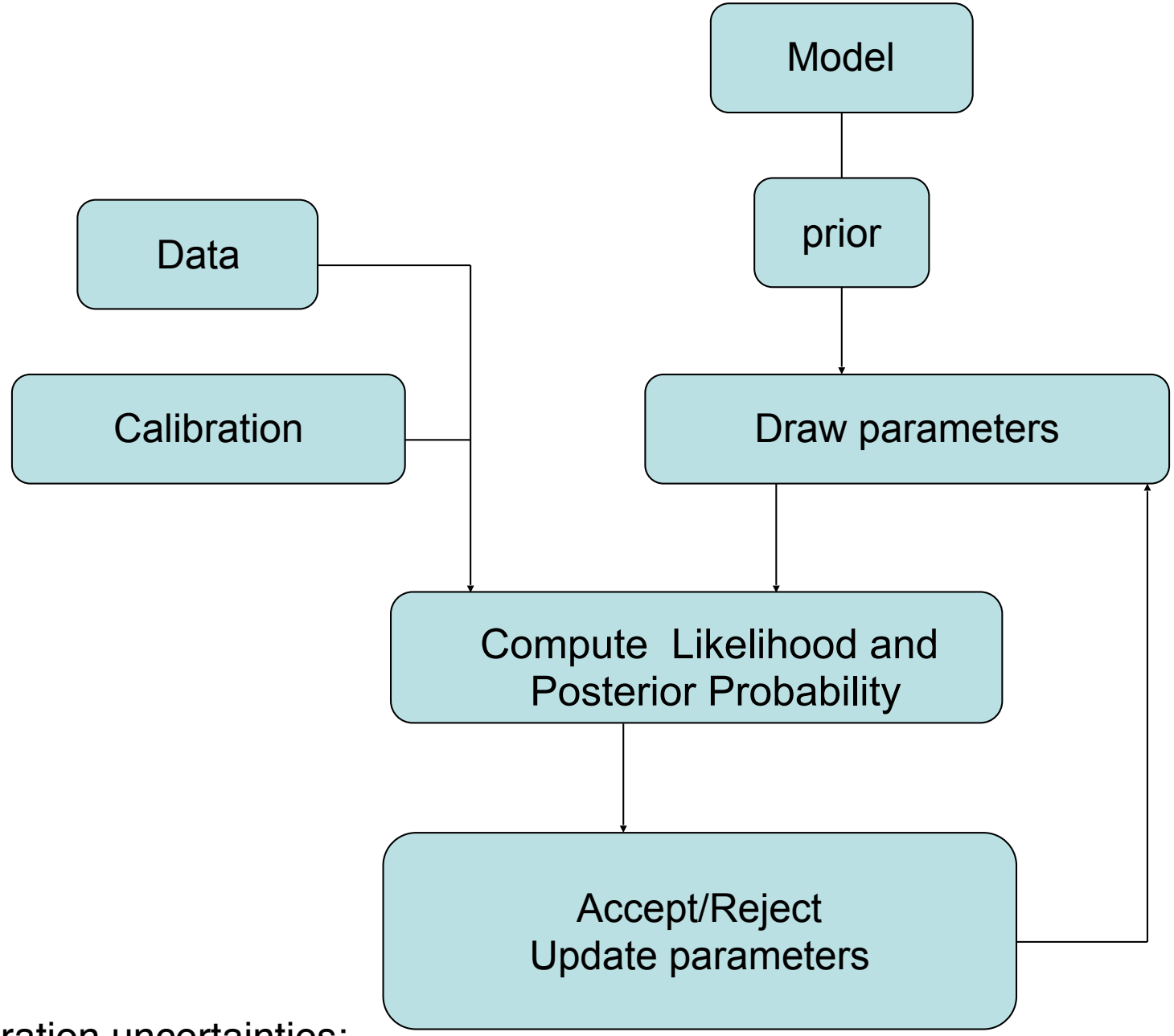


MCMC samplers:

✓ Explore parameter space and summarize the full posterior

✓Simulate replicate data from the posterior predictive distributions

✓Computed parameter uncertainties can include calibration errors.



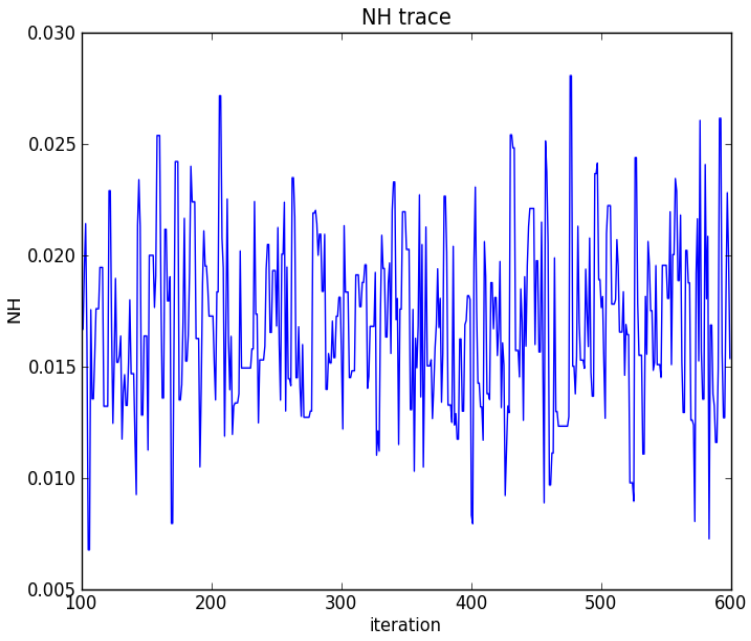
Can Include calibration uncertainties:

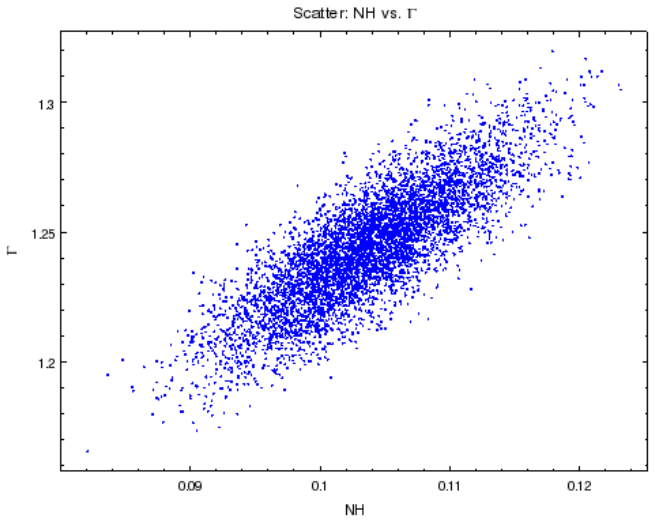
"A Fully Bayesian Method for Jointly Fitting Instrumental Calibration and X-Ray Spectral Models" Xu et al. 2014, ApJ, 794,97

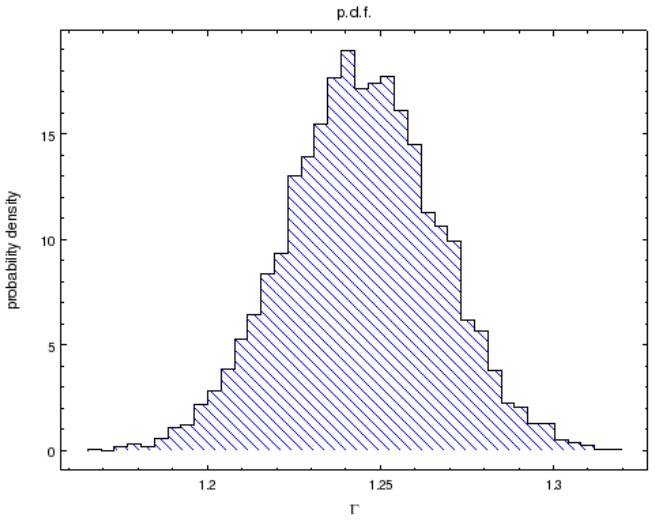
Checking the MCMC Results

Looking at MCMC - trace plots of statistics and parameters:

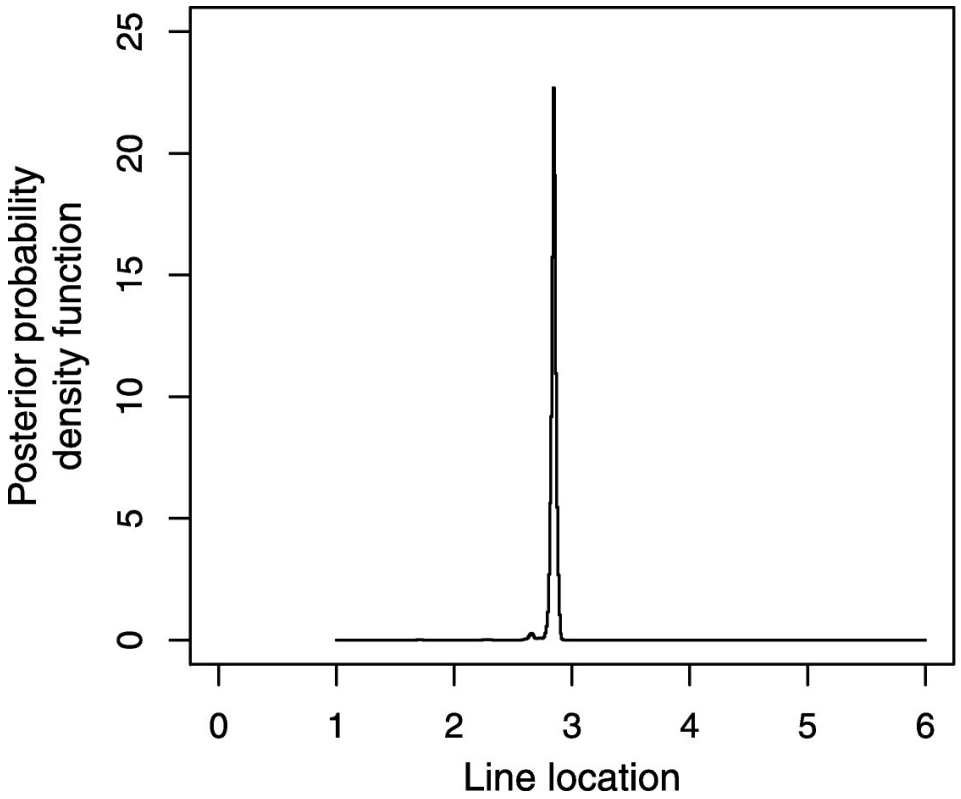
Trace of a parameter during MCMC run

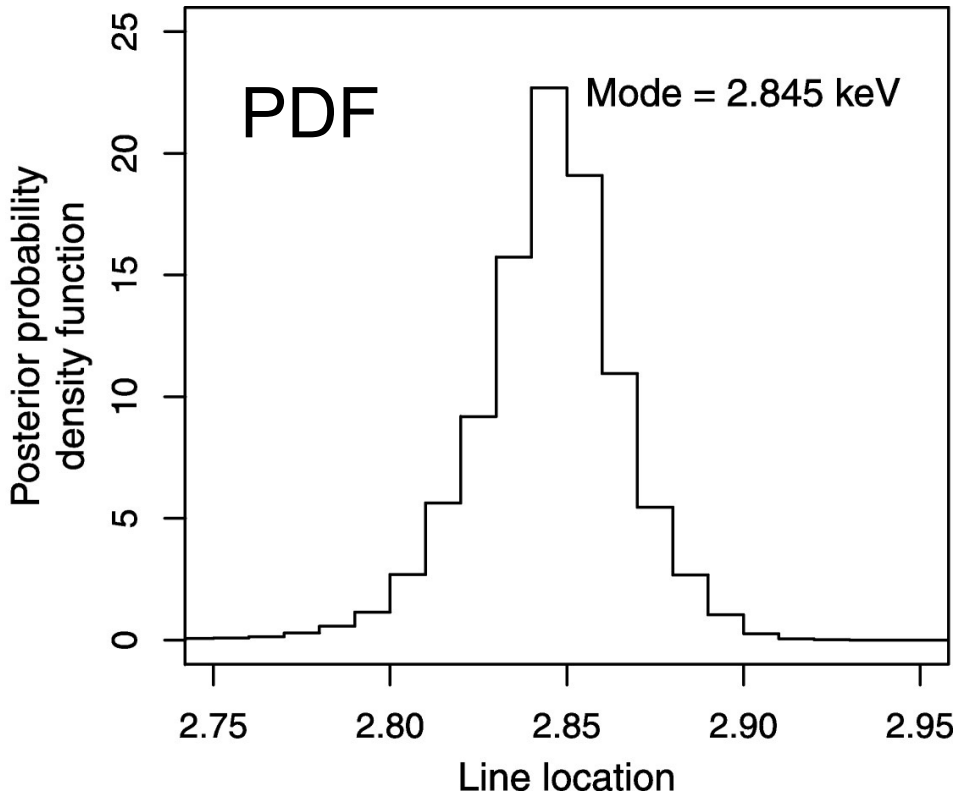






Posterior Probability



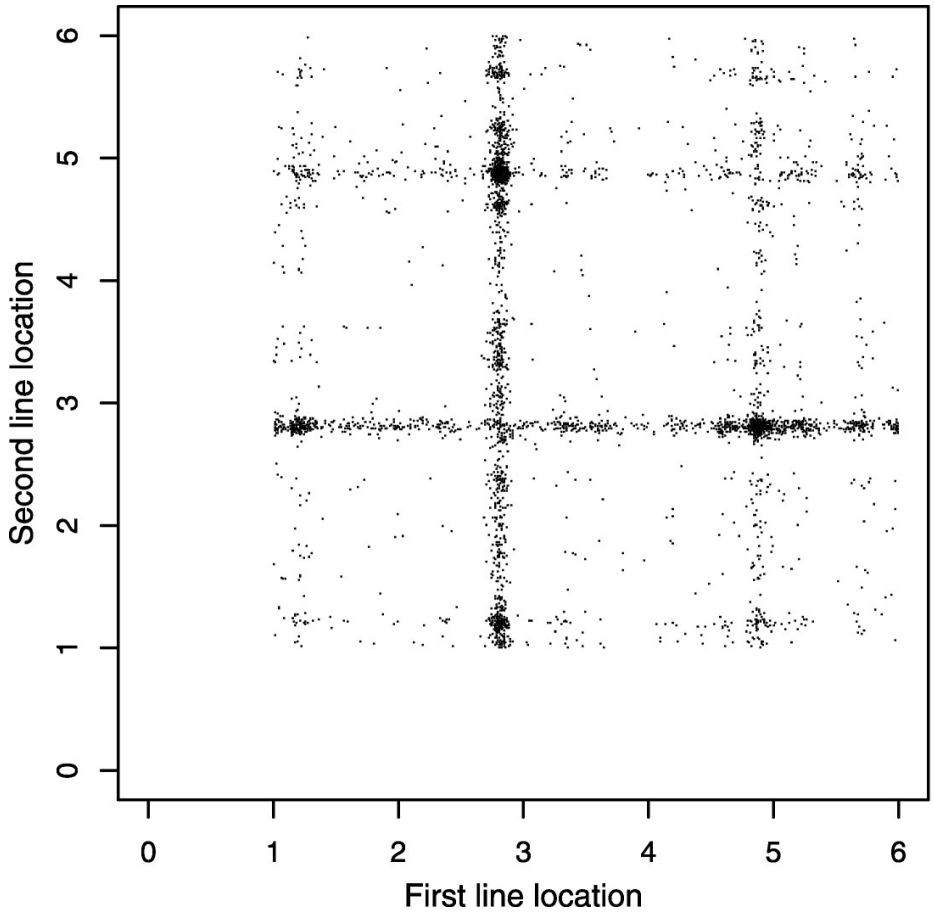


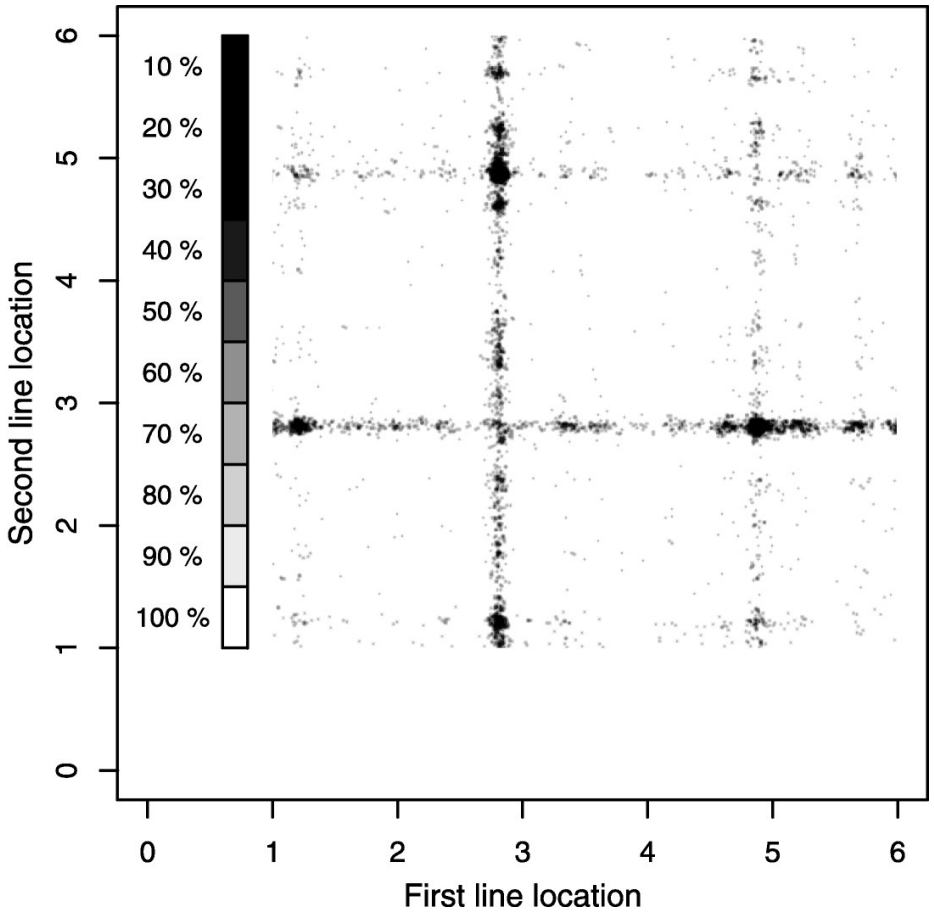
Park et al 2008

Joint Posterior Distribution

2D HPD

smoothed



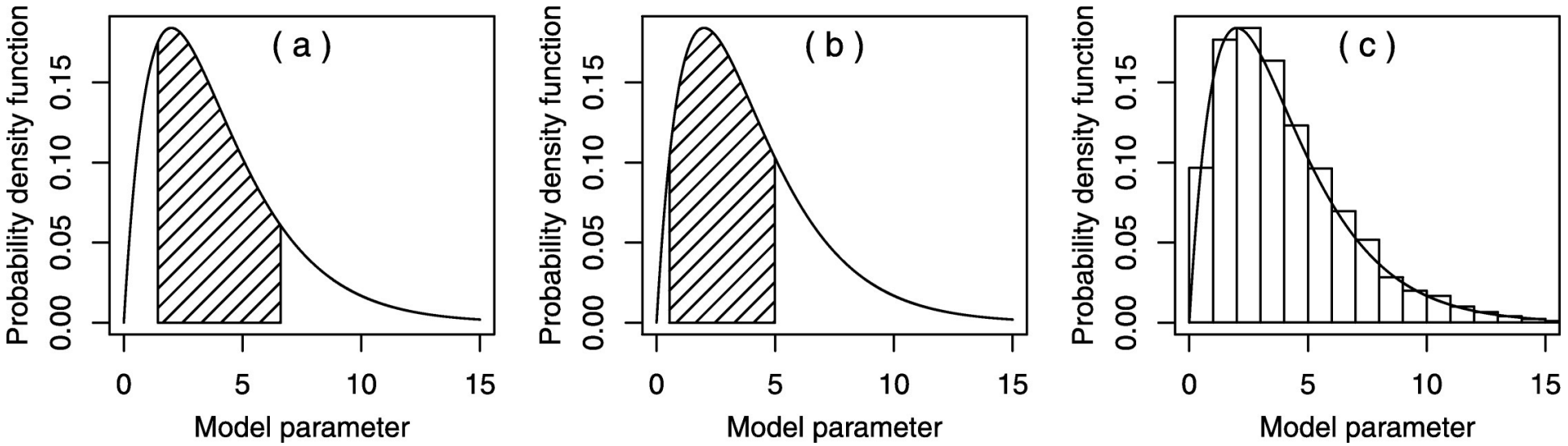


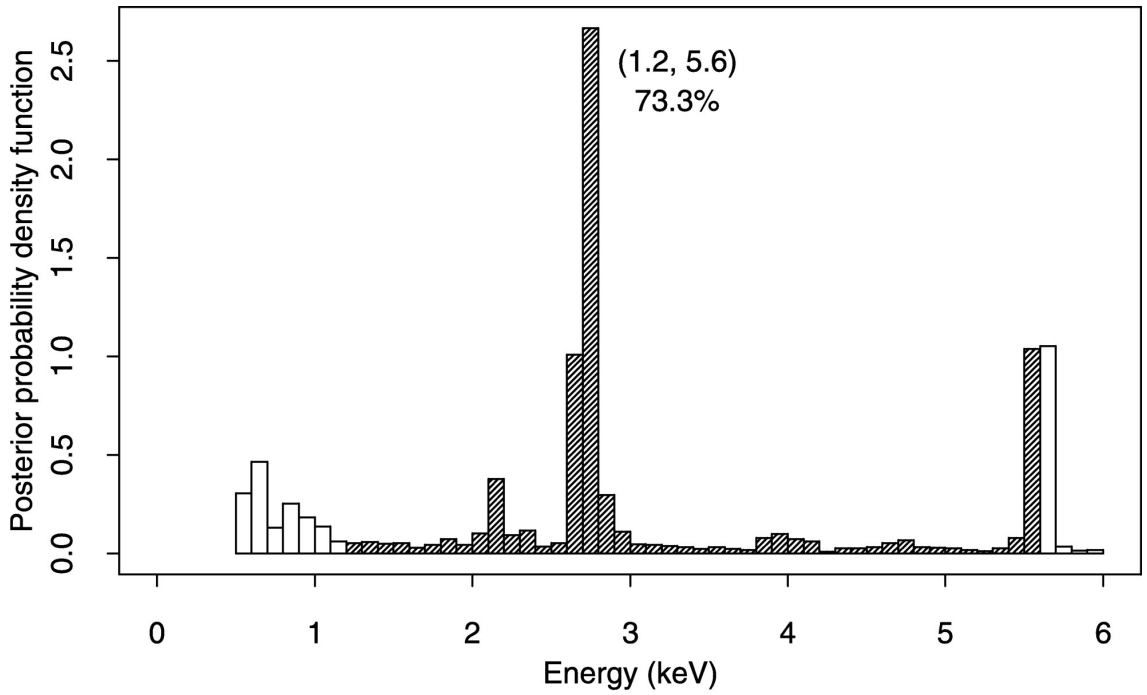
Summary Statistics

Equal tail 68%

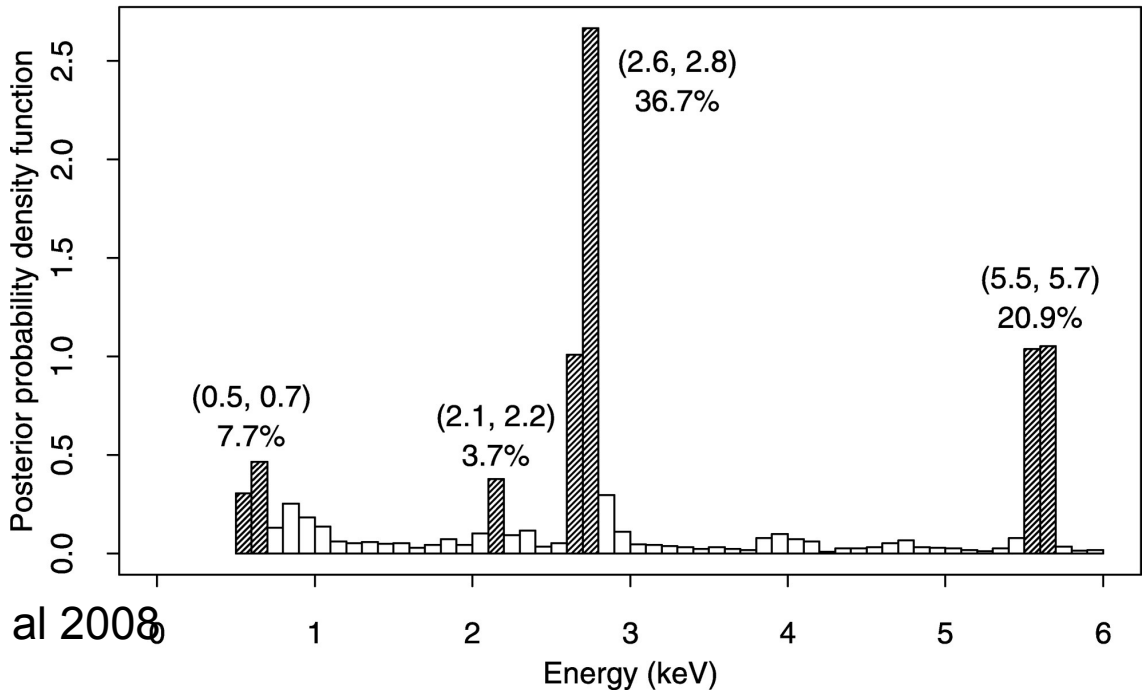
HPD 68%
Highest Posterior Density

Probability Density
and Monte Carlo



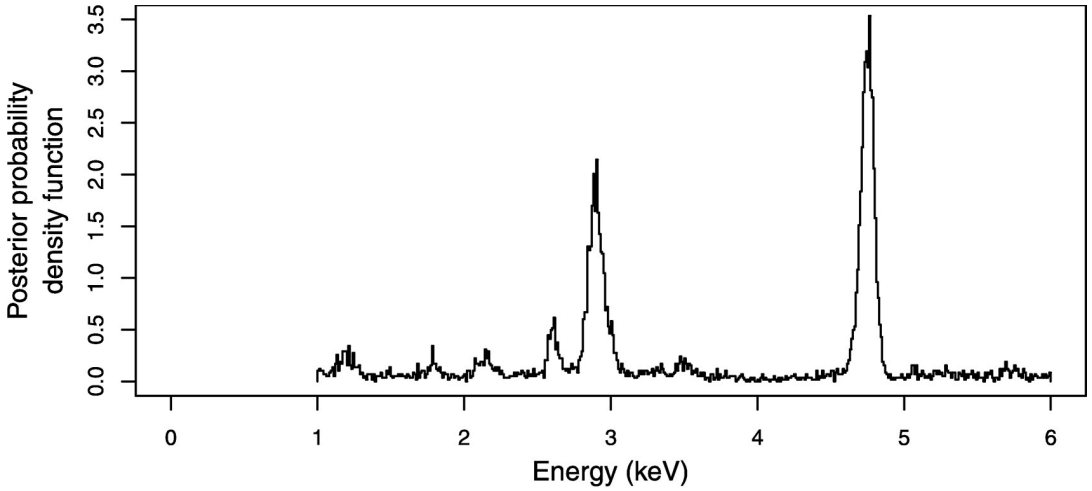


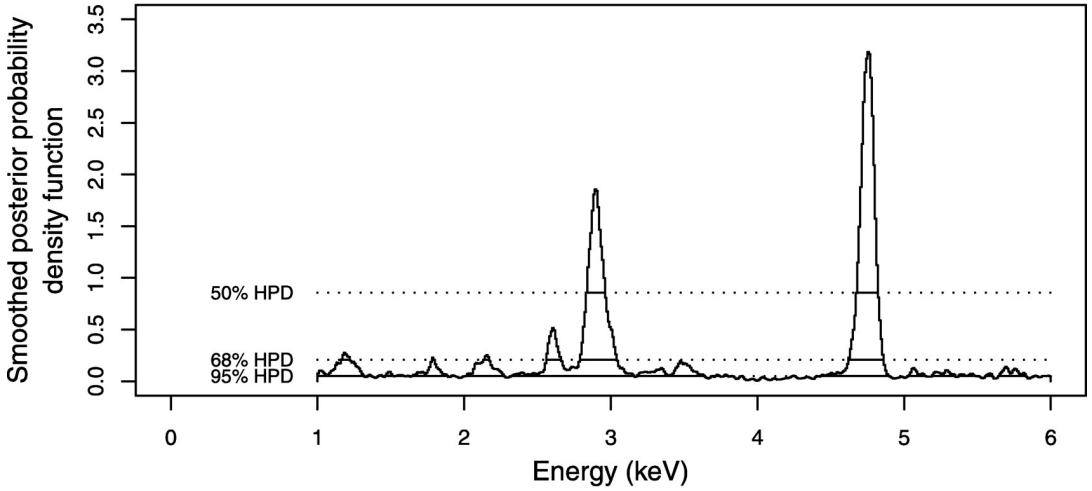
Equal-tail 68% Interval

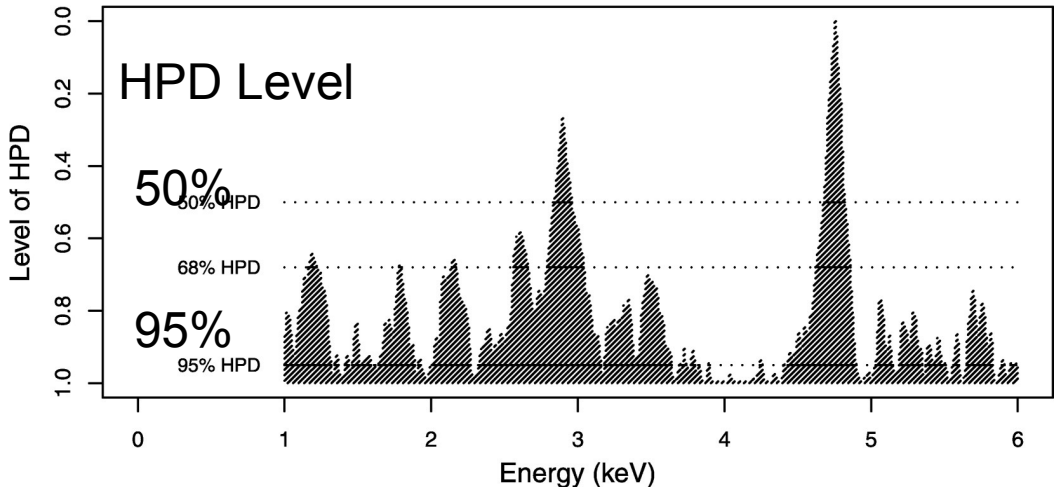


HPD 68% Region

Park et al 2008







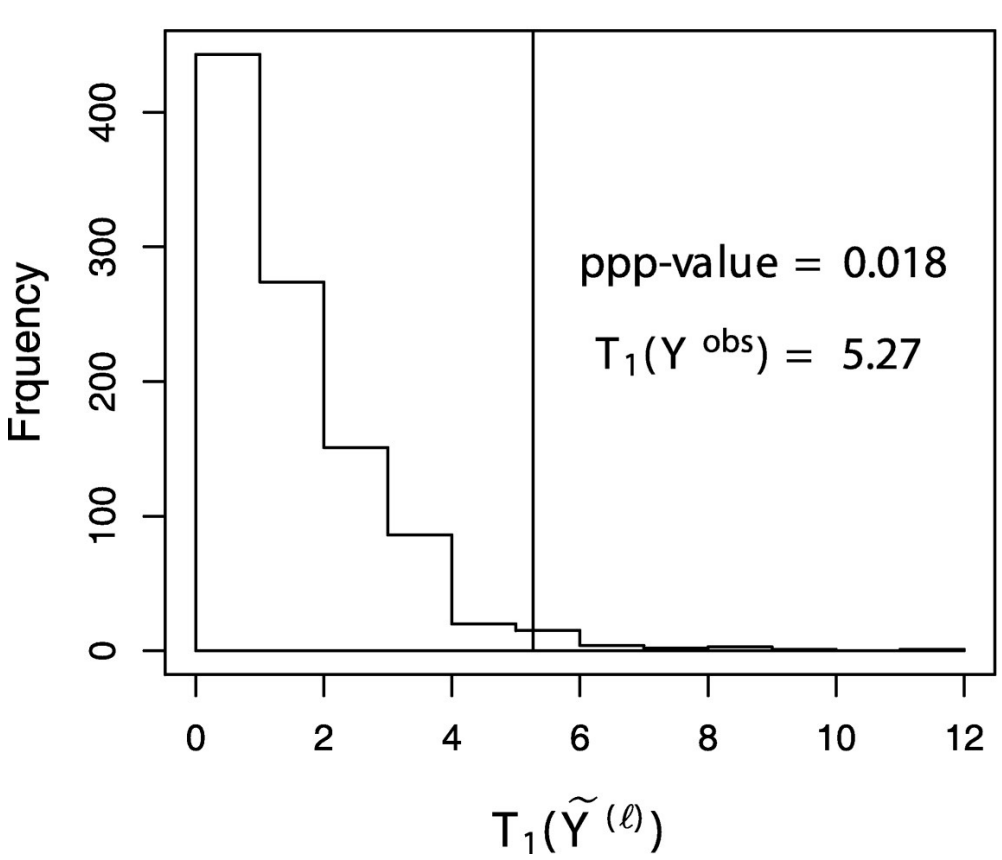
Park et al 2008

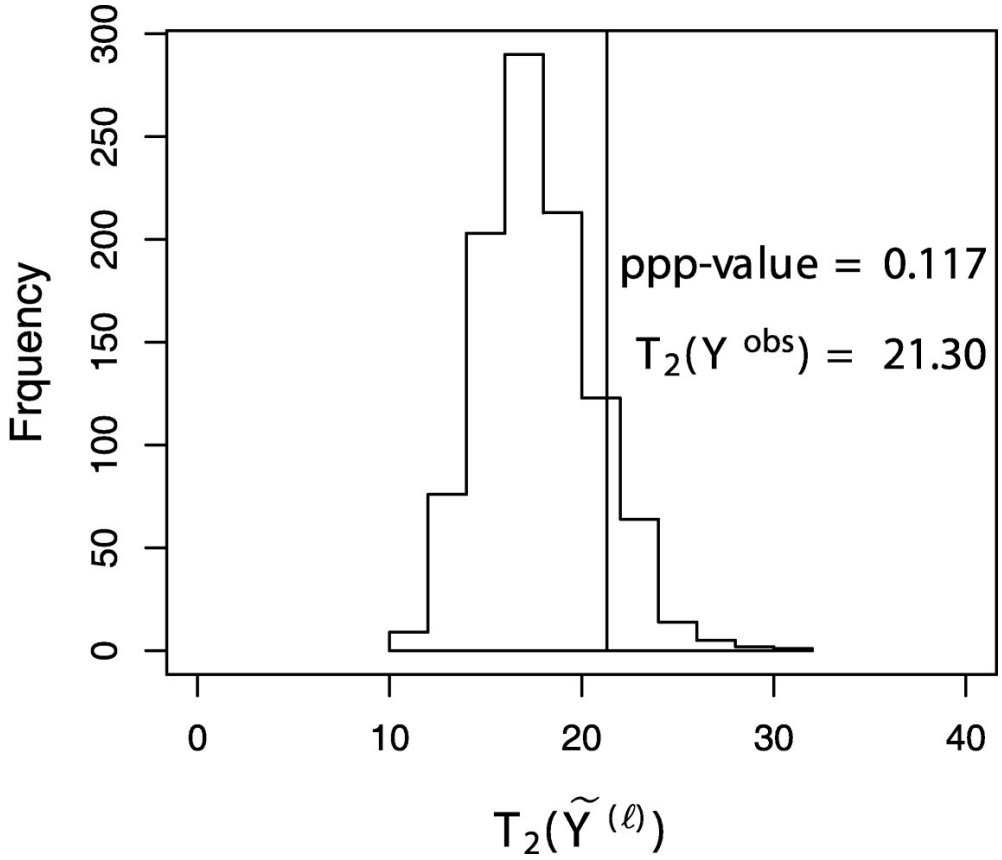
Model Checking

- posterior predictive checking
- posterior predictive p-value
- needs specified Test
- use p-value and magnitude of discrepancy

$$p_B = Pr(T(y^{rep}, \theta) \geq T(y, \theta)|y)$$

Posterior Predictive p-values





Park et al 2008
Protassov et al 2002

ppp - posterior predictive p-values

DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)

FREQUENTIST STATISTICIAN:

BAYESIAN STATISTICIAN:

Model Selections

- χ^2 - goodness of fit test
- F-test
- Likelihood Ratio Tests

Bayesian Model Selection

Bayes' theorem can be applied to model comparison:

$$p(M \mid D) = p(M)\,\frac{p(D \mid M)}{p(D)}.$$

- $p(M)$ is the prior probability for M;
- $p(D)$ is the normalization constant; and
- $p(D \mid M)$ is the average, or global, likelihood, evidence:

$$p(D \mid M) = \int d\theta\, p(\theta \mid M)\, p(D \mid M, \theta)$$
$$= \int d\theta\, p(\theta \mid M)\, \mathcal{L}(M, \theta).$$

It is the (normalized) integral of the posterior distribution over all parameter space.

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

$$DIC = -2\ln p(\mathrm{y}|\theta) + 2\mathrm{p_{DIC}}$$

Var[D(θ)] Model complexity

Bayesian Model Comparison

To compare two models, a Bayesian computes the odds, or odds ratio:

$$O_{21} = \frac{p(M_2|D)}{p(M_1|D)} = \frac{p(M_2)p(D|M_2)}{p(M_1)p(D|M_1)}$$
$$= \frac{p(M_2)}{p(M_1)} B_{21}$$

B_{21} is the **Bayes factor**.

When there is no *a priori* preference for either model, $B_{21} = 1$ indicates that each model is equally likely to be correct, while $B_{21} \geq 10$ may be sufficient to accept the alternative model (although that number should be greater if the alternative model is controversial).

Information Criteria

Defined based on deviance:

log predictive density of the data given the point estimate of the fitted model - $2\log p(y|M_{best})$

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.

$$AIC = -2\ln[L(M)] + 2k + \frac{2k(k+1)}{N-k-1}$$

finite sample correction

$-2\ln[L(M)]$: χ^2 - assuming Normality

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

Example

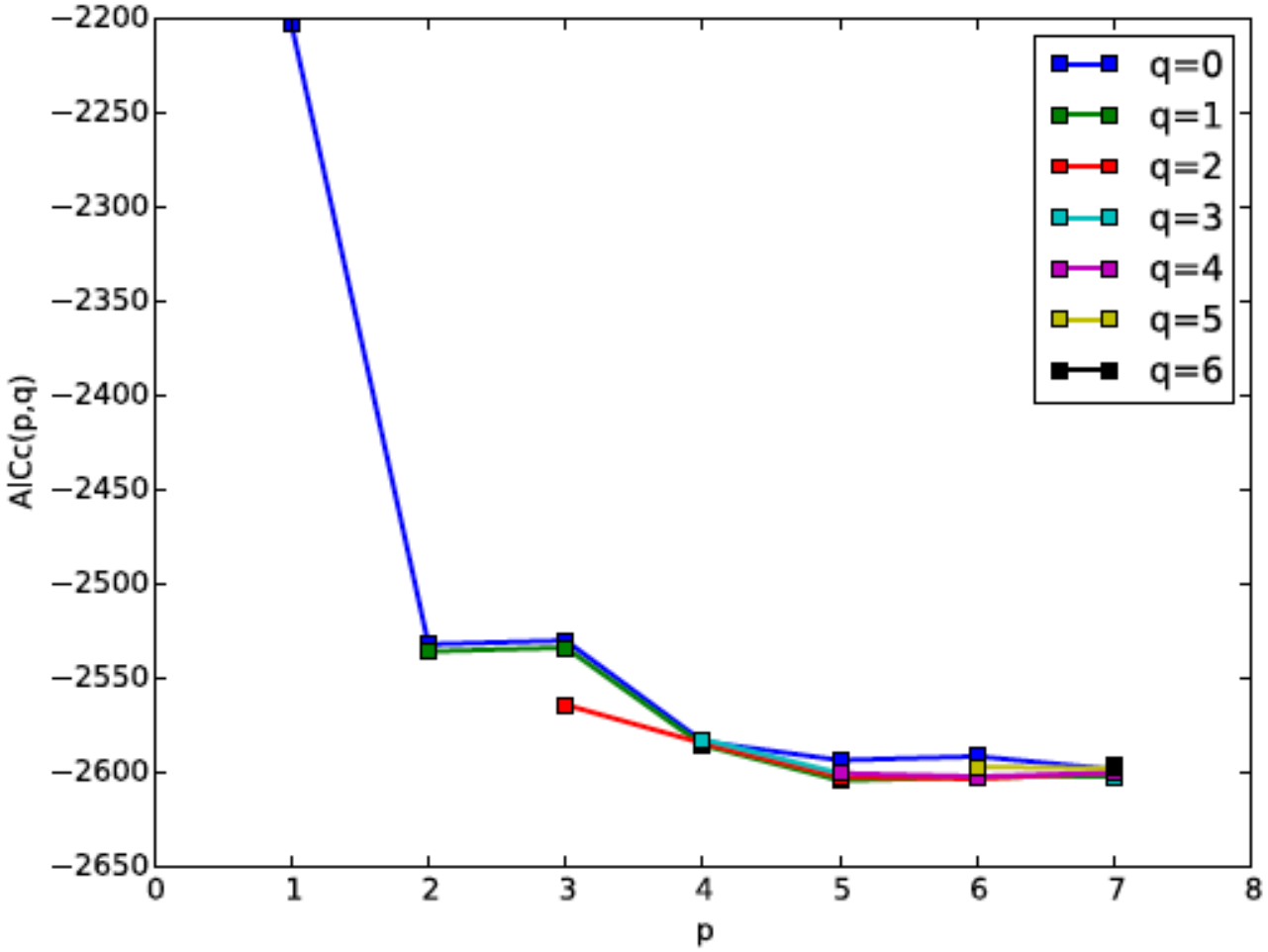


Figure 3. AICc values computed from the simulated light curve shown in Figure 1 for CARMA(p, q) models of order $p \leqslant 7, q < p$. The minimum AICc is achieved for the values $p = 5, q = 1$ although there is little change in the AICc for models of order $p \geqslant 5$.

Kelly et al 2014, ApJ 788,33

Applications Examples

Examples: PyBlocxs

Sherpa - 2D example

Example: LIRA

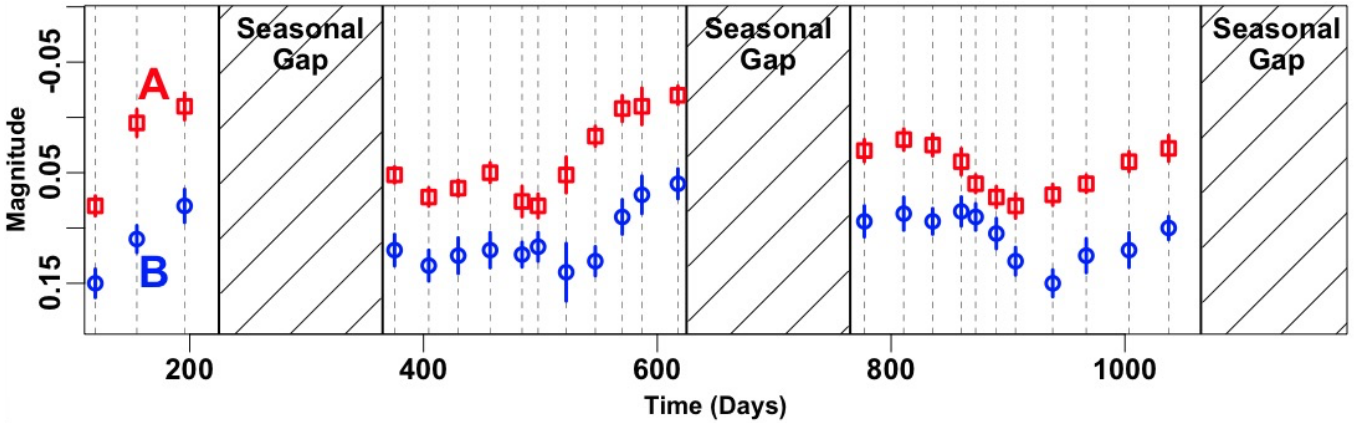
- https://github.com/astrostat/LIRA

Example: Modeling Time-delay

Tak et al 2016,arXiv:1602.01462

Data

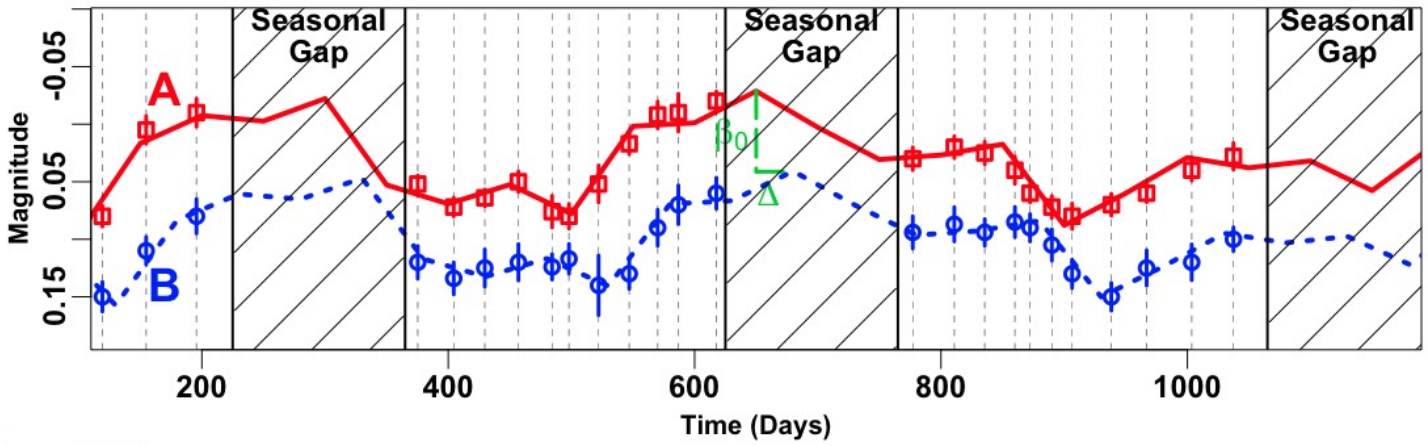
Simulated data of a doubly-lensed quasar



Data comprise of two time series with measurement errors.

- Observation times $\boldsymbol{t} \equiv \{t_1, t_2, \ldots, t_n\}^\top$
- Observed magnitudes $\boldsymbol{x(t)} \equiv \{x(t_1), x(t_2), \ldots, x(t_n)\}^\top$, and $\boldsymbol{y(t)}$
- Measurement errors (SD) $\boldsymbol{\delta(t)} \equiv \{\delta(t_1), \delta(t_2), \ldots, \delta(t_n)\}^\top$ and $\boldsymbol{\eta(t)}$

Our job is to estimate time delay (shift in x-axis) between two time series.



- Assumption 1: $\exists$ latent light curves representing the unobserved true magnitudes in continuous time (red and blue dashed curves).

  $\boldsymbol{X(t)} = (X(t_1), X(t_2), \ldots, X(t_n))^\top$ and $\boldsymbol{Y(t)}$, values on curves at $\boldsymbol{t}$

- Assumption 2 (Curve-Shifted model):

$$\boldsymbol{Y(t)} = \boldsymbol{X(t - \Delta)} + \beta_0,$$

  where the time delay Δ and magnitude offset β_0 are unknown.

Observed data: Independent Gaussian measurement errors

- $x(t_j) \mid X(t_j) \overset{\text{indep.}}{\sim} \mathrm{N}[X(t_j),\ \delta^2(t_j)]$
- $y(t_j) \mid Y(t_j) \overset{\text{indep.}}{\sim} \mathrm{N}[Y(t_j),\ \eta^2(t_j)]$

  $y(t_j) \mid X(t_j - \Delta), \Delta, \beta_0 \overset{\text{indep.}}{\sim} \mathrm{N}[X(t_j - \Delta) + \beta_0,\ \eta^2(t_j)].$

Latent data: Ornstein-Uhlenbeck (O-U)/Damped RW process for $\boldsymbol{X(\cdot)}$

- Kelly+ (2009), Kozlowski+ (2010), MacLeod+ (2010), Zu+ (2013) have supported the O-U process.
- $dX(t) = -\frac{1}{\tau}\big(X(t) - \mu\big)dt + \sigma dB(t)$, where τ is a mean-reversion time, μ is the overall mean, and σ is the short-term variability.
- O-U process is a Gaussian process with a Matern kernel.

Bayesian method

- Prior distributions for the model parameters; $\Delta, \beta_0, \mu, \sigma^2, \tau$
- Metropolis-Hastings within Gibbs sampler
- Pros: Complete investigation on all the model parameters
- Cons: Inefficient when $\exists$ strong correlation and multimodality.

Example: Addition of the Emission Line

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 the related *F*-test, popularized in astrophysics by Eadie and coworkers in 1971, Bevington in 1969, Lampton, Margon, & Bowyer, in 1976, Cash in 1979, and Avni in 1978, do not (even asymptotically) adhere to their nominal χ^2 and *F*-distributions in many statistical tests common in astrophysics, thereby casting many marginal line or source detections and nondetections into doubt. Although the above authors illustrate the many legitimate uses of these statistics, in some important cases it can be impossible to compute the correct false positive rate. For examp use the LRT or the *F*-test to detect a line in a spectral model or a source certain required regularity conditions. (These applications were not or

Need to
Calibrate Test Statistics

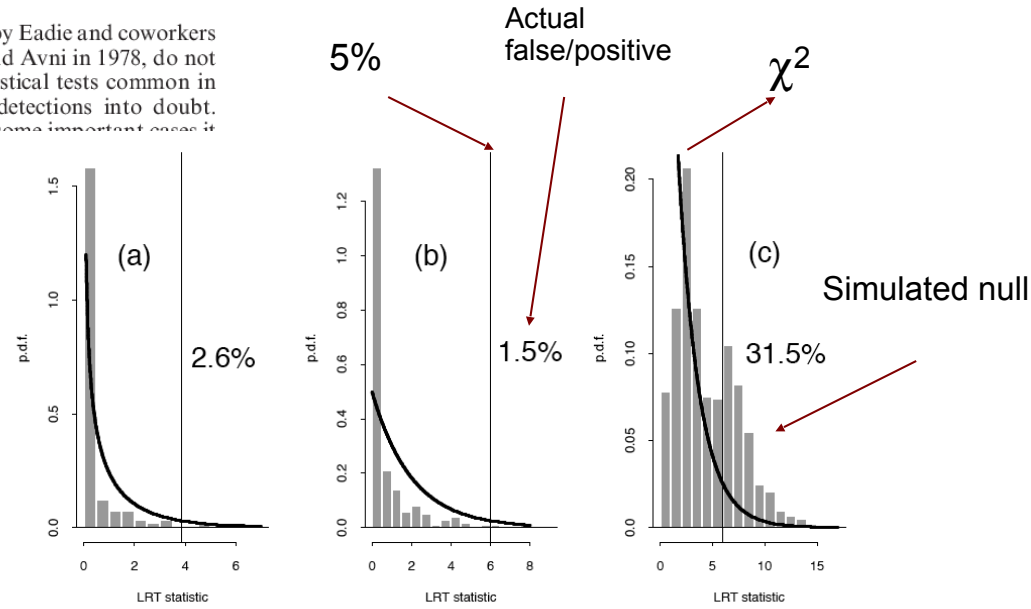


FIG. 1.—Null distribution of the LRT test statistic. The histograms illustrate the simulated null distribution of the LRT statistic in three scenarios and should be compared with nominal χ^2 distributions, which are also plotted. As detailed in § 3.2, the histograms corresponds to (*a*) testing for a 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.

Simulations to Calibrate Test Statistics

Protassov et al 2002, ApJ, 571, 545

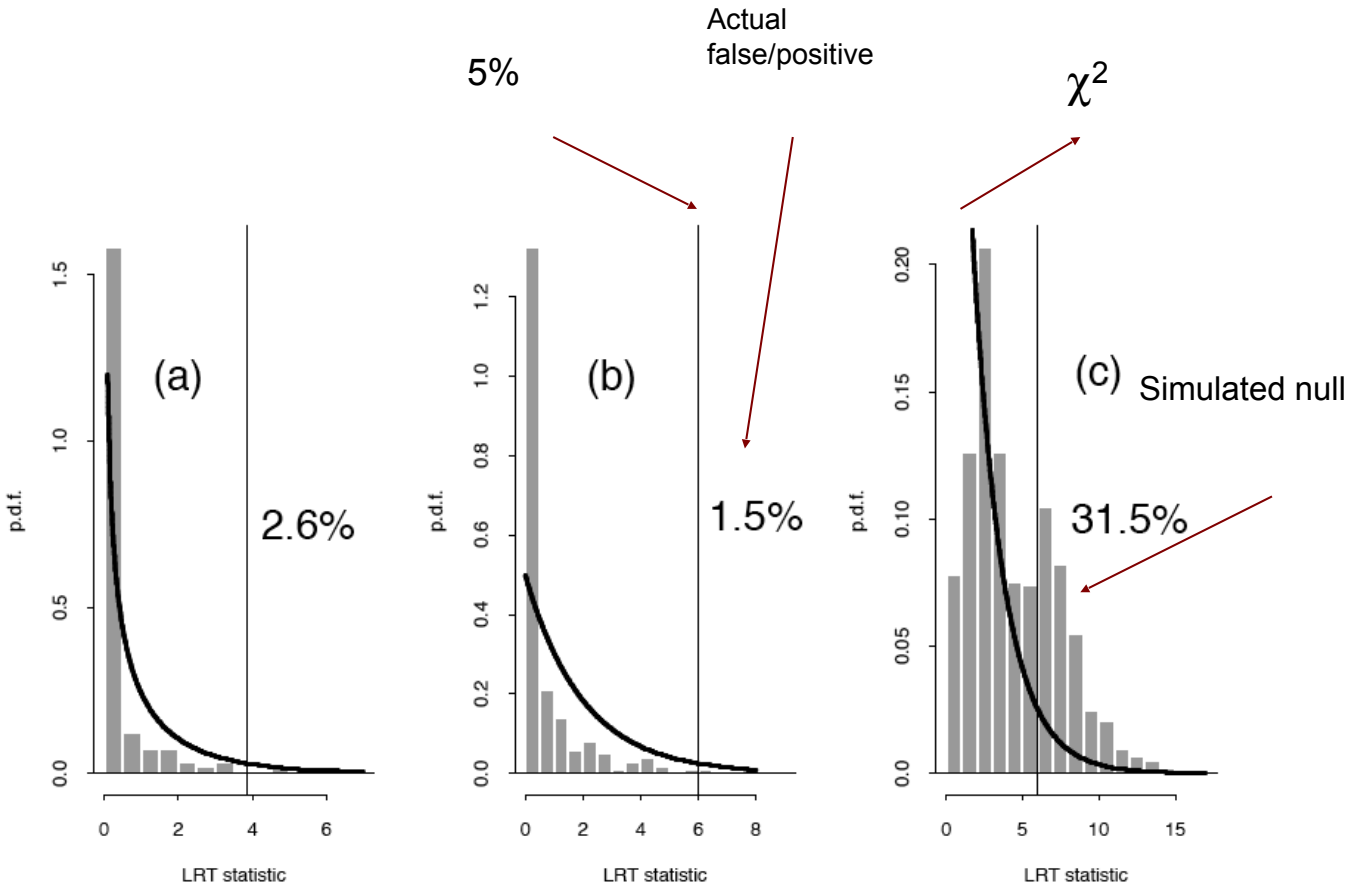
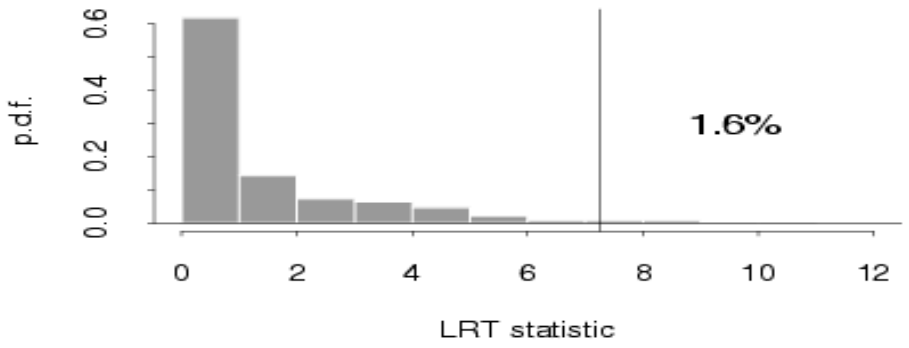


FIG. 1.—Null distribution of the LRT test statistic. The histograms illustrate the simulated null distribution of the LRT statistic in three scenarios and should be compared with nominal χ^2 distributions, which are also plotted. As detailed in § 3.2, the histograms corresponds to (*a*) testing for a 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.

Monte Carlo Simulations

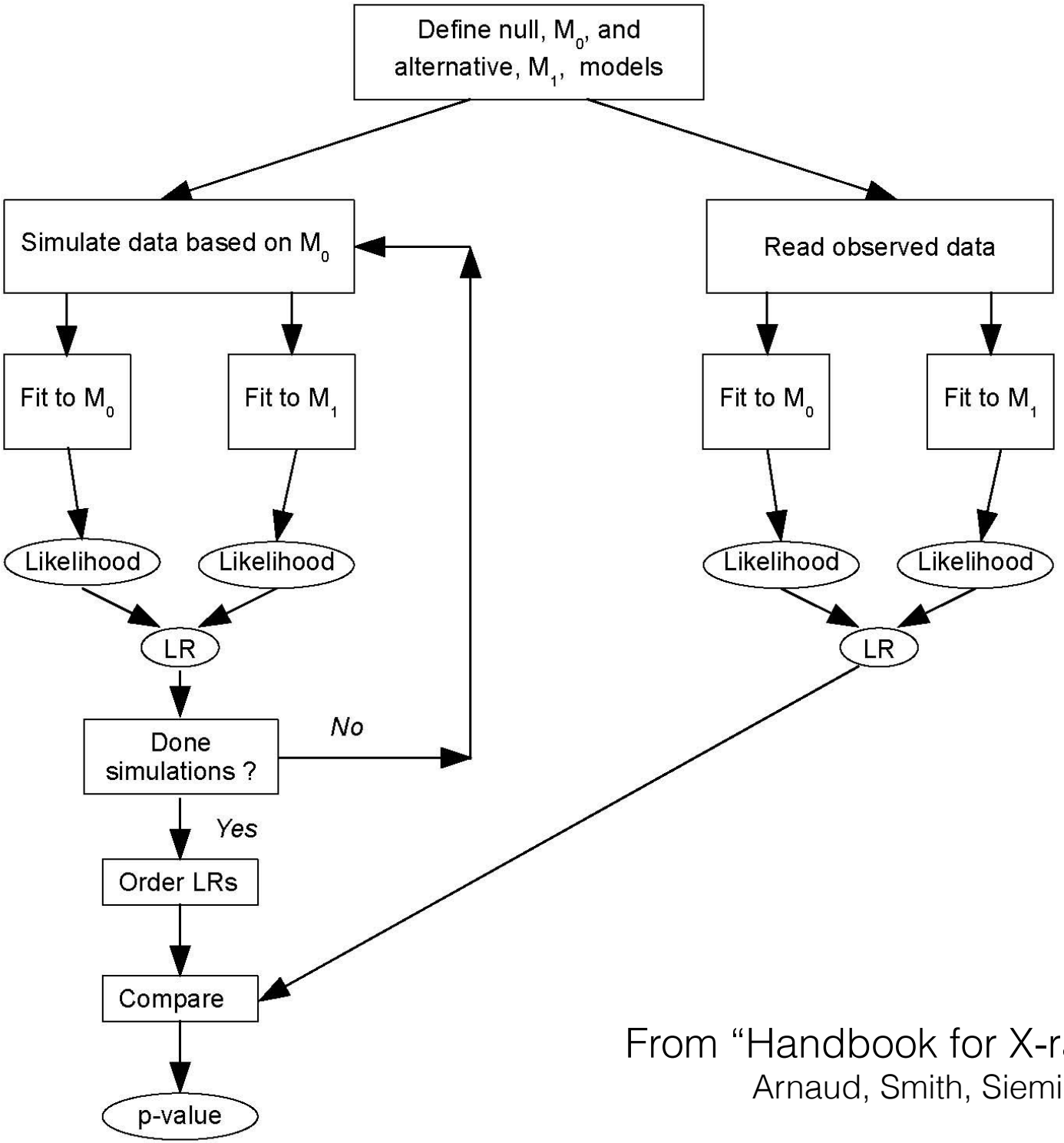
- Simulations to test for more complex models, e.g. addition of an emission line
- Steps:
  - Fit the observed data with both models, M_0, M_1
  - Obtain distributions for parameters
  - Assume a simpler model M_0 for simulations
  - Simulate/Sample data from the assumed simpler model
  - Fit the simulated data with simple and complex model
  - Calculate statistics for each fit
  - Build the probability density for assumed comparison statistics, e.g. LRT and calculate p-value

Example:
Visualization, here accept more complex model, p-value 1.6%



Simulations Steps

1. Fit two models:
   - continuum
   - continuum + line
2. Define the test statistic (LRT) and calculate for the data
3. Simulate the data with a simpler model
4. Fit simulated data with both models
5. Calculate the test statistics
6. Repeat steps 3-5 for large number of iterations
7. Plot the distribution
8. Compare the observed and simulated test statistics.



From “Handbook for X-ray Astronomy”

Arnaud, Smith, Siemiginowska

Steps in Hypothesis Testing

1/ Set up 2 possible exclusive hypotheses - two models:

M0 – null hypothesis – formulated to be rejected
M1 – an alternative hypothesis, research hypothesis

2/ Specify a priori the significance level α

3/ Choose a test which:
- has the required power β
- approximates the conditions
- finds what is needed to obtain the sampling distribution and the region of rejection, whose area is a fraction of the total area in the sampling distribution

3/ Run test: reject *M0* if the test yields a value of the statistics whose probability of occurrence under *M0* is < α

Classical Test Statistics

- Likelihood Ratio Test

  Ratio of likelihood values:

$$LRT = 2[\ln P(D|M_1) - \ln P(D|M_0)]$$

- F-test

  For Gaussian data the statistic follows F distribution

$$F_\chi = \frac{[\chi^2(m) - \chi^2(m+1)]}{[\chi^2(m)/(N-m-1)]} = \frac{\Delta\chi^2}{\chi^2_\nu}$$

- Tests only valid if
  - The models are nested
  - Not on the boundary of the parameter space
  - Asymptotic limit has been reached