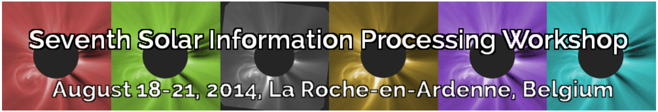
Bayesian Analysis of Power Law Models in High-Energy Astrophysics and in Solar Physics

David A. van Dyk

Statistics Section, Imperial College London



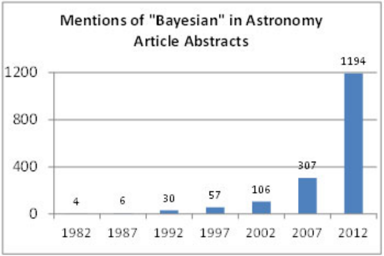
Imperial College London

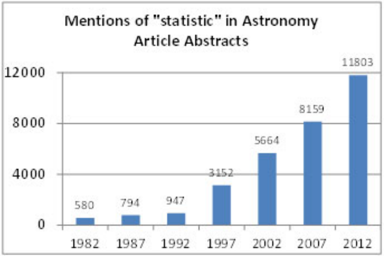
Outline

Outline

Bayesian Renaissance in Astronomy

The use of Statistical Methods in general and Bayesian Methods in particular is growing exponentially in Astronomy.





Source: `http://magazine.amstat.org/blog/2013/12/01/science-policy-intel/`

Imperial College London

Why Use Bayesian Methods?

Advantages of likelihood-based methods:

- Directly model complexities of sources and instruments.
- Allows science-driven modeling. *(Not just predictive modeling.)*
- Combine multiple information sources and/or data streams.
- Allow hierarchical or multi-level structures in data/models.
- Bayesian methods have clear mathematical foundations and can be used to derive principled statistical methods.
- Sophisticated computational methods available.

Challenges:

- Require us to specify “prior distributions” on unknown model parameters.

Bayesian Statistical Analyses: Likelihood

- Many methods based on χ^2 or Gaussian assumptions.
- Bayesian/Likelihood methods easily incorporate more appropriate distributions.
- E.g., for count data, we use a Poisson likelihood:

$$\chi^2 \text{ fitting:} \quad -\sum_{\text{bins}} \frac{(Y_i - \lambda_i)^2}{\sigma_i^2}$$

$$\text{Gaussian Loglikelihood:} \quad -\sum_{\text{bins}} \sigma_i - \sum_{\text{bins}} \frac{(Y_i - \lambda_i)^2}{\sigma_i^2}$$

$$\text{Poisson Loglikelihood:} \quad -\sum_{\text{bins}} \lambda_i + \sum_{\text{bins}} Y_i \log \lambda_i$$

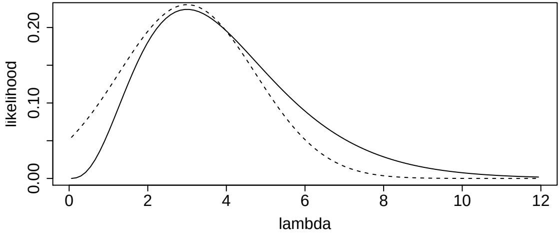
... or Pareto distribution for continuous data following a power law.

Bayesian Statistical Analyses: Likelihood

Likelihood Functions: Distribution of data given model parameters. Single bin detector: $Y \sim \text{Poisson}(\lambda_S)$:

$$\text{likelihood}(\lambda_S) = e^{-\lambda_S}\lambda_S^Y / Y! \qquad \text{loglikelihood}(\lambda_S) = -\lambda_S + Y\log(\lambda_S)$$

Maximum Likelihood Estimation: Suppose $Y = 3$



The likelihood and its normal approximation.

Can estimate λ_S and its error bars.

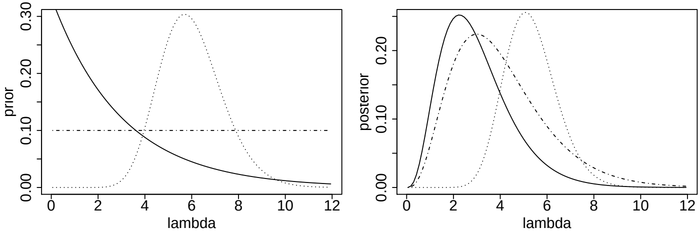
Bayesian Analyses: Prior and Posterior Dist'ns

Prior Distribution: Knowledge obtained *prior* to current data.

Bayes Theorem and Posterior Distribution:

$$\begin{aligned}\text{posterior}(\lambda) &\propto \text{likelihood}(\lambda) \times \text{prior}(\lambda)\\ p(\lambda|Y) &= p(Y|\lambda)p(\lambda)/p(Y)\end{aligned}$$

Combine past and current information:



Bayesian analyses rely on probability theory

Multi-Level Models

Example: Background contamination in a single bin detector

- Contaminated source counts: $Y = Y_S + Y_B$
- Background counts: X
- Background exposure is 24 times source exposure.

A Poisson Multi-Level Model:

LEVEL 1: $Y|Y_B, \lambda_S \overset{\text{dist}}{\sim} \text{Poisson}(\lambda_S) + Y_B$,

LEVEL 2: $Y_B|\lambda_B \overset{\text{dist}}{\sim} \text{Pois}(\lambda_B)$ and $X|\lambda_B \overset{\text{dist}}{\sim} \text{Pois}(\lambda_B \cdot 24)$,

LEVEL 3: specify a prior distribution for λ_B, λ_S.

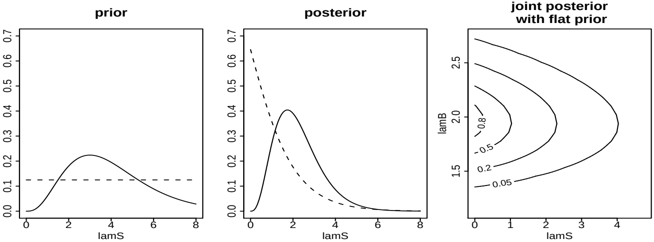
Each level of the model specifies a dist'n given unobserved quantities whose dist'ns are given in lower levels.

Posterior and Marginal Posterior Distributions

Summarizing the posterior distribution:

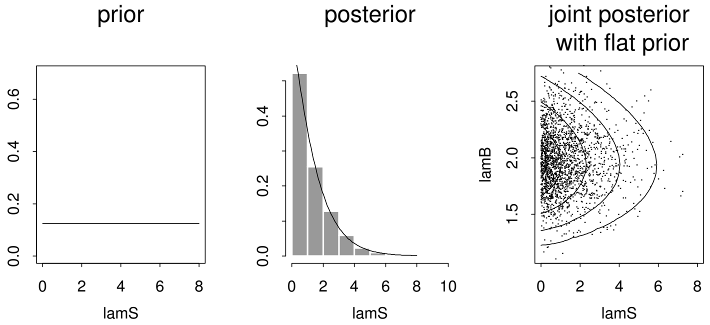
- We can plot the contours of the posterior distribution.
- Plot the marginal distributions of the parameters of interest:

$$p(\lambda_S \mid Y, Y_B) = \int p(\lambda_S, \lambda_B \mid Y, Y_B) d\lambda_B$$



Markov Chain Monte Carlo

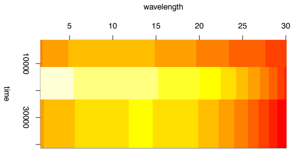
Exploring the posterior distribution via Monte Carlo.

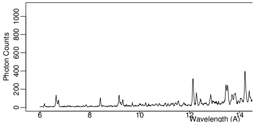


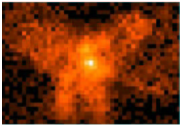
Easily generalizes to higher dimensions.

Outline

Science and Data







The Chandra X-Ray Observatory

- Images $> 30\times$ sharper then any previous X-ray telescope.
- X-rays are produced by multi-millions degree matter, e.g., by high magnetic fields, extreme gravity, explosive forces.

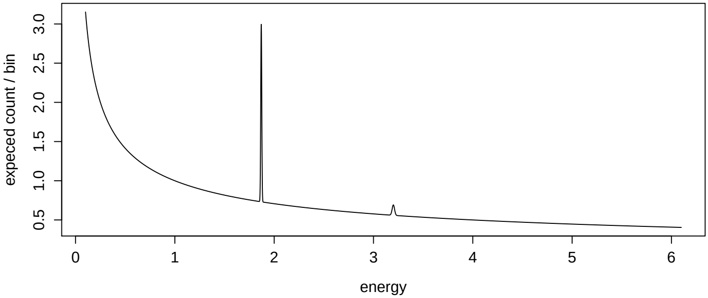
Data is collected for each arriving photon:

- Two-dimensional sky coordinates, energy, and arrival time
- High resolution discrete variables:
  e.g., 4096×4096 spatial and 1024 spectral bins
- Four-way table of photon counts.

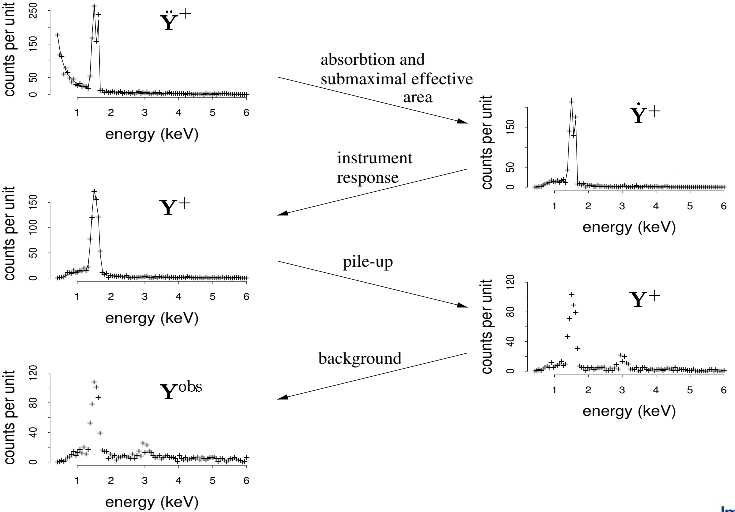
A Basic Spectral Models

Photon counts modeled with Poisson process:

1. The *continuum* indicates the temperature of the source.
2. *Emission* and *absorption lines* gives clues to composition.



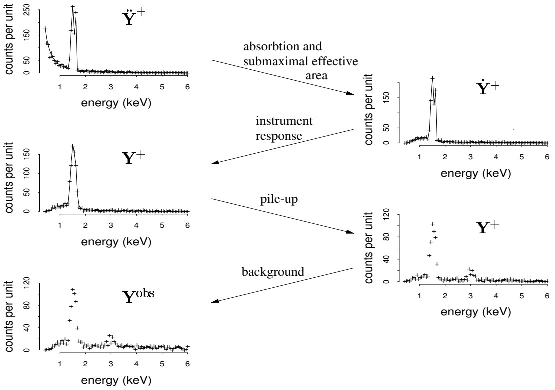
Multi-Level Models: X-ray Spectral Analysis[1]



---

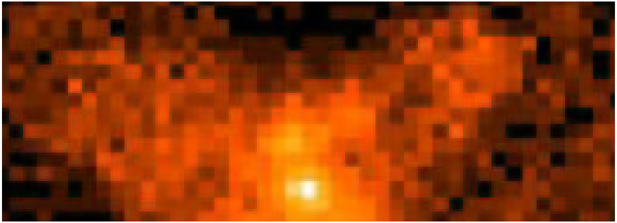
[1] van Dyk, Connors, Kashyap and Siemiginowska (2001). Analysis of energy spectra with low photon counts via Bayesian posterior simulation. *The Astrophysical Journal*, **548**, 224-243.

Modeling Data Collection Mechanism



- *We can separate a complex problem into a sequence of easier-to-solve problems.*
- *Model source, absorption, instrumental effects, and background separately.*

What About Prior Distributions?

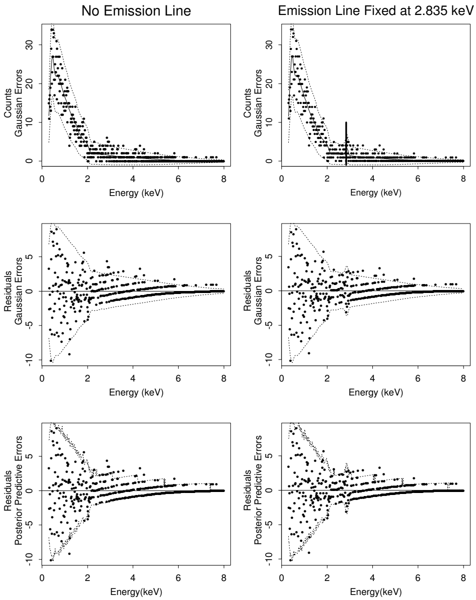


We can often use “objective prior distributions"

1. Priors can be used
   - to incorporate information from outside the data, or
   - to impose structure on the fitted model.[2]
2. Priors offer a principled compromise between “fixing” a parameter & letting it “float free”.
3. The common practice of setting `min` and `max` limits amounts to using a flat prior over a specified range.

[2] Esch, Connors, Karovska, and van Dyk (2004). An image reconstruction technique with error estimates. *The Astrophysical Journal*, **610** 1213-1227.

Model Diagnostics (e.g., van Dyk and Kang, 2004)



Bayesian methods can incorporate specific error characteristics of data models:

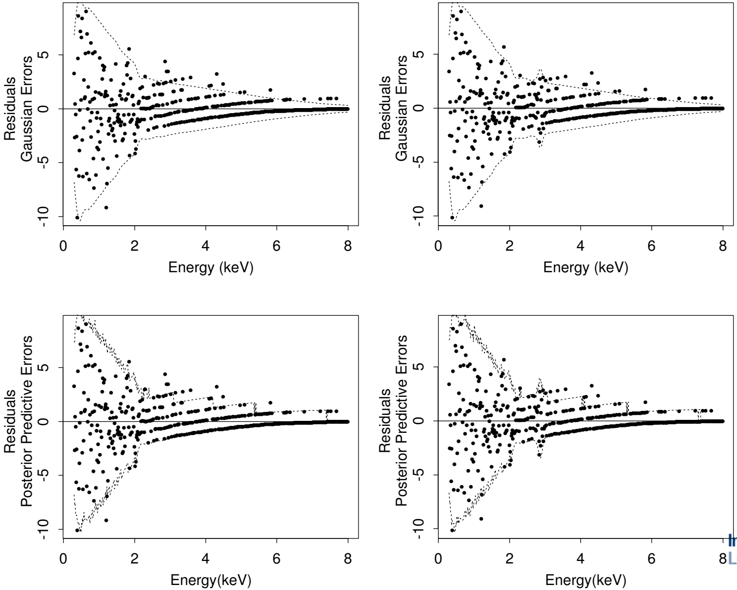
Compare

- Gaussian Errors
- Posterior Predictive Errors.

Posterior Predictive Dist'n:

$$p(Y_{\text{rep}} \mid Y) = \int p(Y_{\text{rep}} \mid \theta)\, p(\theta \mid Y) d\theta$$

Model Diagnostics (e.g., van Dyk and Kang, 2004)



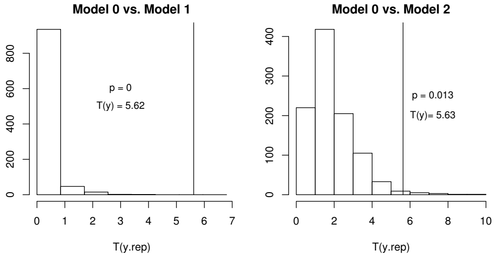
Posterior Predictive Checks: Is there a line?[3]

Model 0: no line **Model 1:** known location **Model 2:** unknown location

- The Likelihood Ratio Test:

$$T(Y_{\text{rep}}) = \log\left\{\frac{\sup_{\theta\in\Theta_i} L(\theta|Y_{\text{rep}})}{\sup_{\theta\in\Theta_0} L(\theta|y_{\text{rep}})}\right\}, \quad i = 1, 2,$$

- Sample Y_{rep} from posterior predictive dist'n under *Model 0*.



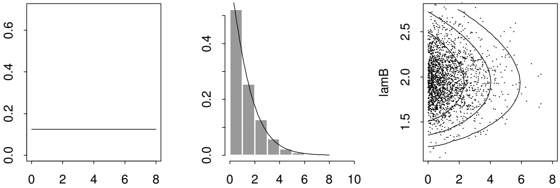
Knowing line location increases strength of evidence.

[3] Protassov, van Dyk, Connors, Kashyap, and Siemiginowska (2002). Statistics: Handle with care — detecting multiple model components with the likelihood ratio test, *The Astrophysical Journal*, **571** 545–559.

Imperial College London

Outline

(Markov Chain) Monte Carlo



- Goal: obtain a sample from the posterior distribution of θ.
- The sample may be independent or dependent.
- Markov chains can be used to obtain a dependent sample.
- Given $\theta^{(0)}$, sample

$$\theta^{(t)} \sim \mathcal{K}(\theta|\theta^{(t-1)}) \text{ for } t = 1, 2, \ldots$$

The Metropolis Sampler

Draw $\theta^{(0)}$ from some starting distribution.

For $t = 1, 2, 3, \ldots$

Sample: $\theta^* = \theta^{(t-1)} +$ random noise

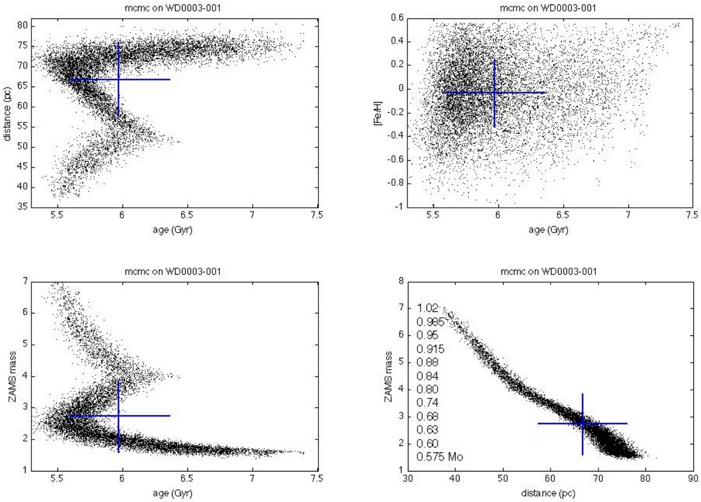
Compute: $r = \frac{p(\theta^* | Y)}{p(\theta^{(t-1)} | Y)}$

Set: $\theta^{(t)} = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \\ \theta^{(t-1)} & \text{otherwise} \end{cases}$

Note

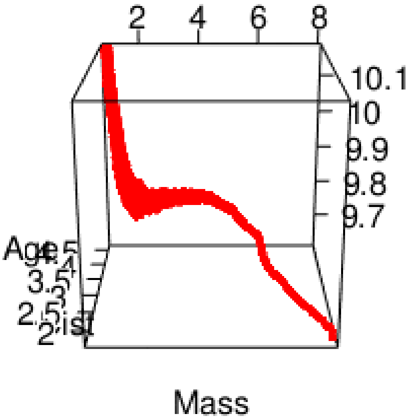
- Random noise must be symmetric,
  e.g., Gaussian or uniform distribution centered at zero.
- If $p(\theta^* | Y) > p(\theta^{(t-1)} | Y)$, *jump!*

Complex Posterior Distributions I



Cannot be summarized with fitted value and error bars.

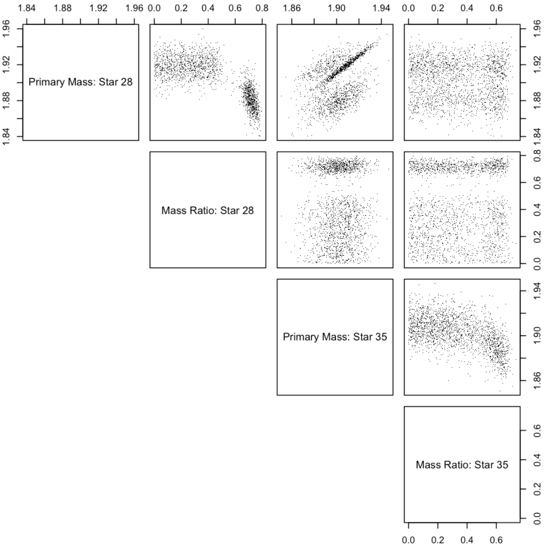
Complex Posterior Distributions I



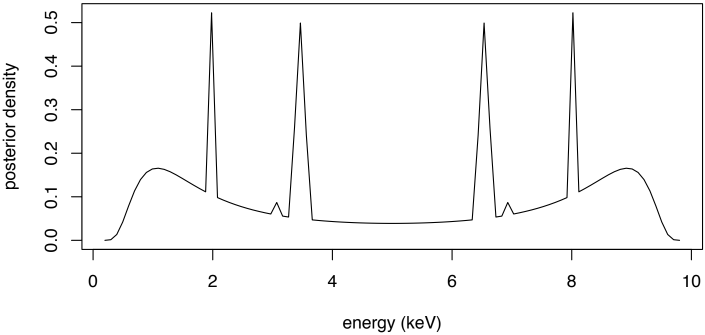
Highly non-linear relationships among parameters.

Imperial College London

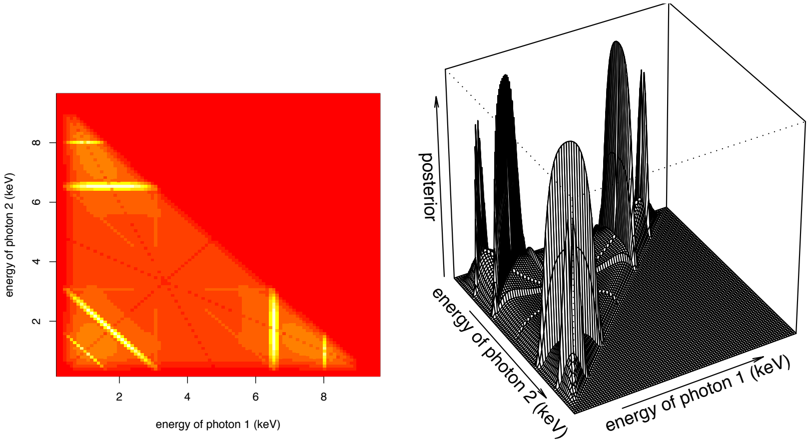
Complex Posterior Distributions II



Complex Posterior Distributions III



Complex Posterior Distributions III



Outline

Multilevel Models

Sequentially account for physical model, errors in recorded energy, selection effects, data contamination, truncation, etc.

- **Model Parameters:** θ.
- **Physical Model:** $p(E|\theta)$ is dist'n of true flare energies.
- **Under-reported Energy:** $p(E_{\text{blur}}|\theta)$.
- **Data Truncation:** $p(E_{\text{trunc}}|\theta)$
- **Data Contamination:** $p(E_{\text{obs}}|\theta)$.

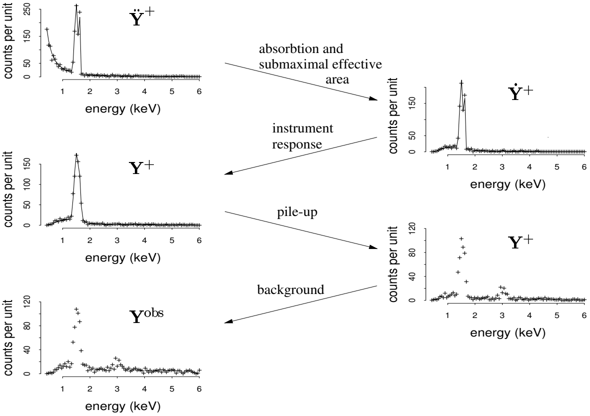
Likelihood:

$$\begin{aligned} p(E_{\text{obs}}|\theta) &= \int p(E_{\text{obs}}, E_{\text{trunc}}, E_{\text{blur}}, E|\theta) dE_{\text{trunc}}\, dE_{\text{blur}} dE \\ &= \int p(E_{\text{obs}}|E_{\text{trunc}}) p(E_{\text{trunc}}|E_{\text{blur}}) p(E_{\text{blur}}|E) p(E) dE_{\text{trunc}}\, dE_{\text{blur}} dE \end{aligned}$$

(Omitting θ in the last line to save space!)

Modeling Data Collection Mechanism

Recall:



Likelihood:

$$p(E_{\text{obs}}|\theta) = \int p(E_{\text{obs}}|E_{\text{trunc}})p(E_{\text{trunc}}|E_{\text{blur}})p(E_{\text{blur}}|E)p(E)dE_{\text{trunc}}\, dE_{\text{blur}}dE$$

(Omitting θ in the last line to save space!)

Power Law for the True Flares Energy

Choice of model:

- If events are are recorded as counts in energy bins, Poisson models are appropriate.
- If continuous energies are recorded, they should be modeled directly:

$$p(E|\theta) = \begin{cases} (\gamma - 1)\left(\frac{E}{E_0}\right)^{-\gamma} E_0^{-1} & \text{for } E > E_0 \\ 0 & \text{otherwise} \end{cases},$$

  where $\gamma > 1$.
- In statistics this is called the *Pareto distribution*.
- Generalization: broken power-law, added features, etc.

Under-Reporting of Energy

Errors in recorded event energies:

- Under-reporting of energies:

$$E_{\text{blur}} = uE, \quad \text{with} \quad u \leq 1$$

- Parnell & Jupp (2000) suggest a Beta($\phi + 1, 1$) distribution:

$$p(u|\theta) = \begin{cases} (\phi + 1)u^{\phi} & \textit{for} \quad 0 < u < 1 \\ 0 & \textit{otherwise} \end{cases},$$

  with $\phi > -1$. *(Larger $\phi \longrightarrow$ less under-reporting.)*

- In principle, any distribution $p(E_{\text{blur}}|E, \theta)$ can be used.
- $p(E_{\text{blur}}|\theta) = \int p(E_{\text{blur}}|E, \phi)p(E|\gamma)dE$. (e.g., skew-Laplace dist'n)

Data Truncation

Selection Effects

- Some events are not observed:

$$Z = \begin{cases} 1 & \text{if event is observed} \\ 0 & \text{otherwise} \end{cases}.$$

- The probability of observation depend on energy:

$$p(Z = 1 | E_{\text{blur}}, \theta) = \Pr(\text{event is observed} | E_{\text{blur}})$$

Full Truncation: Observe if and only if $E_{\min} < E_{\text{blur}} < E_{\max}$.

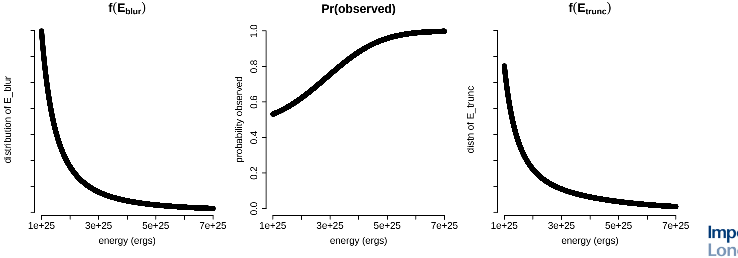
Stochastic Truncation: A probability of observing any event.

Data Truncation

Condition on $Z = 1$ to re-weight $p(E_{\text{blur}}|\theta)$:

$$p(E_{\text{trunc}}|\theta) = p(E_{\text{blur}}|\theta, Z = 1) = \frac{p(E_{\text{blur}}, Z = 1|\theta)}{p(Z = 1|\theta)}$$

$$= \frac{p(E_{\text{blur}}|\theta)p(Z = 1|E_{\text{blur}}, \theta)}{\int p(E_{\text{blur}}|\theta)p(Z = 1|E_{\text{blur}}, \theta)dE_{\text{blur}}}$$



Data Contamination

Two types of selection effects

Truncation Events of interest are not recorded.

Contamination Events are recorded that are not of interest.

$$p(E_{\text{obs}}|\theta) = \alpha p(E_{\text{trunc}}|\theta) + (1-\alpha)p(E_{\text{bkgd}})$$

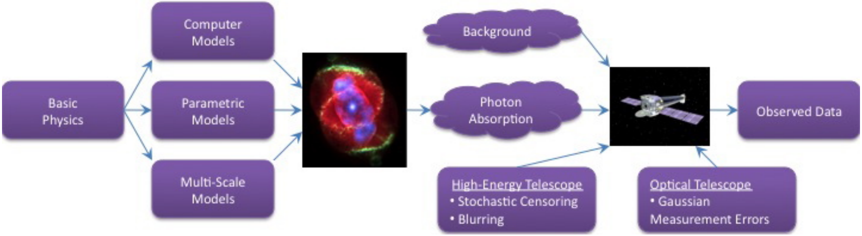
To identify underlying power law, must know something about:

- blurring function, $p(E_{\text{blur}}|E, \theta)$
- probability events of interest are included, $p(Z=1|E_{\text{blur}}, \theta)$.
- distribution of contaminating events, $p(E_{\text{bkgd}}|\theta)$.

After specifying model and obtaining data, fit via MCMC.

Outline

Calibration of X-ray Detectors

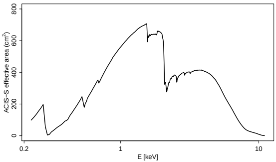


- We must model both
  1. the scientifically interesting source and
  2. instrumental effects.

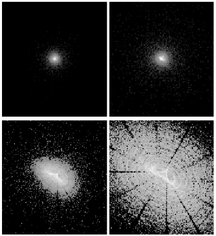
How well are the instruments understood?

Calibration Products

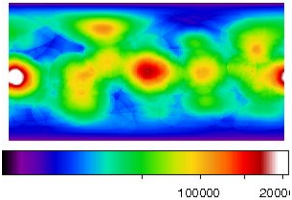
- Analysis is highly dependent on *Calibration Products:*
  - Effective area records sensitivity as a function of energy
  - Energy redistribution matrix can vary with energy/location
  - Point Spread Functions can vary with energy and location
  - Exposure Map shows how effective area varies in an image
- In this talk we focus on uncertainty in the effective area.



A *CHANDRA* effective area.



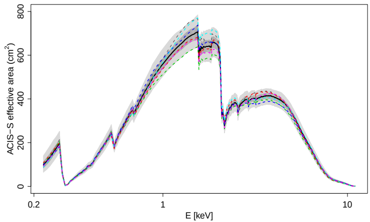
Sample Chandra psf's
(Karovska et al., ADASS X)

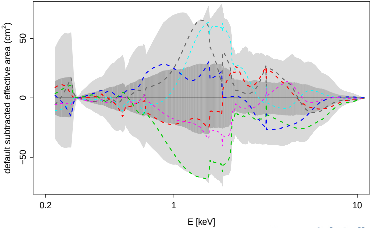


EGERT exposure map
(area × time)

Derivation of Calibration Products

- Effective area records the instrument sensitivity as function of energy
- Aim to capture deterioration of detectors over time.
- Complex computer models of subassembly components.
- Calibration scientists provide a sample representing uncertainty
- *Calibration Sample* is typically of size $M \approx 1000$.





Simple Emulation of Computer Model[4]

We use *Principal Component Analysis* to represent uncertainly:

$$A \sim A_0 + \bar{\delta} + \sum_{j=1}^{m} e_j r_j \boldsymbol{v}_j,$$

A_0: default effective area,

$\bar{\delta}$: mean deviation from A_0,

r_j and $\boldsymbol{v}_j$: first m principle component eigenvalues & vectors,

e_j: independent standard normal deviations.

Capture 95% of variability with $m = 6 - 9$.

[4] Lee, Kashyap, van Dyk, Connors, Drake, Izem, Meng, Min, Park, et al. (2011). Accounting for Calibration Uncertainties in X-ray Analysis: Effective Areas in Spectral Fitting. *The Astrophysical Journal*, **731**, 126–144.

Two Possible Target Distributions[5]

We consider inference under:

A PRAGMATIC BAYESIAN TARGET: $\pi_0(A, \theta) = p(A)p(\theta|A, Y)$.

THE FULLY BAYESIAN POSTERIOR: $\pi(A, \theta) = p(A|Y)p(\theta|A, Y)$.

Concerns:

Statistical Fully Bayesian target is “correct”.

Cultural Astronomers have concerns about letting the current data influence calibration products.

Computational Both targets pose challenges, but pragmatic Bayesian target is easier to sample.

Practical How different are $p(A)$ and $p(A|Y)$?

With MCMC we sample a different effective area curve at each iteration according to its conditional distribution.

[5] Xu, van Dyk, Kashyap, Siemiginowska, Connors, Drake, et al. (2014). A Fully Bayesian Method for Jointly Fitting Instrumental Calibration and X-ray Spectral Models. *The Astrophysical Journal*, to appear.

Implementing the Fully Bayesian Analysis

Direct MH sampling is difficult. (Case-by case tuning of jumping rules.)

Pragmatic Bayesian posterior

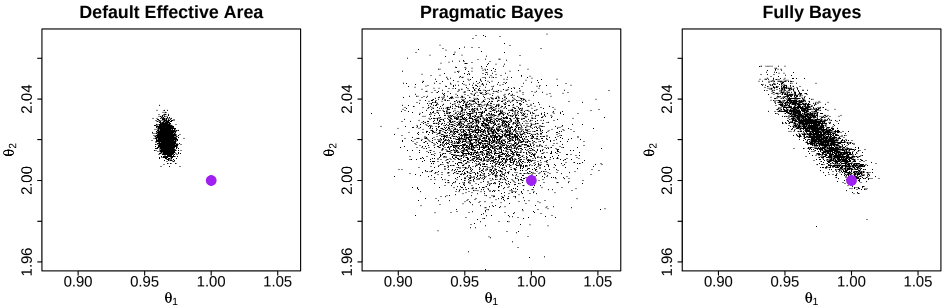
- We can easily sample from $\pi_0(A, \theta)$.
- Well suited proposal dist'n: over-dispersed relative to $\pi(A, \theta)$.
- But $\pi_0(A, \theta)$ cannot be evaluated

$$\pi_0(A, \theta) = p(\theta | Y, A) p(A) = \frac{p(Y|\theta, A) p(\theta)}{p(Y|A)} p(A)$$

This is a doubly intractable distribution.

- We construct a normal approximation ($\sim$ 20 dimensional).
- Use as jumping rule in an independence MH sampler.

Sampling From the Full Posterior

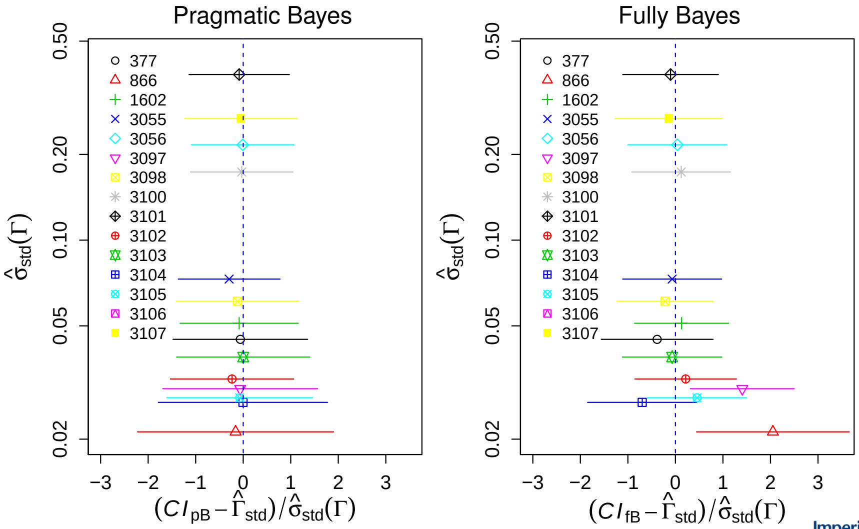


Spectral Model (purple bullet = truth):

$$\text{power law:} \quad \text{mean}(E_j|\theta) = \theta_1 E_j^{-\theta_2}$$

Pragmatic Bayes is clearly better than standard method, but a Fully Bayesian Method is the ultimate goal.

How it Works on a Sample of Radio-Loud Quasars



Imperial College London

For Further Reading I

van Dyk, D., Connors, A., Kashyap, V., and Siemiginowska, A.
Analysis of Energy Spectra with Low Photon Counts via Bayesian Posterior Simulation. *The Astrophysical Journal*, **548**, 224–243, 2001.

Protassov, R., van Dyk, D., Connors, A., Kashyap, V., and Siemiginowska, A.
Statistics: Handle with Care – Detecting Multiple Model Components with the Likelihood Ratio Test. *The Astrophysical Journal*, **571**, 545–559, 2002.

van Dyk, D. and Kang, H.
Highly Structured Models for Spectral Analysis in High-Energy Astrophysics. *Statistical Science*, **19**, 275–293, 2004.

Lee, H., Kashyap, V., van Dyk, D., Connors, A., Drake, J., Izem, R., Min, S., Park, T., Ratzlaff, P., Siemiginowska, A., and Zezas, A.
Accounting for Calibration Uncertainties in X-ray Analysis: Effective Area in Spectral Fitting. *The Astrophysical Journal*, **731**, 126–144, 2011.

Xu, J., van Dyk, D., Kashyap, V., Siemiginowska, A., Connors, A., Drake, J., Meng, X.L., Ratzlaff, P. and Yu, Y.
A Fully Bayesian Method for Jointly Fitting Instrumental Calibration and X-ray Spectral Models. *Astrophysical Journal*, to appear, 2014.