Fully Bayesian Analysis of Calibration Uncertainty In High Energy Spectral Analysis

JIN XU

Department of Statistics, UCI

February 26, 2013

Background

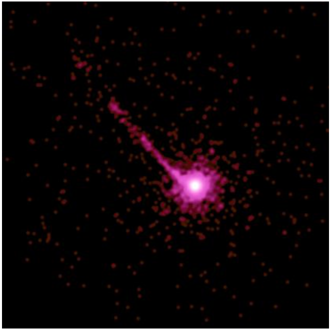
- High-Energy Astrophysics
- Spectral Analysis
- Calibration Products
- Scientific Goals

High-Energy Astrophysics

- Provide understanding into high-energy regions of the Universe.
- Chandra X-ray Observatory is designed to observe X-rays from high-energy regions of the Universe.
- X-ray detectors typically count a small number of photons in each of a large number of pixels.
- Spectral Analysis aims to explore the parameterized pattern between the photon counts and energy.

Quasar

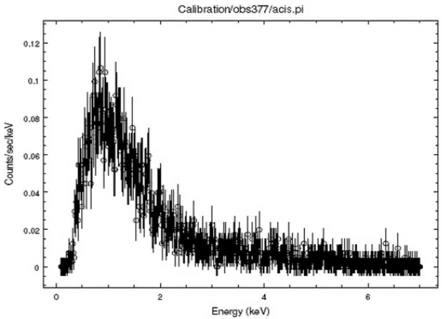
The Chandra X-ray image of the quasar PKS 1127-145, a highly luminous source of X-rays and visible light about 10 billion light years from Earth.



An Example of One Dataset

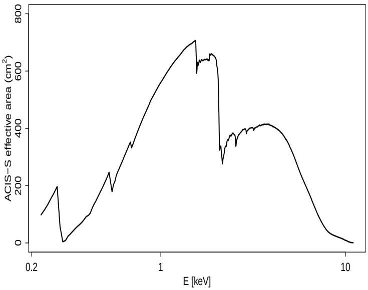
TITLE = EXTENDED EMISSION AROUND A GIGAHERTZ PEAKED RADIO SOURCE

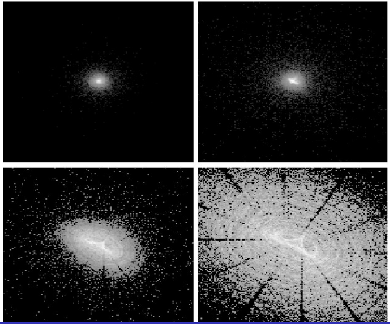
DATE = 2006-12-29 T 16:10:48



Calibration Uncertainty

- 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.





Incorporate Calibration Uncertainty

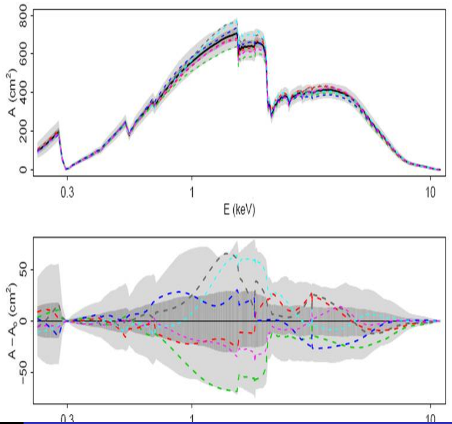
- Calibration Uncertainty in astronomical analysis has been generally ignored.
- No robust principled method is available.
- Our goal is to incorporate the uncertainty by Bayesian Methods.
- In this talk, we will focus on uncertainty in the effective area.

Problem Description

- The true effective area curve can't be observed.
- No parameterized form for the density of effective area curve complicates model fitting.
- Simple MCMC is quite expensive, due to the complexity of the astronomical model.

Generating Calibration Sample

- Drake et al. (2006), suggest generating calibration sample of effective area curves to represent the uncertainty.
- The plot shows the coverage of a sample of 1000 effective area curves, and the default one (A_0) is a black line.
- Calibration Sample: $\{A_1, A_2, A_3, ..., A_L\}$



Three Main Steps

- Use Principle Component Analysis to parameterize effective area curve.
- Model Building, that is combining source model with calibration uncertainty.
- Three Inferencial Models.

A simplified model of telescope

$$E(Y(E_i)) = A(E_i) * S(E_i);\ Y(E_i) \sim Poisson(E(Y(E_i)))$$

$Y(E_i)$: Observed Photon in certain energy bin E_i

$S(E_i)$: True Source Model, we set it as,
$S(E_i) = exp(-n_H * \alpha(E_i)) * a * E_i^{(-\Gamma)} + b$

$A(E_i)$: Effective Area Curve

θ: source parameter, $\theta = \{n_H, a, \Gamma, b\}$

$\alpha(E_i)$: photo-electric cross-section

b: background intensity

Use PCA to represent effective area curve

$$A = A_0 + \bar{\delta} + \sum_{j=1}^{m} e_j r_j v_j$$

A_0 : default effective area,

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

r_j and v_j : first m principle component eigenvalues & vectors,

e_j : independent standard normal deviations.

Capture 95% of uncertainty with m = 6 - 9. (Lee et al. 2011, ApJ)

Three Inferencial Models

- Fixed Effective Area Model(Standard Approach)
- Pragmatic Bayesian Model
  - Original Pragmatic Bayesian Scheme (Lee et al. 2011, ApJ)
  - Efficient Pragmatic Bayesian Scheme
- Fully Bayesian Model
  - Gibbs Sampling Scheme
  - Importance Sampling Scheme

Model One: Fixed Effective Area (Standard Approach)

- Model: $p(\theta|Y, A_0)$
- We assume $A = A_0$, where A_0 is the default affective area curve, and may not be the true one,
- This model doesn't incorporate calibration uncertainty, which is widely used because of its simplicity.
- The estimation may be biased and error bars may be underestimated.
- Only one sampling step involved: $p(\theta|Y, A_0) \propto L(Y|\theta, A_0)\pi(\theta)$
- A mixed approach of Metropolis and Metropolis-hastings is used in the sampling

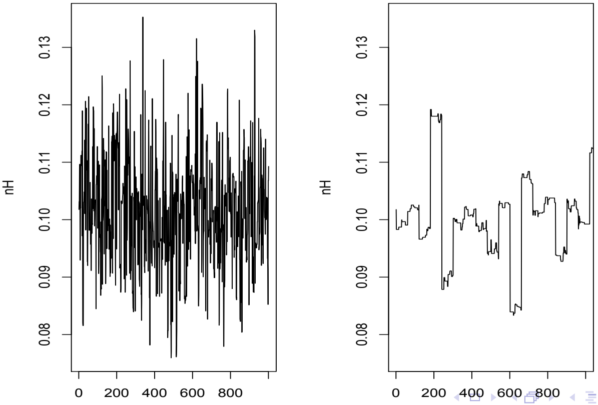
Model Two: Pragmatic Bayesian(Lee et al, 2011, ApJ)

- Model: $P_{prag}(\theta, A|Y) = p(\theta|A, Y) * \pi(A)$
- Doubly-intractable Distribution!
- Main purpose is to reduce complexity of sampling.
- Step One: sample A from $\pi(A)$
- Step Two: sample θ from $p(\theta|Y, A) \propto L(Y|\theta, A)\pi(\theta)$
- A mixed approach of Metropolis and Metropolis-hastings is used in the Step Two

Model Two: Efficient Pragmatic Bayesian

- After each draw of A_i (i from 1 to n) from $\pi(A)$, we have to find the best Metropolis-hastings proposal for $p(\theta|Y, A)$, which costs a long and relatively constant time, say, T_1. (ML involved.)
- Once the proposal distribution is fixed given A_i, each draw of θ from $p(\theta|Y, A)$ costs a rather short time, say, T_2. ($T_1 > T_2$)
- In order to obtain the most effective samples for θ, we sample m θ's given A_i, say, θ_{ij}. (j from 1 to m)

Example


nH
0.08
0.09
0.10
0.11
0.12
0.13
0
200
400
600
800
nH
0.08
0.09
0.10
0.11
0.12
0.13
0
200
400
600
800

Model Two: Efficient Pragmatic Bayesian

- Then this problem could be simplified into one optimization problem.
- Minimize: $Var(\frac{1}{n}\sum_i(\frac{1}{m}\sum_j \theta_{ij}))$
  Subject to: $T = nT_1 + nmT_2$
- T is the total time, and when $m = 1$, the scheme turns into original Pragmatic Bayesian, Lee et al(2011, ApJ)
- We can get simple analytical solution:
  $n = \frac{\sqrt{B}T}{\sqrt{B}T_1 + \sqrt{WT_2T_1}}$; $m = \frac{\sqrt{WT_1}}{\sqrt{BT_2}}$
- Here, $B = \sigma_\theta^2 - \sigma_{\theta|A}^2$, $W = \sigma_{\theta|A}^2$, and we assume θ's given A is independent to each other
- Notice, m is not related to T.

Model Two: Efficient Pragmatic Bayesian

- The assumption that θ's given A is independent to each other can be achieved if we thin the iterations within one A by a big number.
- If we assume θ's given one A are AR(1), neighbor correlation is ρ
- Then $Var(\bar{Y}) = \frac{1}{n}(B + W\frac{1+\rho}{m(1-\rho)})$
- we can get still get similar optimization solutions as above, only need to replace W by $W\frac{1+\rho}{1-\rho}$

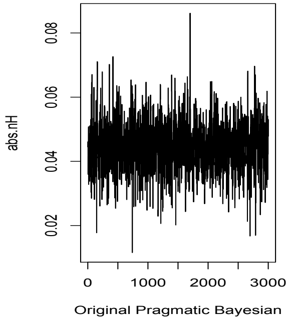
Model Two: Efficient Pragmatic Bayesian Sampling

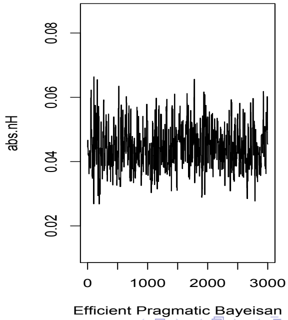
Two ways of Efficient Pragmatic Bayesian Sampling of N θ's

- $(n_1, m_1) \Rightarrow (\hat{B}, \hat{W}) \Rightarrow (\hat{m}) \Rightarrow (\hat{n} = \frac{N - n_1 m_1}{\hat{m}})$
- while $\{n_0 < N\}$
  do { update B and W;
  calculate m;
  sample A;
  sample m θ's from $p(\theta|Y, A)$;
  $n_0 = n_0 + m;\}$
- The second adaptive scheme hasn't been verified yet!

Model Two: Efficient Pragmatic Bayesian Sampling

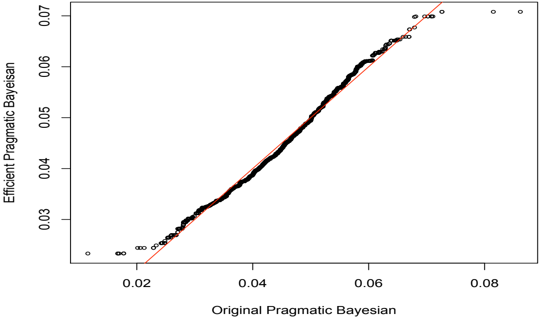
Here are two chains, separately from Pragmatic Bayesian and Efficient Pragmatic Bayesian Samplings for quasar dataset 3104.





Model Two: Efficient Pragmatic Bayesian Sampling

QQ plot of these two chains.



Model Two: Efficient Pragmatic Bayesian Sampling

Results for dataset 3104:
Before Efficient Pragmatic Bayesian Sample, T_1, T_2, B and W are estimated. $T_1 = 7.1$sec, $T_2 = 0.045$sec, B=4.01e-5, W=2.40e-5. Then the optimal $m = 10$, $n = 300$, if we want to draw 3000 θ's.

	$\hat{\mu}_{abs\_nH}$	$\hat{\sigma}^2_{abs\_nH}$	$\hat{\sigma}^2_{\hat{\mu}}$	T
Pragmatic Bayesian	0.0447	6.36e-05	2.12e-08	5.9hrs
Efficient Pragmatic Bayesian	0.0443	6.10e-05	1.13e-07	0.6hrs
Ratio			0.187	10

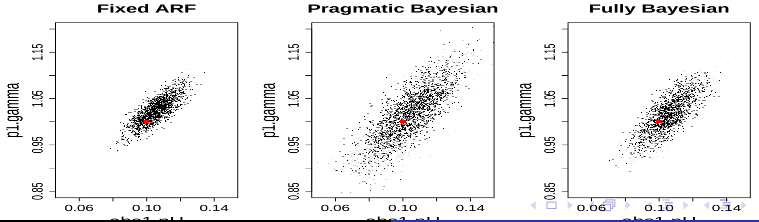
EPB almost doubles effective sample size.

Model Three: Fully Bayesian using Gibbs sampling

- Uses correct Bayesian Approach: $P_{full}(\theta, A|Y) = p(\theta|A, Y) * p(A|Y)$
- This Model allows the current data to influence calibration products,
- Step One: sample A from $p(A|Y, \theta) \propto L(Y|\theta, A)\pi(A)$
- Step Two: sample θ from $p(\theta|Y, A) \propto L(Y|\theta, A)\pi(\theta)$
- A mixed approach of Metropolis and Metropolis-hastings is used in the both steps
- Most difficult approach to sample.

Importance sampling for Fully Bayesian

Fully Bayesian using Gibbs sampling involves a lot of choice of proposal distributions, and the choice of proposal distributions highly influences the performance of the chains. Here, we introduce importance sampling for Fully Bayesian, which takes advantage of the draws from Pragmatic Bayesian Model. (Pragmatic Bayesian Model has larger variance.)



Importance sampling for Fully Bayesian

Steps:

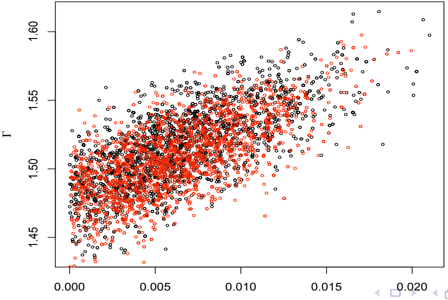
- Get the draws from Pragmatic Bayesian Method
- Approximate $P_{prag}(\theta|A, Y)$ as Multivariate Normal distribution, we call it $P_{new\_prag}(\theta|A, Y)$.
  - $P_{prag}(\theta|A, Y)$ can't be calculated because of doubly-intractable distribution.
  - 18 parameters from A are all independent standard normal.
- Get new draws by sampling $\pi(A)$ and $P_{new\_prag}(\theta|A, Y)$
- Calculate the ratio $r = P_{fully}(A, \theta|Y)/P_{new\_prag}(A, \theta|Y)$
- Use the ratios to do resampling for Fully Bayesian.

Importance sampling for Fully Bayesian

- The great benefit from the new scheme is everything can work automatically, saving the trouble of choosing "nice" proposal distributions in the Fully Bayesian Model using Gibbs sampling.
- The disadvantage is that every time we need to fit Fully Bayesian Model, we have to do Pragmatic Bayesian first. Usually, astronomers would like to use all three Models and to do the comparison.
- The results of Fully Bayesian using Gibbs sampling and Importance sampling are usually identical to each other.

Importance sampling for Fully Bayesian

Here are two scatter plots, separately from Fully Bayesian using Gibbs sampling(Black) and Importance sampling(Red) for quasar dataset 3105.



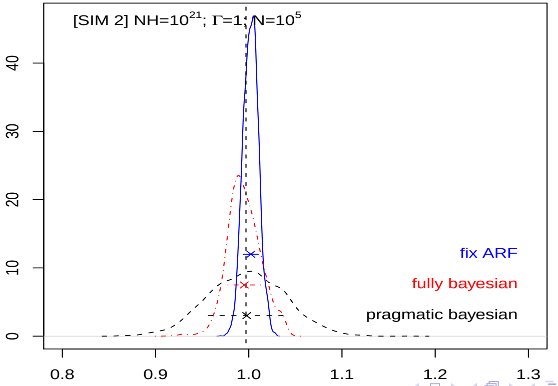
Eight simulated data sets

The first four data sets were all simulated without background contamination using the XSPEC model wabs*powerlaw, nominal default effective area A_0 from the calibration sample of Drake et al. (2006), and a default RMF.

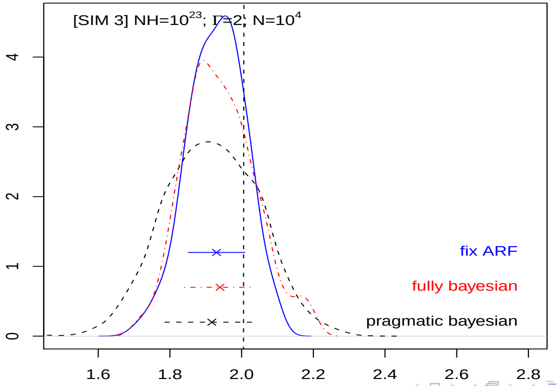
- Simulation 1: $\Gamma = 2, N_H = 2^{23} cm^{-2}$, and 10^5 counts;
- Simulation 2: $\Gamma = 1, N_H = 2^{21} cm^{-2}$, and 10^5 counts;
- Simulation 3: $\Gamma = 2, N_H = 2^{23} cm^{-2}$, and 10^4 counts;
- Simulation 4: $\Gamma = 1, N_H = 2^{21} cm^{-2}$, and 10^4 counts;

The other four data sets (Simulation 5-8) were generated using an extreme instance of an effective area.

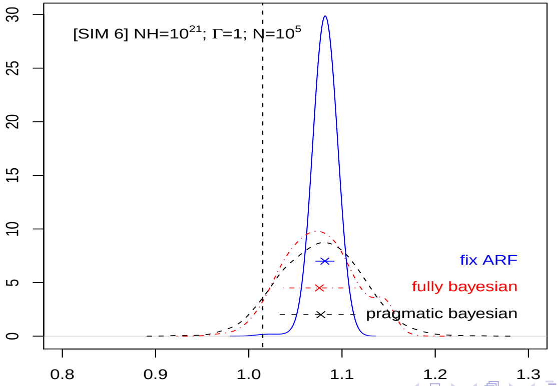
Results for Simulation 2



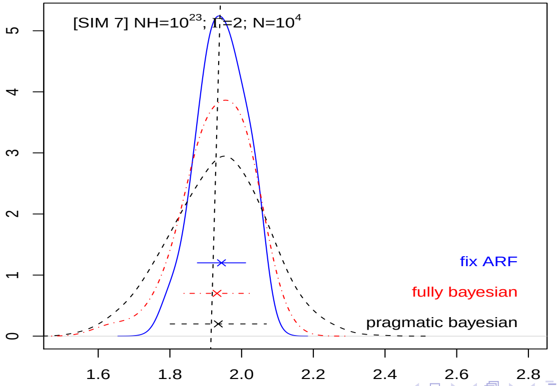
Results for Simulation 3



Results for Simulation 6



Results for Simulation 7

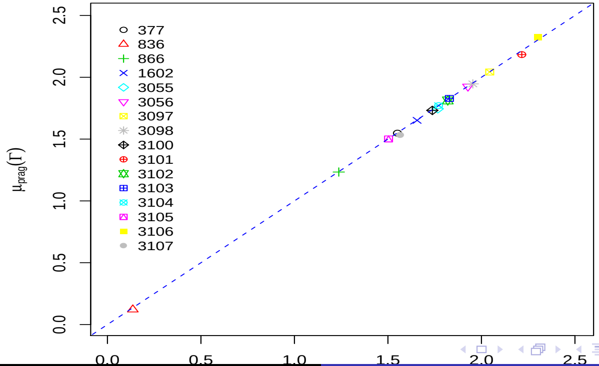


Quasar results

- 16 Quasar data sets were fit by these three models: 377, 836, 866, 1602, 3055, 3056, 3097, 3098, 3100, 3101, 3102, 3103, 3104, 3105, 3106, 3107.
- Most interesting founding for Fully Bayesian model is shift of parameter fitting, besides the change of standard errors.
- Both comparisons of mean and standard errors among three models are shown below.

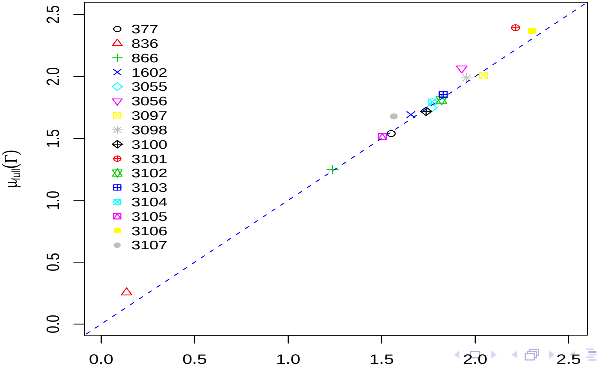
mean: fix-prag

Fixed Effective Area Curve Model has almost the same parameter fitting as Pragmatic Bayesian Model.



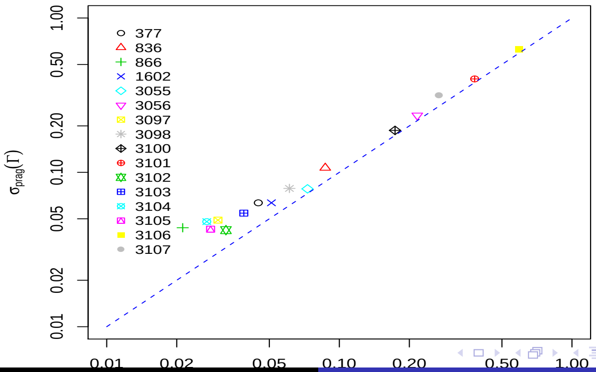
mean: fix-full

Fully Bayesian model shifts the parameter fitting, which mean data itself influence the choice of effective area curve.



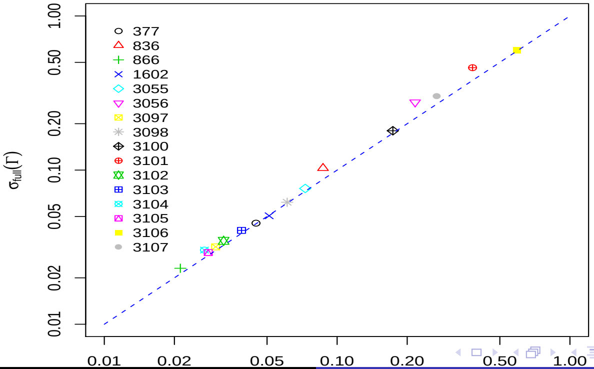
sd: fix-prag

Pragmatic Bayesian has larger parameter standard deviation than Fixed Effective Area Curve Model, especially for large datasets.



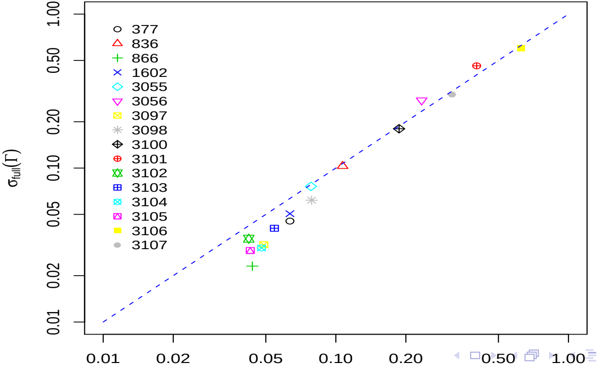
sd: fix-full

Fully Bayesian usually has larger parameter standard deviation than Fixed Effective Area Curve Model, too.



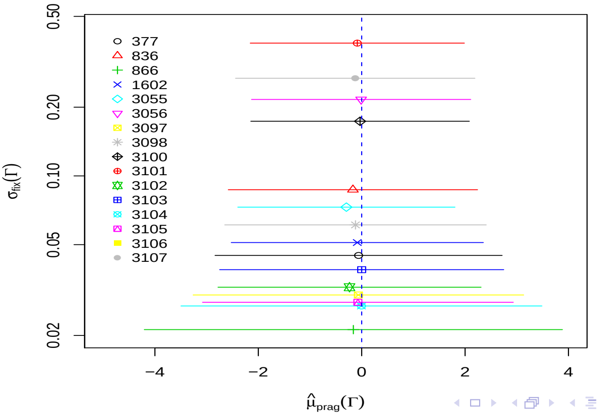
sd: prag-full

It can be observed that generally parameter standard deviation from Fully Bayesian Model is between Pragmatic Bayesian and Fixed Effective Area Curve Models.



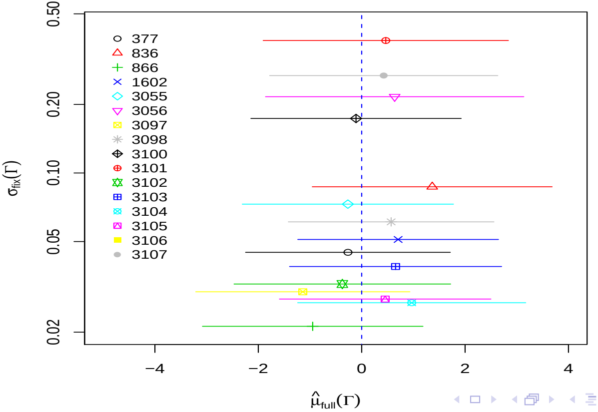
more plots

$\hat{\mu}_{prag}(\Gamma) = \frac{\mu_{prag}(\Gamma) - \mu_{fix}(\Gamma)}{\sigma_{fix}(\Gamma)}$, these lines cover 2 sd.



more plots

$\hat{\mu}_{full}(\Gamma) = \frac{\mu_{full}(\Gamma) - \mu_{fix}(\Gamma)}{\sigma_{fix}(\Gamma)}$, these lines cover 2 sd.



Doubly-intractable Distribution

- Murray et al (2012) defines doubly-intractable distribution this way:

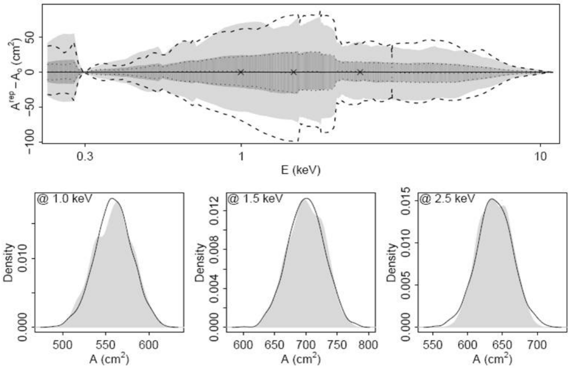
  $p(\theta|y) = (\frac{f(y;\theta)\pi(\theta)}{Z(\theta)})/p(y)$
- Here, $Z(\theta) = \int f(y;\theta)dy$, is called as "unknown constant" and can't be computed.
- Doubly-intractable Distributions widely exist, and most famous method is introduced by Møller, Jesper, et al.(2006), called "Auxiliary Variable Method".
- One of most popular model involving Doubly-intractable Distribution is Ising Model.

Doubly-intractable Distribution

- Our Pragmatic Bayesian Model involves:
  $$P_{prag}(\theta, A|Y) = p(\theta|A, Y) * \pi(A) = \frac{L(Y|\theta,A)*\pi(\theta)}{p(Y|A)} * \pi(A)$$
- Here, $p(Y|A)$ is unknown normalizing constant.
- Interestingly, unlike other Doubly-intractable Distributions, our Pragmatic Bayesian Model can avoid this constant easily and get the posterior draws.
- However, when we use these draws to help sample Fully Bayesian Model by importance sampling, we find out that $ratio = \frac{p(Y|A)}{p(Y)}$, can't avoid the constant!!!
- That's the reason we re-estimate the draws' density by Multivariate Normal, and use new draws from the Multi-Gaussian to do the importance sampling. Avoid the constant again!!!

Other Calibration Uncertainty

For effective area curve, Lee et al. firstly gets 1000 curve samples and then use PCA to parameterize the effective area curve.



Other Calibration Uncertainty

- PCA has its unique advantages: easy to parameterize vector samples; dimension highly reduced; easy to sample afterwards.
- However, there is no principled method of PCA to parameterize matrix samples.
- Other Calibration Uncertainty, rmf and psf are both matrix.
- Our goal is to parameterize these matrix samples, define the probability density of the matrix, and nicely sample new matrix.

Other Calibration Uncertainty

Some possible approaches:

- Diffusion map allows mapping data into a coordinate system that efficiently reveals the geometric structure of data. However, because of its nonlinear transformation, diffusion map is not reversible.
- Delaigle et al (2010) introduced a way to define probability density for a distribution of random functions. However, how to extend it into 2-dimension space is under question.
- Wavelets technique can provide nice way to extract information from 2-dimension data. Besides, complementary wavelets allows us to recover the original information with minimal loss.

Other Calibration Uncertainty

Here is our potential approach:

- Use wavelets to analyze all the matrix samples;
- Use PCA to summarize wavelet parameters from step one;
- By sampling independent standard normal deviations to construct wavelet parameters;
- Use new wavelet parameter to construct new matrix sample.

Citations

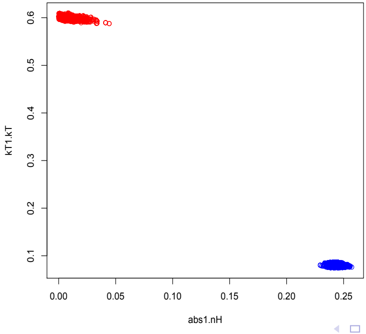
- Lee, Hyunsook, et al. "Accounting for calibration uncertainties in X-ray analysis: effective areas in spectral fitting." The Astrophysical Journal 731.2 (2011): 126.
- Murray, Iain, Zoubin Ghahramani, and David MacKay. "MCMC for doubly-intractable distributions." arXiv preprint arXiv:1206.6848 (2012).
- Delaigle, Aurore, and Peter Hall. "Defining probability density for a distribution of random functions." The Annals of Statistics 38.2 (2010): 1171-1193.
- Lee, Ann B., and Larry Wasserman. "Spectral connectivity analysis." Journal of the American Statistical Association 105.491 (2010): 1241-1255.

1878

- Model: "xsphabs.abs1*(xsapec.kT1+xsapec.kT2)"
- Set kT2.Abundanc = kT1.Abundanc
- Six source parameters: abs1.nH, kT1.kT, kT1.Abundanc, kT1.norm, kT2.kT, kT2.norm
- Sherpa fit() using Neldermead motheds to find MLE fails
- Stuck at boundaries: abs.nH=0 and kT1.Abundanc=5
- Multiple modes

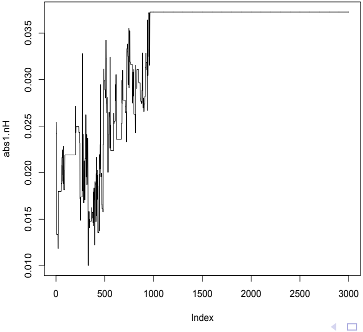
Fixed ARF

The results are highly correlated to starting values. Although the chains might seem converge, it's just stuck in the local mode.



Fully Bayeisan

Similarly, it happens to Fully Bayesian. Once its stuck in the mode, it seems impossible to accept a new ARF.



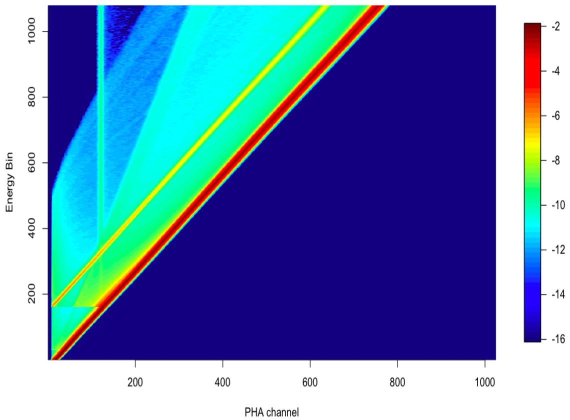
Redistribution Matrix File (RMF)

Matrix consists the probability that an incoming photon of energy E will be detected in the output detector channel I. The matrix has dimension 1078*1024. But the data is stored in a compressed format. For example rmf 0001:
n_grp = UInt64[1078], stores the number of non-zero groups in each row
f_chan = UInt32[1395], stores starting point of each group
n_chan = UInt32[1395], stores the element number of each group
matrix = Float64[384450], stores all the non-zero values

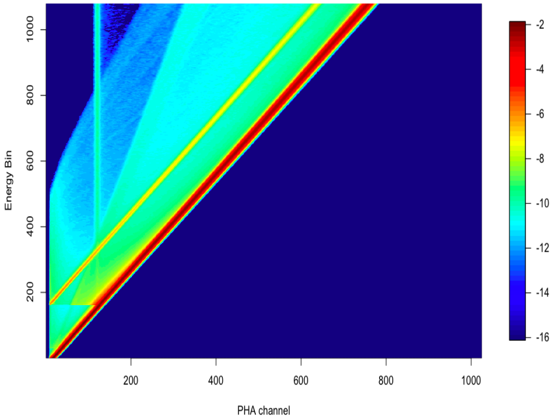
Redistribution Matrix File (RMF)

- The sum of each row equals to one
- sum(n_grp)=1395
- sum(n_chan)=384450
- sum(matrix)=1078

log(Rmf0001+1e-6)



log(Rmf0002+1e-6)



prepare for PCA

- raw data dimension (1000,1078*1024)
- discard bottom-right zeros, dimension (1000, 445380)
- I use python DMP NIPALS algorithm (Nonlinear Iterative Partial Least Squares)
- (1000, 445380) MemoryError!!!
- (100, 445380) MemoryError!!!
- (1000, 100000) MemoryError!!!
- (1000, 10000) 1 hours, I got 5 principal components
- variance: (3.6e-02, 8.8e-03, 1.02e-04, 3.05e-05, 9.74e-06)

simulating rmf