Cosmic Rays and Bayesian Computations

David Ruppert

Cornell University

Dec 5, 2011

Collaborators

- Kunlaya Soiaporn, Graduate Student, ORIE
- Tom Loredo, Research Associate and PI, Astronomy
- Dave Chernoff, Professor, Astronomy
- Ira Wasserman, Professor and Chair, Astronomy

Outline

- Cosmic Rays (CRs)
    - this research is about ultra-high energy cosmic rays
- Active galactic nuclei (AGNs)
    - prime suspects as the source of ultra-high energy cosmic rays
- Association models
    - associate CRs with AGNs
- Bayesian analysis
- Bayesian computation using Markov chain Monte Carlo
- Results
- Future work (time permitting)

What Are Cosmic Rays?

- Cosmic rays are atomic nuclei
- First detected in 1912 by Victor Hess who ascended in a balloon to 5 km
- Range in energy from 10^7 to 10^{20} eV
  - eV = electron volt
- Spectrum is a power law $F \propto E^{-\alpha}$
  - F = flux
  - E = energy
- Detailed look at F versus E (log-log plot) suggests several sources

Where Do Cosmic Rays Originate?

- Cosmic rays are charged particles
  - therefore they are deflected by magnetic fields
  - so it is not obvious where they originate
- Sources of cosmic rays could be
  - supernovae
  - pulsars
  - stars with strong winds
  - black holes
- Active galactic nuclei (AGNs) are prime suspects as sources of cosmic rays at highest energies
  - only AGNs seem capable of accelerating particles to such high energies

Ultra-High Energy Cosmic Rays (UHECRs)

- Our research focuses on cosmic rays of highest energies
- Cosmic ray with E $> 10^{20}$ eV observed in 1962
- 1991: particle with E $\approx 3 \times 10^{20}$ observed
  - same kinetic energy as a baseball at 60 mph
  - over 10 million times more energy than most energetic particles at Large Hadron Collider

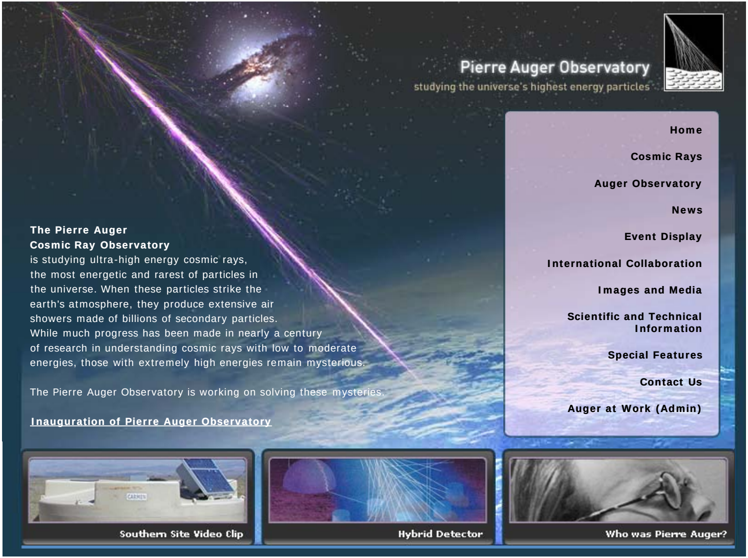
- Not confided to galaxy of origin
- Interact with cosmic microwave background
  - called the Greisen-Zatsepin-Kuz'min (GZK) cutoff
  - So UHERCs must come from within approximately 100 megaparsecs (Mpc)
  - 1 parsec $\approx$ 3.26 light-years
- Closer galaxies are more likely sources
- Flux: 1 particle km^{-2} century^{-1}

Pierre Auger Observatory (PAO):

- Largest and most sensitive cosmic ray detector to date
- In Argentina
- Ultra-high energy cosmic rays create giant air showers of particles
  - first discovered by Pierre Auger (1899–1993)
- PAO uses air fluorescence telescopes and surface detectors
- Operations began in 2008
- Has reported about 70 UHECRs

Pierre Auger Home Page

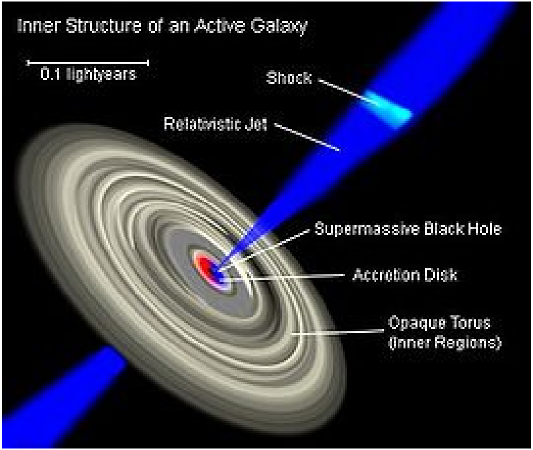
Pierre Auger Observatory



What is an Active Galactic Nucleus (AGN)?

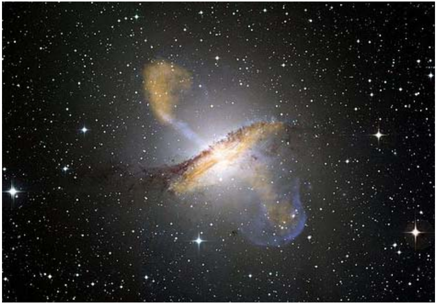
- An AGN is a compact region at the center of a galaxy with high electromagnetic luminosity
  - Example: Quasar
- Activity is believed to come from the accretion of mass by a supermassive black hole
- Our galaxy also harbors a supermassive black hole
  - but the Milky Way does not seem to be active at present

Inner Structure of an AGN



Source: Wikipedia

Radio Galaxy Centaurus A (NGC 5128)



870-micron submillimeter = orange

X-ray = blue

visible light = close to true color

Source: Wikipedia

Our Catalog (AGN data)

- We used the Goulding catalog which contains all AGNs within 15 Mpc (megaparsecs)
  - "volume complete"

PAO Data: Tuning with Period 1

- In Aug 2007 there were 81 UHECRs with E $>$ 40 EeV
- The earliest half of the data (Period 1 $=$ "training data") was used to tune parameters in a test to detect anisotropy
  - These values minimized the p-value of the test.
- The p-value for the second half of the data (Period 2 $=$ "test data") was 1.7×10^{-3}
- Later, after Period 3 data became available, p-value for all data increased to 3×10^{-3}

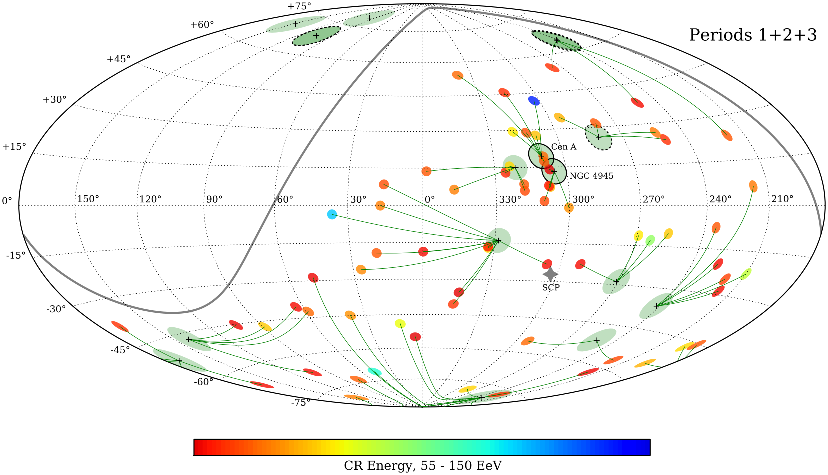
- PAO has reported 69 UHECRs with energy $\geq 5.5 \times 10^{19}$ eV

Period	Dates	Exposure (km^2 sr y)	No. of UHECRs detected
1	01-01-04 – 05-26-06	4390	14
2	05-27-06 – 08-31-07	4500	13
3	09-01-07 – 12-31-09	11480	42

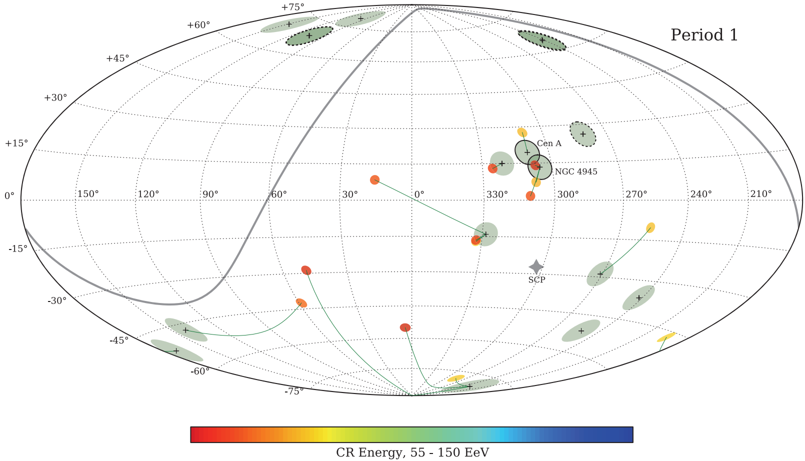
- The CR flux from all 3 periods is

$$(14 + 13 + 42)/(4\pi \times \text{Total Exposure}) = 0.043\ \text{km}^{-2}\text{yr}^{-1}$$

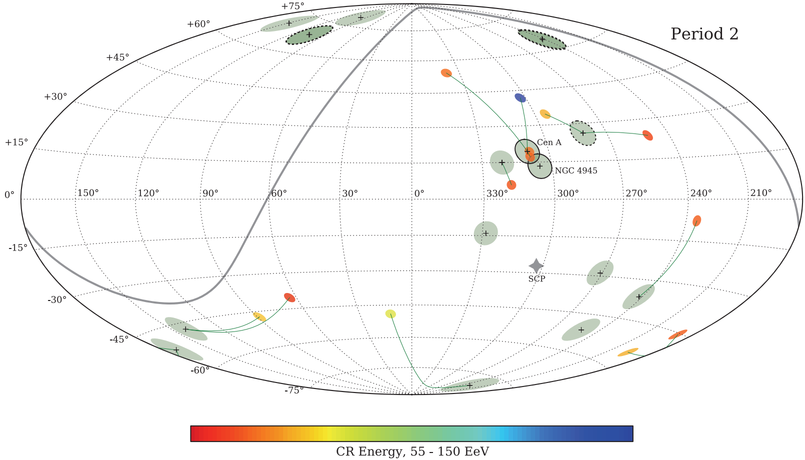
UHECR – AGN Association: Evidence From First 69 CRs



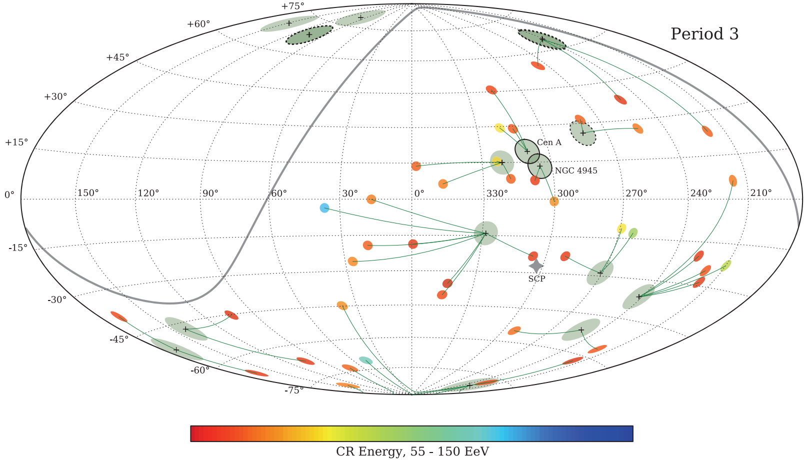
UHECR – AGN Association: Period 1



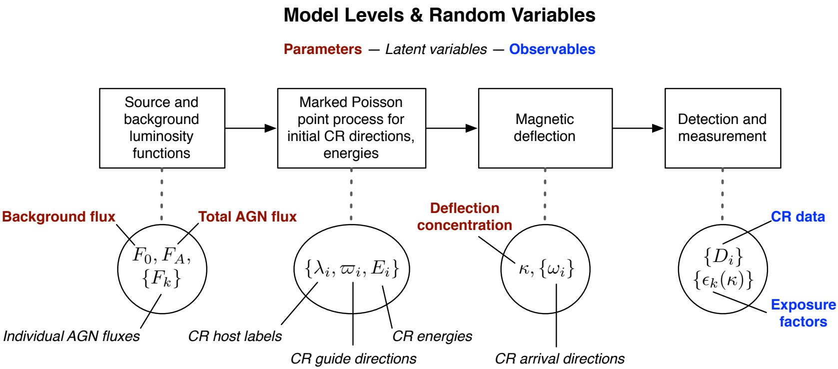
UHECR – AGN Association: Period 2



UHECR – AGN Association: Period 3



Four Levels and Associated Parameters

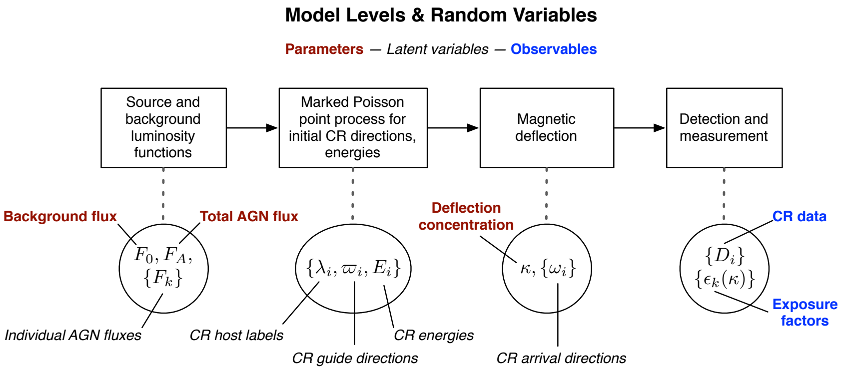


- An isotropic background is included as a“zeroth” source
- 3 different models:
  - M_0 : only isotropic background source
  - M_1: isotropic background source + 17 AGNs
  - M_2: isotropic background source + 2 AGNs: Centaurus A (NGC 5128) and NGC 4945, which are the two closest AGNs

- CR arrivals follow a time-homogeneous Poisson process with rate depending on the fluxes and exposure factors of sources
- $\lambda_i = k$ if the ith CR comes from the k source

- The magnetic deflection of each CR direction is modeled using a Fisher distribution with concentration parameter κ ($\kappa \approx \frac{2.3}{\sigma^2}$ for 2-d Gaussian approximation with standard deviation σ radians)
  - We treat κ as an unknown parameter

- The measurement error of CR direction is modeled using a Fisher distribution with concentration parameter corresponding to angular uncertainty of 0.9°



Overview of the model (again)

Bayesian Analysis: Likelihoods and Priors

$D =$ data (everything known)

$\theta =$ set of all unknown quantities $=$ the "parameters"

Model: $f(D|\theta) =$ probability density function of the data given the parameters (called the likelihood)

Prior density: $\pi(\theta)$ (expresses prior knowledge of θ, if any)

Posterior density: $\pi(\theta|D)$ (expresses knowledge of θ after D has been observed)

Bayesian Analysis: Marginal Likelihood

Joint density of data and parameters: $f(\theta, D) = f(D|\theta)\pi(\theta)$

Marginal density of the data: $f(D) = \int f(D|\theta)\pi(\theta)d\theta$

- called the marginal likelihood and measures how well the model and the prior fit the data
- can be quite sensitive to the prior

From previous slide:

Joint density of data and parameters: $f(\theta, D) = f(D|\theta)\pi(\theta)$

Marginal density of the data: $\int f(D|\theta)\pi(\theta)\,d\theta$

Bayes's Theorem:

$$\text{posterior density} = \pi(\theta|D) = \frac{f(\theta, D)}{f(D)} = \frac{f(D|\theta)\pi(\theta)}{\int f(D|\theta)\pi(\theta)\,d\theta}$$

- typically not sensitive to prior unless the prior is very informative

Data: A bent coin is flipped 5 times and there are D heads

Parameter: $\theta =$ probability of a head on a single toss

Likelihood: $f(D|\theta) = \begin{pmatrix} 5 \\ D \end{pmatrix} \theta^D (1-\theta)^{5-D}$

Prior:

$$\pi(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)},\ 0 < \theta < 1 \text{ (a beta density)}$$

$B(\alpha,\beta) := \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx$ (the beta function)

Here α and β are known "hyperparameters"

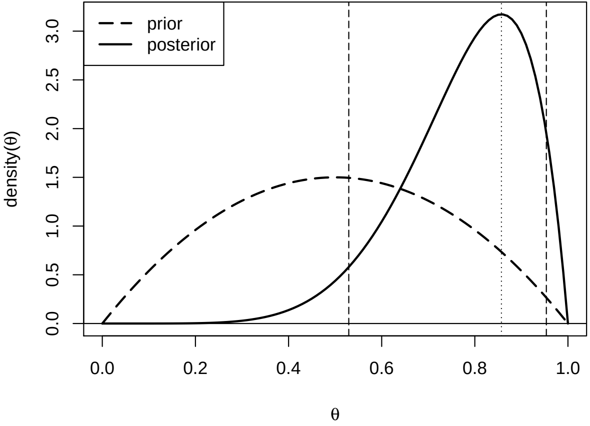
Bayes's Theorem: Example, cont.

Posterior: $\pi(\theta|D) \propto \pi(\theta)f(D|\theta) \propto \theta^{D+\alpha-1}(1-\theta)^{5-D+\beta-1}$ so the posterior is the Beta$(D+\alpha, 5-D+\beta)$ distribution

Marginal likelihood:

$$f(D) = \int f(D|\theta)\pi(\theta)\,d\theta = \binom{5}{D}\frac{B(5+\alpha, 5-D+\beta)}{B(\alpha,\beta)}$$

Example: Prior and Posterior Densities



$\alpha = \beta = 2$ and $D = 5$

Suppose there are N hypotheses (models), $H_1, \ldots, H_N$

Let M_k be the marginal likelihood for the kth hypothesis

Then

$$B_{jk} := \frac{M_j}{M_k}$$

is called the Bayes factor for M_j and against M_k.

Let $P(H_j)$ be the prior probability that H_j is true. Then

$$\frac{P(H_j|D)}{P(H_k|D)} = B_{jk}\frac{P(H_j)}{P(H_k)}$$

Stated differently,

$$\text{Posterior odds} = \text{Bayes factor} \times \text{Prior odds}$$

So the Bayes factor is the evidence from the data for H_j and against H_k.

Caveat: Bayes factors depend heavily upon the priors for the parameters under the models.

Bayes's Theorem:

$$\pi(\theta|D) = \frac{f(\theta, D)}{f(D)} = \frac{f(D|\theta)\pi(\theta)}{\int f(D|\theta)\pi(\theta)\,d\theta}$$

Often:

- The integral $\int f(D|\theta)\pi(\theta)\,d\theta$ cannot be evaluated analytically
- θ is of high dimension so quadrature will not work
- a solution to this problem is Monte Carlo

Modern Bayesian computation usually starts with a Monte Carlo sample from the posterior. Then, for example,

- the posterior means, standard deviations, and correlations are estimated by sample means, standard deviations, and correlations
- if $h(\theta)$ is a scalar function of the parameters, then a 95% (for example) posterior interval ranges from the 2.5 to the 97.5 sample percentiles.

Ideal: independent sample from the posterior

- not often possible

Usual practice: the sample is a Markov Chain with stationary distribution equal to the posterior.

$X_1, X_2, \ldots$ is a Markov Chain if $\mathcal{L}(X_n | X_1, \ldots, X_{n-1}) = \mathcal{L}(X_n | X_{n-1})$. ("$\mathcal{L}(X)$" is the distribution of the random variable X.)

$\mathcal{D}$ is the stationary distribution of this Markov Chain if $X_n \sim \mathcal{D}$ implies that $X_{n+1} \sim \mathcal{D}$. ("$\sim$" means "is distributed as")

Suppose $\theta = (\theta_1, \theta_2)$ and it is easy to sample from

I: $\pi(\theta_1|\mathcal{D}, \theta_2)$

and

II: $\pi(\theta_2|\mathcal{D}, \theta_1)$

The Gibbs sampler iterates between sampling I and II

Often θ is partitioned into more than two elements.

MCMC is a “game changer”

Gibbs sampling is one of many MCMC methods

The Metropolis-Hastings algorithm is quite general and includes Gibbs sampling as a special case

Before MCMC: Bayesian analysis was feasible only for the simplest problems
Now: MCMC can handle models that would be difficult or impossible without MCMC

Our model for UHECRs is a case that seems infeasible without MCMC

Application to our AGN – CR Association Models

We used Gibbs sampling

κ (the deflection parameter) was held fixed during Gibbs sampling

Marginal likelihoods were computed from the Gibbs output by Chib's (1995) method

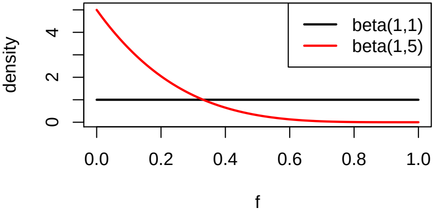
- We compare models 1 and 2 to model 0. The Bayes factors are computed as

$$\mathsf{BF}_{10} = \frac{\ell_1}{\ell_0} \ , \ \mathsf{BF}_{20} = \frac{\ell_2}{\ell_0}$$

- We computed the Bayes factors as function of the amount of magnetic deflection, which is determined by κ
- Later, we marginalized over κ

f = proportion of CRs from AGNs

$f = 0 \iff$ isotropic model

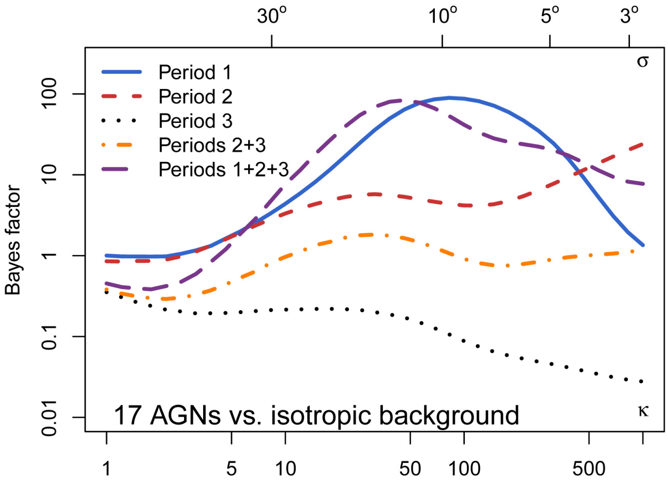


For other models:

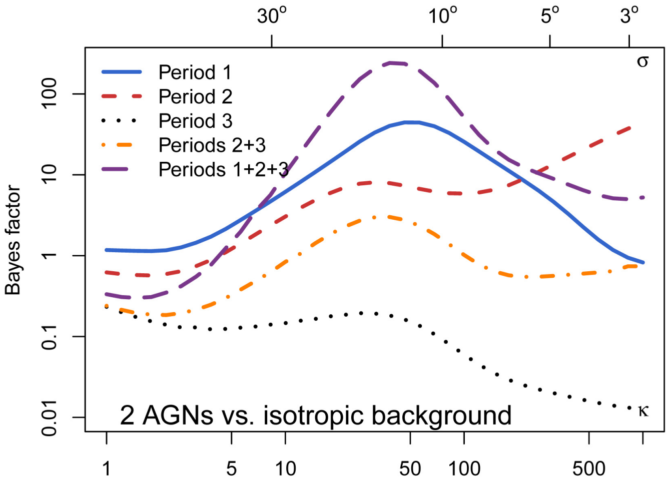
The beta(1,1) was used for most results.

The beta(1,5) prior was used to test sensitivity to the prior.

Bayes Factor Plot – 17 AGNs versus isotropic model



Bayes Factor Plot – 2 AGNs versus isotropic model



Overall Bayes Factors for log-flat prior on κ over [1,1000]

Priors for f	Model	Bayes factors				
		Period 1	Period 2	Period 3	Periods 2&3	Periods 1&2&3
beta(1,1)	17 AGNs	30.53	6.51	$0.15 = 1/6.67$	0.99	25.90
	2 AGNs	14.78	9.89	$0.11 = 1/9.09$	1.06	50.67
beta(1,5)	17 AGNs	39.27	15.12	$0.52 = 1/1.92$	3.39	78.69
	2 AGNs	31.97	27.97	$0.42 = 1/2.38$	4.08	176.65

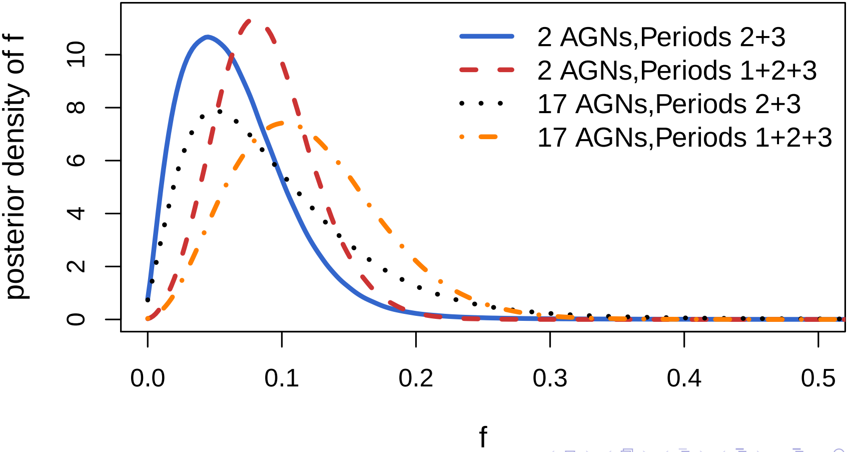
TABLE 1

Kass and Raftery's (1995) rules of thumb:

3 and 20 is "positive" evidence

20 and 150 is "strong" evidence.

Posterior density of $f := F_A/(F_A + F_0)$



Astrophysics Suggests No Period Effect

The true fluxes should not vary over the time scales involved

- A cosmic ray takes millions of years to reach earth from another galaxy.
- Magnetic deflections will cause variation in the paths taken
- Even a burst of cosmic rays will arrive over the period of thousands of years due to the variation in the lengths of their paths

Bayes Factors Also Suggest No Period Effects

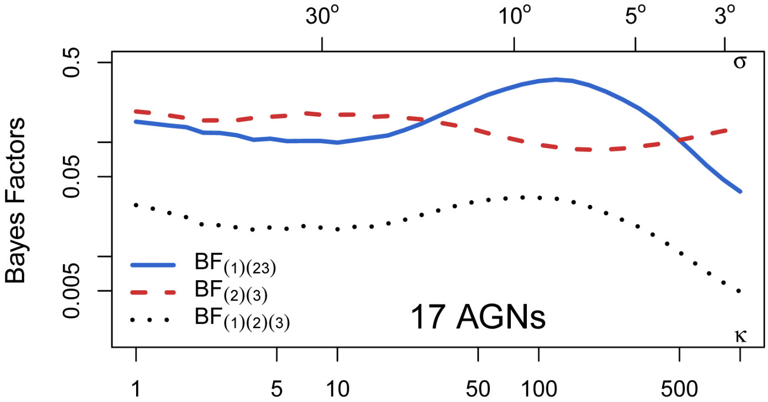
We used Bayes factors to investigate evidence of a period effect

- we found no evidence
- this will be seen in the next frames

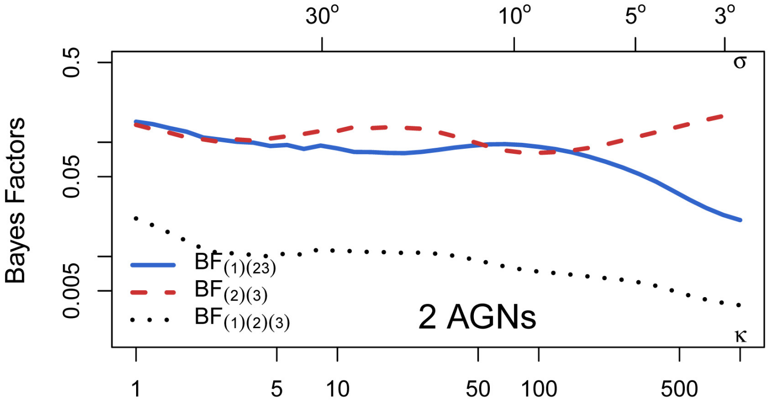
$$
\begin{aligned}
B_{(1)(23)} &= \frac{\mathcal{L}_1 \mathcal{L}_{23}}{\mathcal{L}_{123}} \\
B_{(2)(3)} &= \frac{\mathcal{L}_2 \mathcal{L}_3}{\mathcal{L}_{23}} \\
B_{(1)(2)(3)} &= \frac{\mathcal{L}_1 \mathcal{L}_2 \mathcal{L}_3}{\mathcal{L}_{123}}
\end{aligned}
$$

$\mathcal{L}_{i_1,\ldots,i_q}$ is the marginal likelihood computed for the cosmic rays from periods $i_1, \ldots, i_q$

Bayes Factors for Period Effects: 17 AGNs



Bayes Factors for Period Effects: 2 AGNs



A Puzzle

We found the the Bayes factors for M_1 and M_2 versus M_0 vary greatly between periods

- Yet the Bayes factors just examined indicate that the parameters do not vary between periods
- Could the Bayes factor variation be due to chance?
  - An extensive computer simulation study indicates "yes"
- There is still the worry that the large Bayes factors for period 1 are due to tuning
  - tuning could be investigated by a Monte Carlo study

Summary:

- Three models:
  - M_1: Sources are 17 closest AGNs and isotropic source
  - M_2: Sources are 2 closest AGNs and isotropic source
  - M_0: All CRs come from the isotropic source
- Using all three periods, Bayes factors provide
  - strong evidence for either M_1 or M_2 against M_0
  - little evidence for or against M_1 versus M_2.

Between-period variation in Bayes factors

We observed large between-period variation in the Bayes factors

- Simulation study shows that this could be expected
- Bayes factor do not support hypothesis that parameters vary between periods

Future Work: Other luminosity functions

- We assumed that the fluxes are inversely proportional to squared distances from earth
- Other luminosity functions are plausible
- An example is a model where some AGNs are emitting CRs and others are not
  - This model would use latent indicator variables of emitting AGNS
  - This would mean even more parameters and an even greater need for MCMC

Future Work: Other deflection models

- Deflections could be modeled to decrease with the energy
- A "Radiant" model allows CRs from a single source to have a shared deflection history
    - the shared history would be modeled by a "guide" direction drawn from a Fisher distribution with concentration κ_g, say, and centered at the direction to the source
    - individual CRs would have arrival directions drawn from a Fisher distribution centered at the guide direction

Future Work: Other deflection models, cont.

- The magnetic deflections depend on the proton numbers of the CRs
    - The atomic species are unknown and could be hydrogen, silicon, iron, or other elements
    - A mixture model seems appropriate

Future Work: Other Catalogs

We used a volume-complete catalog to 15 megaparsecs

Could instead use

- larger catalog, or
- flux limited catalog

Need for More Data

A fuller investigation of the sources of UHECRs requires that the PAO disclose more data, e.g.,

- untuned data from period 1
- data collected since period 3 ended in 2009.