Using Bayes Factors for Model Selection in High-Energy Astrophysics

Shandong Zhao

Department of Statistic, UCI

April, 2013
Model Comparison in Astrophysics

- Nested models (line detection in spectral analysis):

  ![Graph showing photon counts vs energy (keV)]

- Non-nested models:
  - *Powerlaw vs Bremsstrahlung* for the red curve.

- Bottom line: need more than a confidence interval on “nesting parameter” to formally compare or select a model.
Spectral Analysis in High Energy Astrophysics

- **Goal:** Study the distribution of the energy of photons originating from a source (We use a Poisson model)

- The photon detector
  1. Counts photons into energy bins, with energy $E_1, \ldots, E_J$.
  2. May misclassify photons into wrong energy bins. (**Redistribution Matrix, M**)
  3. Has sensitivity that varies with energy. (**effective area, d**)
  4. Is subject to **background contamination**, $\theta^B$

- Mathematically: $\Xi(E_i) = \sum_{j \in J} M_{ij} \Lambda(E_j) d_j + \theta^B_i$

- We ignore 2-4 in our initial simulations.
The spectral model can often be formulated as a finite mixture model. A simple form consists of a continuum and an emission line: \( \Lambda(E_i) = \alpha E_i^\beta + \omega I_{\mu=i} \)

The line detection problem:

\[
H_0 : \quad \Lambda(E_i) = \alpha E_i^\beta \\
H_a : \quad \Lambda(E_i) = \alpha E_i^\beta + \omega I_{\mu=i}
\]
A naive method is to use the likelihood ratio test. However, the standard asymptotics of the LRT statistic do not apply.

- $\mu$ has no value under $H_0$.
- $\omega$ must be non-negative under $H_a$ while its target tested value under $H_0$, zero, is on the boundary of the parameter space.

For "precise null hypotheses", $p$-values bias inference in the direction of false discovery.

- When compared to BF or $Pr(H_0|Y)$, $p$-values vastly overstate the evidence for $H_1$ (even using the prior most favorable to $H_1$)
- Computed given data as extreme or more extreme than $Y$, which is much stronger evidence for $H_1$.

Protassov et al. (ApJ, 2002) address the first set of concerns by simulating the null dist’n of the Likelihood ratio statistic and use posterior predictive $p$-values (PPP) instead.
Bayesian Model Selection

- **Bayesian Evidence:** The average likelihood over the prior distribution of the parameters under a specific model choice:

\[ p(Y|M) \equiv \int p(Y|M, \theta)p(\theta|M)d\theta \]

where \( Y, \theta \) and \( M \) are the observed data, parameters, and underlying models respectively.

- **Bayes Factor (BF):** The ratio of candidate model’s Bayesian Evidence:

\[ B_{01} \equiv \frac{p(Y|M_0)}{p(Y|M_1)} \]
Interpretation of BF

- BF and posterior probability ratio.

\[
\frac{p(M_0 \mid Y)}{p(M_1 \mid Y)} = B_{01} \frac{p(M_0)}{p(M_1)}
\]

- Interpretation against the Jeffreys’ scale.

<table>
<thead>
<tr>
<th>BF</th>
<th>Strength of evidence (toward ( M_0 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 \sim 3</td>
<td>Barely worth mentioning</td>
</tr>
<tr>
<td>3 \sim 10</td>
<td>Substantial</td>
</tr>
<tr>
<td>10 \sim 30</td>
<td>Strong</td>
</tr>
<tr>
<td>30 \sim 100</td>
<td>Very strong</td>
</tr>
<tr>
<td>&gt; 100</td>
<td>Decisive</td>
</tr>
</tbody>
</table>
Disadvantage of the Bayes Factor

- Assumes that one of the two models is true.
- Computation could be hard.
- Sensitive to prior specification.
  
  \textit{How does the prior dependency of BF compare to that of PPP?}

- BF is ill-defined with an improper prior.
  
  \textit{Non-informative prior for parameters in common?}
The Computation of BF

- Task is to compute \( p(Y|M) \equiv \int p(Y|M, \theta)p(\theta|M)d\theta \).

  - **Gaussian Approximation.**
    - If the posterior dist’n is approximately Gaussian.

  - **Monte Carlo Method.**
    - If could get a sample from either the prior or posterior dist’n.

  - **Nested Sampling.**

- None of the method is perfect for spectral analysis.
  - The joint posterior dist’n has many local modes.
  - Most Monte Carlo methods are inefficient.
  - Nested Sampling has bias up to 25% in simulation studies.
A New Method

On the other hand, \( B_{01} = \frac{p(M_0 | Y)}{p(M_1 | Y)} / \frac{p(M_0)}{p(M_1)} \)

Computing the ratio of the posterior probability is not easy.

Challenge is to sample from \((I_{M_0}, \Theta_0, I_{M_1}, \Theta_1)\), where \(\Theta_0\) and \(\Theta_1\) might have different parameter settings and dimensions.

**Example**: \(\Theta_0\) for Powerlaw while \(\Theta_1\) for Bremsstrahlung.

It’s usually straightforward, however, to sample from \(p(\Theta_0 | M_0, Y)\) and \(p(\Theta_1 | M_1, Y)\), seperately.
Jump between the Parameter Space

Assume we run $2K$ MCMC chains with half of them starting from $\Theta_0$ and $\Theta_1$ respectively. The parameter space for each chain is $(I_M, \Theta_M)$.

1. Run usual M-H algorithm for each chain with $q_0(\theta_0^{\text{old}}, \theta_0^{\text{new}})$ and $q_1(\theta_1^{\text{old}}, \theta_1^{\text{new}})$ being the proposal dist’n for sampling within $p(\Theta_0|M_0, Y)$ and $p(\Theta_1|M_1, Y)$, respectively.

2. For chain $i$, randomly pick one of the other chains, $j$, and propose a new draw based on its corresponding proposal dist’n. Doing so is equivalent to use the proposal dist’n of:

$$\frac{1}{K-1} \sum_{j \neq i} q_j^i(\theta^i, \theta^{\text{new}}), \text{ where } q_j^i(\theta^i, \theta^{\text{new}}) = 0 \text{ if } I_M(\theta^j) \neq I_M(\theta^{\text{new}})$$

3. Combine all the chains, compute the ratio of $I_{M_0}/I_{M_1}$ as the Monte Carlo estimate of the posterior probability ratio.
The parallel MCMC algorithm was first introduced to help MCMC chain jump between modes.

For step 2, the acceptance rate is

\[
p(\theta^{\text{new}} | M(\theta^{\text{new}}), Y) \frac{\sum_{j \neq i} q^j(\theta^i, \theta^{\text{new}})}{\sum_{j \neq i} q^j(\theta^j, \theta^i)}
\]

Challenge now is to find a good local proposal dist’n.
Is Improper Prior Always Improper?

- If $\theta^*$ only shows up in $M_1$, using improper prior for $\theta^*$ is improper.

$$p(Y|M_1) \equiv \int p(Y|\theta^*, \tilde{\theta})p(\tilde{\theta}|\theta^*)p(\theta^*)d\tilde{\theta}d\theta^*, \ \Theta^1 = (\theta^*, \tilde{\theta})$$

- What if $\theta^*$ is one of the parameters in common?

In the line detection problem with $\beta, \mu$ being fixed and assuming $p(\omega/\alpha) \sim U(0, \eta)$,

$$H_0 : \Lambda(E_i) = \alpha E_i^\beta \ \text{vs} \ \ H_a : \Lambda(E_i) = \alpha E_i^\beta + \omega I_{\mu=\mu_i}$$

The BF$s$ under the prior of $p(\alpha) \sim U(0, N)$ converge as $N \to \infty$, to the BF under the prior of $p(\alpha) \propto 1$.

- What about the priors for $\omega$ and $\mu$?
The Example

If \( p(\alpha) \propto 1 \),

\[
BF = \eta \bigg/ \int_0^{\eta} \frac{(1 + \tilde{\omega}/E_{\mu}^{-\beta})^{Y_{\mu}}}{(1 + \tilde{\omega}/\Sigma E_i^{-\beta})^{\Sigma Y_i + 1}} \, d\tilde{\omega}
\]

If \( p(\alpha) \sim U(0, N) \),

\[
BF_N = \eta \bigg/ \int_0^{\eta} \frac{(1 + \tilde{\omega}/E_{\mu}^{-\beta})^{Y_{\mu}}}{(1 + \tilde{\omega}/\Sigma E_i^{-\beta})^{\Sigma Y_i + 1}} \cdot \frac{\Pr(\tilde{z} \leq N)}{\Pr(z \leq N)} \, d\tilde{\omega}
\]

where \( z \sim Gamma(\Sigma Y_i + 1, \frac{1}{\Sigma E_i^{-\beta}}) \), \( \tilde{z} \sim Gamma(\Sigma Y_i + 1, \frac{1}{\Sigma E_i^{-\beta} + \tilde{\omega}}) \)
Because

\[
\frac{\left(1 + \frac{\tilde{\omega}}{E_\mu^{-\beta}}\right)^\mu}{\left(1 + \frac{\tilde{\omega}}{\Sigma E_i^{-\beta}}\right)^{\Sigma Y_{i+1}}} \cdot \Pr(\tilde{Z} \leq N) \leq \frac{\left(1 + \frac{\tilde{\omega}}{E_\mu^{-\beta}}\right)^\mu}{\left(1 + \frac{\tilde{\omega}}{\Sigma E_i^{-\beta}}\right)^{\Sigma Y_{i+1}}}
\]

\[
\lim_{N \to \infty} BF_N = \lim_{N \to \infty} \int_0^\eta \frac{\left(1 + \frac{\tilde{\omega}}{E_\mu^{-\beta}}\right)^\mu}{\left(1 + \frac{\tilde{\omega}}{\Sigma E_i^{-\beta}}\right)^{\Sigma Y_{i+1}}} \cdot \Pr(\tilde{Z} \leq N) \, d\tilde{\omega} \Bigg/ \lim_{N \to \infty} \Pr(z \leq N)
\]

\[
= \int_0^\eta \lim_{N \to \infty} \frac{\left(1 + \frac{\tilde{\omega}}{E_\mu^{-\beta}}\right)^\mu}{\left(1 + \frac{\tilde{\omega}}{\Sigma E_i^{-\beta}}\right)^{\Sigma Y_{i+1}}} \cdot \Pr(\tilde{Z} \leq N) \, d\tilde{\omega}
\]

\[
= BF
\]

where the second "\(\equiv\)" holds by Lebesgue dominated convergence theorem.
How to Assign a Proper Prior

- Compared to $\alpha$ and $\beta$, priors for $\omega$ and $\mu$ have much more influence on the BF. And they have to be proper.

- Different priors on $\omega$ and $\mu$ can totally change your decision based on BF. For example, with everything else held the same, under $p(\mu) \sim N(\mu_0, \sigma_1)$, BF supports $M_0$ 
  under $p(\mu) \sim N(\mu_0, \sigma_2)$, BF can’t distinguish btwn the models 
  under $p(\mu) \sim N(\mu_0, \sigma_3)$, BF supports $M_1$

- Is the prior dependency always a problem?

- How does the prior influence of BF compare to that of the PPP?
**Simulation Study Design**

- **Simulation Models:** We compare a power law continuum with one delta function emission line model, with 1000 energy bins equally spaced between 0.3 to 7(keV).

  \[H_0 : \Lambda(E_i) = \alpha E_i^\beta\]
  \[H_a : \Lambda(E_i) = \alpha E_i^\beta + \omega I_{\mu=i}\]

  with \(i = 1 \sim 1000\) and \(\alpha = 50, \beta = 1.69\).

- The prior influence of \(\alpha\) and \(\beta\) are negligible compared to that of \(\omega\) and \(\mu\). Thus, they will be fixed in the simulation study.

- Assume:

  \[\omega \sim U(0, \eta); \mu \sim \text{discrete}[N(\mu_0, \sigma^2)]\]

  *Using a Gamma prior for \(\omega\) will have similar results.*
The ordinary Gibbs breaks down here because the subchain for $\mu$ does not move from its starting value, regardless of what it is. We use the \textbf{PCGS} to draw posterior samples.

5000 posterior draws with $\alpha = 50, \beta = 1.69, \omega = 10, \mu = 150$. 

\begin{itemize}
  \item $\eta=100, \mu_0=150, \sigma=3$
  \item $\eta=10, \mu_0=500, \sigma=3$
\end{itemize}
To Study The Prior Influence

- Fix $\alpha$ and $\beta$ throughout. Calculate BF by numerical integration.
- The “true” emission line is set at bin 150, or $\mu = 1.3$ keV.
- The intensity from the continuum in this bin is 32.
- We control the strength of data support toward $H_a$ by altering the observed counts at 1.3 keV.
Prior Settings

- Recall $\omega \sim U(0, \eta)$. We control its strength by changing its upper range $\eta$.
  - $\eta$ will range from 10 to 108 with a step size of 2.

- For $\mu$, because $\mu \sim \text{discrete}[N(\mu_0, \sigma^2)]$, we control both its mode $\mu_0$ and s.d $\sigma$.
  - We use two different value for $\mu_0$, 1.3keV and 1.97keV respectively (150 and 250 in terms of bin number).
  - For $\sigma$, it will range from 1 to 99 (bin width) with a step size of 2.
We will plot the heatmap of $\log(BF)$ against $\eta$, $\mu_0$, and $\sigma$ on the simplified Jeffrey’s scale.

<table>
<thead>
<tr>
<th>BF</th>
<th>$\log(BF)$</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>$&gt;30$</td>
<td>$&gt;1.5$</td>
<td>Very strong to overwhelming for $H_0$</td>
</tr>
<tr>
<td>[3, 30]</td>
<td>[0.5, 1.5]</td>
<td>Substantial to strong for $H_0$</td>
</tr>
<tr>
<td>$[-3, 3]$</td>
<td>$[-0.5, 0.5]$</td>
<td>Not worth mentioning</td>
</tr>
<tr>
<td>$[-30, -3]$</td>
<td>$[-1.5, -0.5]$</td>
<td>Substantial to strong for $H_a$</td>
</tr>
<tr>
<td>$&lt;-30$</td>
<td>$&lt;-1.5$</td>
<td>Very strong to overwhelming for $H_a$</td>
</tr>
</tbody>
</table>
Results: A Weak Spectral Line

\[ Y(E = 1.3) \text{ is about 3 s.d above null model intensity.} \]

\[ Y(E=1.3)=32+17, \quad \mu_0=1.3\text{keV} \]

\[ Y(E=1.3)=32+17, \quad \mu_0=1.97\text{keV} \]

Diffuse or misplaced priors weaken evidence
Results: A Stronger Spectral Line

\( Y(E = 1.3) \) is about 5 s.d above null model intensity.

Diffuse or misplaced priors could completely change the decision.
Results: Stronger Prior

We use a stronger prior for $\mu$: uniform prior with a span of $11 \sim 51$ bin width centered at the true location.
Take Home Messages

- If the data is dominantly strong, we probably don’t need BF.
- The priors can reflect different scientific questions
  - $p(\mu)$: where to look for the lines
  - $p(\omega)$: how strong are the lines that we’re looking for
- Even for likelihood ratio test, looking for lines
  - at a fixed bin location,
  - within a restricted region,
  - over the whole energy range
  will return tests with varied strength of the evidence.
- How does the prior dependency of BF compared to the PPP?
**Compare BF with P-values**

ppp-values (based on 1000 MC samples)

<table>
<thead>
<tr>
<th>Y(E=1.3 keV)</th>
<th>32+17</th>
<th>32+22</th>
<th>32+28</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_A$: known line location</td>
<td>0.008</td>
<td>0.002</td>
<td>0.000</td>
</tr>
<tr>
<td>$H_A$: fitted line location (0.3-7.0 keV)</td>
<td>0.539</td>
<td>0.184</td>
<td>0.006</td>
</tr>
</tbody>
</table>
Prior on line intensity: $\omega \sim U(0, \eta)$ and $\mu \sim U(1.3 \pm \kappa)$.

$H_A$: known line location
- ppp-value = 0.002.

$H_A$: Unknown line location
- ppp-value = 0.184.

minimum Bayes Factor = 0.044 (span=0.07, $\eta = 30$)

Both ppp-value and Bayes Factor depend on where we look for line.

Can we calibrate the dependence?
Assuming $P(M_0)/P(M_1) = 1$, we plot the PPP against $P(M_0|Y)$.

**Evidence decreases with more diffuse prior, for both.**

**BFs are more conservative.**

Prior on $\mu$

- *let’s decide where to look,*
- *penalize us for looking too many place. i.e., look elsewhere effect*
- *Sensitivity of BF to prior for $\mu$ is sensible.*
Bayes Factors for Detection:

\[ BF = \frac{p_0(Y)}{p_A(Y)} = \frac{\int p(Y|\theta, \omega = 0)p(\theta)d\theta}{\int p(Y|\theta, \mu, \omega)p(\theta, \mu, \omega)d\theta d\mu d\omega} \]

Setting priors

\[ \theta = (\alpha, \beta) : \quad \text{Non-informative / diffuse priors.} \]
\[ \mu : \quad \text{Where we want to look for the line.} \]
\[ \omega : \quad \text{How strong of a line do we want to look for?} \]

Narrower prior ranges yield stronger results.

If strong lines are easy to see, maybe we can confine attention to weak lines.