Discussion of “Cosmological Bayesian Model Selection: Recent Advances and Open Challenges”

David A. van Dyk

Department of Statistics, University of California, Irvine
Statistics Section, Imperial College London

SCMA V, June 2011

Outline

Outline

The Model Selection & Checking Problems

1. Typically begin with baseline, default, or presumed model:
   Null Hypothesis: The Universe is “Flat”
   - *Model Checking:* Is the model consistent with the data?
   - If not, characterize inconsistency, improve model, recheck.
2. May have another model that we suspect or hope is better:
   Alternative Hypothesis: The Universe is “Hyperbolic”
   - *Model Selection / Comparison:* Decide between or weigh the evidence for the two (or more?) models.
3. These are surprisingly subtle problems:
   - No consensus exists on how to proceed.
   - Disagreement between Bayesian and Frequentist methods.

Neyman-Pearson

Model Selection:

H_0 The Universe is Flat: $\Omega_\kappa = 0$

H_A The Universe is not Flat: $\Omega_\kappa \neq 0$.

- Need test statistic, T, with known distribution under H_0.
- Threshold $T^\star$ is the smallest value such that

$$\Pr(T > T^\star | \Omega_\kappa = 0, \text{other parameters}) \leq \alpha,$$

If $T > T^\star$ sufficient evidence to declare non-flat.

Assessment?

Pro: Frequency properties: Bounded Pr(false positive).

Con: No characterization of the strength of evidence. How to find T??

Bayes Factors and Posterior Probabilites

Bayesian methods have no trouble with unknown parameters

- The prior predictive distribution:

$$p_i(x) = \int p_i(x|\theta)p_i(\theta)d\theta$$

- How likely is X under model i (likelihood + prior dist'n).
- Compare two models with the *Bayes Factor:*

$$\text{Bayes Factor} = \frac{p_0(x)}{p_A(x)}.$$

or the *posterior probability of H_0:*

$$\Pr(H_0|x) = \frac{p_0(x)\pi_0}{p_0(x)\pi_0 + p_A(x)(1-\pi_0)}.$$

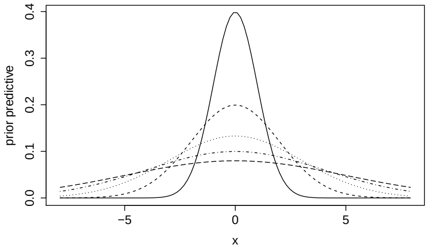
The Choice of Prior Dist'n Matters!

Example:

Likelihood: $X \sim \mathrm{N}(\mu, 1)$.

Prior Dist'n: $\mu \sim \mathrm{N}(0, \tau^2)$.

Prior Pred.: $X \sim \mathrm{N}(0, 1+\tau^2)$.



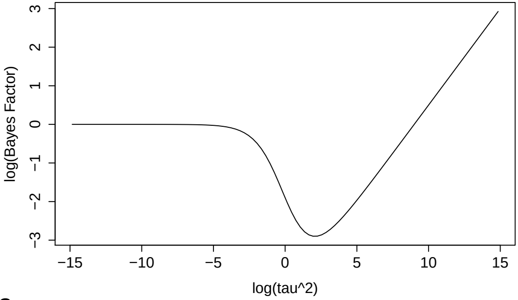
Value of $p_A(x)$ depends on τ^2!

Must think hard about choice of prior and report!

The Choice of Prior Dist'n Matters!

Bayes Factor:

$$H_0: \quad X \sim \mathrm{N}(0, 1).$$
$$H_A: \quad X \sim \mathrm{N}(0, 1 + \tau^2).$$



Assessment of Bayes Factors.

Cons: Bayes Factor depends heavily on the prior scale.

Pros: Probability based principled method, answers right question, no problem with nuisance parameters.

How to Choose the Prior Dist'n.

- Unlike with parameter inference, prior must be proper.
  - Prior Predictive Distribution is improper with improper prior!
- There is no default prior distribution.
- Possible Solutions
  1. Minimize Bayes Factor over a class of priors (see below).
  2. Use a subjective prior distribution.
- Subjective prior distributions are especially elusive:
  What are likely values a parameters in a possible model?
- Problem is even more complicated when:
  - Parameter space is large.
  - H_0 and H_A have different (non-nested) parameters.

Prior Distributions in Cosmology

Prior distributions:

- 1. "Astronomer's Prior:" $\Omega_\kappa \sim \mathrm{Unif}(-1, 1)$
  2. "Curvature Scale Prior:" $\log|\Omega_\kappa| \sim \mathrm{Unif}(-5, 0)$
  3. Inflationary Model: "little if anything is known *a priori* about the free parameter Ψ..."
  4. "Typical priors are uniform on the log of this parameter."
  5. "Non-linear transformations ... in general change ... the model comparison results"
- These appear to be priors on convenience...
- Bayes Factors based on such priors are questionable.

P-values

Recall our Example of Neyman-Pearson:

$$H_0 : \Omega_\kappa = 0 \qquad \text{versus} \qquad H_A : \Omega_\kappa \neq 0.$$

Threshold $T^\star$ is the smallest value such that

$$\Pr(T > T^\star | \Omega_\kappa = 0, \text{other parameters}) \leq \alpha,$$

If $T \leq T^\star$ we accept $H_0 : \Omega_\kappa = 0$.
If $T > T^\star$ we reject $H_0 : \Omega_\kappa = 0$.

To quantify the degree of evidence, *p-value* is often reported:

$$\text{p-value} = \Pr(T > T^\star | \Omega_\kappa = 0, \text{other parameters}).$$

A Dangerous Method....

Although the use of p-values is endemic in data analysis, they are not easily interpreted (for a precise H_0[1]):

1. When compared to Bayes Factors or $\Pr(H_0|\text{data})$, p-values *vastly overstate the evidence for* H_1.
   - Even using the prior most favorable to H_1 (in a large class).
2. Computed given data as extreme or more extreme than X.
   - This is *much stronger evidence* for H_1 than X.
   - Agree with Bayes measures given “as/more extreme’.
3. P-values cannot be easily calibrated with Bayes Measures
   - Depends on sample size, model, and precision of H_0.

P-values bias inference in the direction of false discovery.

[1] Berger & Delampady, *Testing Precise Hypotheses*, Stat. Sci., 1987

Not a Frequentist Method...

> *"... a rough rule known to astronomers, i.e., that differences up to twice standard error usually disappear when more or better observations become available, and that thoes of three or more times usually persist."* [2]

- Suppose over time, H_0 is true about half the time.
- Looking back over results with $1.96 <$ p-value < 2.00, the astronomer might find H_0 to be true 30% of the time.
- The absolute minimum limiting proportion is 22%.
- Compare with "5% significance" associated with p-value.

Why are p-values so popular?

[2] Jeffrey (1980) in Berger & Delampady (1987)

Why are p-values so popular?

Maybe it is just a bad habit....

Assessment of P-values

Cons: Biased toward (false!!) discovery and uninterpretable.

Pros: Everyone is doing it...

Posterior Predictive P-values

Hybrid Methods: Recall the definition of the p-value:

$$\text{p-value} = \Pr(T > T^{\rm obs} | H_0).$$

How do we compute p-value with unknown param's under H_0?

1. Careful choice of T, dist'n may not depend on unknowns.
2. Use estimates of unknowns under H_0.
3. Average over the posterior dist'n of unknowns under H_0:

$$\text{ppp-value} = \int \Pr(T > T^{\rm obs} | H_0) p(\theta | x) d\theta.$$

ppp-values may be very weak with poor choice of T. Use LRT!

Example

Spectral Analysis in High Energy Astrophysics:
Quasar PG1637+706.

MODEL 0. There is no emission line.

MODEL 1. There in an emission line with fixed location in the spectrum, but unknown intensity.

MODEL 2. There is an emission line with unknown location and intensity.

To fit Model 2 under H_0 we use multiple starting values...
and use the *same* starts with the real data.

Results

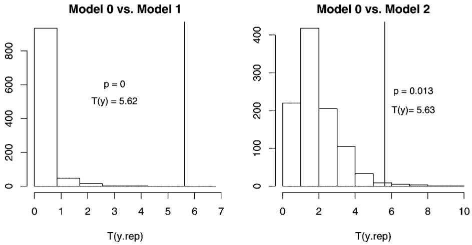


FIG. 4. *The posterior predictive check. The two histograms compare the observed likelihood ratio test statistics (vertical lines) with* 1000 *simulations from the posterior predictive distribution. The left plot is the comparison between Model* 0 *and Model* 1, *and the right plot is the comparison between Model* 0 *and Model* 2. *Both model checks indicate strong evidence for including the emission line.*

Assessment of ppp-values

Pros: Can handle nuisance parameters.

Cons: They look like p-values!

Other Methods

There are *Many* other methods....

1. Bayesian Model Averaging
   - Pros: Bayesian, but less dependent on the choice of prior.
   - Cons: More appropriate for prediction than model selection.
2. Decision Theory
   - Pros: Derives rules tailored to specific scientific goals.
   - Cons: Sensitive to choice of Loss Function and Prior.
3. Information Criteria (e.g., AIC, BIC, etc.)
   - Pros: Simple to compute with an intuitive form!
   - Cons: Ad hoc—with questionable statistical properties.
4. Conditional Error Probabilities
   - Pros: Bayesian methods with frequency interpretation!
   - Cons: Frequency conditional prob's make eyes glaze over.

Other Methods

There are *Many* other methods....

5. "Default Bayes Factors"
   - Pros: Derive a proper prior dist'n based on training sample.
   - Cons: Result depends on the choice of training sample.

These are all useful methods!

.... But they all must be handled with care with an understanding of their pros and cons.

Conditional Error Probabilities

1. Define (Berger, Brown, and Wolpert, AoS, 1994)
   - p_0: The p-value as we have defined it.
   - p_1: The p-value with H_0 and H_A interchanged.
   - S The maximum of p_0 and p_1.
2. Reject H_0 if $p_0 < p_1$ accept H_0 and accept otherwise.
3. Report the conditional error probabilities:
   - $\alpha(s)$: Probability of Type 1 error given $S = s$.
   - $\beta(s)$: Probability of Type 2 error given $S = s$.
4. Note $\alpha(s) = \Pr(H_0|X)$ and $\beta(s) = \Pr(H_A|X)$ with $\pi_0 = 0.5$.

Example of the use of conditioning to improve the properties of statistical procedures.

Conditional Error Probabilities

Assessment of conditional methods

- Pros: Bayesian methods with frequency interpretation!
- Cons: Frequency conditional probabilities make eyes glaze over.

Decision Theory

A decision theoretic approach begins with a "Loss" Function, perhaps with $c << C$.

	Decision	
Truth	H_0	H_A
H_0	0	C
H_A	c	0

Derive *decision rule*, for example minimizing the *Bayes Risk*:

$$\text{Bayes Risk} = \pi_0 \, \mathrm{E}(\text{Loss}|\text{decision}, \mathrm{H}_0) + (1-\pi_0) \, \mathrm{E}(\text{Loss}|\text{decision}, \mathrm{H}_1)$$

Assessment of Decision Theory

Pros: Derives rules tailored to specific scientific goals.

Cons: Sensitive to choice of Loss Function and Prior.

Can we *abandon formal model selection* all together?

- Nested Models:

$$H_0\text{: } \Omega_\kappa = 0 \text{ (a special case of } H_A)$$
$$H_A\text{: } \Omega_\kappa \neq 0 \text{ or } \Omega_\kappa > 0 \text{ or } \Omega_\kappa < 0$$

  1. Fit the larger model and give an interval for θ: ***No Testing!***

- Does this answer the larger question?
  1. Is null value a *special value*?
  2. Should extra weight be put on default / presumed model?
     - *If not* an interval may suffice.
     - *If yes* some sort of formal model selection may be needed.

- “Nested models are fairly common in cosmology”
  1. “flat or near flat universe is predicted by inflation"
  2. testing for infinite universe, $\Omega_\kappa \leq 0$.

Outline

Are Bayesian Methods Best??

1. Why use Bayesian Methods?
2. Bayesian methods *require* a prior distribution—and for model selection the prior distribution really matters.
3. Bayes Factors require an Alternative Hypothesis.
   - Might we just be interested in validity of proposed model?
   - Yes, but any test statistic has an implicit alternative.
   - Practically speaking, there is always an alternative.
   - Formalizing H_A, leads to a *much larger* toolbox.

I view these as disadvantages of Bayesian Methods.

Outline

Model Selection & Model Checking are not for the faint of heart...

- Approach Model Selection with humility.
- If possible it should simply be avoided...
- This seems possible in cosmology—at least in some cases.

If model comparison is necessary.....

1. It is hard to justify p-values—they are simply not calibrated

   We feel that the correct interpretation of a P-value, although perhaps objective, is nearly meaningless, and that the actual meaning usually ascribed to a P-value by practitioners contains hidden and extreme bias. — J. Berger and M. Delampady (Stat Sci., 1987).

2. Bayes Factors are *highly* dependent on choice of prior.

   Bayesians address the question everyone is interested in by using assumptions no one believes, while frequntists use impeccable logic to deal with an issue not of interest to anyone. — L. Lyons (via R. Trotta).

If model comparison is necessary.....

1. At least the Bayesian can clearly identify the assumptions.
2. So... I prefer Bayes Factors—but with:
    1. Careful choice of prior distribution.
    2. Clearly identified prior distribution.
    3. Comprehensive analysis of sensitivity to prior.
3. If no informative prior is available, identify classes of prior distribution that lead to one choice or the other.

As Always: Try several methods and compare results!!!