



Current Challenges in Bayesian Model Choice: Comments



William H. Jefferys

Department of Astronomy
University of Texas at Austin

and

Department of Statistics
University of Vermont



The Problem

- This paper thoroughly discusses modern Bayesian model choice, both theory and experience
- ★
- ★ • Context: Exoplanets group of 2006 SAMSI Astrostatistics Program
- ★
 - Group met weekly to discuss various statistical aspects of exoplanet research
 - From the beginning, model choice was deemed to be a major challenge
 - Is a signal a real planet or not?
 - Given a real planet, is the orbit degenerate (essentially zero eccentricity) or not?



The Problem

- Similar problems arise throughout astronomy, e.g.,
 - ★ ● Given a class of empirical models (e.g., truncated polynomials, splines, Fourier polynomials, wavelets), which model(s) most parsimoniously but adequately represent the signal?
 - ★
 - ★
 - ★
 - E.g., fitting a light/velocity curve to Cepheid variable data
 - Problems of this type are best viewed as model averaging problems, but the techniques used are related and similar



The Problem

- Similar problems arise throughout astronomy, e.g.,
 - ★ • Given a CMD of a cluster (with perhaps other information), is a given star best regarded as
 - ★ – A cluster member or a field star?
 - ★ – A single or a binary cluster member?



The Problem

- Similar problems arise throughout astronomy, e.g.,



- From WMAP data, can we infer that $n_s < 1$?





Dimensionality

- In all these examples, we are comparing models of *differing dimensionality*, i.e., the number of parameters in different models is different



$$\theta_i \in \Theta_m, \quad m \in M, \quad i = 1, \dots, N_m,$$

- The models may be nested, i.e.,

$$\Theta_1 \subset \Theta_2 \subset \dots \subset \Theta_m \subset \dots$$

or not.

- If the models are not nested, the parameters in one model need not bear any physical relationship to those in another model



Frequentist Approaches

- A frequentist approach will typically pick a particular model as the “null model,” and calculate the tail area of the probability density under that model that lies beyond the observed data (a “p-value,” for example).
- Working scientists often want to interpret a p-value as “the probability that the null hypothesis is true,” or “the probability that the results were obtained by chance”
 - Neither of these interpretations is correct
- Tests such as likelihood ratio tests also consider an alternative hypothesis, but the interpretation of the test statistic is equally problematic



Bayesian Approaches

- Bayesian approaches are required to consider all models; no particular model is distinguished as “the null model”
- ★ • Bayesian approaches have the advantage that their interpretation is more natural and is the one that most working scientists would like to have
 - Thus, a posterior probability of model i is interpreted as the probability that model i is the true model, given the data observed (and the models, including priors)
- They naturally handle both nested and unnested models



Bayesian Approaches

- But priors must be explicitly displayed; and other information must also be provided by the scientist (e.g., if a decision is to be made to act as if a particular state of nature is the true one, a loss function must be provided.)





Priors

- As the paper points out, a critical aspect of Bayesian model selection problems under variable dimensionality is the choice of prior under each model
 - ★
 - ★
- This problem is much more difficult than in parameter fitting
 - ★
 - In different models the “same” parameter may have a different interpretation and generally will need a different prior
 - Improper priors, popular in parameter fitting problems, are generally disallowed in model selection problems, because they are only defined up to an arbitrary multiplicative constant...this results in marginal likelihoods and Bayes factors that also contain arbitrary multiplicative constants



Priors

- There are no general prescriptions for prior choice in such problems, although some useful rules are available in special situations
 - ★
 - ★
 - In linear models, the Zellner-Siow prior often works well
 - Various other methods mentioned, such as Intrinsic Bayes Factors, Fractional Bayes Factors, Expected Posterior Priors, use a “training sample” of the data to produce priors that may work well, by calibrating the prior in a “minimal” fashion.
 - There is a danger of using the data twice, so this must be compensated for
 - In any case, great care must be taken



Computation

- The other major difficulty is computational

- ★ • We reduce the problem to evaluating integrals of the form



$$m(x) = \int_{\Theta_m} f(x | \theta_m) \pi(\theta_m) d\theta_m$$



where f is the likelihood, π the prior, and the integral is over the space Θ_m , which is in general of very large dimension

- This comes from writing Bayes' theorem in the form

$$m(x)\pi(\theta_m | x) = f(x | \theta_m)\pi(\theta_m)$$

and integrating over Θ_m , noting that the posterior $\pi(\theta_m | x)$ is normalized



Computation

- Because of the high dimensionality, computing this integral in many real problems can be quite challenging (or, unfortunately, even unfeasible)





Computation

- The curse of dimensionality
 - ★ • A number of appealing methods work well only in lower dimensions, e.g., cubature methods and importance sampling. Where they work, they can work quite well. In an exoplanet problem with just a few planets in the system, these methods can be quite effective.
 - However, beyond about 20 dimensions these approaches begin to fail



Computation

- Since we may have already spent a good deal of time producing an MCMC sample from the posterior distribution under each model, the first thing that comes to mind is, can't we somehow bootstrap this information into corresponding estimates of the required marginal likelihoods?
 - ★
 - ★
 - ★
- This appealing idea turns out to be more difficult than it appears at first glance



Computation

- Thus, Bayes' theorem again, in a different form:



$$\frac{\pi(\theta_m)}{m(x)} = \frac{\pi(\theta_m | x)}{f(x | \theta_m)}$$



Integrating,



$$\frac{1}{m(x)} = \int_{\Theta_m} \frac{\pi(\theta_m | x)}{f(x | \theta_m)} d\theta_m$$

leading to the estimate

$$\frac{1}{m(x)} \cong \left\langle \frac{1}{f(x | \theta_m)} \right\rangle_{\theta_m^*}$$

where the average is over the sample $\{\theta_m^*\}$

- This “harmonic mean” idea suffers from having infinite variance



Computation

- Gelfand and Dey proposed writing Bayes' theorem as

$$\frac{q(\theta_m)}{m(x)} = \frac{q(\theta_m)\pi(\theta_m | x)}{f(x | \theta_m)\pi(\theta_m)}$$

with q a proper density, and integrating to get

$$\frac{1}{m(x)} = \int_{\Theta_m} \frac{q(\theta_m)\pi(\theta_m | x)}{f(x | \theta_m)\pi(\theta_m)} d\theta_m$$

and estimate

$$\frac{1}{m(x)} \cong \left\langle \frac{q(\theta_m)}{f(x | \theta_m)\pi(\theta_m)} \right\rangle_{\theta_m^*}$$

- This is difficult to implement in practice since it is not easy to choose a reasonable tuning function q , particularly in high dimensional cases. q needs to have thin tails.



Computation

- Jim Berger proposed “Crazy Idea No. 1”. It defines an importance function derived as a mixture over (a subset of) the MCMC samples (with t_4 kernels).



$$q(\theta_m) = \frac{1}{M} \sum_{j=1}^M t_4(\theta_m | \theta_{m,j}^*, V_j)$$

- Then by drawing a sample $\{\theta_m^*\}$ from q we can approximate the marginal likelihood as the average

$$m(x) \cong \left\langle \frac{f(x | \theta_m) \pi(\theta_m)}{q(\theta_m)} \right\rangle_{\theta_m^*}$$

- Worked for low dimensions but expected to fail in high dimensions. We want q to have “fat tails,” hence the t_4 density



Computation

- Another family of approaches using the MCMC output is based on Chib's idea of solving Bayes' theorem for the marginal likelihood and evaluating at any point θ_m^* in the sample space, i.e.,



$$m(x) = \frac{f(x | \theta_m^*) \pi(\theta_m^*)}{\pi(\theta_m^* | x)}$$

- This requires an accurate estimate of the posterior density at the point of evaluation
- It may be difficult in multi-modal problems such as those that arise in exoplanet problems because many discrete periods for the planet may fit the data more or less well



Computation

- Jim Berger's "Crazy Idea #2" starts with the Chib-like identity



$$m(x)\pi(\theta_m | x)q(\theta_m) = f(x | \theta_m)\pi(\theta_m)q(\theta_m)$$



- Integrating and dividing through yields



$$m(x) = \frac{\int f(x | \theta_m)\pi(\theta_m)q(\theta_m)d\theta_m}{\int \pi(\theta_m | x)q(\theta_m)d\theta_m}$$

- The upper integral is approximated by averaging $f\pi$ over a draw from q , and the lower integral by averaging q over a subsample of the MCMC draws from the posterior (preferably an independent subsample from that used to define q).
- But it pay too much to the mode(s), since each integrand is approximately the square of the posterior density.



Computation

- Sensitivity to proposals
 - ★ ● Some methods, such as reversible-jump MCMC (RJ-MCMC) can work in high dimensions, but they can also be sensitive to the proposal distributions on the parameters.
 - ★ ● Poor proposal distributions may lead to the sampler getting “stuck” on particular models, and thus poor mixing
 - ★ ● Multimodal distributions are difficult
 - Parallel tempering, by running models in the background that mix more easily, can help to alleviate both problems
 - E.g., Phil Gregory’s automatic code, written in Mathematica (see poster paper #12 at this conference)



Computation

- Sensitivity to tuning parameters
 - ★ ● Our group spent some time on Skilling's "nested sampling" idea
 - ★ ● This reduces a high-dimensional problem to a one-dimensional problem and is in theory, at least, a very attractive approach
 - ★ ● However, our experiments showed the method to be very sensitive to the choice of tuning parameters, and we found the nested sampling step (which is conditional on sampling only where the posterior probability is larger than the most recent evaluation) to be problematic, and disappointing
 - Often the results were way off from the known values



The Bottom Line

- Bayesian model selection is easy in concept but difficult in practice
- ★ • Great care must be taken in choosing priors
- ★ • There are no “automatic” methods of getting accurate calculations of the required marginal likelihoods
 - Computational methods should be chosen on a case-by-case basis
 - It is useful to compare results of several methods